KIVA: KID-INSPIRED VISUAL ANALOGIES FOR TESTING LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

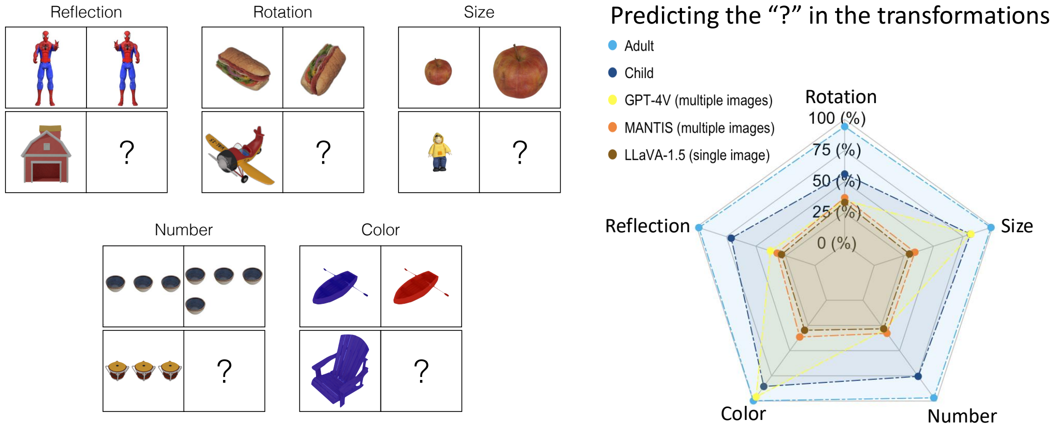
This paper investigates visual analogical reasoning in large multimodal models (LMMs) compared to human adults and children. A "visual analogy" is an abstract rule inferred from one image and applied to another. While benchmarks exist for testing visual reasoning in LMMs, they require advanced skills and omit basic visual analogies that even young children can make. Inspired by developmental psychology, we propose a new benchmark of 1,400 visual transformations of everyday objects to test LMMs on visual analogical reasoning and compare them to children and adults. We structure the evaluation into three stages: identifying *what* changed (e.g., color, number, etc.), *how* it changed (e.g., added one object), and *applying the rule* to new scenarios. Our findings show that while models like GPT-4V, LLaVA-1.5, and MANTIS identify the "what" effectively, they struggle with quantifying the "how" and extrapolating this rule to new objects. In contrast, children and adults exhibit much stronger analogical reasoning at all three stages. Additionally, the strongest tested model, GPT-4V, performs better in tasks involving simple surface-level visual attributes like color and size, correlating with quicker human adult response times. Conversely, more complex tasks such as number, rotation, and reflection, which necessitate extensive cognitive processing and understanding of extrinsic spatial properties in the physical world, present more significant challenges. Altogether, these findings highlight the limitations of training models on data that primarily consists of 2D images and text. [1]

1 INTRODUCTION

What is visual cognition? Humans make countless visual inferences everyday from observing objects and scenes, quickly detecting even subtle visual changes. We generalize common patterns about changes from different observations and use these insights to solve new problems. If we put a wool sweater in the washing machine and it comes out smaller, we might infer that the wash shrinks wool and avoid washing wool coat in the future. If cookies disappear, we might infer that someone is eating our treats and and proceed to hide the chocolate elsewhere. This ability to draw parallels between situations and apply learned patterns to a new scenario is known as *analogical reasoning*. Formally defined, an analogy is a systematic comparison between structures that uses the properties and relations of objects in a source structure to infer properties and relations of objects in a target structure (Mitchell, 2021; Schunn & Dunbar, 1996). Analogical reasoning is a hallmark of human intelligence and learning (Gentner, 1983; Holyoak, 2012; Mitchell, 2021; Sternberg, 1977). It is what enables us to be flexible, adaptive and robust learners across a wide variety of settings, finding meaning in patterns and making out-of-distribution generalizations (Chollet, 2019; Mitchell, 2021). Analogical reasoning is already available to young children (Goddu et al., 2020; Goswami, 2013; Sternberg & Rifkin, 1979), and is crucial for human problem-solving in various contexts, from building scientific models to appreciating metaphors to formulating legal arguments.

Today, large multimodal (LMMs) have made significant progress, but they remain data-hungry and require substantial human effort to adapt to new contexts (Chollet, 2019; Reizinger et al., 2024). As analogical reasoning is instrumental for general-purpose and adaptive machines, it is crucial to examine whether current models have such capabilities. Critically, examining analogical capabilities does not permit models to "cheat" by merely depending on their training data because it requires

[1] Benchmark (code, data, models) is available at: https://anonymous.4open.science/r/KiVA-5CCF



(a) Visual analogy domains.

(b) Extrapolation accuracy.

Figure 1: **KiVA: Kid-inspired Visual Analogies.** **(a)** 5 visual analogy domains examined in KiVA. The top rows show example transformations, while the bottom rows prompt subjects to extrapolate the identified transformations to a novel object by analogy (see Figure 3 for the full task format). **(b)** Performance of children, adults & LMMs in extrapolating a transformation rule to a novel object.

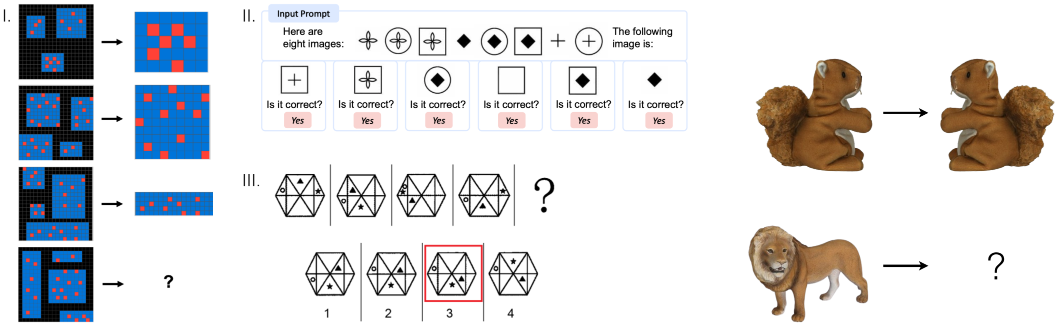
context-dependent abstraction beyond general object recognition. In KiVA, the same object may undergo different kinds of transformations, requiring models to combine familiar elements in new, trial-specific ways. Reasoning about analogies involves first classifying *relationships* between object characteristics, specifying similarities and differences, then extrapolating the *same relationship* to new objects. This paper focuses on visual analogies, testing models' ability to reason abstractly about visual observations. See Figure 1 for a summary of the KiVA benchmark and results.

There is a growing body of work examining visual analogical reasoning capabilities in large multimodal models (Ahrabian et al., 2024; Huang et al., 2024; Moskvichev et al., 2023; Petersen & van der Plas, 2023; Webb et al., 2023). Existing benchmarks of visual analogies include (a) ConceptARC (Mitchell et al., 2023; Moskvichev et al., 2023), (b) variations of Raven's Progressive Matrices (Huang et al., 2024) and (c) abstract spatial reasoning (Ahrabian et al., 2024) (see prior benchmarks in Figure 2). These prior benchmarks all have several critical limitations. First, they rely on abstract shapes and grids, lacking real-world relevance. This abstraction of stimuli neither aligns with the training data of large multimodal models nor effectively mimics the complexity and variability found in everyday visual tasks, making it less suitable for assessing how well AI models can perform analogical reasoning in practical contexts. Second, the transformations examined involve conjunctions of visual concepts such as extracting *and* transposing pixels according to some arbitrary rule, which does not tap into basic visual cognition. Humans do not require the ability to solve these specific tasks to function effectively in their daily lives nor to demonstrate their capacity for visual analogical reasoning. Third, while we know that models often perform poorly on these benchmarks, where they fail in the reasoning process needs to be clarified since existing evaluations focus solely on prediction accuracy rather than the reasoning approach or what is perceived.

We propose a Kid-inspired Visual Analogies (KiVA) benchmark founded on developmental psychology (Figure 1 (left)) (Goddu et al., 2020; Lehmann et al., 2014). We focus our analysis on basic visual analogical capabilities that are present early in human development and are important for understanding the physical world. KiVA isolates the following fundamental capabilities that emerge early in human development: detecting changes in **color** (Ross-sheehy et al., 2003; Wang & Goldman, 2016) and **size** (Day & McKenzie, 1981; Wang & Goldman, 2016), changes that involve **rotation** and **reflection** (Frick et al., 2013; Quaiser-Pohl, 2003), and changes in small **numbers** of objects (Cherian et al., 2023; Levine et al., 1992). KiVA stands out in the following ways:

First, our dataset utilizes *real-world*, *physically grounded* objects curated from established 3D datasets of common household items (Downs et al., 2022) and toys that are familiar to human children (Stojanov et al., 2021), which align more with the training distribution of computer vision models and visual data of humans more than other visual analogical reasoning datasets (Figure 2).

Second, our approach is inspired by *developmental psychology*, specifically how children learn to perform analogical reasoning not abstractly, but from simple objects in grounded contexts (Christie



(a) Prior benchmarks.

(b) Our benchmark.

Figure 2: **Prior benchmarks versus KiVA for visual analogies.** **(a)** Prior benchmarks such as **I.** ConceptARC (Moskvichev et al., 2023), **II.** Raven's Progressive Matrices (Huang et al., 2024), and **III.** CCSE Reasoning (Ahrabian et al., 2024) involve arbitrary changes of abstract shapes and grids. **(b)** By contrast, KiVA examines basic visual transformations that even three-year-olds can solve.

& Gentner, 2010; Gentner, 1983; Goddu et al., 2020). We propose a similar approach for large multimodal models, investigating if they can perform like children on basic visual analogical reasoning tasks related to color, size, orientation, and number – as already reported in child development journals Coates et al. (2023); Goddu et al. (2020; 2025). Starting with simple, real-world relevant tasks in child development allows models to develop robust reasoning abilities before tackling more advanced tasks, providing a clearer pathway for evaluating and improving cognitive functions in AI.

Third, we break down our evaluation to examine the *different steps* involved in analogical reasoning to determine which steps a model can perform and where it may fail: *1)* classifying the domain of a visual transformation, *2)* specifying the transformation rule, and *3)* extrapolating the inferred rule to a new item. This three-stage evaluation (Figure 3) gives us insights into models' reasoning processes beyond simply selecting a correct or incorrect response at the end.

Results from KiVA demonstrate that state-of-the-art large multimodal models, i.e., GPT-4V OpenAI (2023a), LLaVA-1.5 (Liu et al., 2024) and MANTIS (Jiang et al., 2024a), cannot solve visual analogies like humans can. These models do not match even the capabilities of a three-year-old child (Figure 1b). While LMMs can detect some object transformations, they cannot make extrapolations about those transformations to new objects. In particular, GPT-4V outperforms LLaVA-1.5 and MANTIS but also demonstrates weaker performance in orientation and number changes than in size and color changes which are processed more quickly by humans, at an earlier age (Slater et al., 1990; Wang & Goldman, 2016), and in a more primary region of the visual cortex (Zeki et al., 1991; Zeng et al., 2020).

