# BabyVision: Visual Reasoning Beyond Language

Liang Chen[1,6]*, Weichu Xie[1,6]*, Yiyan Liang[1,6]*, Hongfeng He[1]*, Hans Zhao[1]*, Zhibo Yang[3], Zhiqi Huang[4], Haoning Wu[4], Haoyu Lu[4], Y. charles[4], Yiping Bao[4], Yuantao Fan[5], Guopeng Li[5], Haiyang Shen[1,6], Xuanzhong Chen[1,7], Wendong Xu[1], Shuzheng Si[7], Zefan Cai[8], Wenhao Chai[9], Ziqi Huang[10], Fangfu Liu[7], Tianyu Liu[6], Baobao Chang[6], Ming Wu[11], Xiaobo Hu[2], Kaiyuan Chen[2], Yixin Ren[2], Yang Liu[2], Yuan Gong[2], Kuan Li[1]

[1]UniPat AI [2]xbench [3]Alibaba Group [4]MoonShot AI [5]StepFun [6]Peking University [7]Tsinghua University [8]University of Wisconsin–Madison [9]Princeton University [10]Nanyang Technological University [11]0G Labs

## Abstract

While humans develop core visual skills long before acquiring language, contemporary Multimodal LLMs (MLLMs) still rely heavily on linguistic priors to compensate for their fragile visual understanding. We uncovered a crucial fact: state-of-the-art MLLMs consistently fail on basic visual tasks that humans, even 3-year-olds, can solve effortlessly. To systematically investigate this gap, we introduce BABYVISION, a benchmark designed to assess core visual abilities independent of linguistic knowledge for MLLMs. BABYVISION spans a wide range of tasks, with 388 items divided into 22 subclasses across four key categories. Empirical results and human evaluation reveal that leading MLLMs perform significantly below human baselines. Gemini3-Pro-Preview scores 49.7, lagging behind 6-year-old humans and falling well behind the average adult score of 94.1. These results show despite excelling in knowledge-heavy evaluations, current MLLMs still lack fundamental visual primitives. Progress in BABYVISION represents a step toward human-level visual perception and reasoning capabilities. We also explore solving visual reasoning with generation models by proposing BABYVISION-GEN and automatic evaluation toolkit. Our code and benchmark data are released at UniPat-AI/BabyVision for reproduction.

HOMEPAGE: https://unipat.ai/blog/BabyVision

arXiv:2601.06521v2 [cs.CV] 7 Jul 2026

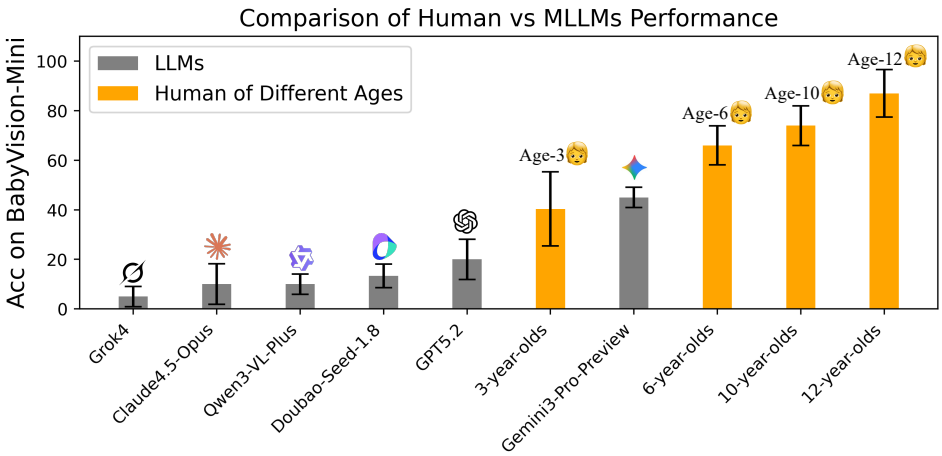



Figure 1: Performance on BABYVISION among MLLMs and human of different ages.

*Equal Core Contributors. Correspondence: Liang Chen <liangchen@unipat.ai>, Kuan Li <kuanli@unipat.ai>

## 1 Introduction

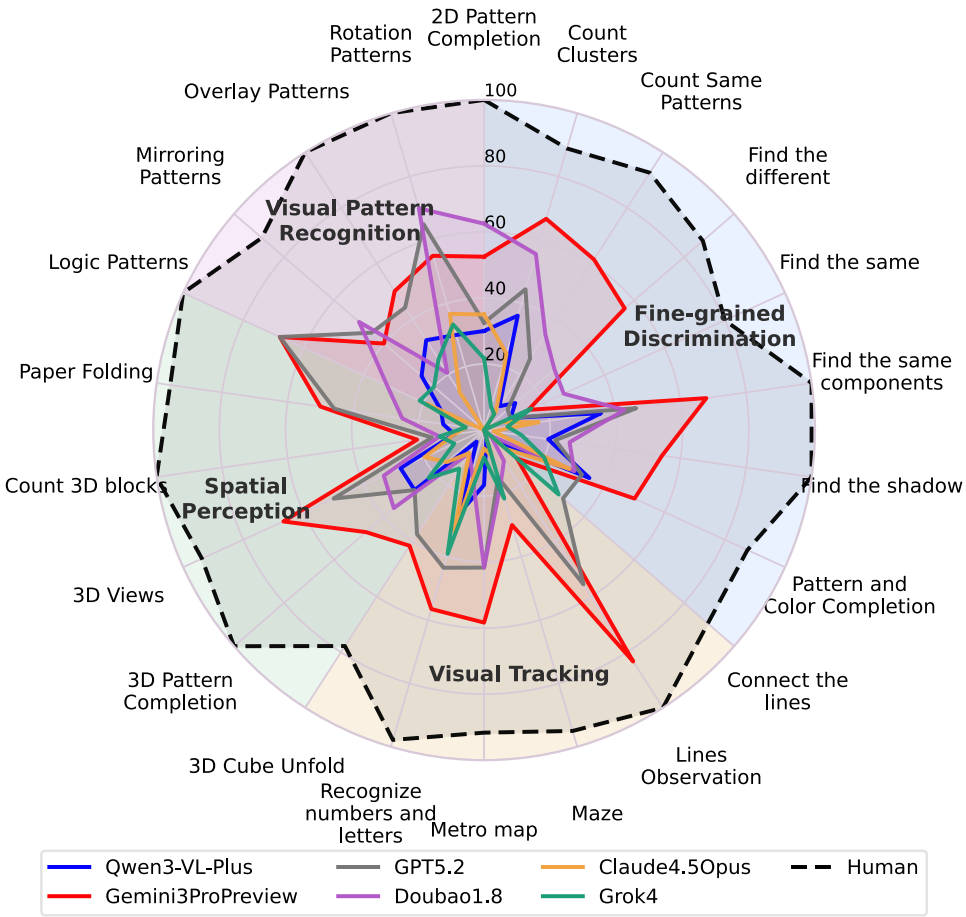



Figure 2: Fine-grained performance analysis on the full BABYVISION benchmark.

Human visual understanding emerges remarkably early in life—well before children acquire language or formal symbolic reasoning. Within the first months, infants can already discriminate shapes and textures, track moving objects, infer occlusion and depth, and form expectations about simple physical events (Johnson, 2010; Braddick & Atkinson, 2011; Kellman et al., 2006).

These *early-vision abilities* form a foundation for later cognitive functions—including language, abstract reasoning, and motor planning—by supporting structured representations of the physical world.

In contrast, contemporary Multimodal Large Language Models (MLLMs) (Team et al., 2023; OpenAI, 2025; Bai et al., 2025) appear to exhibit an inverted profile of competence. They achieve strong performance on many high-level, knowledge-intensive benchmarks that typically demand substantial education or domain expertise (e.g., HLE (Team, 2025b), MMMU (Yue et al., 2024)), including multi-step geometric and mathematical reasoning (e.g., MathVista (Lu et al., 2024), MathVision (Wang et al., 2024)), large-scale recognition of people and places (e.g., MME (Fu et al., 2024a)), and even medically styled question answering (e.g., DrVD-Bench (Zhou et al., 2025)). Yet, despite these successes, our studies reveal a consistent weakness in *basic* visual tasks that children (ages 3–12) can solve with little or no language mediation—such as visual discrimination, visual tracking, spatial understanding, and simple visual pattern recognition. Even the strongest model we tested, Gemini3-Pro-Preview, remains approximately 20% behind 6-year-old children in our pilot study, as shown in Figure 1. This discrepancy suggests that current MLLMs still lack the atomic visual competencies that underlie human vision from its earliest stages. However, existing evaluations rarely target these **beyond-language visual understanding** abilities in a systematic way.

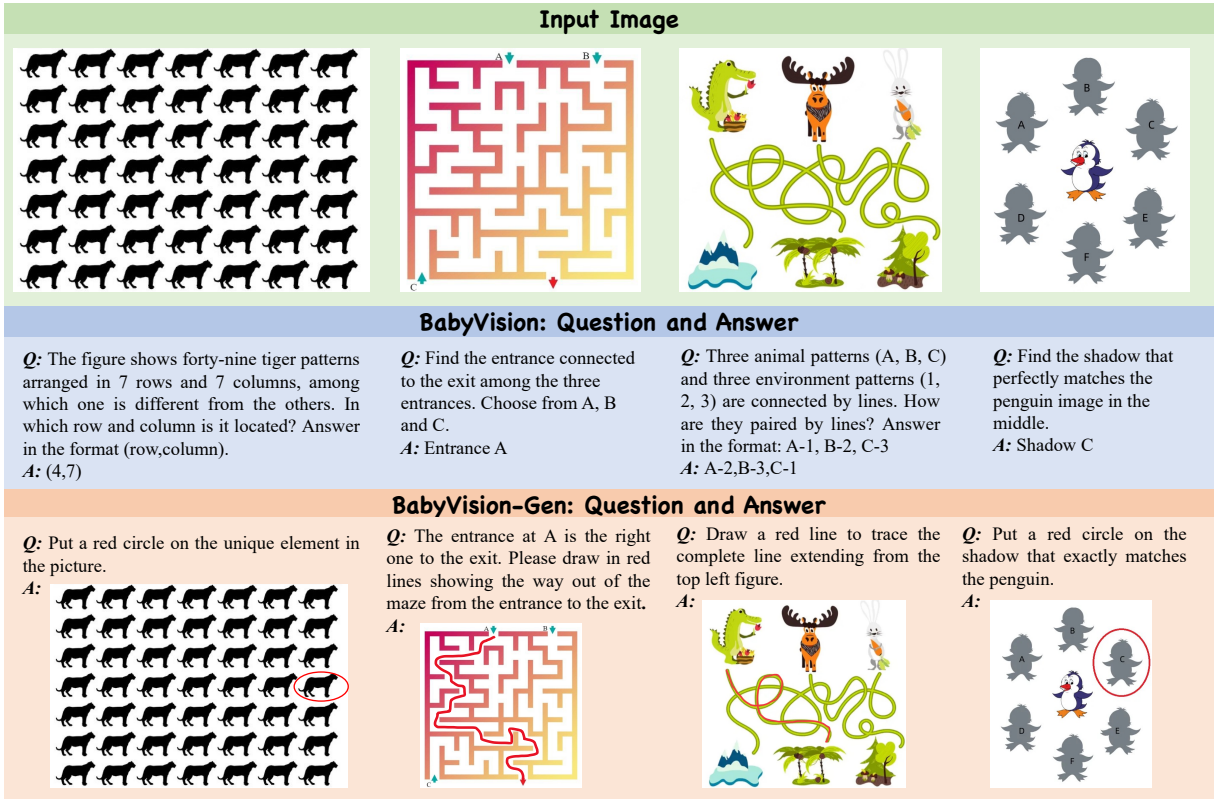



Figure 3: Examples of BABYVISION and BABYVISION-GEN. While BABYVISION evaluates visual understanding through language output, BABYVISION-GEN evaluates visual reasoning through image generation.