Finally, we include two more challenging versions of the benchmark. We present *KiVA-adults* (dataset and results described in Appendix E), which includes 2,900 transformations and requires deeper generalization from training examples. These transformations are solvable by adults but not by children under seven years old, providing the next benchmark for models to surpass after KiVA. We also release in our project page code for *KiVA-compositionality*, which combines multiple objects and transformations together to probe even more complex compositional reasoning. Taken together, KiVA not only mirrors the natural progression of human cognitive development, but also provides a more structured and comprehensive framework for evaluating the capabilities and growth of LMMs.

2 Related Work

Evaluating human visual analogical reasoning. There is a variety of tasks designed in Developmental Psychology to examine human visual analogical reasoning early on in life. Children are asked to compare simple object and relational matches (Christie & Gentner, 2010; Goddu et al., 2020; Kuwabara & Smith, 2012) along dimensions such as color (Milewski & Siqueland, 1975; Ross-sheehy et al., 2003), number (Cherian et al., 2023; Levine et al., 1992), size (Day & McKenzie, 1981; Slater et al., 1990) and spatial orientation (Frick et al., 2013; Quaiser-Pohl, 2003). Older children and adults are evaluated on Raven's Progressive Matrices (RPMs) (Carpenter et al., 1990; Lovett & Forbus, 2017; Raven & Court, 1938) and Bongard Problems (Bongard, 1970; Weitnauer et al., 2023). Even though they tend to be the most representative and largest testbeds for testing advanced visual analogical reasoning, RPMs and Bongard problems use abstract geometric shapes and test recognition of arbitrary patterns that (1) cannot be solved by children before the age of 6 and (2) are not critical to everyday visual processing. KiVA is the first visual analogical reasoning benchmark that includes common real-world objects and more natural visual cognition skills such as counting and spatial transformations — tasks that even a three-year-old child can handle (Goddu et al., 2020). We also examine where people and models fail with more fine-grained evaluation.

Evaluating visuo-linguistic reasoning in AI models. Several proposals for evaluating modern AI systems' visuo-linguistic reasoning capabilities followed the recent successes of large multimodal models. Many concentrate on a narrow, isolated set of tasks for detecting object properties like size estimation (Chen et al., 2024; Liu et al., 2022), color perception (Abdou et al., 2021; Samin et al., 2024), counting objects (Liang et al., 2023; Paiss et al., 2023), object viewpoint/pose and chirality (Kapelyukh et al., 2023; Lin et al., 2020; Chen et al., 2024) and visuo-linguistic compositionality (Thrush et al., 2022; Kamath et al., 2023; Liu et al., 2023). Typically, the objective of these tasks is to evaluate models' ability to report a correct property about objects in an image. They lack the depth to probe pattern abstraction and generalization involved in visual analogical reasoning.

Broader benchmarks, such as visual question answering setups (Antol et al., 2015; Goyal et al., 2017), attempt to investigate the models' understanding of various visual concepts. One approach taken by (Bubeck et al., 2023; Yang et al., 2023) was to try and push the envelope on various tasks to capture anecdotal and qualitative observations regarding the performance of GPT-4. Perception Test (Pătrăucean et al., 2023) proposed a second approach: a visual video-based benchmark including developmentally-inspired tasks such as object permanence, object tracking, spatial relations, etc. Recently, the BLINK benchmark was introduced to show that core visual perception tasks, easily solvable by humans "within a blink," remain challenging for large multimodal models due to their resistance to language-based mediation (Fu et al., 2024). However, all these benchmarks fall short in evaluating the deeper, more complex aspects of visual analogical reasoning and generalization.

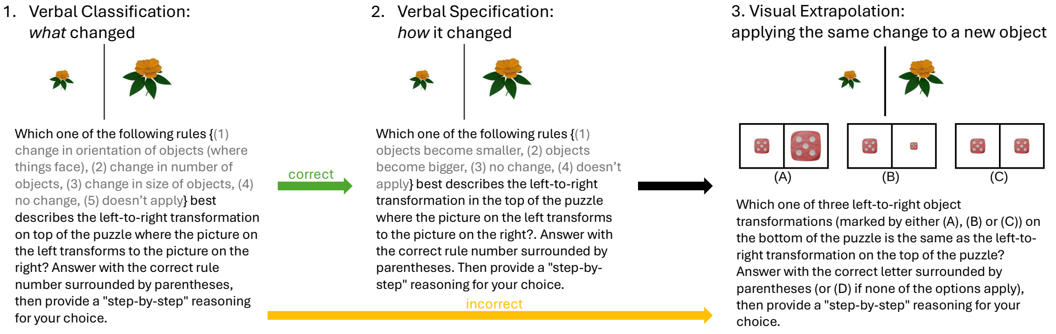
Another specific class of benchmarks tests generalization and reasoning within abstract puzzle grids. These include Abstraction and Reasoning Corpus (ARC) (Chollet, 2019; Moskvichev et al., 2023; Mitchell et al., 2023); a direct translation of RPMs-based human evaluation has previously been applied to models by (Ahrabian et al., 2024) and (Huang et al., 2024) (also see prior benchmarks (b) and (c) in Figure 2). However, the stimuli are simple, monotonic shapes like squares and circles, lacking real-world complexity and variability. Moreover, they primarily focus on convoluted pattern recognition and logical sequencing without incorporating real-world contexts, thereby neglecting more fundamental visual cognition domains that even human children are capable of.

3 The KiVA Benchmark for Visual Analogical Reasoning

We introduce KiVA, a Kid-inspired Visual Analogies benchmark, wherein real-world objects undergo common transformations necessary for everyday visual cognition. We focus on isolating and testing basic visual transformations that even a three-year-old child understands Goddu et al. (2020). As we show in Figure 1, we examine noticing **color changes** Ross-sheehy et al. (2003); Milewski & Siqueland (1975), **size changes** Day & McKenzie (1981); Slater et al. (1990), **rotation**, **reflection** Quaiser-Pohl (2003); Frick et al. (2013), and **number changes** such as addition and subtraction of a small number of objects Cherian et al. (2023); Levine et al. (1992).

3.1 A Three-Stage Experimental Paradigm

We use our proposed dataset to benchmark computational models' and human subjects' visual analogical reasoning capabilities. We utilize the same testing procedure (Figure 3) for both kinds of subjects. In each trial, we start by presenting a given transformation of an object that changes by a specific rule, following the experimental paradigm of other analogical reasoning benchmarks for humans and computational models (Moskvichev et al., 2023; Bongard, 1970; Goddu et al., 2020). Inspired by the component processes model of analogical reasoning (Sternberg, 1977), we evaluate the subject's ability to determine *what* changed (*Verbal Classification*) *how* it changed (*Verbal Specification*), and apply the the same transformation rule to predict the outcome of a new object—i.e., a *Visual Extrapolation*. We break the question down into these three steps to test the different cognitive processes involved in analogical reasoning. The first two assess the necessary prerequisites



Which one of the following rules {(1) change in orientation of objects (where things face), (2) change in number of objects, (3) change in size of objects, (4) no change, (5) doesn't apply} best describes the left-to-right transformation on top of the puzzle where the picture on the left transforms to the picture on the right? Answer with the correct rule number surrounded by parentheses, then provide a "step-by-step" reasoning for your choice.

Which one of the following rules {(1) objects become smaller, (2) objects become bigger, (3) no change, (4) doesn't apply} best describes the left-to-right transformation in the top of the puzzle where the picture on the left transforms to the picture on the right?. Answer with the correct rule number surrounded by parentheses. Then provide a "step-by-step" reasoning for your choice.

Which one of three left-to-right object transformations (marked by either (A), (B) or (C)) on the bottom of the puzzle is the same as the left-to-right transformation on the top of the puzzle? Answer with the correct letter surrounded by parentheses (or (D) if none of the options apply), then provide a "step-by-step" reasoning for your choice.

Figure 3: **An example of a trial in KiVA.** Models and humans are first asked to classify a given transformation (left). If the classification is correct (green arrow), humans and models are further evaluated on their verbal specification of the transformation (middle) and then on visual extrapolation (right). Otherwise, humans and models skip to make a visual extrapolation (yellow arrow).

for accurate analogical reasoning, while the last step represents the core visual analogy task. Critically, KiVA retains the core nonverbal extrapolation task (last step) from previous benchmarks and the verbal questions *do not replace* the core nonverbal tasks. Even without correct verbal responses, humans and models can still tackle the independently-assessed visual extrapolation tasks. Thus, KiVA doesn't require specific language skills but provides a window into the analogical reasoning process of humans and models in reaching their final solutions. The first two verbal questions were further paraphrased by developmental psychologists so that it is comprehensible to a three-year-old child (Appendix A.3); models and adults did not benefit from the child-appropriate prompting so the original prompt in Figure 3 was preserved. We pose all questions in a multiple-choice format for human children, adults and models, which enables automatic scoring. Option labels for correct responses were randomized such that LMMs' option label bias does not correlate with task accuracy. Furthermore, we provided the opportunity to select "Doesn't apply" to accommodate responses that the provided choices may not cover. Excluding the "Doesn't apply" option, chance level is 25% for Verbal Classification (4 possible choices were provided) and 33% for Verbal Specification and Visual Extrapolation (3 possible choices were provided). Please refer to Figure 3 for the three-stage query pipeline and Appendix A.2 for more details of the specific prompts.

Verbal classification of transformation ("what"). We first evaluate if the model or human can detect what changed in a given transformation and classify it in the correct visual domain, such as size or number (see Figure 3). We randomly sample incorrect multiple-choice options from other possible transformation domains. "No change" and "Doesn't apply" are always included as options to accommodate for alternative forms ofreasoning that are not covered by the choices. Suppose the model fails to identify basic changes, such as distinguishing a numerical change from a color change. It will be unable to predict how new objects change based on the given transformations. This is an inadequacy of existing visual analogical reasoning benchmarks (Moskvichev et al., 2023; Mitchell et al., 2023; Ahrabian et al., 2024; Huang et al., 2024), which focus solely on advanced predictions without ensuring fundamental change detection capabilities.

Verbal specification of transformation ("how"). If a subject correctly classifies the transformation, we ask them to further specify also in the form of multiple-choice the transformation (see green arrow in 3). This step is crucial because it ensures the subject can accurately specify the rule governing the transformation before extrapolating it to a new object. If they fail to identify the specific change, any attempt at extrapolation would more likely be incorrect (see Table 4 in Appendix B.3 for evidence in models). By pinpointing where reasoning fails, we can better understand models' and humans' limitations and improve their analogical reasoning capabilities.

Visual Extrapolation of transformation. Finally, we proceed to the step captured by other benchmarks: presenting a new image and asking the model to extrapolate how it will change based on the previously identified transformation (see Figure 3 and other extrapolation examples of other visual domains in Appendix A.1). We ask models to visually extrapolate independent of their performance in verbal change identification to account for the possibility that models may engage in visual analogical reasoning separately from verbal reasoning and can, therefore, perform well in visual tasks even if they struggle with the prior verbal descriptions. This approach helps us determine if a model's visual reasoning can function independently of its verbal reasoning skills. It provides a more nuanced evaluation of its cognitive capabilities and identifies specific areas for improvement.