To fill this gap, we introduce BABYVISION, a benchmark designed with a scientific and rigorous data curation pipeline to probe the atomic visual skills humans develop early in life. BABYVISION aims to minimize reliance on linguistic knowledge by emphasizing tasks driven by perception and pattern regularities rather than textual reasoning or semantic priors. The benchmark contains 388 unique questions spanning 22 subclasses across four domains: *Fine-grained Discrimination*, *Visual Tracking*, *Spatial Perception*, and *Visual Pattern Recognition*. Together, these domains cover a broad spectrum of early-vision competencies with substantial diversity in visual conditions and task structure.

As shown in Figure 2, we evaluate state-of-the-art MLLMs on BABYVISION, covering both proprietary and open models (e.g., Gemini3-Pro-Preview, GPT-5.2, Qwen3VL-235B-Thinking). In general, we observe a substantial gap between current MLLMs and human performance on early-vision competencies. Specifically, the best-performing model achieves an overall accuracy of **49.7**%, whereas adult human testers reach **94.1**% (a **44.4**% absolute gap), with consistent deficits across all four domains. The largest failures appear in *Visual Tracking* and *Spatial Perception*, where models frequently exhibit errors such as a *loss of manifold identity* (e.g., losing track of curves through intersections) and a *failure of spatial imagination* (e.g., the inability to mentally transform 3D structures). These results indicate that strong high-level multimodal reasoning does not imply robust foundational visual competence, and that BABYVISION exposes a distinct failure mode not captured by existing benchmarks.

Beyond textual evaluation, we observe that many questions from BABYVISION are most naturally solved visually: humans typically draw on the image—by tracing paths, completing patterns, or marking spatial relations—rather than verbalizing their reasoning. Motivated by this, we introduce BABYVISION-GEN, a generative counterpart to BABYVISION that evaluates visual reasoning through visual generation **beyond language output**. We also develop an automatic evaluation tool achieving a 96% agreement rate

with human judgments. Experiments with frontier image and video generators (e.g., Nano-Banana-Pro, Sora-2) show promising gains on tasks that are challenging for MLLMs, especially visual tracking and fine-grained discrimination, though overall reliability and solution consistency remain limited.

In summary, this paper makes three main contributions: (1) We introduce BABYVISION, a refined benchmark for evaluating **beyond-language** visual reasoning abilities, containing **388** questions across **22** subclasses in four domains. (2) Motivated by the observation that many tasks are most naturally solved via visual generation, we further propose BABYVISION-GEN, an extension of BABYVISION that evaluates visual reasoning ability for generation models **beyond language output**, along with a reliable automatic evaluation toolkit. (3) We conduct a comprehensive evaluation of state-of-the-art MLLMs and generative models, revealing a new performance landscape marked by a substantial human–model gap (44.4%) and pronounced inter-model disparities (e.g., Gemini3-Pro-Preview outperforming the runner-up by approximately 15 points). We further provide fine-grained analyses of failure patterns at both the domain and subclass levels, and investigate how RLVR training and generative modeling contribute to improved visual reasoning performance.

## 2 Related Work

### 2.1 Benchmarks for Multimodal Large Language Models

The rapid progress of Multimodal Large Language Models (MLLMs) has driven the development of diverse evaluation benchmarks. A prominent line of work targets expert-level, knowledge-intensive multimodal reasoning. MMMU (Yue et al., 2024) assembles 11.5K college-level questions spanning 30 disciplines, while HLE (Team, 2025b) pushes further toward "humanity's last exam" with expert-level problems. Mathematical reasoning benchmarks such as MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), and MathVerse (Zhang et al., 2024) focus on geometric and quantitative reasoning grounded in images, while MME (Fu et al., 2024a) and MMEvalPro (Huang et al., 2025) provide comprehensive evaluations across perception and cognition dimensions. On these benchmarks, leading models achieve impressive results—Gemini 3 Pro attains 92.8% on MMMU—demonstrating the increasingly effective integration of visual inputs with domain knowledge.

However, strong results on semantic and knowledge-driven benchmarks do not necessarily translate into robust visual perception. This inverted competence profile—where models excel at expert-level tasks but struggle with basic perception—has been documented across several studies. BLINK (Fu et al., 2024b) reveals a substantial gap between humans (95.7%) and leading MLLMs on classical vision problems that require no specialized knowledge. MMStar (Chen et al., 2024) demonstrates that models can achieve 42.9% on MMMU *without* visual input, exposing severe language leakage and over-reliance on textual shortcuts. MMVP (Tong et al., 2024) further probes visual limitations and finds that models fail to distinguish images with clear perceptual differences. These findings reveal a recurring blind spot: existing benchmarks predominantly draw from tasks designed for adult experts and rely on semantic recognition rather than perceptual primitives. BABYVISION addresses this gap by targeting pre-linguistic visual abilities—the foundational competencies humans develop before language acquisition—and comparing model performance directly to children aged 3–12.

### 2.2 Early Vision and Developmental Foundations

The design of BABYVISION draws on developmental psychology research establishing that humans acquire core visual competencies well before language. Infants demonstrate object permanence by 3–4 months (Baillargeon et al., 1985), track objects through occlusion (Johnson, 2010), and discriminate depth and shape before acquiring language (Kellman et al., 2006; Braddick & Atkinson, 2011). The core knowledge hypothesis (Spelke, 2000) posits innate systems for representing objects, space, numbers, and agents—capacities that are independent of linguistic mediation. By ages 3–6, children perform complex

visual discrimination, understand spatial transformations, and recognize patterns, often demonstrating competence through non-verbal responses (Karmiloff-Smith, 1994; Golomb, 2003).

This developmental trajectory motivates key design choices in BABYVISION. We select tasks solvable by young children, ensuring they test foundational rather than expert-level abilities. We minimize verbal reasoning demands by focusing on perceptual regularities rather than semantic knowledge. We compare model performance directly to children at different developmental stages, revealing not just whether models fail, but *how far* they fall below human visual development. Developmental research also shows that children externalize visual reasoning through drawing before verbalizing solutions (Kellogg, 1969)—they trace paths, complete patterns, and mark spatial relations rather than describing them. This observation motivates BABYVISION-GEN: evaluating models solely on verbal answers may underestimate visual competence that could be demonstrated through visual output, suggesting generation-based evaluation as a complementary paradigm.

## 3 BabyVision

In this section, we present BABYVISION, a benchmark designed to evaluate the early-vision capabilities of Multimodal Large Language Models (MLLMs). Unlike previous benchmarks that focus on complex semantic reasoning or domain knowledge, BABYVISION targets foundational visual skills that humans acquire during early development—abilities that emerge before or alongside language acquisition. We split vision-centric reasoning into four core categories: *Fine-grained Discrimination* (detecting subtle visual differences; 8 subtypes), *Visual Tracking* (following paths, lines, and trajectories; 5 subtypes), *Spatial Perception* (understanding 3D structures and relationships; 5 subtypes), and *Visual Pattern Recognition* (identifying logical and geometric patterns; 4 subtypes). Together, these comprise 22 basic subtypes, each targeting a specific fundamental visual capability. Through a rigorous data curation pipeline, we construct 388 questions spanning a wide diversity of visual reasoning tasks.

### 3.1 Data Curation Pipeline

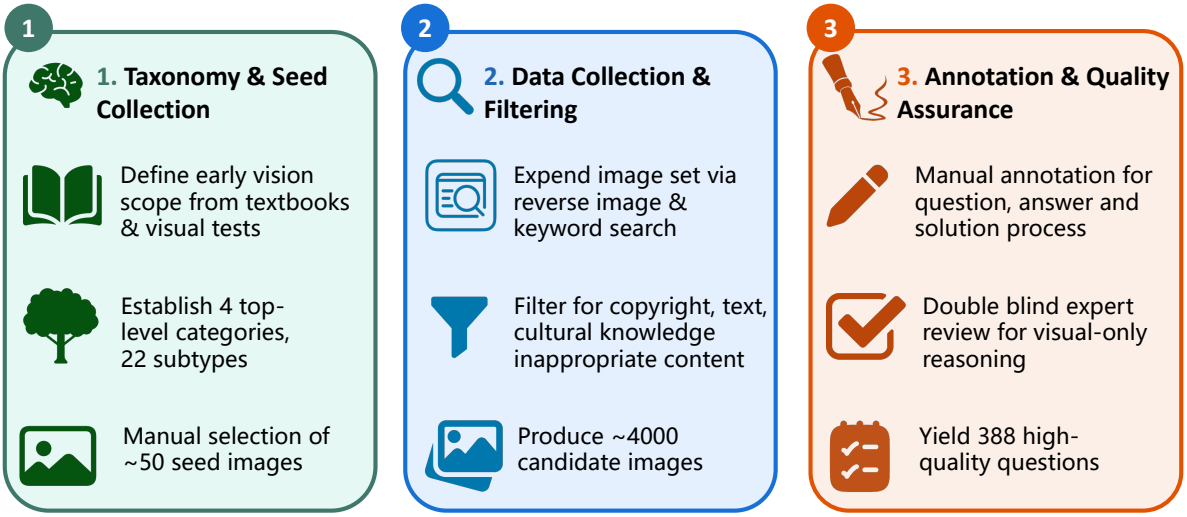



Figure 4: Overview of the multi-stage data collection and curation pipeline for BABYVISION: taxonomy design & seed selection, data augmentation & filtering, and annotation & quality assurance.

To ensure high quality and validity, we employ a rigorous multi-stage data collection and curation process, as shown in Figure 4. The whole process includes the following three steps:

**Taxonomy Definition and Seed Collection.** We first define the scope of "early vision" grounded in developmental psychology. Consulting textbooks designed for children under 12 and standardized visual