3.2 A Dataset of Visual Analogies

We create a dataset of stimuli using everyday objects that better represent real-world visual data and better match the training data of computer vision models (and humans). We take 3D models of household objects from Downs et al. (2022) and objects commonly encountered by infants and children from Stojanov et al. (2021). Each object in the dataset is handpicked by developmental psychology experts to ensure that they are child-friendly. To set up the dataset, we perform five basic visual transformation domains: changing the size, color, and number of objects, rotating and reflecting the objects along different axes (see Figure 1 for the transformation domains examined). Our benchmark includes code allowing users to perform these transformations on any object image, enabling infinite expansion of the benchmark. We select these five types of object transformations because they are crucial for object and scene recognition, (e.g., Diwadkar & McNamara (1997); Gevers & Smeulders (1999)), scene segmentation (e.g., Chattopadhyay et al. (2017)), and detecting significant changes in the environment (Hatfield & Allred, 2012; Duh & Wang, 2014). Other visual properties, such as depth (Chen et al., 2016), spatial compositionality (Jiang et al., 2022; Thrush et al., 2022), and physical affordances (Jiang et al., 2023; Sawatzky et al., 2019) are also crucial for such purposes; however, we prioritized these five transformations for our benchmark in particular because young children can solve these visual analogies, as already shown in developmental psychology literature (Goddu et al., 2020; Harris et al., 2013). Below, we outline the five visual transformation domains, each consisting of subdomains that specify an object transformation. There are 100 object transformations for each subdomain of transformation, totaling 1,400 object transformations.

Color changes. Noticing color changes can signal alterations in an object's state or presence, which is essential for tasks like identifying ripe fruit or detecting hazards (Maule et al., 2023). In KiVA, the general transformation rule for color is that input objects change to a single color as in (Goddu et al., 2020). The subdomains of color examined are red, green, and blue.

Size changes. Size perception allows individuals to understand and interact with their environment accurately, guiding tasks like identifying objects, planning actions, navigating spaces, and avoiding obstacles (Giudice, 2018). In KiVA, objects undergo transformations in two subdomains: they turn bigger or smaller (in both height and width) as in (Goddu et al., 2020) by a factor of 2.

Rotation. Mental rotation is the ability to recognize and map different views of the same object (Shepard & Metzler, 1971). This is essential for identifying objects despite changes in perspective, which is vital for navigation, object manipulation, and spatial orientation (Pinto et al., 2008). KiVA takes inspiration from psychometric studies probing mental rotation in humans (e.g., (Bodner & Guay, 1997; Quaiser-Pohl, 2003)), featuring object rotation in 2D space by the subdomains of 90 degrees (clockwise or counterclockwise) or 180 degrees.

Reflection. Reflection aids in understanding object symmetry and chirality, essential for distinguishing left and right shoes or gloves, for example (Holmes et al., 2018). Chiral objects cannot be rotated and translated into alignment with their reflection, so there is a visual difference between the object and its reflection (Lin et al., 2020). We reflect chiral objects in two subdomains: along the x-axis or the y-axis as in (Goddu et al., 2020).

Number changes. Accurately monitoring and comparing quantities is essential in various fields like economics and science; it is also important in daily life activities like shopping, cooking, and managing resources effectively, as in caching and rationing (Chattopadhyay et al., 2017; Cohen, 2005). Transformations in this domain reflect basic mathematical operations over the number of objects in an image. Subdomains of number transformations include addition $(+1, +2)$ and subtraction $(-1, -2)$. We restrict the number of objects in an input or output image to under 8.

4 Comparing Analogical Reasoning in LMMs and Humans

Evaluating Large Multimodal Models. We test several LMMs: 1) GPT4-V (OpenAI API model: gpt-4-vision-preview) (OpenAI, 2023b): an extension of the language-only GPT-4 (OpenAI, 2023a) incorporating computer vision capabilities, 2) LLaVA-1.5 (Liu et al., 2024): an open-source model that integrates a vision encoder with a language model, specifically designed to enhance general-purpose visual and language understanding, 3) MANTIS (Jiang et al., 2024a) which builds on modified architectures from notable models like LLaVA to support interleaved multi-image input. We combine the given transformation with the choices of new object transformations at the extrapolation step into a single composite input image for LLaVA-1.5, which is limited to processing a single image, but present these as separate images to MANTIS, which is fine-tuned to manage interleaved multi-image conversations. We evaluate GPT-4V under both multi-image and single-image presentations. For all models, the temperature is set to 1 and the maximum token size is set to 300. We randomize each experiment over three seeds and run each trial (Figure 3) on a model three times. We score correct choices as 1 and incorrect choices as 0. We calculate the mean score across its three seeds. Subsequently, to evaluate the performance for each transformation domain, we calculate the overall mean and standard error for the average scores of all averaged trials within each domain. GPT-4V, LLaVA-1.5, and MANTIS complete the entire benchmark, featuring 1,400 transformations. Open-source models ran on an A6000 48 GB single GPU for under 12 hours.

Evaluating Humans. A corresponding visual analogies task, developed using JsPsych (De Leeuw, 2015), was administered to two groups of human participants. All methods were approved by IRB (protocol 2020-10-13755) prior to testing both child and adult participants. We recruited 250 adults (21 to 40 years old) on Prolific (Prolific) to complete 14 trials randomly sampled from the benchmark such that every trial was annotated by 3-13 adults. We recruited 50 children (aged 3 to 7 years, $mean$ = 4.43 years, se = 0.16 years) from local early childhood centers and ChildrenHelpingScience (Science) to complete a random subset of 10 trials (we randomly sampled 2 trials for each of the 5 transformation domains), leading to a total of 400 responses. Participants completed a practice trial with an "unrelated" transformation (adding a dot to geometric shapes) and received feedback to ensure understanding. Participants who failed within three attempts were excluded. Those who succeeded proceeded to test trials without feedback, and were told that rewards depended on their performance. Adults were paid at least $12/hour with a bonus of $0.01 per correct response, while children received stickers based on their performance. For every object set that was randomly sampled and tested in children, children scored at least an average of 60.5% at visual extrapolation; for every object set in the full benchmark, at least 2 out of 3 adults selected the correct response at visual extrapolation, confirming that the benchmark is fully solvable by humans.

Mirroring model evaluation, we calculate average scores and standard errors across trials per domain. The standard error reflects the variation in performance across trials completed by all participants.

5 Results

Models get worse with increasing reasoning complexity, unlike humans. Overall, LMMs can detect transformations and identify the general visual domain of the transformations (e.g., color vs. size), as indicated by the blue bars labeled "Verbal Classification" in Figure 4 and Figure 5, with GPT-4V and MANTIS generally outperforming LLaVA-1.5. GPT-4V and MANTIS outperform human children in recognizing when an object changes color or spatial orientation (rotation and reflection). That said, all models, unlike humans, hallucinate transformations in trials that actually involve no change (Appendix B.2). Furthermore, performance generally declines when the models are asked to further specify the transformation within the correctly identified visual domain (e.g., becoming bigger or smaller if size is the correctly identified domain), as reflected by the orange bars labeled "Verbal Specification". Performance for visual extrapolation declines even more, as

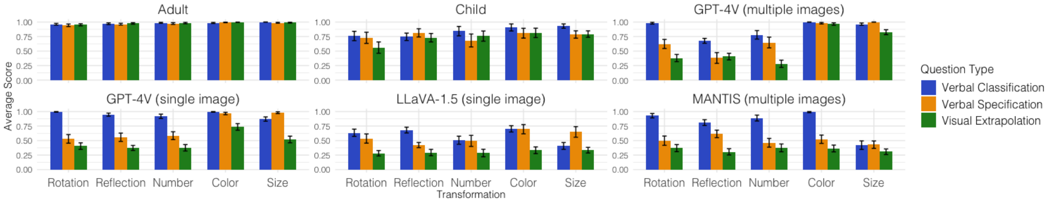


Figure 4: **Human and model performance in the benchmark samples annotated by children, sorted by Transformation and Question Type.** Error bars represent standard errors of performance across object variations within the same transformation. Chance level is 25% for Verbal Classification (blue) and 33% for Verbal Specification (orange) and Visual Extrapolation (green).

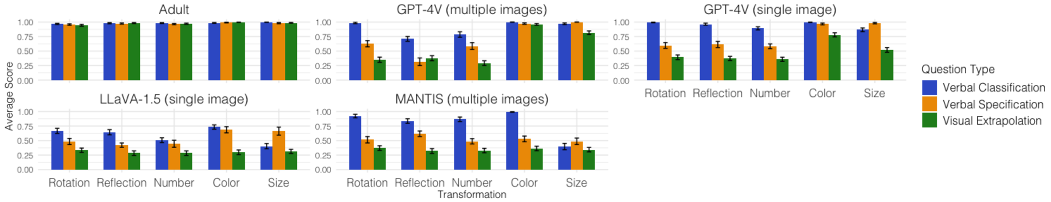


Figure 5: **Human adult and model performance on the full benchmark.** A similar pattern is observed in human adult and model performance as in Figure 4.

illustrated by the green bars labeled “Visual Extrapolation”. In other words, models’ success in verbally describing transformations does not guarantee their success in extrapolation. We may attribute part of the models’ failure in analogical reasoning to an inability to correctly recognize the given transformation. However, another part of the model’s failure lies in extrapolating the correctly identified transformation to a novel object and predicting the corresponding outcome. Even when given the correct verbal specification of the transformation, models still fail to solve extrapolation in different visual domains (Appendix B.4). By contrast, children and adults show robust performance from verbal classification to verbal specification to visual extrapolation, as shown in Figure 4. Even young children can verbally describe the transformations as reflected by their significantly-above-chance performance in verbal classification and verbal specification, and can then use their selected verbal descriptions to extrapolate the visual transformations to new objects. Specific numerical results can be found in Appendix B.1, with similar pattern of results for KiVA-adults (Appendix E.2).

Model performance depends on the visual domain and correlates with human response times. Overall, models are better at detecting and describing color and size transformations than transformations in other domains. GPT-4V performs better at detecting and extrapolating color and size transformations, which involve more discrete and local processing than the other domains (Zeki et al., 1991; Zeng et al., 2020). However, LLaVA-1.5 and MANTIS do not perform well on generalizing size transformations relative to other types of transformations, even when all the possible relative sizes of objects are presented to LLaVA in a single image. Overall, all three models are less able to tell what changed within the visual domains of rotation, reflection, and number, as demonstrated by the orange bars labeled “Verbal Specification” in Figure 4 and Figure 5, and consequently also did not perform well in extrapolations for those domains. In contrast, children and adults generally perform well across all visual domains in Figure 4, with children performing slightly worse on rotation, suggesting greater difficulty in appreciation of spatial orientation compared to other domains. While human adults perform about equally well in all visual domains, their response times correlate with GPT-4V’s error scores (1-Accuracy) across the domains, as demonstrated in Figure 6. What is cognitively demanding to humans is perhaps also more computationally challenging for GPT-4V.

Models are inconsistent in their responses within trials and across reasoning steps. We measured model response consistency (1) within repeated trials (Figure 7) and (2) from verbal classification to visual extrapolation (Figure 8). Models demonstrate strongest response consistency in Verbal Classification and least consistent responses in Visual Extrapolation. Only GPT-4 (multiple images)

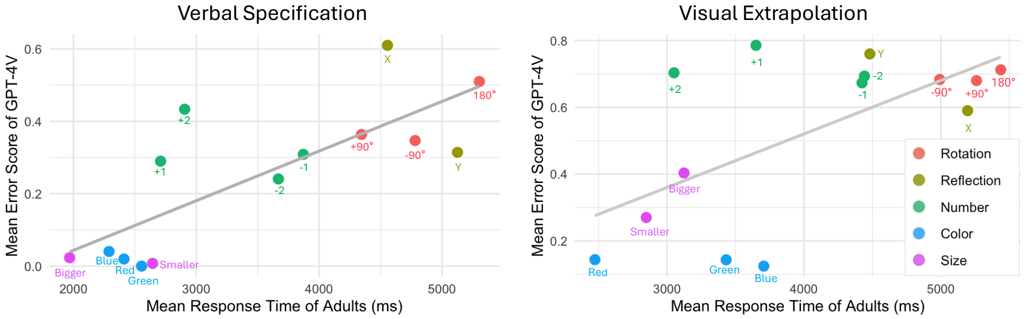


Figure 6: **Mean error scores of GPT-4V (single image and multiple images) positively correlate with mean response times of adults in verbal specification (left) and visual extrapolation (right).** Each dot is labeled by the specific transformation). In both specification and extrapolation, the longer adults took to respond correctly, the more errors GPT-4V made, r = .78, p = .0011.

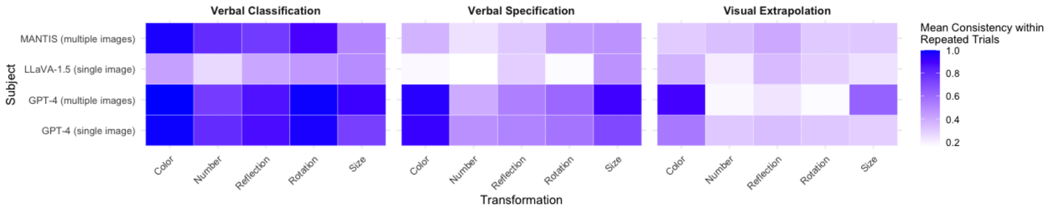


Figure 7: **Proportion of model consistent responses within repeated trials.** Each model was queried three times on the same trial and was scored as consistent if it selected the same choice across the repeated trials and inconsistent otherwise. The heat map shows the proportion of consistent trials (out of total number of trials), broken down by model, transformation domain, and question type.

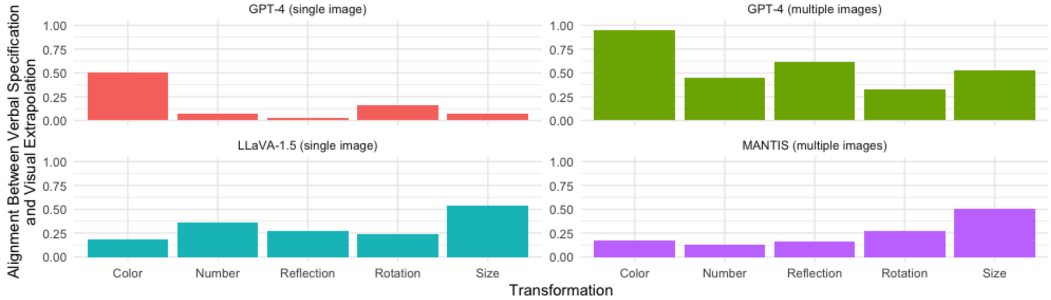


Figure 8: **Proportion of Aligned Model Responses from Verbal Specification to Visual Extrapolation.** A trial was scored as aligned if the chosen visual extrapolation matched the chosen verbal specification and unaligned otherwise. The barplot shows the proportion of aligned trials (out of the total number of trials), broken down by model and transformation domain.

in the color domain shows strong consistency throughout the three-step query. Furthermore, models exhibit a disconnect between their verbally specifying a transformation rule and visually extrapolating to a new object. While GPT-4V shows the strongest alignment between verbal and visual reasoning for color transformations, LLaVA-1.5 and MANTIS perform best with size transformations, aligning only half the time. This underscores a key limitation in the visual analogical reasoning of LMMs: the ability to articulate a rule does not reliably translate to applying that rule in a new visual context.

Verbal questions facilitate visual extrapolation in humans and GPT-4V. We tested another 200 adults, 20 children and the best-performing model, GPT-4V (single image and multiple images), on a visual-extrapolation-only task, removing the verbal questions to replicate previous visual analogy benchmarks. Without verbal questions, adults demonstrated similar accuracy but significantly slower response times, children performed significantly worse in extrapolation within all domains but rotation, while GPT-4V performed worse in extrapolating color and resize transformations — its strongest domains in the original setup—without the verbal questions (it was at chance level for the rest of the domains) (Figure 15 in Appendix C). This indicates that our three-step query pipeline with verbal questions facilitates humans and models to subsequently perform visual extrapolations.

In-context learning and prompt engineering did not improve model performance. We explore whether model performance improves through careful prompt engineering (Appendix A), which has shown promising results on various tasks (Wei et al., 2022; Qin & Eisner, 2021). We consider four different prompt engineering methods: *1) Reasoning through code* (Sharma et al., 2024): We first prompt the model to generate code snippets describing each transformation in the task, then rephrase the task question to incorporate the generated code. *2) Reasoning after Reflection* (Valmeekam et al., 2023): We ask the model to reflect on its answers two times for each question in the task. *3) Reasoning through instruction*: inspired by Wei et al. (2022), which shows that chain-of-thought reasoning is more effective on several benchmarks, we prompt the model to generate step-by-step instructions on how to answer each question, then use the instructions to generate an answer. *4) In-Context Learning* (Dong et al., 2022): We give the model two randomly sampled examples with solutions for each concept before displaying the task. Apart from text prompt engineering, we experiment with different visual prompting for LLaVA-1.5. Recent works (Bai et al., 2023; Bar et al., 2022; Wang et al., 2023) show that visual model performance is sensitive to the alterations in color and size of the visual input. We apply two visual prompting approaches: *1) Color*: we alter the image background color (initially transparent) into black and white (Bai et al., 2023). *2) Size*: we apply a center crop to the images, varying the image size between 0.9 and 1. None of these approaches improve performance, which points to the challenging nature of our benchmark.

6 Discussion

Despite the scale of image and text data used to train GPT-4V, LLaVA-1.5, and MANTIS, these models fail to reason about visual analogies like young children can. GPT-4V outperforms the other models but still does not approach child performance in rotation, reflection and number. It performs better in transformations related to color and size than those related to number and spatial orientation, but like other models its performance substantial declines with verbal specification leading up to visual extrapolation, whereas humans do not show much decline from generalizing changes to making extrapolations about new objects. Even when models successfully infer (Appendix B.3) or are given (Appendix B.4) the correct verbal specification, they can still fail visual extrapolation. This highlights a deeper issue with visual analogical reasoning that goes beyond recognizing what transformation it is, but more about specifically mapping the transformation from the source object to the visual features of the target object while preserving the relational structure (Gentner, 1983).

Unlike simply observing discrete feature-level changes such as color or size, appreciating reflection, rotation, and numerical changes tend to occur when actively engaging with the environment: determining number changes involves sequentially keeping track of numbers, whereas reflection and rotation necessitate visualizing and mentally manipulating objects in space. Our domain-dependent findings also align with prior research showing that large multimodal models (LMMs) often struggle with spatial reasoning (e.g., poor performance in rotation and reflection tasks) (Wang et al., 2024; Rahmanzadehgervi et al., 2024) and counting (consistent with our results in number change tasks) (Jiang et al., 2024b; Rahmanzadehgervi et al., 2024)]. These tasks are not only more cognitively complex but are also not solvable merely through 2D image-text correlations, which are often used in training LMMs. Furthermore, compared to spatial and numerical transformations, changes in size and color tend to be processed much earlier in the visual pathway of the brain (Zeki et al., 1991; Zeng et al., 2020) and in human development (Day & McKenzie, 1981; Milewski & Siqueland, 1975; Ross-sheehy et al., 2003; Slater et al., 1990). Adults take longer to process visual analogies related to changes in number and orientation, suggesting that it may be more cognitively challenging.

Overall, KiVA is designed to assess basic visual change detection and analogical reasoning capabilities commonly studied in children by psychologists (Goddu et al., 2020). It aligns with the documented capabilities of young children, as evidenced by the successful completion of our task by individuals as young as three years old. We observe that LMMs perform worse than human participants without marked improvement through in-context learning and visual or textual prompting techniques. Further research should explore practical strategies for improving model performance, potentially through symbolic visual vocabulary and Bayesian inference (Depeweg et al., 2024). As models develop and improve in performance, we provide KiVA-adults and KiVA-compositionality beyond KiVA to set up a curriculum for probing more sophisticated analogical reasoning capabilities. Our benchmark does not capture the full spectrum of visual cognition, but it represents a foundational effort to systematically evaluate visual analogical reasoning fundamentally and grounded using everyday real-world objects and established principles from developmental and cognitive psychology.

7 Conclusion

Our findings suggest that large pretrained multimodal models are still less capable than humans in visual analogical reasoning. While these models can classify changes in images, their performance diminishes when tasked with specifying changes and further declines when extrapolating these identified changes to a new, unseen object. Among the three models evaluated, GPT-4V performs the best, particularly in tasks involving changes in color and size - surface-level features that do not alter the object's properties when viewed from different angles and spatial configurations. However, it struggles with more challenging analogies of spatial and quantitative properties that perhaps require a greater understanding of the 3D physical world. These rely less on simple image-text correlations. In contrast, humans, including young children, can recognize and interpret a wide range of object relations and transformations, as noted in existing studies (Goddu et al., 2020; Mitchell et al., 2023).