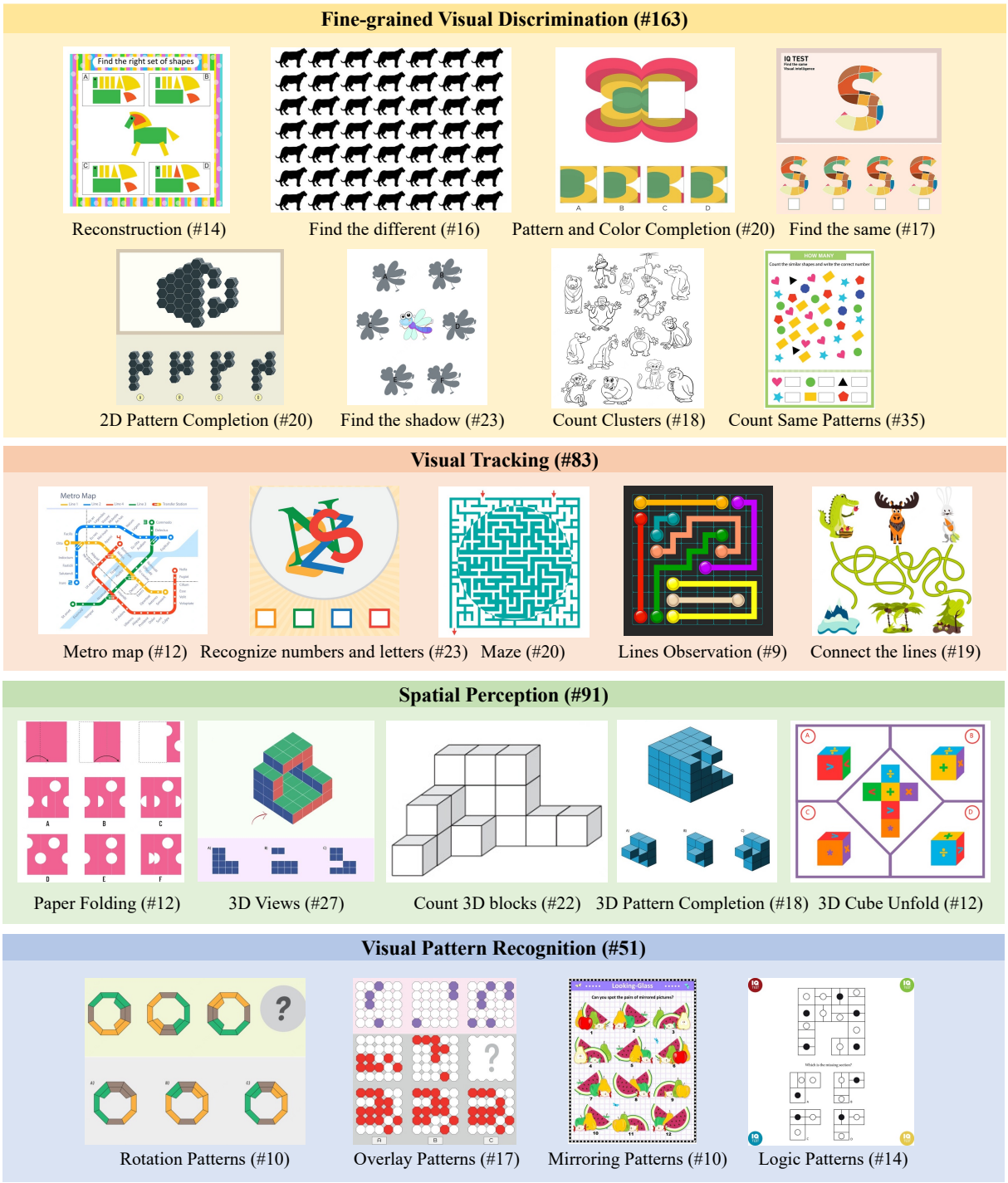



Figure 5: Example questions and the number of examples (#) from BABYVISION spanning the four core categories and 22 types.

development tests, we identify core visual competencies and establish four top-level categories with 22 specific subtypes. For each subtype, we manually select 4–5 "seed images" that exemplify the visual task, yielding approximately 100 high-quality seed examples that serve as prototypes for data expansion.

**Data Collection and Filtering.** Using the seed images, we expand the dataset through reverse image search and keyword-based retrieval, crawling similar images from the internet. We strictly adhere to copyright regulations, retaining only images permitted for academic use. Images containing substantial text, requiring cultural knowledge, or depicting inappropriate content are filtered out. This stage produces

approximately 4,000 candidate images.

**Annotation and Quality Assurance.** Each candidate image undergoes manual annotation by trained annotators. Annotators first determine whether the image can support a meaningful visual question aligned with our taxonomy—images that rely heavily on text reading or require cultural background knowledge are discarded. For valid images, annotators write specific questions and answers and, crucially, provide a detailed "solution process" proving the answer's correctness.

All annotations then pass through a double-blind review: two independent experts verify that the answer is unambiguous and derivable largely from visual analysis rather than language. A question is included only if both experts agree on the answer and reasoning logic. Disagreements are returned to the original annotator for revision; questions that remain contentious after revision are permanently discarded.

### 3.2 Dataset Statistics

The final BABYVISION benchmark comprises 388 carefully curated questions, each paired with a unique image. The dataset includes 135 multiple-choice questions (34.8%) and 253 fill-in-the-blank questions (65.2%), with balanced answer distributions to mitigate position bias. The questions are concise (averaging 25.9 words) and tightly grounded in visual content, minimizing opportunities for language-based shortcuts. The distribution of task categories is shown in Figure 5. In addition, we collect 1,400 training examples following the same construction pipeline but sourced more broadly, enabling an investigation of how training influences model performance on BABYVISION, which is discussed in Section 6.2.

### 3.3 Evaluation

For inference, we use a unified prompt template across all models: `question Think about the question and give your final answer in format.` This encourages models to reason through the problem before answering and facilitates automated answer extraction. By default, we select the highest reasoning effort of the tested models if adjustable.

We evaluate model responses using an LLM-as-judge approach. Given a model's output answer and the ground-truth answer, we query Qwen3-Max to determine semantic equivalence, which shows 100% consistency with human evaluators. The full judge prompt is provided in Appendix A.

## 4 BabyVision-Gen

A key observation from BABYVISION is that many tasks admit solutions that are most naturally expressed *visually*. When humans solve maze navigation or pattern completion problems, they often draw trajectories or complete shapes rather than verbalizing step-by-step reasoning (Kellogg, 1969; Golomb, 2003). Similarly, in find-the-difference tasks, solutions are typically conveyed by directly circling the differing regions in the image. This observation motivates BABYVISION-GEN, a generative extension that evaluates whether models can perform visual reasoning through image (or video) generation, providing an alternative pathway that bypasses the verbalization bottleneck.

### 4.1 Test Reasoning in Generation Models

BABYVISION-GEN adapts a subset of BABYVISION questions for visual output and alters the prompt to support generation. For each question, we provide a generation prompt instructing the model to produce an image that demonstrates the solution—tracing a path, completing a pattern, or marking spatial relationships. This formulation tests whether visual generation models can leverage visual reasoning capabilities that may exceed their ability to verbalize answers. We show some examples of BABYVISION-GEN in Figure 3.

**Statistics.** BABYVISION-GEN comprises 280 questions across 21 subtypes organized into the same four categories as BABYVISION: Fine-grained Discrimination (128 questions, 8 subtypes), Visual Tracking (55 questions, 4 subtypes), Spatial Perception (59 questions, 5 subtypes), and Visual Pattern Recognition (38 questions, 4 subtypes). Note that one subtype from BABYVISION is excluded as it does not naturally admit a generative solution format. Each instance includes the original image, a generation prompt (averaging 22.9 words), and a reference solution image for evaluation, reflecting the consensus of three human annotators. The detailed distribution across categories and subtypes is provided in Table 4 in Appendix B.4.

### 4.2 Evaluation

Evaluating BABYVISION-GEN requires determining whether a model's generated image expresses the correct visual solution. We standardize inference by instructing models to annotate the original input image with minimal overlays (e.g., circles, lines, arrows, or text), preserving the underlying visual context while making the predicted answer explicit.

For automatic evaluation, we employ Gemini-3-Flash as a judge. The judge is given three images: (i) the input image, (ii) a human-annotated ground-truth solution, and (iii) the model-generated output. It returns a binary decision indicating whether the generated annotation matches the ground truth under subtype-specific criteria (e.g., identical selected option; same traced route for mazes). Full prompts and criteria are provided in Appendix B.

To further assess the reliability of automatic evaluation, we perform human evaluation on all Nano-Banana-Pro outputs using PhD-level annotators. Automatic and human judgments agree on **96.1%** of instances (269/280; F1=0.924), supporting the use of the automatic judge for scalable evaluation. Subtype-wise agreement rates and the confusion matrix are reported in Appendix B.

## 5 Main Experiments

In this section, we present a comprehensive evaluation on the BabyVision benchmarks. We first describe the experimental setup in Section 5.1, followed by the results on BABYVISION in Section 5.2 and on BABYVISION-GEN in Section 5.3.

### 5.1 Evaluation Setup

**Models.** Our evaluation encompasses 11 frontier models, comprising 5 proprietary and 6 open-source systems. The proprietary models include Gemini3-Pro-Preview (Team et al., 2023), GPT-5.2 (OpenAI, 2025), Qwen3-VL-Plus (Bai et al., 2025), Doubao-1.8 (Seed, 2025), and Claude-4.5-Opus (Anthropic, 2025). The open-source selection consists of Qwen3VL-235B-Thinking (Bai et al., 2025), InternVL3.5-241B (Wang et al., 2025), Step3 (StepFun et al., 2025), KimiVL-A3B (Team et al., 2025b), MimoVL-7B-RL (Team et al., 2025a), and GLM4.6V (Team, 2025a).

Furthermore, to examine the influence of model scaling and reasoning modes (thinking vs. non-thinking), we evaluate additional variants of the Qwen3VL family, including Qwen3VL-235B (Thinking/Instruct) and the smaller Qwen3VL-32B/8B/4B-Thinking models.

**Evaluation Settings.** We report the Avg@3 results for all evaluated models. For proprietary models, we utilize the official APIs without modifying default parameters, with the exception of reasoning effort, which is set to the maximum level where adjustable. For open-source models, we adhere to the inference temperatures recommended in their official papers; unless otherwise specified, we employ a default sampling temperature of 1.0. The inference and evaluation prompts for BABYVISION and BABYVISION-GEN are detailed in Section 3.3 and Section 4.2, respectively.

| Type | Sub-Type | Human | Gemini3-Pro-Preview | | GPT-5.2 | | Doubao-1.8 | | Qwen3-VL-Plus | | Claude-4.5-Opus | | Grok-4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ |
| Fine-grained Discrimination | 2D Pattern Completion | 100 | 52.5 | 2.5 | 32.5 | 7.5 | **62.5** | 2.5 | 30.0 | 5.0 | 35.0 | 4.1 | 21.7 | 4.7 |
| | Count Clusters | 88.9 | **66.7** | 0.0 | 44.4 | 5.6 | 55.6 | 5.6 | 36.1 | 2.8 | 24.1 | 2.6 | 7.4 | 2.6 |
| | Count Same Patterns | 92.7 | **61.4** | 4.3 | 25.7 | 2.9 | 34.3 | 5.7 | 8.6 | 2.9 | 6.7 | 1.4 | 5.7 | 0.0 |
| | Find the Different | 87.5 | **56.3** | 6.3 | 9.4 | 3.1 | 28.1 | 3.1 | 12.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Find the Same | 80 | 14.7 | 2.9 | 8.8 | 2.9 | **26.5** | 2.9 | 8.8 | 8.8 | 5.9 | 0.0 | 15.7 | 2.8 |
| | Reconstruction | 100 | **67.9** | 3.6 | 46.4 | 3.6 | 42.9 | 0.0 | 35.7 | 0.0 | 16.7 | 3.4 | 7.1 | 5.8 |
| | Find the Shadow | 100 | **54.4** | 6.5 | 21.7 | 0.0 | 26.1 | 4.4 | 19.6 | 10.9 | 2.9 | 2.1 | 11.6 | 7.4 |
| | Pattern and Color Completion | 87.5 | **50.0** | 10.0 | 32.5 | 7.5 | 30.0 | 5.0 | 35.0 | 0.0 | 28.3 | 2.4 | 20.0 | 4.1 |
| | **Sub-Type Overall Results** | 92.3 | **46.2** | 5.5 | 27.3 | 0.9 | 39.2 | 0.0 | 21.8 | 2.2 | 14.3 | 0.3 | 11.0 | 2.7 |
| Visual Tracking | Connect the Lines | 89.5 | 13.2 | 2.6 | **31.6** | 0.0 | 2.6 | 2.6 | 5.3 | 5.3 | 8.8 | 6.6 | 29.8 | 2.5 |
| | Lines Observation | 100 | **83.3** | 5.6 | 55.6 | 0.0 | 11.1 | 11.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Maze | 95.0 | **30.0** | 5.0 | 15.0 | 0.0 | 17.5 | 2.5 | 2.5 | 2.5 | 8.3 | 2.4 | 21.7 | 4.7 |
| | Metro Map | 91.7 | **58.3** | 8.3 | 41.7 | 0.0 | 41.7 | 8.3 | 16.7 | 0.0 | 5.6 | 3.9 | 8.3 | 6.8 |
| | Recognize Numbers and Letters | 97.8 | **56.5** | 4.4 | 43.5 | 0.0 | 13.0 | 0.0 | 26.1 | 4.4 | 31.9 | 2.1 | 39.1 | 0.0 |
| | **Sub-Type Overall Results** | 94.6 | **43.4** | 2.4 | 34.9 | 0.0 | 15.7 | 1.2 | 11.5 | 0.6 | 13.7 | 2.8 | 24.1 | 1.0 |
| Spatial Perception | 3D Cube Unfold | 77.8 | **41.7** | 8.3 | 37.5 | 4.2 | 8.3 | 0.0 | 4.2 | 4.2 | 8.3 | 6.8 | 13.9 | 10.4 |
| | 3D Pattern Completion | 100.0 | **47.2** | 13.9 | 27.8 | 0.0 | 36.1 | 8.3 | 27.8 | 5.6 | 14.8 | 2.6 | 22.2 | 7.9 |
| | 3D Views | 93.8 | **66.7** | 7.4 | 50.0 | 1.9 | 33.3 | 3.7 | 27.8 | 5.6 | 19.8 | 6.3 | 9.9 | 4.6 |
| | Count 3D Blocks | 100 | **20.5** | 2.3 | 15.9 | 6.8 | 13.6 | 4.6 | 9.1 | 4.6 | 9.1 | 3.7 | 13.6 | 3.7 |
| | Paper Folding | 95.8 | **50.0** | 0.0 | 45.8 | 12.5 | 25.0 | 0.0 | 12.5 | 4.2 | 5.6 | 3.9 | 5.6 | 3.9 |
| | **Sub-Type Overall Results** | 94.7 | **53.7** | 1.5 | 35.2 | 2.2 | 24.7 | 0.6 | 18.1 | 1.7 | 12.8 | 1.4 | 13.2 | 3.2 |
| Visual Pattern Recognition | Logic Patterns | 100 | **67.9** | 3.6 | **67.9** | 3.6 | 32.1 | 3.6 | 14.3 | 7.1 | 19.1 | 3.4 | 21.4 | 5.8 |
| | Mirroring Patterns | 88.9 | 40.0 | 10.0 | 45.0 | 5.0 | **50.0** | 0.0 | 25.0 | 5.0 | 0.0 | 0.0 | 20.0 | 0.0 |
| | Overlay Patterns | 100 | **50.0** | 2.9 | 44.1 | 2.9 | 20.6 | 2.9 | 32.4 | 14.7 | 13.7 | 5.6 | 25.5 | 2.8 |
| | Rotation Patterns | 100 | 55.0 | 15.0 | 65.0 | 15.0 | **70.0** | 0.0 | 30.0 | 10.0 | 36.7 | 12.5 | 33.3 | 9.4 |
| | **Sub-Type Overall Results** | 97.8 | 53.9 | 1.0 | **54.9** | 5.9 | 37.7 | 1.5 | 25.5 | 3.9 | 17.0 | 3.7 | 24.8 | 2.5 |
| All | **Overall Results** | 94.1 | **49.7** | 2.6 | 34.4 | 1.7 | 30.2 | 0.3 | 19.2 | 0.6 | 14.2 | 1.2 | 16.2 | 1.3 |