REFERENCES

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021. 4

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. The curious case of nonverbal abstract reasoning with multi-modal large language models. *arXiv preprint arXiv:2401.12117*, 2024. 2, 3, 4, 5

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015. 4

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023. 9, 10

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 9

George M Bodner and Roland B Guay. The purdue visualization of rotations test. *The chemical educator*, 2(4):1–17, 1997. 6

M. M. Bongard. *Pattern Recognition*. Spartan Books, New York, 1970. 3, 4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 4

Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990. 3

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1135–1144, 2017. 6

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. 4

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 6

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10834–10844, 2023. 2, 3, 4

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 1, 4

Stella Christie and Dedre Gentner. Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11(3):356–373, 2010. 2, 3

Nicole Coates, Max Siegel, Josh Tenenbaum, and Laura Schulz. Representations of abstract relations in early childhood. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023. 3

I Bernard Cohen. *The triumph of numbers: How counting shaped modern life*. WW Norton & Company, 2005. 6

RH Day and BE McKenzie. Infant perception of the invariant size of approaching and receding objects. *Developmental Psychology*, 17(5):670, 1981. 2, 3, 4, 10

Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47:1–12, 2015. 7

Stefan Depeweg, Contantin A Rothkopf, and Frank Jäkel. Solving bongard problems with a visual language and pragmatic constraints. *Cognitive Science*, 48(5):e13432, 2024. 10

Vaibhav A Diwadkar and Timothy P McNamara. Viewpoint dependence in scene recognition. *Psychological science*, 8(4):302–307, 1997. 6

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 9

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022. 2, 6

Shinchieh Duh and Su-hua Wang. Infants detect changes in everyday scenes: The role of scene gist. *Cognitive Psychology*, 72:142–161, 2014. 6

Andrea Frick, Melissa A Hansen, and Nora S Newcombe. Development of mental rotation in 3-to 5-year-old children. *Cognitive Development*, 28(4):386–399, 2013. 2, 3, 4

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. *arXiv preprint arXiv:2404.12390*, 2024. 4

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170, 1983. 1, 3, 10

Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3): 453–464, 1999. 6

Nicholas A Giudice. Navigating without vision: Principles of blind spatial cognition. In *Handbook of behavioral and cognitive geography*, pp. 260–288. Edward Elgar Publishing, 2018. 6

Mariel K Goddu, Tania Lombrozo, and Alison Gopnik. Transformations and transfer: Preschool children understand abstract relations and reason analogically in a causal task. *Child development*, 91(6):1898–1915, 2020. 1, 2, 3, 4, 6, 10

Mariel K Goddu, Eunice Yiu, and Alison Gopnik. Causal relational problem solving in toddlers. *Cognition*, 254:105959, 2025. 3

Usha Goswami. *Analogical reasoning in children*. Psychology Press, 2013. 1

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017. 4

Justin Harris, Kathy Hirsh-Pasek, and Nora S Newcombe. Understanding spatial transformations: Similarities and differences between mental rotation and mental folding. *Cognitive processing*, 14: 105–115, 2013. 6

Gary Hatfield and Sarah Allred. *Visual experience: sensation, cognition, and constancy*. Oxford University Press, 2012. 6

Corinne A Holmes, Nora S Newcombe, and Thomas F Shipley. Move to learn: Integrating spatial information from multiple viewpoints. *Cognition*, 178:7–25, 2018. 6

Keith J Holyoak. Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, pp. 234–259, 2012. 1

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 5

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024a. 3, 7

Guangyuan Jiang, Chuyue Tang, Yuyang Li, and Yu Liu. Bongard-tool: Tool concept induction from few-shot visual exemplars. In *PKU 22Fall Course: Cognitive Reasoning*, 2023. 6

Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19056–19065, 2022. 6

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024b. 10

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 4

Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2023. 4

Megumi Kuwabara and Linda B Smith. Cross-cultural differences in cognitive development: Attention to relations and objects. *Journal of experimental child psychology*, 113(1):20–35, 2012. 3

Jennifer Lehmann, Claudia Quaiser-Pohl, and Petra Jansen. Correlation of motor skill, mental rotation, and working memory in 3-to 6-year-old children. *European Journal of Developmental Psychology*, 11(5):560–573, 2014. 2

Susan Cohen Levine, Nancy C Jordan, and Janellen Huttenlocher. Development of calculation abilities in young children. *Journal of experimental child psychology*, 53(1):72–103, 1992. 2, 3, 4

Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. Crowdclip: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2893–2903, 2023. 4

Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12295–12303, 2020. 4, 6

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 4

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 6

Zixu Liu, Qian Wang, and Fanlin Meng. A benchmark for multi-class object counting and size estimation using deep convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 116:105449, 2022. 4

Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60, 2017. 3

John Maule, Alice E Skelton, and Anna Franklin. The development of color perception and cognition. *Annual Review of Psychology*, 74:87–111, 2023. 6

Allen E Milewski and Einar R Siqueland. Discrimination of color and pattern novelty in one-month human infants. *Journal of Experimental Child Psychology*, 19(1):122–136, 1975. 3, 4, 10

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021. 1

Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*, 2023. 2, 4, 5, 10

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023. 2, 3, 4, 5

OpenAI. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023a. 3, 6

OpenAI. Gpt-4v(ision) technical work and authors. 2023b. https://cdn.openai.com/contributions/gpt-4v.pdf. 6

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023. 4

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2305.13786*, 2023. 4

Molly R Petersen and Lonneke van der Plas. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. *arXiv preprint arXiv:2310.05597*, 2023. 2

Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27, 2008. 6

Prolific. Prolific: Online platform for participant recruitment. https://www.prolific.com. Accessed: 2024-06-05. 7

Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021. 9

Claudia Quaiser-Pohl. The mental cutting test" schnitte" and the picture rotation test-two new measures to assess spatial ability. *International Journal of Testing*, 3(3):219–231, 2003. 2, 3, 4, 6

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024. 10

John C Raven and JH Court. *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA, 1938. 3

Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Understanding llms requires more than statistical generalization. *arXiv preprint arXiv:2405.01964*, 2024. 1

Shannon Ross-sheehy, Lisa M Oakes, and Steven J Luck. The development of visual short-term memory capacity in infants. *Child development*, 74(6):1807–1822, 2003. 2, 3, 4, 10

Ahnaf Mozib Samin, M Firoz Ahmed, and Md Mushtaq Shahriyar Rafee. Colorfoil: Investigating color blindness in large vision and language models. *arXiv preprint arXiv:2405.11685*, 2024. 4

Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should i use?-task driven object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7605–7614, 2019. 6

Christian D Schunn and Kevin Dunbar. Priming, analogy, and awareness in complex reasoning. *Memory & Cognition*, 24(3):271–284, 1996. 1

Children Helping Science. Children helping science: The online directory of research studies for children. https://www.childrenhelpingscience.com. Accessed: 2024-06-05. 7

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. *arXiv preprint arXiv:2401.01862*, 2024. 9

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971. 6

Alan Slater, Anne Mattock, and Elizabeth Brown. Size constancy at birth: Newborn infants' responses to retinal and real size. *Journal of experimental child psychology*, 49(2):314–322, 1990. 3, 4, 10

Robert J Sternberg. Component processes in analogical reasoning. *Psychological review*, 84(4):353, 1977. 1, 4

Robert J Sternberg and Bathsheva Rifkin. The development of analogical reasoning processes. *Journal of experimental child psychology*, 27(2):195–232, 1979. 1

Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1798–1808, 2021. 2, 6

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. 4, 6

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*, 2023. 9

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *arXiv preprint arXiv:2406.14852*, 2024. 10

Su-hua Wang and Elizabeth J Goldman. Infants actively construct and update their representations of physical events: Evidence from change detection by 12-month-olds. *Child Development Research*, 2016, 2016. 2, 3

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023. 9

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023. 2

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 9

Erik Weitnauer, Robert L Goldstone, and Helge Ritter. Perception and simulation during concept learning. *Psychological Review*, 2023. 3

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 4

Semir Zeki, JD Watson, CJ Lueck, Karl J Friston, C Kennard, and RS Frackowiak. A direct demonstration of functional specialization in human visual cortex. *Journal of neuroscience*, 11(3): 641–649, 1991. 3, 8, 10

Hang Zeng, Gereon R Fink, and Ralph Weidner. Visual size processing in early visual cortex follows lateral occipital cortex involvement. *Journal of Neuroscience*, 40(22):4410–4417, 2020. 3, 8, 10

A Visual Analogical Reasoning Prompts

A.1 Stitched visual extrapolation examples for each domain

Visual Extrapolation. As the final step of the querying process, we presented an image of a new object and ask the model to predict what the object will look like if it goes through the same change as the given transformation.

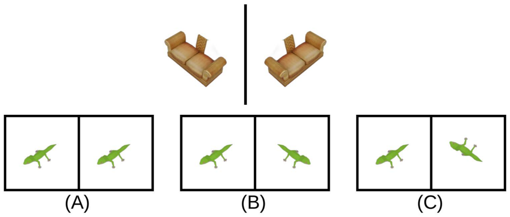


Figure 9: Example of a visual extrapolation trial involving a **reflection**.

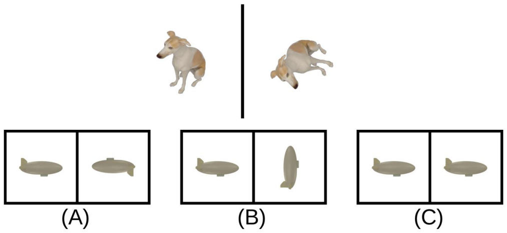


Figure 10: Example of a visual extrapolation trial involving an angular **rotation**.

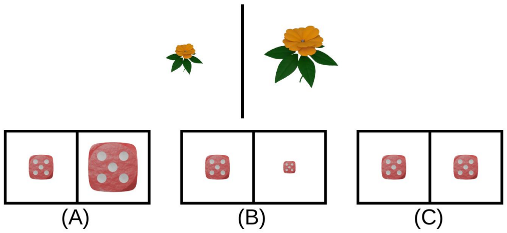


Figure 11: Example of a visual extrapolation trial involving a **size change**.

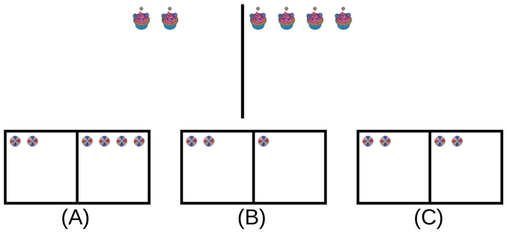


Figure 12: Example of a visual extrapolation trial involving a **number change**.

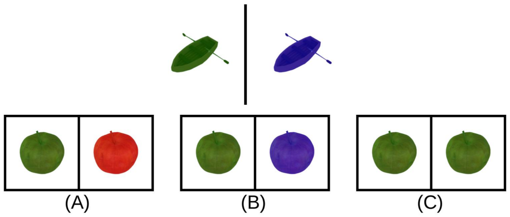


Figure 13: Example of a visual extrapolation trial involving a **color change**.

A.2 Prompting of Models and Human Adults

We first include a system prompt to orient the models for visual analogical reasoning. *You are an excellent visual puzzle solver! You will be given a visual puzzle that requires using visual analogical* *reasoning.* For models, we include a chain-of-thought prompt. *You will think "step-by-step" and carefully examine the visual evidence before providing an answer.* For human adults, we additionally include the following prompt to motivate their participation. *At the end of the experiment, you will see the total number of correct answers you provided. Each correct answer will convert to $0.01 additional compensation for your study participation.* Then we provide an initial instruction prompt: *You are given a visual puzzle. The puzzle features a left-to-right transformation of an object on top and three left-to-right transformations of a different object on the bottom marked by (A) or (B) or (C). The transformations involve a change of either the size, orientation, number, or color of an object.*

1. **Verbal Classification ("*what*").**

   "Which one of the following rules best describes the left-to-right transformation on top of the puzzle where the picture on the left transforms to the picture on the right? Answer with the correct rule number. Surrounded by parentheses, then provide a "step-by-step" reasoning for your choice."

2. **Verbal Specification ("*how*").**

   "Which one of the following rules best describes the left-to-right transformation in the top of the puzzle where the picture on the left transforms to the picture on the right?. Answer with the correct rule number surrounded by parentheses. Then provide a "step-by-step" reasoning for your choice."

3. **Visual Extrapolation.**

   "Which one of the three left-to-right object transformations (marked by either (A), (B) or (C)) on the bottom of the puzzle is the same as the left-to-right transformation on the top of the puzzle? Answer with the correct letter surrounded by parentheses (or (D) if none of the options apply), then provide a a "step-by-step" reasoning for your choice."

A.3 PROMPTING HUMAN CHILDREN

All verbal instructions are read out loud to children by a human experimenter. We first provide a context to motivate children's participation in the experiment. *You are on a mission as a picture detective. You will see how different pictures change. Your job as a picture detective is to figure out how the pictures change, and to guess how a new picture would change based on that. These pictures can change in size, where they face, number, or color. Every time you answer correctly, you will get a coin. You won't find out how many coins you get until the end of the game. At the end of the game, you will see the total number of coins you win. The more coins you get, the more stickers you win.*

1. **Verbal Classification ("*what*").**

   "Here are two pictures separated by a black line in the middle. The picture on the left turns into the picture on the right. Do you think there is a change? What do you think the change is?"

2. **Verbal Specification ("*how*").**

   "Can you say more about the change from the left to the right?"

3. **Visual Extrapolation.**

   "Here is another picture that goes through the same change from the left to right. Can you find the box that shows the same change?"

Note that the prompt used for children did not improve model or human adult performance.

A.4 PROMPTING MODELS THROUGH REFLECTION AND SELF-CRITIQUE

1. **Verbal Classification ("*what*").**

   "Which one of the following rules best describes the left-to-right transformation on top of the puzzle where the picture on the left transforms to the picture on the right? Answer with the correct rule number surrounded by parentheses, then provide a "step-by-step" reasoning for your choice. Please reflect on your answer and provide a revised response if necessary."

   (repeat three times following model output) *Start your response with your updated answer.*

2. **Verbal Specification (“*how*”).**

   “Which one of the following rules best describes the left-to-right transformation in the top of the puzzle where the picture on the left transforms to the picture on the right?. Answer with the correct rule number surrounded by parentheses, then provide a "step-by-step" reasoning for your choice. Please reflect on your answer and provide a revised response if necessary."

   (repeat three times following model output) *Start your response with your updated answer.*

3. **Visual Extrapolation.**

   “Which one of three left-to-right object transformations (marked by either (A), (B) or (C)) on the bottom of the puzzle is the same as the left-to-right transformation on the top of the puzzle? Answer with the correct letter surrounded by parentheses (or (D) if none of the options apply), then provide a "step-by-step" reasoning for your choice. Please reflect on your answer and provide a revised response if necessary."

   (repeat three times following model output) *Start your response with your updated answer.*

A.5 Prompting models through instructions

1. **Verbal Classification (“*what*”).**

   “Which one of the following rules best describes the left-to-right transformation on top of the puzzle where the picture on the left transforms to the picture on the right? Answer with the correct rule number surrounded by parentheses, then provide a “step-by-step” reasoning for your choice."

2. **Verbal Specification (“*how*”).**

   “Provide brief instructions on how to establish if a transformation involves an object rotates 90 degrees or 180 degrees. Use the instructions form before to answer the following question: Which one of the following rules best describes the transformation in the top of the puzzle where the picture on the left transforms to the picture on the right?. Answer with the correct rule number surrounded by parentheses, then provide a “step-by-step” reasoning for your choice."

3. **Visual Extrapolation.**

   “Provide brief instructions on how to determine which one of three left-to-right object transformations (marked by either (A), (B) or (C)) on the bottom of the puzzle is the same as the left-to-right transformation on the top of the puzzle? Use the instructions from before to determine which one of three left-to-right object transformations (marked by either (A), (B) or (C)) on the bottom of the puzzle is the same as the left-to-right transformation on the top of the puzzle? Answer with the correct letter surrounded by parentheses (or (D) if none of the options apply), then provide a step-by-step reasoning for your choice."

A.6 Prompting models through code

1. **Verbal Classification (“*what*”).**

   “Which one of the following rules best describes the left-to-right transformation on top of the puzzle where the picture on the left transforms to the picture on the right? Answer with the correct rule number surrounded by parentheses, then provide a "step-by-step" reasoning for your choice."

2. **Verbal Specification (“*how*”).**

   “Generate python code using the package pillow that takes in the left image in the left-to-right transformation on top and outputs the right image. Denote this snippet as training snippet using the insights from the training code snippet, which one of the following rules best describes the left-to-right transformation in the top of the puzzle where the picture on the left transforms to the picture on the right?. Answer with the correct rule number surrounded by parentheses, then provide a "step-by-step" reasoning for your choice."

3. **Visual Extrapolation.**

   “Generate a brief code snippet using python and the pillow package for each left-to-right transformation in the bottom. Each snippet takes in the left picture of the transformation and outputs the right one. Now Which one of three code snippets is the same as the training

code snippet you have produced before. Answer with the correct snippet letter ((A) or (B) or (C)) surrounded by parentheses (or (D) if none of the options apply), then provide a "step-by-step" reasoning for your choice."

B Additional comparative statistics

B.1 Breakdown of performance on KiVA

We present a breakdown of performance organized by transformation visual domain, question type, and model or human: Table 1 includes comparisons on a sample of KiVA completed by children, whereas Table 2 presents adult and model performance on the full KiVA benchmark.

	Rotation			**Reflection**			**Number**			**Color**			**Size**		
	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation
Human adults	**97.7%** (1.48%)	**95.4%** (3.00%)	**94.3%** (3.44%)	**97.8%** (1.84%)	**96.5%** (2.06%)	**97.2%** (2.05%)	**99.4%** (0.41%)	**97.07%** (1.44%)	**97.2%** (2.12%)	**97.0%** (1.73%)	**100%** (0%)	**100%** (0%)	**100%** (0%)	**97.4%** (1.53%)	**98.8%** (1.25%)
Human children	**83.0%** (9.21%)	**73.5%** (14.56%)	**67.4%** (13.34%)	**74.9%** (10.16%)	**89.3%** (5.69%)	**80.2%** (9.46%)	**90.1%** (9.90%)	**78.2%** (14.79%)	**77.1%** (12.4%)	**93.1%** (5.77%)	**95.5%** (4.55%)	**93.9%** (4.07%)	**95.0%** (5.00%)	**79.4%** (10.36%)	**88.3%** (7.01%)
GPT-4V (single image)	**98.6%** (1.39%)	**50.2%** (10.78%)	**39.9%** (8.05%)	**96.2%** (3.23%)	**56.8%** (10.29%)	**38.6%** (6.49%)	**93.4%** (5.21%)	**64.4%** (8.57%)	**40.6%** (7.63%)	**100%** (0%)	**96.2%** (2.46%)	**75.4%** (8.11%)	**88.9%** (4.68%)	**97.1%** (2.86%)	**50.0%** (8.79%)
GPT-4V (multiple images)	**97.2%** (2.78%)	**64.2%** (12.2%)	**31.2%** (7.49%)	**64.8%** (5.93%)	**47.7%** (14.15%)	**44.3%** (7.39%)	**79.9%** (11.30%)	**57.8%** (11.95%)	**26.4%** (8.63%)	**100%** (0%)	**98.1%** (1.94%)	**98.0%** (1.36%)	**96.7%** (2.41%)	**100%** (0%)	**84.4%** (5.85%)
LLaVA-1.5 (single image)	**65.3%** (1.01%)	**51.4%** (11.0%)	**33.7%** (7.57%)	**69.6%** (6.24%)	**41.2%** (5.93%)	**30.2%** (7.57%)	**49.3%** (10.57%)	**53.2%** (12.03%)	**30.6%** (9.32%)	**69.7%** (8.06%)	**71.8%** (10.71%)	**32.4%** (8.26%)	**35.6%** (8.72%)	**76.7%** (9.33%)	**40.0%** (6.43%)
MANTIS (multiple images)	**92.7%** (4.27%)	**52.5%** (12.4%)	**37.8%** (8.28%)	**82.2%** (6.56%)	**57.4%** (9.21%)	**33.2%** (7.00%)	**90.3%** (7.14%)	**44.3%** (9.87%)	**35.1%** (8.83%)	**99.1%** (0.93%)	**45.9%** (10.70%)	**37.2%** (8.21%)	**45.6%** (10.5%)	**47.6%** (7.16%)	**26.7%** (6.56%)

Table 1: **Mean performance of humans and models on the child sample of KiVA benchmark, sorted by question type and transformation domain.** Standard errors are in parentheses.

	Rotation			**Reflection**			**Number**			**Color**			**Size**		
	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation
Human adults	**97.5%** (1.43%)	**95.0%** (2.00%)	**92.4%** (2.43%)	**98.8%** (1.25%)	**98.7%** (1.27%)	**98.7%** (1.27%)	**95.6%** (1.62%)	**94.4%** (1.83%)	**97.5%** (1.25%)	**99.2%** (0.83%)	**95.8%** (1.83%)	**99.2%** (0.84%)	**98.8%** (1.25%)	**100%** (0%)	**97.5%** (0.18%)
GPT-4V (single image)	**99.3%** (0.53%)	**59.7%** (5.00%)	**39.3%** (4.03%)	**96.3%** (1.76%)	**61.3%** (5.29%)	**37.3%** (3.58%)	**89.2%** (2.70%)	**58.2%** (4.14%)	**35.8%** (3.58%)	**99.8%** (0.22%)	**96.7%** (1.52%)	**78.0%** (3.52%)	**86.7%** (3.13%)	**98.1%** (1.17%)	**52.0%** (4.11%)
GPT-4V (multiple images)	**98.4%** (1.21%)	**62.7%** (5.37%)	**35.1%** (4.43%)	**71.0%** (4.12%)	**31.6%** (6.54%)	**37.7%** (4.36%)	**78.5%** (4.71%)	**58.7%** (5.95%)	**29.2%** (4.03%)	**100%** (0%)	**97.3%** (1.20%)	**96.0%** (1.51%)	**97.0%** (1.45%)	**100%** (0%)	**81.7%** (3.11%)
LLaVA-1.5 (single image)	**66.7%** (4.42%)	**48.4%** (5.25%)	**33.3%** (3.78%)	**64.3%** (4.48%)	**42.1%** (3.77%)	**28.3%** (3.92%)	**50.7%** (4.16%)	**44.4%** (6.08%)	**28.3%** (3.93%)	**73.3%** (3.69%)	**68.6%** (5.53%)	**29.8%** (4.04%)	**40.3%** (4.49%)	**66.2%** (6.98%)	**31.3%** (3.54%)
MANTIS (multiple images)	**92.2%** (3.08%)	**51.4%** (5.47%)	**37.1%** (3.99%)	**83.7%** (3.99%)	**61.7%** (4.78%)	**32.0%** (4.14%)	**86.7%** (3.90%)	**48.7%** (4.61%)	**32.8%** (3.81%)	**99.3%** (0.53%)	**52.9%** (4.91%)	**36.4%** (3.89%)	**39.7%** (5.46%)	**48.8%** (5.57%)	**34.3%** (3.83%)

Table 2: **Mean performance of humans and models on the full KiVA benchmark, sorted by question type and transformation domain.** Standard errors are included in parentheses.