Table 1: **Performance (Avg@3) of Close Source MLLMs on BabyVision.** The best results for each question type are marked in **bold**. Reported values represent the average Pass@1 accuracy across three random runs, accompanied by the standard deviation. All models are in thinking model with highest reasoning budget.

**Human Testers.** In our primary experiment, we conducted a comparative study using 20 representative samples (BabyVision-Mini) spanning diverse categories. We recruited participants across four age groups—3, 6, 10, and 12 years old—with 20 individuals per group[1]. The study was conducted with full permission from school authorities and parents, as well as the voluntary assent of the participants. Each group was allotted one class period (45 minutes) to complete the test. In our comprehensive evaluation of BABYVISION, 16 adult participants, each holding at least a bachelor's degree, completed the entire benchmark (388 questions). Human testers are not included in the evaluation of BABYVISION-GEN.

## 5.2 BabyVision Results

**Proprietary Model Performance.** The strongest proprietary model performance remains substantially below the human baseline (94.1%). Among proprietary systems, Gemini3-Pro-Preview achieves the highest overall score (49.7%), followed by GPT-5.2 (34.4%) and Doubao-1.8 (30.2%), while the remaining models trail by a wide margin (e.g., Qwen3-VL-Plus at 19.2%, Grok-4 at 16.2%, Claude-4.5-Opus at 14.2%). The large gap to human performance is consistent across all four task families, with no single category dominating the error profile. This pattern suggests that current systems still lack foundational visual competencies, pointing to a systemic limitation rather than an isolated weakness.

At the subtype level, several tasks prove challenging for nearly all systems. In particular, *Count 3D Blocks* yields uniformly low accuracy (best: 20.5%), indicating that compositional 3D counting under occlusion and viewpoint variation remains difficult. Similarly, *Find the Same* reaches only 26.5% at best (Doubao-1.8), despite a human baseline of 80%. Such failures are informative because they reflect deficiencies in structured scene representations (e.g., object permanence and depth-aware composition), rather than shortcomings in recognition or superficial pattern matching.

**Open-source Model Performance.** For open-source models, the best reported system (Qwen3VL-235B-Thinking) attains 22.2% overall, with most alternatives falling between 12% and 19%. Test-time thinking consistently provides measurable gains: within the Qwen3VL family, the Thinking variants outperform

[1]Testers aged 3-12 are from a single school population. Results may vary across different populations, educational backgrounds, or cultural contexts.

| Type | Sub-Type | Qwen3VL-235B-Thinking | | InternVL3.5-241B | | GLM4.6V | | MimoVL-7B-RL | | Step3 | | KimiVL-A3B | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ |
| Fine-grained Discrimination | 2D Pattern Completion | 25.0 | 10.8 | 11.7 | 6.2 | **40.0** | 10.8 | 30.0 | 10.8 | 33.3 | 9.4 | 25.0 | 8.2 |
| | Count Clusters | **27.8** | 0.0 | 20.4 | 9.4 | 20.4 | 2.6 | 11.1 | 4.5 | 9.3 | 6.9 | 3.7 | 2.6 |
| | Count Same Patterns | **15.2** | 3.6 | 10.5 | 2.7 | 6.7 | 2.7 | 9.5 | 1.4 | 3.8 | 2.7 | 1.9 | 1.4 |
| | Find the Different | 16.7 | 3.0 | **20.8** | 3.0 | 14.6 | 3.0 | 2.1 | 3.0 | 0.0 | 0.0 | 6.3 | 0.0 |
| | Find the Same | **23.5** | 8.3 | 9.8 | 5.6 | 9.8 | 2.8 | 7.8 | 2.8 | 2.0 | 2.8 | 3.9 | 5.6 |
| | Reconstruction | 33.3 | 3.4 | **38.1** | 3.4 | 26.2 | 3.4 | 9.5 | 3.4 | 11.9 | 8.9 | 14.3 | 5.8 |
| | Find the Shadow | 15.9 | 5.4 | **17.4** | 3.6 | 7.3 | 2.1 | 7.3 | 2.1 | 10.1 | 2.1 | 7.3 | 5.4 |
| | Pattern and Color Completion | **38.3** | 9.4 | 31.7 | 6.2 | 25.0 | 7.1 | 35.0 | 7.1 | 21.7 | 6.2 | 13.3 | 8.5 |
| | **Sub-Type Overall Results** | **23.3** | 0.9 | 18.6 | 2.3 | 17.4 | 2.9 | 14.1 | 2.7 | 11.3 | 2.3 | 8.8 | 1.5 |
| Visual Tracking | Connect the Lines | 5.3 | 4.3 | 15.8 | 4.3 | 14.0 | 2.5 | 3.5 | 2.5 | 7.0 | 2.5 | **24.6** | 2.5 |
| | Lines Observation | **7.4** | 5.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Maze | 11.7 | 6.2 | **23.3** | 2.4 | 21.7 | 2.4 | 8.3 | 2.4 | 21.7 | 2.4 | **23.3** | 4.7 |
| | Metro Map | **13.9** | 3.9 | 2.8 | 3.9 | 0.0 | 0.0 | **13.9** | 10.4 | 2.8 | 3.9 | 5.6 | 3.9 |
| | Recognize Numbers and Letters | 36.2 | 7.4 | 39.1 | 3.6 | 30.4 | 6.2 | 14.5 | 2.1 | **53.6** | 4.1 | 8.7 | 3.6 |
| | **Sub-Type Overall Results** | 16.9 | 4.3 | 20.5 | 2.0 | 16.9 | 1.7 | 8.8 | 2.1 | **22.1** | 1.5 | 14.5 | 2.0 |
| Spatial Perception | 3D Cube Unfold | **19.4** | 7.9 | 8.3 | 6.8 | 8.3 | 6.8 | 8.3 | 6.8 | 11.1 | 10.4 | 2.8 | 3.9 |
| | 3D Pattern Completion | 29.6 | 6.9 | 22.2 | 4.5 | 33.3 | 7.9 | **37.0** | 2.6 | 25.9 | 9.4 | 31.5 | 14.6 |
| | 3D Views | **29.6** | 8.0 | 18.5 | 3.0 | 22.2 | 3.0 | 23.5 | 1.8 | 9.9 | 4.6 | 19.8 | 4.6 |
| | Count 3D Blocks | **10.6** | 2.1 | 6.1 | 2.1 | 6.1 | 4.3 | **10.6** | 4.3 | 3.0 | 2.1 | 6.1 | 4.3 |
| | Paper Folding | 8.3 | 6.8 | 11.1 | 3.9 | 16.7 | 0.0 | 13.9 | 3.9 | 11.1 | 7.9 | **19.4** | 3.9 |
| | **Sub-Type Overall Results** | **20.9** | 2.7 | 13.9 | 2.9 | 18.0 | 1.4 | 19.8 | 0.9 | 11.7 | 1.4 | 16.5 | 2.4 |
| Visual Pattern Recognition | Logic Patterns | 19.1 | 6.7 | **26.2** | 3.4 | 11.9 | 3.4 | 9.5 | 3.4 | 14.3 | 0.0 | 4.8 | 3.4 |
| | Mirroring Patterns | **26.7** | 9.4 | 23.3 | 4.7 | 23.3 | 4.7 | **26.7** | 4.7 | 6.7 | 4.7 | 6.7 | 9.4 |
| | Overlay Patterns | **33.3** | 7.3 | 25.5 | 2.8 | 13.7 | 7.3 | 21.6 | 2.8 | 31.4 | 2.8 | 21.6 | 14.0 |
| | Rotation Patterns | 40.0 | 0.0 | **43.3** | 12.5 | 33.3 | 4.7 | 26.7 | 12.5 | 16.7 | 4.7 | 16.7 | 9.4 |
| | **Sub-Type Overall Results** | **29.4** | 5.8 | 28.8 | 1.9 | 19.0 | 2.5 | 20.3 | 3.3 | 19.0 | 0.9 | 13.1 | 4.9 |
| All | **Overall Results** | **22.2** | 1.0 | 19.2 | 0.7 | 17.6 | 1.8 | 15.1 | 1.3 | 14.7 | 0.8 | 12.4 | 1.7 |

Table 2: **Performance (Avg@3) of Open Source MLLMs on BabyVision.** The best results for each question type are marked in **bold**. Reported values represent the average Pass@1 accuracy across three random runs, accompanied by the standard deviation.

their Instruct counterparts (e.g., 22.2% vs. 19.5% at 235B), suggesting that explicit intermediate reasoning can partially mitigate visual uncertainty once relevant evidence is available. Nevertheless, even the largest open-source model remains far behind the top proprietary system, highlighting a persistent capability gap for the current open-source models on these fundamental visual abilities.

**Impact of Model Scaling.** We further investigate how model size and reasoning strategy affect performance within the Qwen3VL family as shown in Table 6. The Thinking variant consistently outperforms the Instruct variant at the same scale (e.g., 22.2% vs. 19.5% at 235B), confirming that explicit intermediate reasoning provides measurable gains on visual tasks. Across model sizes, we observe that performance generally improves with scale: 235B-Thinking (22.2%) outperforms 32B-Thinking (17.4%), which in turn exceeds 8B-Thinking (13.1%). However, the 4B-Thinking model (14.6%) slightly exceeds 8B-Thinking, suggesting that scaling alone does not guarantee monotonic improvement on these fundamental perceptual tasks. These results indicate that improvements in early-vision competencies may require architectural or training innovations beyond simply increasing model parameters.

**Comparison with Young Humans.** As BabyVision-Mini is built for meaningful developmental comparison, its tasks are strictly vision-centric, minimizing language and prior-knowledge demands so that scores reflect visual reasoning rather than text-based inference. Its small size also makes it practical to complete within a single class period for young children.

Under this lens, the gap is striking as shown in the comparison figure (see Figure 1). Most frontier MLLMs perform well below the average 3-year-old, despite their PhD-level results on language benchmarks. Gemini3-Pro-Preview is the notable outlier—the only model consistently above the Age-3 band—yet it still lags typical 6-year-olds by about 20 points.

This highlights a core limitation: the issue is not solving "hard problems," but struggling with pre-language visual primitives—the early perceptual and spatial abilities humans acquire before language becomes the main reasoning tool.

**Fine-grained Analysis.** To identify which visual primitives are most challenging for current MLLMs, we analyze performance at the subtype level across all four domains.

1. Fine-grained Discrimination. Within this category, models show highly variable performance across subtypes. Gemini3-Pro-Preview achieves relatively strong results on *Reconstruction* (67.9%) and *Count Clusters* (66.7%), but all models struggle with *Find the Same* (best: 26.5%) and *Find the Different* (best: 56.3%). These tasks require detecting subtle visual differences among highly similar patterns—a capability that demands high-fidelity perceptual encoding rather than semantic abstraction.

2. Visual Tracking. This category exposes a fundamental weakness in maintaining continuous identity across spatial trajectories. The *Lines Observation* task is particularly diagnostic: Gemini3-Pro-Preview achieves 83.3%, but most other models score near zero, indicating a complete inability to trace curves through intersections. Similarly, *Connect the Lines* and *Maze* tasks reveal that models frequently "switch tracks" at crossings—errors that are immediately obvious to human observers.

3. Spatial Perception. The *Count 3D Blocks* subtype is uniformly challenging across all models, with the best accuracy at only 20.5% (Gemini3-Pro-Preview). This task requires inferring hidden volume and maintaining a coherent 3D mental model—capabilities that cannot be verbalized without information loss. In contrast, *3D Views* (66.7% for Gemini3-Pro-Preview) proves more tractable, likely because the viewpoint transformation can be partially reasoned through geometric rules.

4. Visual Pattern Recognition. Models show comparatively stronger performance in this category, with both Gemini3-Pro-Preview and GPT-5.2 achieving 67.9% on *Logic Patterns* and Doubao-1.8 reaching 70.0% on *Rotation Patterns*. These results suggest that when patterns involve discrete, rule-based transformations (rotation, reflection), models can partially leverage their symbolic reasoning capabilities. However, *Overlay Patterns* (best: 50.0%) and *Mirroring Patterns* (best: 50.0%) remain challenging compared to the near-perfect human baselines (100% and 88.9% respectively), indicating that continuous spatial transformations are still problematic.

### 5.3 BabyVision-Gen Results

We evaluate three frontier visual generation models on BABYVISION-GEN: NanoBanana-Pro, GPT-Image-1.5, and Qwen-Image-Edit. Results are presented in Table 3. Note that since BABYVISION-GEN uses a different evaluation methodology (visual output assessment via image generation) and covers a subset of tasks adapted for generative evaluation, the results are not directly comparable to the MLLM experiments on BABYVISION.

**Overall Performance.** NanoBanana-Pro achieves the highest overall accuracy at 18.3% ($\pm$2.2), substantially outperforming GPT-Image-1.5 (9.8% $\pm$1.0) and Qwen-Image-Edit (4.8% $\pm$0.7). The relatively low standard deviations across runs suggest consistent behavior, though overall accuracy remains limited.

The performance distribution across categories reveals distinct patterns. NanoBanana-Pro achieves its strongest results on *Fine-grained Discrimination* (24.5%) and *Visual Pattern Recognition* (22.8%), while *Spatial Perception* (13.0%) shows moderate performance with higher variance ($\pm$6.0). In contrast, *Visual Tracking* remains extremely challenging (6.7% for the best model), with all models struggling on maze navigation and line tracing tasks. Notably, tasks requiring continuous trajectory maintenance (*Maze*, *Connect the Lines*) yield 0% accuracy across all generation models, indicating that current architectures cannot maintain spatial coherence over extended generation sequences.

**Subtype Analysis.** Several subtypes reveal the characteristics of visual generation approaches. On *Find the Different*, NanoBanana-Pro achieves 35.4% ($\pm$2.9), and *Count Same Patterns* reaches 31.2% ($\pm$6.1). These tasks share a common characteristic: solutions are naturally expressed by drawing or marking on the image. For *Spatial Perception* tasks such as *Count 3D Blocks* (26.7%) and *Paper Folding* (20.0%), NanoBanana-Pro demonstrates non-trivial performance, though with considerable variance across runs. In *Visual Pattern Recognition*, the models show mixed results: *Logic Patterns* (30.3%) and *Overlay Patterns* (24.4%) yield moderate accuracy, while *Mirroring Patterns* proves challenging (0% for NanoBanana-Pro,

| Type | Sub-Type | # | NanoBanana-Pro | | GPT-Image-1.5 | | Qwen-Image-Edit-2511 | |
|---|---|---|---|---|---|---|---|---|
| | | | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ |
| Fine-grained Discrimination | 2D Pattern Completion | 19 | **24.6** | 4.8 | 12.3 | 2.5 | 14.0 | 6.6 |
| | Count Clusters | 9 | **22.2** | 9.1 | 3.7 | 5.2 | 7.4 | 5.2 |
| | Count Same Patterns | 31 | **31.2** | 6.1 | 0.0 | 0.0 | 2.2 | 3.0 |
| | Find the Different | 16 | **35.4** | 2.9 | 12.5 | 8.8 | 2.1 | 2.9 |
| | Find the Same | 10 | 3.3 | 4.7 | **3.3** | 4.7 | 0.0 | 0.0 |
| | Reconstruction | 13 | **30.8** | 6.3 | 10.3 | 3.6 | 0.0 | 0.0 |
| | Find the Shadow | 14 | **9.5** | 3.4 | 7.1 | 5.8 | 2.4 | 3.4 |
| | Pattern and Color Completion | 16 | 22.9 | 2.9 | **31.3** | 0.0 | 8.3 | 5.9 |
| | **Category Average** | 128 | **24.5** | 1.8 | 9.6 | 1.2 | 4.7 | 0.8 |
| Visual Tracking | Connect the Lines | 12 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Maze | 18 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Metro Map | 12 | **11.1** | 3.9 | 5.6 | 3.9 | 0.0 | 0.0 |
| | Recognize Numbers and Letters | 13 | **17.9** | 3.6 | 5.1 | 3.6 | 0.0 | 0.0 |
| | **Category Average** | 55 | **6.7** | 2.1 | 2.4 | 1.0 | 0.0 | 0.0 |
| Spatial Perception | 3D Cube Unfold | 12 | 2.8 | 3.9 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 3D Pattern Completion | 18 | 5.6 | 0.0 | **14.8** | 2.6 | 13.0 | 5.2 |
| | 3D Views | 19 | **21.1** | 8.6 | 17.5 | 2.5 | 10.5 | 4.5 |
| | Count 3D Blocks | 5 | **26.7** | 9.4 | 20.0 | 0.0 | 0.0 | 0.0 |
| | Paper Folding | 5 | **20.0** | 16.3 | 6.7 | 9.4 | 0.0 | 0.0 |
| | **Category Average** | 59 | **13.0** | 6.0 | 12.4 | 2.6 | 7.3 | 3.5 |
| Visual Pattern Recognition | Logic Patterns | 11 | **30.3** | 8.6 | 12.1 | 8.6 | 12.1 | 11.2 |
| | Mirroring Patterns | 3 | 0.0 | 0.0 | **44.4** | 15.7 | 11.1 | 15.7 |
| | Overlay Patterns | 15 | **24.4** | 16.5 | 8.9 | 2.9 | 0.0 | 0.0 |
| | Rotation Patterns | 9 | 18.5 | 5.2 | **25.9** | 5.2 | 14.8 | 10.5 |
| | **Category Average** | 38 | **22.8** | 10.0 | 16.7 | 5.5 | 7.9 | 2.6 |
| All | **Overall** | 280 | **18.3** | 2.2 | 9.8 | 1.0 | 4.8 | 0.7 |

Table 3: **Performance (%) of Visual Generation Models on BabyVision-Gen.** The best results for each subtype are marked in **bold**. # denotes the number of questions in each subtype. Reported values represent the average accuracy across three random runs, accompanied by the standard deviation.

though GPT-Image-1.5 achieves 44.4% with high variance). These results suggest that visual generation offers a promising but currently limited pathway for visual reasoning tasks.

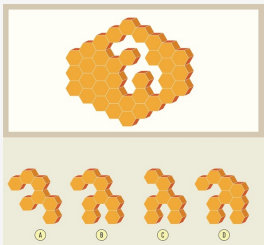
## 6 Discussion

In this section, we discuss the implications of these findings and outline potential directions for future research.

### 6.1 Why Do Frontier Models Fail on Seemingly Simple Tasks?

Our evaluation reveals a striking inversion: state-of-the-art MLLMs that excel at PhD-level reasoning benchmarks struggle with visual tasks that young children solve effortlessly. Qualitative analysis characterizes four systematic failure modes where the visual modality gap acts as an information bottleneck. Figure 6 illustrates four representative failure cases from Gemini3-Pro-Preview that motivate the analysis below.

**Loss of Fine-Grained Detail.** A pervasive weakness across BabyVision is the degradation of fine-grained visual information, where MLLMs fail to distinguish candidates relying on sub-semantic cues like specific curvature or pixel-level boundary alignment. While humans solve these tasks via direct shape matching—a parallel geometric operation that verifies congruence without intermediate description—MLLMs rely on implicit semantic compression, attempting to discretize continuous shapes into lossy linguistic tokens. This process creates a resolution gap where fine spatial structure is flattened into semantic space, rendering micro-differences indistinguishable and forcing the model to reason through a

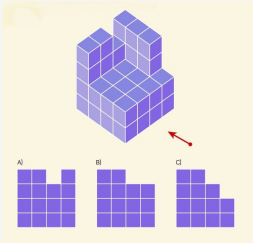


**Question:** Find the missing piece that fits into the empty white space within the hexagonal structure.
**Answer:** B **Model Answer:** D
**Reasoning:** Bottom island implies split legs; check right then left flank.
Option A has a hexagon sticking out to the right... Option B also has a "shoulder"... Option C has a straight right edge... Option D has an empty right side (a narrow "neck")...
Option C is flat on the left... Option B is compact... Option D has a left protrusion...
**Why wrong:** Over-verbalized geometry; misses the exact contour.

(a) Non-verbal fine-grained perception failure

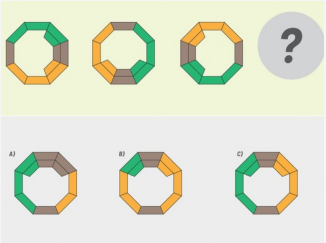
**Question:** Which colored trash bin will the plastic cup, old newspaper, and apple core each go into?
**Answer:** A-blue, B-yellow, C-green
**Model Answer:** A-green, B-yellow, C-blue
**Reasoning:** For plastic cup (A), start at its dot; Follow the line to the left. It curves down and forms the outermost path on the left side of the maze. It continues down, runs along the bottom, and goes up to green (A-green). For apple core (C), the top path goes left, down the second vertical line, and into blue (C-blue). The remaining line links the newspaper (B) to yellow.
**Why wrong:** Track-switching at crossings swaps endpoints.