B.2 Models' performance on trials involving no change

For each type of transformation, 10% of the positive transformation trials were randomly sampled and reassigned with training and test transformations that involved no change. In these cases, the original correct and incorrect transformation options were treated as distractor options, while the transformation involving “no change” was designated as the correct answer. Among the models, only GPT-4V (multiple images) performed significantly above chance when correctly identifying "no change" in the verbal classification stage, and this was limited to the size domain. In contrast, GPT-4V (single images), LLaVA-1.5, and MANTIS consistently "hallucinated" a change in 100% of the no-change trials during verbal classification. GPT-4V (single images) often inferred changes in orientation or size, while LLaVA-1.5 and MANTIS made errors equally across unrelated domains. None of the models showed significantly-above-chance visual extrapolation accuracy (i.e., identifying a new object that also underwent no change) above chance level across all domains.

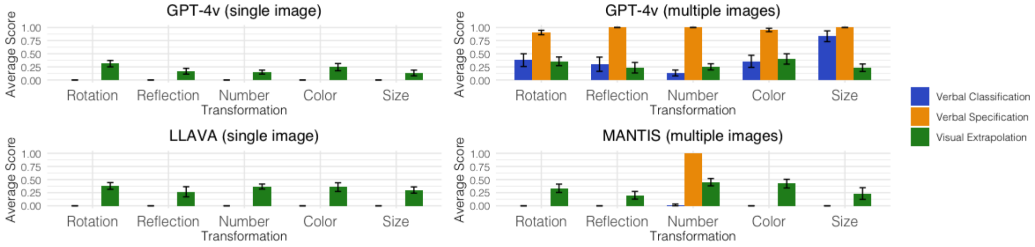


Figure 14: **Model performance on no-change trials.** Note that Verbal Specification is skipped if Verbal Classification is incorrect (as described in Figure 3).

B.3 Models' extrapolation performance conditional on previous verbal reasoning steps

Furthermore, we report models' extrapolation performance *conditional* on succeeding (Table 3) or failing (Table 4) at the previous steps of verbal reasoning. Only GPT-4V (multiple images and single images) demonstrated extrapolation accuracy that is significantly above chance in the color and size domains. Focusing on the GPT-4V model and these domains, we found that extrapolation accuracy is significantly higher when correct verbal classification and / or specification is correct (note that verbal specification is only asked if verbal classification is correct), suggesting that the likelihood of successful extrapolation is contingent on solving verbal classification or specification correctly.

	Rotation		**Reflection**		**Number**		**Color**		**Size**	
	Correct Verbal Classification	Correct Verbal Specification	Correct Verbal Classification	Correct Verbal Specification	Correct Verbal Classification	Correct Verbal Specification	Correct Verbal Classification	Correct Verbal Specification	Correct Verbal Classification	Correct Verbal Specification
GPT-4V (single image)	**36.0%** (2.35%)	**39.6%** (3.34%)	**33.8%** (3.00%)	**40.0%** (4.15%)	**34.6%** (2.21%)	**40.4%** (2.97%)	**73.5%** (2.67%)	**80.2%** (2.04%)	**58.2%** (3.41%)	**52.7%** (3.24%)
GPT-4V (multiple images)	**29.3%** (2.27%)	**38.0%** (3.46%)	**28.7%** (3.21%)	**67.6%** (6.71%)	**30.0%** (2.38%)	**35.0%** (3.56%)	**94.5%** (1.18%)	**97.4%** (0.76%)	**77.4%** (3.14%)	**84.2%** (2.72%)
LLaVA-1.5 (single image)	**34.5%** (2.17%)	**30.6%** (2.87%)	**30.0%** (2.59%)	**29.9%** (3.71%)	**30.2%** (2.05%)	**23.8%** (2.79%)	**29.1%** (2.04%)	**27.5%** (2.38%)	**36.3%** (2.25%)	**30.7%** (4.91%)
MANTIS (multiple images)	**37.6%** (2.44%)	**44.1%** (3.79%)	**31.9%** (3.32%)	**24.3%** (3.35%)	**32.5%** (2.17%)	**37.4%** (3.35%)	**36.3%** (2.25%)	**36.4%** (3.32%)	**36.3%** (4.60%)	**45.3%** (7.00%)

Table 3: **Mean extrapolation performance of models following *Correct* verbal classification / specification, sorted by transformation domain.** Standard errors are in parentheses.

	Rotation		**Reflection**		**Number**		**Color**		**Size**	
	Incorrect Verbal Classification	Incorrect Verbal Specification	Incorrect Verbal Classification	Incorrect Verbal Specification	Incorrect Verbal Classification	Incorrect Verbal Specification	Incorrect Verbal Classification	Incorrect Verbal Specification	Incorrect Verbal Classification	Incorrect Verbal Specification
GPT-4V (single image)	**31.5%** (6.86%)	**36.3%** (3.91%)	**26.5%** (7.80%)	**31.5%** (4.32%)	**19.5%** (4.28%)	**26.3%** (2.79%)	**24.4%** (6.88%)	**32.1%** (1.24%)	**27.9%** (6.44%)	**43.3%** (8.00%)
GPT-4V (multiple images)	**33.3%** (9.80%)	**28.0%** (4.06%)	**33.8%** (6.39%)	**30.4%** (3.62%)	**13.9%** (2.85%)	**24.9%** (2.81%)	**50.0%** (1.21%)	**43.3%** (15.8%)	**20.8%** (11.45%)	**16.7%** (1.05%)
LLaVA-1.5 (single image)	**29.0%** (2.77%)	**35.5%** (2.13%)	**22.8%** (3.18%)	**27.0%** (2.45%)	**28.8%** (2.07%)	**28.4%** (1.65%)	**36.0%** (3.31%)	**33.6%** (2.47%)	**27.8%** (2.63%)	**28.5%** (2.02%)
MANTIS (multiple images)	**18.3%** (6.83%)	**33.9%** (3.43%)	**30.9%** (7.16%)	**38.5%** (4.50%)	**30.0%** (5.57%)	**31.0%** (2.74%)	**66.7%** (33.3%)	**37.4%** (3.74%)	**30.5%** (3.79%)	**30.6%** (3.21%)

Table 4: **Mean extrapolation performance of models following *Incorrect* verbal classification / specification, sorted by transformation domain.** Standard errors are in parentheses.

B.4 Models' subsequent performance when given correct previous verbal reasoning step

10% of existing transformation trials were randomly sampled(evenly distributed across the five visual domains) to evaluate if model performance would improve when given the correct answer to the previous reasoning step. In one experiment, we provided the correct verbal classification answer and evaluated the models' verbal specification. In another experiment, we provided the correct verbal specification answer and evaluated the models' visual extrapolation. The data are outlined in Table 5. Overall, having the ground truth for the preceding verbal reasoning step did not guarantee success in the subsequent verbal specification or visual extrapolation tasks, with performance varying by model and domain. For example, the verbal specification performance of LLaVA-1.5 and MANTIS improved when given the correct verbal classification in the color domain, while the visual extrapolation performance of GPT-4V (multiple images) improved when given the correct verbal specification in the rotation and reflection domains. However, other models and domains did not show noticeable improvement when provided with the ground truth of the previous step.

	Rotation		**Reflection**		**Number**		**Color**		**Size**	
	Verbal Specification Given Correct Classification	Visual Extrapolation Given Correct Specification	Verbal Specification Given Correct Classification	Visual Extrapolation Given Correct Specification	Verbal Specification Given Correct Classification	Visual Extrapolation Given Correct Specification	Verbal Specification Given Correct Classification	Visual Extrapolation Given Correct Specification	Verbal Specification Given Correct Classification	Visual Extrapolation Given Correct Specification
GPT-4V (single image)	**44.4%** (8.40%)	**33.3%** (7.97%)	**63.7%** (8.53%)	**33.3%** (6.40%)	**51.7%** (7.04%)	**40.0%** (7.09%)	**97.8%** (2.22%)	**73.3%** (8.10%)	**73.3%** (10.89%)	**56.7%** (8.68%)
GPT-4V (multiple images)	**62.2%** (10.72%)	**48.9%** (7.18%)	**43.3%** (14.10%)	**66.7%** (12.17%)	**38.3%** (9.45%)	**38.3%** (6.52%)	**97.8%** (2.22%)	**93.3%** (4.82%)	**93.3%** (4.44%)	**83.3%** (11.39%)
LLaVA-1.5 (single image)	**53.3%** (10.18%)	**37.8%** (7.18%)	**50.0%** (16.67%)	**43.3%** (8.68%)	**46.7%** (5.62%)	**25.0%** (6.79%)	**86.7%** (5.44%)	**31.1%** (6.06%)	**60.0%** (9.69%)	**23.3%** (7.11%)
MANTIS (multiple images)	**55.6%** (9.58%)	**37.8%** (5.51%)	**50.0%** (16.67%)	**36.7%** (11.6%)	**28.3%** (7.36%)	**31.7%** (7.44%)	**91.1%** (6.88%)	**33.3%** (7.97%)	**53.3%** (8.89%)	**26.7%** (8.31%)

Table 5: **Mean subsequent performance of models when given Correct verbal classification or Correct verbal specification, sorted by transformation domain.** Standard errors are in parentheses.

C Evaluating humans and GPT-4V only on the visual extrapolation of KiVA

We presented 20 children, 200 adults, and GPT-4V (single and multiple images) with only the visual extrapolation task, removing the verbal questions and replicating previous visual analogy benchmarks. Without verbal reasoning, adults demonstrated similar accuracy but significantly slower response times ($t > 3$, $p < .01$ for all domains), children performed significantly worse in reflection, number, color and size domains ($t > 2.5$, $p < .05$), while GPT-4V (single and multiple images) performed worse ($p < .01$) in abstracting color and resize transformations — its strongest domains in the original setup—without the verbal questions (it was at chance level for the rest of the domains) (Figure 15). This indicates that the prior verbal reasoning steps enable humans and models to subsequently perform visual extrapolations. Without the full three-step query, humans and models would perform worse.