(b) Failure of manifold identity tracking



**Question:** Which is the correct view when looking at the 3D structure from the direction indicated by the arrow?
**Answer:** A **Model Answer:** C
**Reasoning:** Arrow indicates a right-side view; project the silhouette by columns. Columns 1–2 are tallest (tower behind the seat), column 3 is lower (seat), and column 4 is the step. Its height is 2. There is nothing taller behind it. Therefore, the silhouette height is 2. So the profile becomes 4, 4, 3, 2 and the model picks C.
**Why wrong:** Verbal height logic ignores the true 3D structure and occlusion.

(c) Breakdown in 3D spatial imagination



**Question:** Among the three given options, choose the most suitable one to fill in the question mark, so that it shows a certain pattern.
**Answer:** C **Model Answer:** B
**Reasoning:** Each octagon is composed of 8 equal segments, with two green, two brown, and four orange segments. The brown pair shifts two positions counterclockwise each step, so the next should place brown at the top; B matches that. Option C is rejected because brown appears at the top left and bottom.
**Why wrong:** Focuses on color placement rather than the actual structural transform.

(d) Failure in abstract visual pattern induction

Figure 6: Four classic vision-centric challenges for MLLMs. All examples highlight failures caused by compressing perceptual reasoning into language.

low-fidelity proxy rather than end-to-end perceptual comparison.

**Loss of Manifold Identity.** We observe a critical failure in topological consistency, where MLLMs struggle to maintain the identity of a continuous manifold (e.g., a winding line) as it interacts with others. Unlike humans, who utilize primitive contour integration to "lock onto" a curve and track it through occlusions, MLLMs attempt to map these 1D manifolds into discrete instruction sequences. Without a persistent, distinct representation of the specific curve, the model faces combinatorial branching at every intersection; consequently, it often "switches tracks" or hallucinates endpoints, revealing an inability to separate overlapping signals or preserve perceptual identity across extended spatial trajectories.

**Failure of Spatial Imagination.** A third bottleneck is the inability to perform mental affine transformations—constructing and manipulating a stable 3D mental model from 2D input (e.g., rotation, projection). While humans solve these tasks via non-verbal mental simulation that retains geometric fidelity, MLLMs attempt to approximate 3D states through descriptive logic: summarizing components and inferring views via language. This fails because language acts as an insufficient coordinate system for volumetric constraints; by substituting physics-compliant rendering with probabilistic text generation, models inevitably hallucinate hidden structures or deduce impossible projections.

**Failure of Visual Pattern Induction.** The fourth challenge is abstract rule acquisition, or the ability to induce generalized transformation rules from sparse examples. Humans readily disentangle abstract relational structures (such as rotation or nesting) from specific visual attributes, but MLLMs frequently conflate style with structure. By approaching induction through attribute enumeration rather than relational mapping, models fixate on spurious surface correlations (e.g., specific colors) instead of the underlying transformation logic, failing to treat objects as variables in a logical operation and missing the compositional generalization required for these tasks.

**The Verbalization Bottleneck** A unifying theme across all four failure modes is what we term the *verbalization bottleneck*. Current MLLMs process visual inputs by translating them into language representations before reasoning. While this approach leverages the powerful reasoning capabilities of large language models, it introduces a fundamental limitation: visual information that cannot be faithfully expressed in language is lost.

Consider the distinction between *describable* and *undescribable* visual properties. Semantic content ("a red car on a road") translates well to language, but geometric relationships (the precise curvature of a boundary, the exact position of an intersection point) resist verbalization. BabyVision specifically targets these undescribable properties, which explains why models that excel at language-heavy benchmarks fail here.

This analysis suggests that progress on early-vision tasks may require architectural innovations that preserve visual information throughout the reasoning process, rather than compressing it into a linguistic bottleneck.

### 6.2 Insight from Training: Can RLVR Help Visual Reasoning?

We have identified large performance gaps in BabyVision, not only among human and frontier models, but also across closed- and open-source models. We are further interested in: how can we develop stronger visual reasoning skills and achieve better scores on BabyVision with open models?

As Reinforcement Learning with Verifiable Rewards (RLVR) has recently demonstrated strong gains in language-reasoning performance for LLMs, we conduct an init study to investigate whether RLVR can similarly improve visual abilities measured by BabyVision. We use Qwen3-VL-8B-Thinking as the base model and apply RLVR fine-tuning. For data collection, we adopt a BabyVision-style pipeline but draw from larger image sources and remove duplicates, yielding 1,400 training examples. The collected data

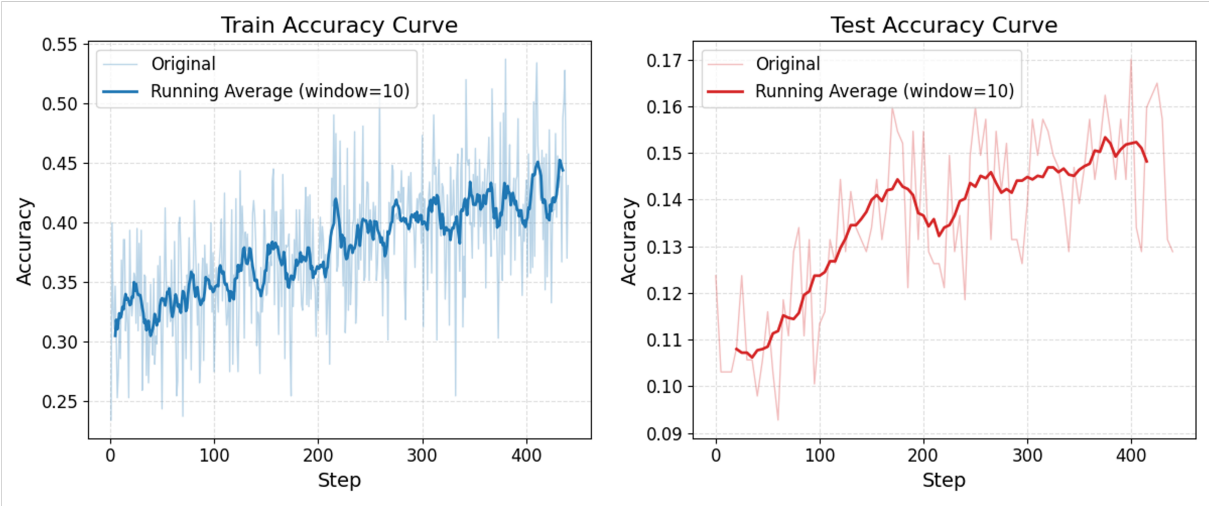



Figure 7: GRPO training dynamics for Qwen3-VL-8B-Thinking. Both training accuracy and held-out test accuracy steadily improve during training.

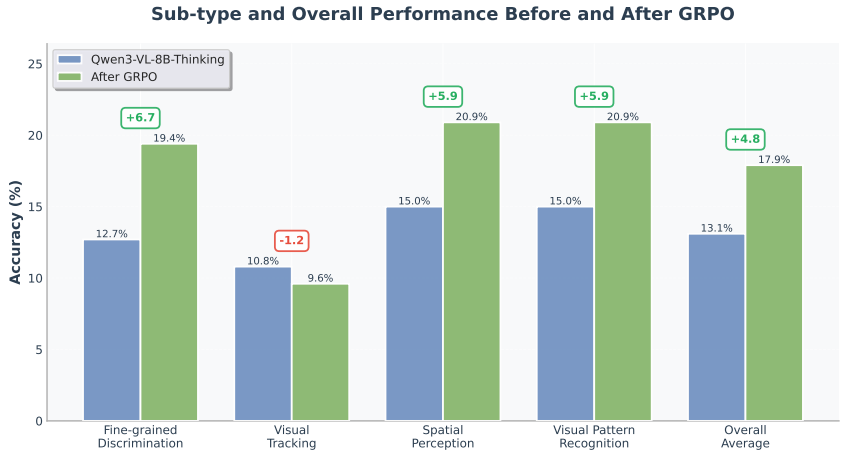



Figure 8: BabyVision accuracy before vs. after RLVR fine-tuning. RLVR yields a +4.8 overall accuracy gain and improves most subtypes.

covers all four major BabyVision task families, yet its difficulty distribution is not completely aligned with BabyVision: the model achieves 34.2% initial accuracy on the RLVR training set, but only 13.1% on BabyVision, when evaluated with the same base model.

**Training Details.** We train Qwen3-VL-8B-Thinking using GRPO on 8 H800 GPUs for approximately 3 days over 18 epochs. To allow for sufficient exploration, we configure the rollout n to 10, the max response length to 16384, and set the clip range (clip higher) to 0.28. Both the global batch size and rollout batch size are set to 64 within the EasyR1 (Zheng et al., 2025; Sheng et al., 2024) framework. We employ an LLM judge to determine the consistency between the answer extracted from `\boxed{}` label and the ground truth, using it as the reward signal to guide the model. We observe that RLVR is effective on the collected training dataset: both training accuracy and held-out test accuracy consistently improve over the course of training as shown in Figure 7.

**Results and Analysis.** The BabyVision performance of Qwen3-VL-8B-Thinking before and after RL fine-tuning is reported in Figure 8. The model achieves a +4.8-point overall accuracy improvement after RLVR training. We also observe consistent gains across most task subtypes, with the sole exception of visual tracking, for which RL fine-tuning yields little to even negative improvement. We think that

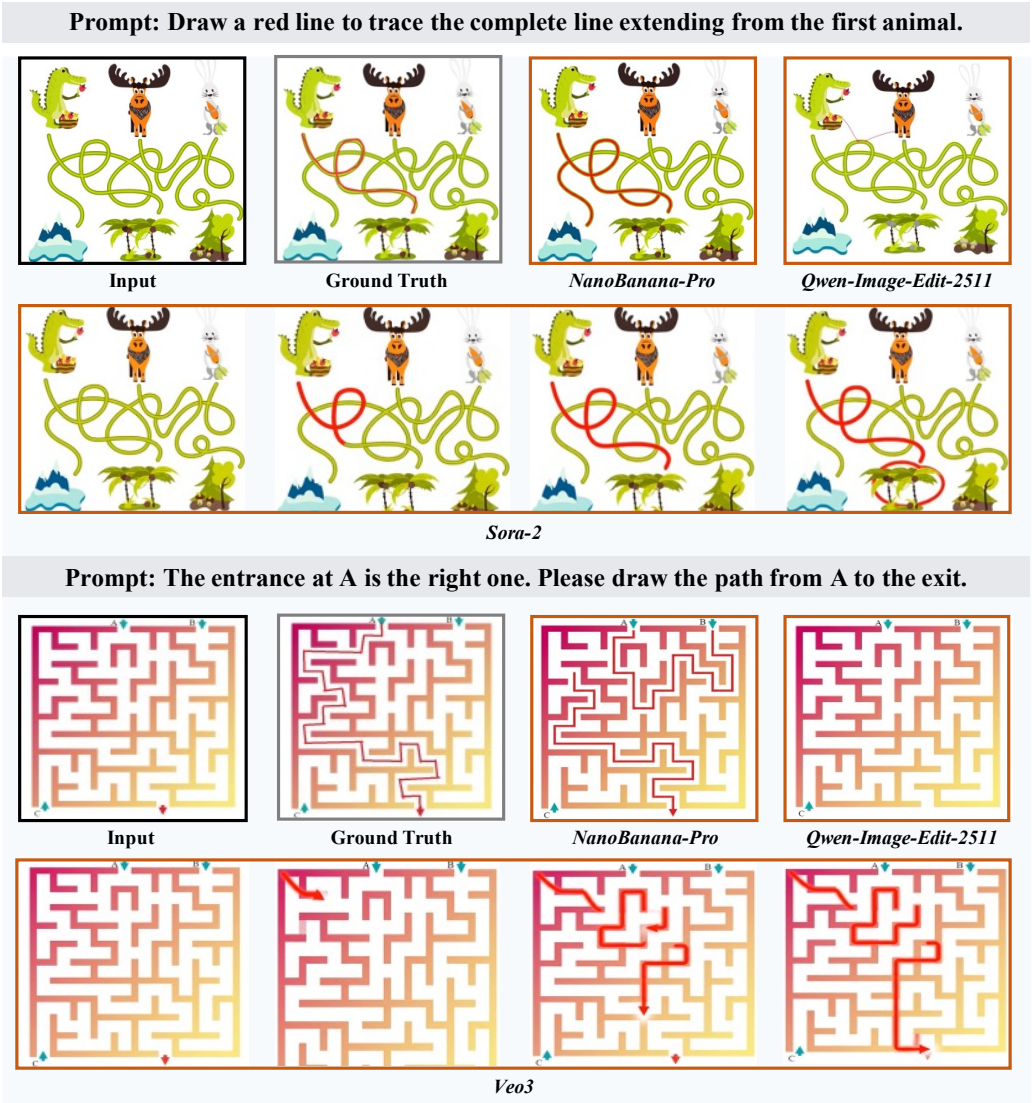



Figure 9: Representative examples of visual reasoning results from different image/video generation models evaluated on BABYVISION-GEN.

this is because visual tracking is the least amenable to verbalization; since RLVR primarily enhances performance by encouraging longer and more structured "thinking-token" reasoning, it provides less benefit on tasks that depend on continuous perceptual tracking rather than language-mediated reasoning. Detailed performance comparison is listed in Table 7 in the Appendix.