D Divergence between Human Adults' and Models' Visual Extrapolation Performance on KiVA

We conducted a KL divergence analysis comparing the distribution of human adult choices to model choices across the multiple-choice options in visual extrapolation, the core of any visual analogy task. First, we found that divergences primarily stemmed from adults consistently selecting the correct option while models chose incorrect ones. Notably, there were no trials in which both models and adults scored 0%. Figure 16 illustrates the domains and models with the greatest divergence from adults in extrapolation choices, with GPT-4V (multiple images) aligning most closely with adults in the color domain. Second, we observed that the greatest divergences occurred in trials or object

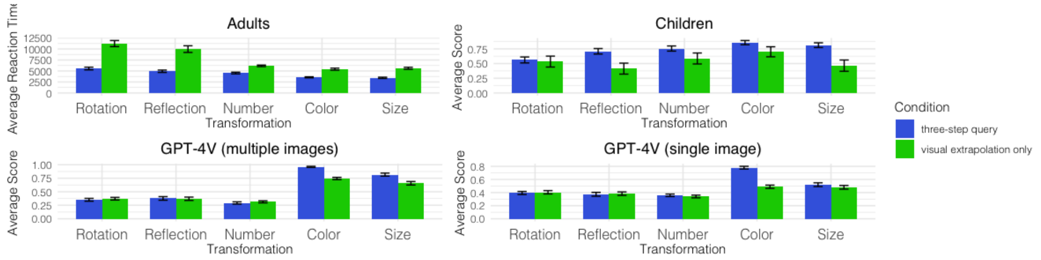


Figure 15: **Adults' Mean Response Times, Children's and GPT-4V's Mean Accuracy in Extrapolation with and without the three-step query.**

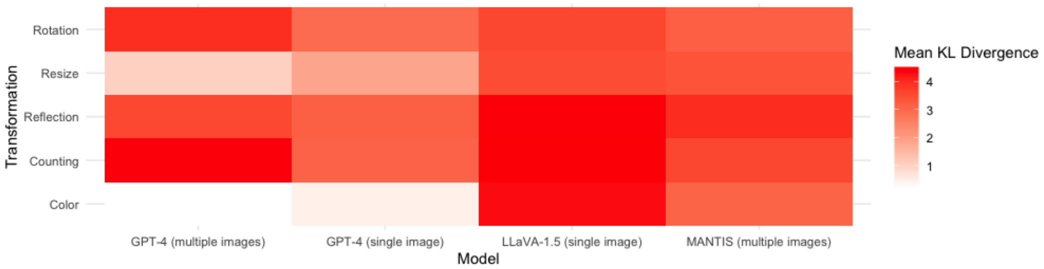
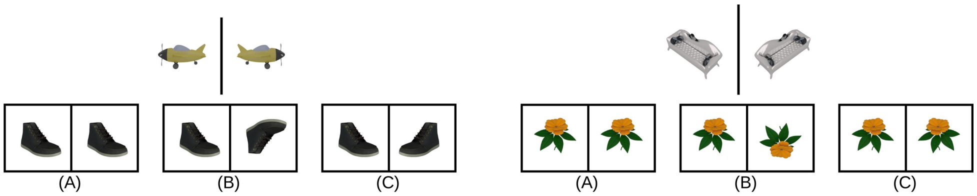


Figure 16: **Heatmap of Mean KL Divergence Between Human Adults and Models.**

sets where models only chose incorrect options (scored 0%) and adults only chose correct options (scored 100%). Crucially, there were no trials where this pattern was reversed. Depending on the transformation type and the model, 0% to 50% of trials showed models scoring 0% while adults scored 100%. Across the entire benchmark, there were only two trials where all models scored 0% but adults scored 100%. Both of these trials involved reflection along the Y-axis and are shown in Figure 17.



(a) Trial 18 of Reflection along the y-axis, annotated correctly by 8/8 human adults.

(b) Trial 30 of Reflection along the y-axis, annotated correctly by 6/6 human adults.

Figure 17: **The two trials in KiVa wherein models scored 0% but human adults scored 100%.** The marigold in (b) features a subtle reflection that is noticeable in the stem and petal orientation.

E KiVA-ADULTS

E.1 KiVA-ADULTS COMPARED TO KiVA

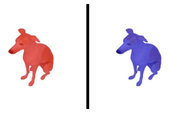
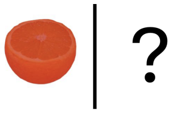
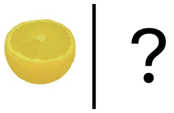
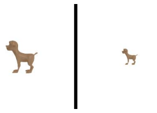
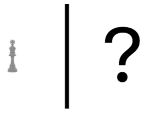
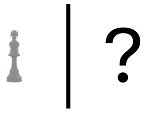
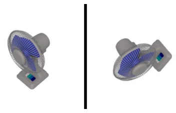
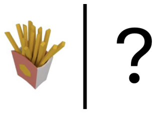
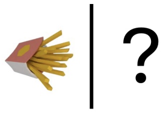
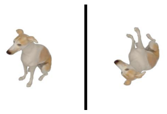
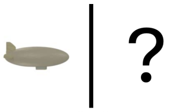
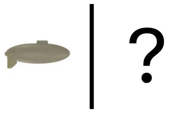
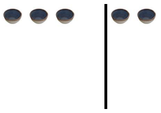
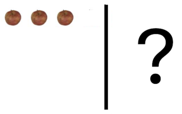
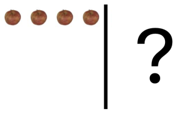
	Given Transformation	KiVA Extrapolation	KiVA-adults Extrapolation
Color			
Size			
Rotation			
Reflection			
Number			

Table 6: **Examples of transformation input values tested on KiVA vs. KiVA-adults.** Unlike KiVA, KiVA-adults' extrapolation input values are different from the given transformation input values.

KiVA preserves the input and output values between the given transformation and test transformation for extrapolation. As shown in Table 6, only the object undergoing transformation varies between training (left column) and test (middle column). For example, in a numerical transformation, if the given transformation involves turning *three* bowls into two bowls, test asks for the result of transforming *three* apples (answer: two apples). KiVA-adults, on the other hand, changes the input value at test (right column). For instance, if the given transformation shows *three* bowls converting into two bowls again, the test asks for the transformation of *four* apples (answer: three apples). This requires not only generalizing a transformation across different objects but also abstracting a *function* to generalize over different *input values*.

	KiVA	KiVA-adults
Color	Red, Green, Blue	Red, Green, Blue, Yellow, Grey
Size	$1/2\times$ and $2\times$ height & width	$1/2\times$ and $2\times$ height & width, or separately
Rotation	$\pm 90°$, $180°$	$\pm 45°$, $\pm 90°$, $\pm 135°$, $180°$
Reflection	reflect along the x or y axes	reflect along x or y axes, or both axes together
Number	$+1, +2, -1, -2$	$+1, +2, -1, -2, \times 2, \times 3, \div 2, \div 3$

Table 7: **More visual subdomains are involved in KiVA-adults compared to KiVA. Color**: KiVA-adults includes two more colors transformations (yellow, grey), non-uniform scaling of size (independently halving or doubling either width or height), two additional angular values ($\pm 45°$ and $\pm 135°$), rotating objects along both axes, as well as division and multiplication of objects by 2 or 3.

Furthermore, KiVA-adults includes more transformation subdomains than the original task. The subdomains in KiVA are more distinguishable from each other (see Table 7 for the subdomains

	Rotation			Reflection			Number			Color			Size		
	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation	Verbal Classification	Verbal Specification	Visual Extrapolation
Human adults	**100%** (0%)	**85.1%** (2.82%)	**88.8%** (2.49%)	**99.2%** (0.83%)	**94.2%** (2.43%)	**84.2%** (3.35%)	**98.8%** (0.88%)	**93.7%** (1.94%)	**97.5%** (1.24%)	**100%** (0%)	**98.0%** (0.98%)	**98.0%** (0.99%)	**98.3%** (1.16%)	**95.8%** (1.85%)	**98.3%** (1.16%)
GPT-4V (single image)	**96.4%** (0.55%)	**47.2%** (1.43%)	**37.6%** (1.18%)	**93.8%** (1.35%)	**48.2%** (2.87%)	**37.1%** (2.24%)	**90.9%** (0.68%)	**62.6%** (1.36%)	**35.5%** (0.98%)	**99.5%** (0.27%)	**95.4%** (0.91%)	**73.1%** (1.70%)	**70.4%** (1.33%)	**81.7%** (1.30%)	**49.2%** (1.35%)
GPT-4V (multiple images)	**97.2%** (0.78%)	**42.9%** (2.06%)	**30.0%** (1.51%)	**79.8%** (2.85%)	**27.2%** (2.92%)	**32.9%** (2.27%)	**82.0%** (1.26%)	**54.4%** (1.65%)	**26.7%** (1.03%)	**99.7%** (0.27%)	**94.4%** (1.19%)	**92.7%** (1.19%)	**75.8%** (1.69%)	**70.9%** (2.01%)	**53.7%** (1.86%)
LLaVA-1.5 (single image)	**68.7%** (1.20%)	**30.6%** (1.34%)	**33.9%** (1.13%)	**65.9%** (1.89%)	**55.3%** (2.55%)	**31.5%** (1.82%)	**50.9%** (1.03%)	**35.1%** (1.16%)	**28.7%** (0.93%)	**75.0%** (1.23%)	**56.6%** (1.73%)	**31.2%** (1.18%)	**24.2%** (1.21%)	**35.7%** (1.69%)	**31.7%** (1.10%)
MANTIS (multiple images)	**93.6%** (1.02%)	**49.5%** (2.27%)	**40.0%** (1.53%)	**86.6%** (2.18%)	**51.0%** (2.72%)	**33.3%** (2.27%)	**86.5%** (1.41%)	**39.4%** (1.67%)	**33.2%** (1.38%)	**95.6%** (1.04%)	**54.8%** (2.51%)	**34.4%** (1.71%)	**38.3%** (1.99%)	**48.1%** (2.18%)	**33.1%** (1.55%)

Table 8: **Mean performance of humans and models on the KiVA-adults benchmark, sorted by question type and transformation domain.** Standard errors are in parentheses.

evaluated in the two tasks respectively). In KiVA-adults, we also have to add additional constraints in the sampling for the transformation domain involving number changes to avoid ambiguity between multiplication and addition (both involving an increase in number), or between subtraction and division (both involving a decrease in number).

We run GPT-4V, LLaVA-1.5 and MANTIS on KiVA-adults which consists of 2,900 transformations. We also test 10 children and 40 adults (of the same age ranges as those recruited for KiVA) on KiVA-adults. Each adult completes 19 trials that are randomly sampled from the 29 subdomains, totaling 760 responses. Each child is assigned to complete 10 randomly sampled trials, but none of them succeed on completing them with passing accuracy.

E.2 Models' and human adults' performance on KiVA-adults

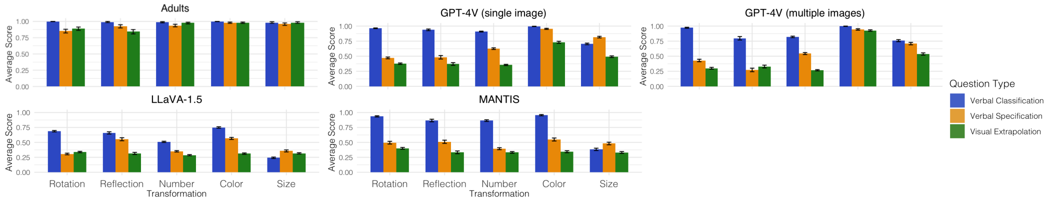


Figure 18: **Performance of Human Adults, GPT-4V, LLaVA-1.5 and MANTIS on KiVA-adults.** Performance is organized by question type and transformation domain. Error bars represent standard errors of performance across object variations within the same transformation domain.