### 6.3 Beyond Language: Visual Externalization

A key observation motivating BABYVISION-GEN is that many BabyVision tasks are most naturally solved by *visual externalization*—drawing, marking, or tracing on the image. When humans solve maze or line-tracking problems, they often point or draw trajectories rather than verbalizing directions. This suggests an alternative paradigm: instead of reasoning in language space and outputting a verbal answer, models could reason in visual space and output a visual solution.

Our experiments with generation models (image and video) on BABYVISION-GEN reveal promising signals. As shown in Figure 9, we present examples of visual reasoning results from generation models[2]. Models like Sora-2 and NanoBanana-Pro exhibit human-like visual thinking processes, explicitly drawing

[2]More output images and videos are available at https://unipat.ai/blog/BabyVision

trajectories along paths. However, their solutions still contain noticeable errors, indicating that visual generation alone is insufficient—the generation process must be guided by robust visual understanding.

This points to a compelling research direction: *visual generation models as multimodal reasoners*. If models can learn to manipulate images directly—tracing paths, completing patterns, transforming shapes—they may bypass the verbalization bottleneck entirely. Recent unified multimodal models that natively integrate visual understanding and generation, such as Bagel (ByteDance Seed, 2025), offer a new path forward. Unlike traditional MLLMs that compress visual inputs through a language bottleneck, these natively multimodal architectures maintain visual representations throughout the reasoning process, enabling them to explicitly think in visual space—sketching intermediate steps, highlighting regions, or drawing solution trajectories. Similarly, frontier video generation models like Sora 2 (OpenAI, 2025) and Veo 3 (Google DeepMind, 2025) demonstrate emergent capabilities in modeling physical dynamics and spatial relationships, allowing the generative process itself to serve as a form of visual reasoning (Cai et al., 2025).

## 7 Conclusion

We introduce BabyVision, a benchmark for evaluating early-vision abilities that humans develop before language acquisition. Our evaluation of state-of-the-art MLLMs reveals a striking gap: even the best model (Gemini3-Pro-Preview, 49.7%) falls far short of human performance (94.1%), with consistent deficits across fine-grained discrimination, visual tracking, spatial perception, and pattern recognition. Error analysis identifies four failure modes—loss of fine-grained detail, loss of manifold identity, failure of spatial imagination, and confusion of appearance with structure—all stemming from a verbalization bottleneck that discards undescribable visual information.

These findings suggest that strong performance on language-heavy benchmarks does not imply robust visual foundations. Progress will likely require architectural innovations that preserve visual fidelity throughout reasoning, rather than compressing perception into language. We also introduce BABYVISION-GEN, a generative extension enabling evaluation through visual outputs, and show preliminary evidence that generation models can exhibit human-like visual thinking. BabyVision provides a diagnostic tool for measuring progress toward multimodal systems with genuinely grounded visual intelligence.

## 8 Acknowledgements

We would like to thank Xiaotao Gu (Zhipu AI), Junyang Lin (Alibaba Group), Shuai Bai (Alibaba Group), and Shuhuai Ren (Xiaomi MiMo) for their valuable discussions and insightful feedback throughout this project.

# A Evaluation Details for BabyVision

We employ Qwen3-Max as the judge model to evaluate semantic equivalence between model outputs and ground-truth answers. The judge receives the following prompt:

**EVALUATION PROMPT FOR BABYVISION**

```
You are a careful and strict evaluator.  You will be given:
1.  QUESTION
2.  GROUND TRUTH ANSWER (correct answer)
3.  MODEL OUTPUT (answer from another model)
YOUR GOAL: Determine if the Model Output accurately matches the Ground Truth Answer in
meaning.
MATCHING MEANS: the facts, entities, and key details are equivalent, even if phrasing
differs.
NOT MATCHING MEANS: the Model Output is wrong, incomplete, contains extra incorrect facts,
or changes the meaning.
PROCESS (INTERNAL REASONING):
    1. Read and understand the Question, Ground Truth Answer, and Model Output.
    2. Ignore small wording differences, formatting, or synonyms.
    3. If all factual content matches, conclude True.  Otherwise, conclude False.
IMPORTANT: Think through your decision step-by-step internally before responding.  In
your final output, return only True or False, with no extra text or explanation.
OUTPUT FORMAT: True or False
INPUT:
Question:  {Question}
Ground Truth Answer:  {Groundtruth}
Model Output:  {Model Output}
```

# B Evaluation Details for BabyVision-Gen

## B.1 Inference Protocol

For BABYVISION-GEN, we instruct visual generation models to annotate the original image with their solution. This ensures that the original visual context is preserved while allowing models to demonstrate their reasoning through visual marks. We use the following prompt template:

**Generation Prompt Template**

```
CRITICAL INSTRUCTION: You are a visual annotation assistant.  Your task is to add ONLY
the requested annotations (circles, lines, arrows, text labels) to mark the answer on the
image.
IMPORTANT RULES:
    1. DO NOT modify, redraw, or alter ANY part of the original image content
    2. DO NOT change colors, shapes, positions, or any visual elements of the original
       image
    3. ONLY add overlay annotations (circles, lines, arrows, text) on TOP of the original
       image
    4. The original image must remain 100% intact and unchanged
    5. Use bright, visible colors (red, green, blue) for your annotations so they stand
       out
    6. Keep annotations minimal and precise - only mark what is asked
YOUR TASK: {Question}
REMINDER: Keep the original image EXACTLY as it is.  ONLY add annotation marks (circles,
lines, arrows, or text) to indicate your answer.  Do not redraw or modify any part of the
original image.
```

## B.2 Automatic Evaluation Protocol

We employ Gemini-3-Flash as the judge model to evaluate whether generated images correctly solve the visual reasoning task. The evaluation is conducted by comparing three images: (1) the original input image, (2) the ground-truth solution image annotated by human experts, and (3) the model-generated

image. The judge determines whether the generated solution matches the ground truth.

We design type-specific evaluation criteria for each subtype to ensure accurate assessment. The general prompt structure is:

**Automatic Evaluation Prompt**

```
You are evaluating an AI-generated image for a visual reasoning task.
TASK TYPE: {task_type}
SUBTYPE: {subtype}
GENERATION INSTRUCTION: "{generation_prompt}"
You are provided with THREE images:
      • Image 1 (Input): The original question/puzzle image
      • Image 2 (Ground Truth): The CORRECT answer showing what the result SHOULD look
        like
      • Image 3 (Generated): The AI-generated result to be evaluated
Compare Image 3 (Generated) with Image 2 (Ground Truth) to determine if they show the
SAME answer.
[Type-Specific Criteria Inserted Here]
DECISION RULES:
      • TRUE: Generated image shows the EXACT SAME answer as Ground Truth
      • FALSE: Generated shows a DIFFERENT answer, NO answer, or UNCLEAR answer
IMPORTANT:
      • Focus ONLY on whether the ANSWER matches, ignore style differences
      • A marking on a DIFFERENT element/option = FALSE
      • A path taking a DIFFERENT route = FALSE
      • A DIFFERENT number or character = FALSE
      • Missing required answer = FALSE
Respond with ONLY one word: "True" or "False"
```

## B.3 Type-Specific Evaluation Criteria

We customize evaluation criteria for each task subtype. The complete criteria are provided below:

### B.3.1 Fine-grained Discrimination

**Find the Different**

```
TASK: Find the unique/different element among many similar elements.
CRITERIA: Is the circle on the SAME grid position as ground truth?
- Circle on DIFFERENT element = FALSE
- No circle visible = FALSE
```

**Find the Same**

```
TASK: Find identical elements or matching figures.
CRITERIA: Are ALL marked elements the same as in ground truth?
- Circles on DIFFERENT elements = FALSE
- Missing circles = FALSE
```

**Find the Shadow**

```
TASK: Find the shadow/silhouette that matches the colored figure.
CRITERIA: Is the SAME option circled?
```

**Reconstruction / 2D Pattern Completion / Pattern and Color Completion**

```
TASK: Find/select the correct option.
CRITERIA: Is the SAME option circled/selected?
```

**Count Same Patterns / Count Clusters**

```
TASK: Count patterns or fill in numbers.
CRITERIA: Do the markings/numbers match ground truth exactly?
```

#### B.3.2 Visual Tracking

**Maze**

```
TASK: Draw a path through the maze.
CRITERIA: Does the path follow the EXACT SAME route as ground truth?
- Different route = FALSE
- No visible path = FALSE
```

**Connect the Lines**

```
TASK: Trace a line following the continuous path.
CRITERIA: Does the traced line follow the SAME path as ground truth?
```

**Metro Map**

```
TASK: Draw the shortest path between metro stations.
CRITERIA: Does the path follow the EXACT SAME route as ground truth?
```

**Recognize Numbers and Letters**

```
TASK: Fill in letters/numbers in blanks.
CRITERIA: Are the EXACT SAME characters filled in each blank?
```

#### B.3.3 Spatial Perception

**3D Views / 3D Cube Unfold / Paper Folding / 3D Pattern Completion**

```
TASK: Select the correct option for spatial reasoning.
CRITERIA: Is the SAME option circled?
```

**Count 3D Blocks**

```
TASK: Count cubes in a 3D structure.
CRITERIA: Is the EXACT SAME number written?
```

#### B.3.4 Visual Pattern Recognition

**Logic Patterns / Mirroring Patterns / Overlay Patterns / Rotation Patterns**

```
TASK: Identify pattern and select correct option.
CRITERIA: Is the SAME option circled?
```

### B.4 BabyVision-Gen Statistics

Table 4 presents the detailed distribution of questions across categories and subtypes in BABYVISION-GEN.

### B.5 Human Evaluation and Validation

To validate the reliability of our automatic evaluation pipeline, we conducted comprehensive human evaluation on all 280 NanoBanana-Pro outputs. PhD-level annotators independently judged each

| Category | Subtype | # |
|---|---|---|
| Fine-grained Discrimination | 2D Pattern Completion | 19 |
| | Count Clusters | 9 |
| | Count Same Patterns | 31 |
| | Find the Different | 16 |
| | Find the Same | 10 |
| | Reconstruction | 13 |
| | Find the Shadow | 14 |
| | Pattern and Color Completion | 16 |
| | *Subtotal* | *128* |
| Visual Tracking | Connect the Lines | 12 |
| | Maze | 18 |
| | Metro Map | 12 |
| | Recognize Numbers and Letters | 13 |
| | *Subtotal* | *55* |
| Spatial Perception | 3D Cube Unfold | 12 |
| | 3D Pattern Completion | 18 |
| | 3D Views | 19 |
| | Count 3D Blocks | 5 |
| | Paper Folding | 5 |
| | *Subtotal* | *59* |
| Visual Pattern Recognition | Logic Patterns | 11 |
| | Mirroring Patterns | 3 |
| | Overlay Patterns | 15 |
| | Rotation Patterns | 9 |
| | *Subtotal* | *38* |
| **Total** | | **280** |

Table 4: Distribution of questions across categories and subtypes in BABYVISION-GEN. One subtype from BABYVISION is excluded as it does not naturally admit a generative solution format.

generated image against the ground truth.

**Overall Agreement.** The automatic scorer achieves **96.1%** agreement with human judgments (269/280 samples). The confusion matrix is shown in Table 5.

**Evaluation Metrics.** The key metrics are summarized below:

- **Human Evaluation Accuracy:** 75/280 (26.8%)
- **Auto Evaluation Accuracy:** 70/280 (25.0%)
- **Precision** (auto correct when human correct): 0.957
- **Recall** (finds human correct cases): 0.893
- **F1 Score:** 0.924

| | Human: Correct | Human: Incorrect | Total |
|---|---|---|---|
| **Auto: Correct** | 67 | 3 | 70 |
| **Auto: Incorrect** | 8 | 202 | 210 |
| **Total** | 75 | 205 | 280 |

Table 5: Confusion matrix for automatic vs. human evaluation on NanoBanana-Pro outputs. Green cells indicate agreement; red cells indicate disagreement.

# C Additional Experimental Results

| Type | Sub-Type | Qwen3VL-235B-Thinking | | Qwen3VL-235B-Instruct | | Qwen3VL-32B-Thinking | | Qwen3VL-8B-Thinking | | Qwen3VL-4B-Thinking | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ | Avg ($\mu$) ↑ | Std ($\sigma$) ↓ |
| Fine-grained Discrimination | 2D Pattern Completion | 25.0 | 10.8 | 31.7 | 10.3 | **38.3** | 4.7 | 23.3 | 8.5 | 20.0 | 8.2 |
| | Count Clusters | 27.8 | 0.0 | 16.7 | 4.5 | **35.2** | 5.2 | 20.4 | 2.6 | 16.7 | 4.5 |
| | Count Same Patterns | **15.2** | 3.6 | 11.4 | 0.0 | 11.4 | 2.3 | 2.9 | 2.3 | 3.8 | 1.4 |
| | Find the Different | **16.7** | 3.0 | 4.2 | 3.0 | 6.3 | 5.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Find the Same | **23.5** | 8.3 | 11.8 | 4.8 | 7.8 | 2.8 | 3.9 | 2.8 | 11.8 | 4.8 |
| | Reconstruction | **33.3** | 3.4 | 21.4 | 5.8 | 23.8 | 6.7 | 19.1 | 3.4 | 19.1 | 3.4 |
| | Find the Shadow | 15.9 | 5.4 | **24.6** | 2.1 | **24.6** | 5.4 | 11.6 | 2.1 | 8.7 | 6.2 |
| | Pattern and Color Completion | **38.3** | 9.4 | 10.0 | 4.1 | 25.0 | 7.1 | 26.7 | 4.7 | 23.3 | 6.2 |
| | **Sub-Type Overall Results** | **23.3** | 0.9 | 16.4 | 1.3 | 21.1 | 1.8 | 12.7 | 0.6 | 12.1 | 2.3 |
| Visual Tracking | Connect the Lines | 5.3 | 4.3 | 10.5 | 7.4 | 10.5 | 4.3 | **12.3** | 2.5 | 10.5 | 0.0 |
| | Lines Observation | **7.4** | 5.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Maze | 11.7 | 6.2 | 11.7 | 4.7 | 5.0 | 4.1 | 20.0 | 0.0 | **28.3** | 6.2 |
| | Metro Map | 13.9 | 3.9 | 13.9 | 3.9 | **19.4** | 3.9 | 5.6 | 3.9 | 5.6 | 7.9 |
| | Recognize Numbers and Letters | **36.2** | 7.4 | 24.6 | 2.1 | 17.4 | 3.6 | 8.7 | 0.0 | 17.4 | 3.6 |
| | **Sub-Type Overall Results** | **16.9** | 4.3 | 15.3 | 0.6 | 8.8 | 2.5 | 10.8 | 1.0 | 14.9 | 1.1 |
| Spatial Perception | 3D Cube Unfold | **19.4** | 7.9 | 13.9 | 3.9 | 16.7 | 6.8 | 5.6 | 3.9 | 8.3 | 6.8 |
| | 3D Pattern Completion | 29.6 | 6.9 | **31.5** | 5.2 | 24.1 | 5.2 | 22.2 | 4.5 | 25.9 | 9.4 |
| | 3D Views | **29.6** | 8.0 | **29.6** | 3.0 | 11.1 | 8.0 | 21.0 | 8.7 | **29.6** | 3.0 |
| | Count 3D Blocks | 10.6 | 2.1 | **21.2** | 9.3 | 7.6 | 2.1 | 7.6 | 5.7 | 12.1 | 4.3 |
| | Paper Folding | 8.3 | 6.8 | 11.1 | 10.4 | 5.6 | 3.9 | **13.9** | 7.9 | 8.3 | 6.8 |
| | **Sub-Type Overall Results** | 20.9 | 2.7 | **23.4** | 2.7 | 12.8 | 1.9 | 15.0 | 3.2 | 19.1 | 1.4 |
| Visual Pattern Recognition | Logic Patterns | 19.1 | 6.7 | **21.4** | 5.8 | **21.4** | 5.8 | 11.9 | 3.4 | 14.3 | 5.8 |
| | Mirroring Patterns | 26.7 | 9.4 | 20.0 | 0.0 | **30.0** | 8.2 | 6.7 | 4.7 | 10.0 | 8.2 |
| | Overlay Patterns | **33.3** | 7.3 | **33.3** | 2.8 | 19.6 | 5.6 | 7.8 | 5.6 | 11.8 | 0.0 |
| | Rotation Patterns | 40.0 | 0.0 | 43.3 | 12.5 | **46.7** | 9.4 | 40.0 | 14.1 | 23.3 | 12.5 |
| | **Sub-Type Overall Results** | **29.4** | 5.8 | **29.4** | 4.2 | 27.5 | 3.2 | 15.0 | 1.9 | 14.4 | 2.5 |
| All | **Overall Results** | **22.2** | 1.0 | 19.5 | 1.4 | 17.4 | 2.0 | 13.1 | 1.1 | 14.6 | 1.2 |

Table 6: **Performance (Avg@3) of Qwen3VL Instruct/Thinking on BabyVision.** The best results for each question type are marked in **bold**. Reported values represent the average Pass@1 accuracy across three random runs, accompanied by the standard deviation.

| Category | Subtype | Qwen3-VL-8B-Thinking (%) | After GRPO (%) | Δ (%) |
|---|---|---|---|---|
| **Fine-grained Discrimination** | 2D Pattern Completion | 23.3 | **45.0** | +21.7 |
| | Count Clusters | **20.4** | **20.4** | +0.0 |
| | Count Same Patterns | 2.9 | **5.7** | +2.8 |
| | Find the different | 0.0 | **2.1** | +2.1 |
| | Find the same | 3.9 | **15.7** | +11.8 |
| | Reconstruction | 19.1 | **23.8** | +4.7 |
| | Find the shadow | 11.6 | **17.4** | +5.8 |
| | Pattern and Color Completion | 26.7 | **33.3** | +6.6 |
| | **Sub-Type Overall Results** | 12.7 | **19.4** | +6.7 |
| **Visual Tracking** | Connect the lines | **12.3** | 3.5 | -8.8 |
| | Lines Observation | **0.0** | **0.0** | +0.0 |
| | Maze | **20.0** | 15.0 | -5.0 |
| | Metro map | 5.6 | **13.9** | +8.3 |
| | Recognize numbers and letters | 8.7 | **11.6** | +2.9 |
| | **Sub-Type Overall Results** | **10.8** | 9.6 | -1.2 |
| **Spatial Perception** | 3D Cube Unfold | 5.6 | **16.7** | +11.1 |
| | 3D Pattern Completion | 22.2 | **37.0** | +14.8 |
| | 3D Views | 21.0 | **25.9** | +4.9 |
| | Count 3D blocks | 7.6 | **9.1** | +1.5 |
| | Paper Folding | **13.9** | 11.1 | -2.8 |
| | **Sub-Type Overall Results** | 15.0 | **20.9** | +5.9 |
| **Visual Pattern Recognition** | Logic Patterns | 11.9 | **26.2** | +14.3 |
| | Mirroring Patterns | 6.7 | **20.0** | +13.3 |
| | Overlay Patterns | 7.8 | **13.7** | +5.9 |
| | Rotation Patterns | **40.0** | 26.7 | -13.3 |
| | **Sub-Type Overall Results** | 15.0 | **20.9** | +5.9 |
| **Overall Average Accuracy** | | 13.1 | **17.9** | +4.8 |

Table 7: GRPO results by subtype (in %). Vertical rules separate category, subtype, and metric groups. For each row, the larger value between the original and GRPO model is bolded.

## References


Anthropic. Introducing claude opus 4.5, 2025. URL https://www.anthropic.com/news/claude-opus-4-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Renée Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

Oliver Braddick and Janette Atkinson. Development of human visual function. *Vision research*, 51(13): 1588–1609, 2011.

ByteDance Seed. Bagel: Unified model for image understanding and generation, 2025. URL https://bagel-ai.org/. Accessed: 2025-01-09.

Zefan Cai, Haoyi Qiu, Tianyi Ma, Haozhe Zhao, Gengze Zhou, Kung-Hsiang Huang, Parisa Kordjamshidi, Minjia Zhang, Wen Xiao, Jiuxiang Gu, Nanyun Peng, and Junjie Hu. Mmgr: Multi-modal generative reasoning, 2025. URL https://arxiv.org/abs/2512.14691.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024a.

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. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024b.

Claire Golomb. *The child's creation of a pictorial world*. Psychology Press, 2003.

Google DeepMind. Veo 3: Google's next-generation video model, 2025. URL https://deepmind.google/technologies/veo/. Accessed: 2025-01-09.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation, 2025. URL https://arxiv.org/abs/2407.00468.

Scott P Johnson. Development of visual perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5): 529–541, 2010.

By A Karmiloff-Smith. Beyond modularity: A developmental perspective on cognitive science. *European journal of disorders of communication*, 29(1):95–105, 1994.

Philip J Kellman, ME Arterberry, W Damon, RM Lerner, D Kuhn, RS Siegler, et al. Infant visual perception. 2006.

Rhoda Kellogg. *Analyzing Children's Art*. Mayfield Publishing Company, 1969.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

OpenAI. Update to gpt-5 system card: Gpt-5.2, 2025. URL https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf.

OpenAI. Sora 2: Advanced video generation, 2025. URL https://openai.com/sora. Accessed: 2025-01-09.

ByteDance Seed. Seed1.6 tech introduction, 2025. URL https://seed.bytedance.com/en/seed1_6.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Elizabeth S Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000.

StepFun, :, Bin Wang, Bojun Wang, Changyi Wan, et al. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding, 2025. URL https://arxiv.org/abs/2507.19427.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, et al. Mimo-vl technical report, 2025a. URL https://arxiv.org/abs/2506.03569.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

GLM-V Team. Glm-4.6v: Open source multimodal models with native tool use, 2025a. URL https://z.ai/blog/glm-4.6v.

HLE Team. Humanity's last exam, 2025b. URL https://arxiv.org/abs/2501.14249.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report, 2025b. URL https://arxiv.org/abs/2504.07491.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *arXiv preprint arXiv:2505.24173*, 2025.