# KIDVIS: BENCHMARKING FUNDAMENTAL VISUAL PRIMITIVES IN MULTIMODAL LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Despite the remarkable proficiency of Multimodal Large Language Models (MLLMs) in high-level reasoning, it remains questionable whether they possess human-like foundational visual abilities. To investigate this question, we present KidVis, a diagnostic benchmark grounded in cognitive developmental psychology rather than standard semantic recognition. KidVis deconstructs visual perception into six atomic primitives, including Visual Closure, Spatial Orientation, and Visual Tracking, and instantiates them through controlled, low-semantic, motor-free tasks. We evaluate 22 leading MLLMs against an age-specific human reference baseline. Our study uncovers a fundamental disconnect: while children achieve near-perfect accuracy of 95.31%, the strongest evaluated model, GPT-5, reaches 67.33%. We further observe a visual scaling paradox: within the evaluated model families, larger parameter counts do not consistently translate into stronger foundational visual perception, and may even coincide with semantic interference on low-level tasks. These findings suggest that current MLLMs still face systematic limitations in basic visual grounding, highlighting the need for evaluation and model design beyond semantic-heavy benchmarks and simple parameter scaling.

## 1 INTRODUCTION

With the rapid evolution of Multimodal Large Language Models, artificial intelligence appears to be approaching the frontier of Artificial General Intelligence (AGI) (OpenAI, 2023; Team et al., 2025a; Bubeck et al., 2023). Contemporary MLLMs demonstrate expert-level proficiency not only in fine-grained image descriptions and open-domain Visual Question Answering (VQA) but also in high-level cognitive tasks, such as parsing complex chart logic, assisting in medical diagnosis, and performing spatiotemporal planning as embodied agents (Driess et al., 2023; Li et al., 2023a; Xia et al., 2023). However, a fundamental paradox underlies these developments: while MLLMs can articulate the philosophical connotations of art or generate code from sketches via Chain-of-Thought, their robustness in elementary perceptual tasks remains surprisingly fragile (Li et al., 2025; Tong et al., 2024). This phenomenon echoes Moravec's Paradox: high-level semantic reasoning has become computationally accessible, yet sensorimotor tasks that are instinctive to human children continue to be significant obstacles for machines. This raises a key issue: Do current large-scale models truly perceive physical reality, or do they merely mimic perceptual behaviors driven by high-dimensional statistical correlations?

To investigate these limitations, substantial research efforts have studied the failure modes of MLLMs. For instance, studies on Object Hallucination attribute the generation of non-existent objects to an over-reliance on linguistic priors learned during pre-training, rather than faithful visual grounding (Li et al., 2023c; Guan et al., 2024). Similarly, experiments reveal that models exhibit severe biases in basic recognition tasks when visual stimuli conflict with inherent knowledge bases (Vo et al., 2025). While domain-specific benchmarks (Lu et al., 2024) and comprehensive evaluation frameworks (Xu et al., 2025) have exposed structural bottlenecks in mathematical reasoning and general multimodal understanding, they primarily focus on "adult-level" high-level semantic reasoning or provide fragmented diagnostic perspectives on specific symptoms. Crucially, existing works often fail to systematically decompose vision into constitutive cognitive primitives or benchmark these deficits against the foundational stages of human visual development.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this study, we hypothesize that MLLM failures in complex scenarios stem from deficits in core visual primitives—capabilities that are typically basically developed in 6-7-year-old children (Siu & Murphy, 2018). Drawing on cognitive development theories (Carroll, 1993; Colarusso & Hammill, 1972), we categorize these capabilities into six dimensions: Discrimination, Memory, Concentration, Tracking, Spatial, and Closure, as illustrated in Figure 1. To test this, we introduce KidVis, a benchmark comprising ten categories of child-level visual tasks. Unlike existing datasets rich in semantic context, these tasks serve as diagnostic baselines: they minimize the interference of motor skills and semantic reasoning, strictly demanding precise visual processing mechanisms. We conduct an extensive zero-shot evaluation of 22 mainstream MLLMs, encompassing both proprietary models and open-source series, comparing them against an age-specific human reference established by 6-7-year-old children under standardized evaluation conditions.

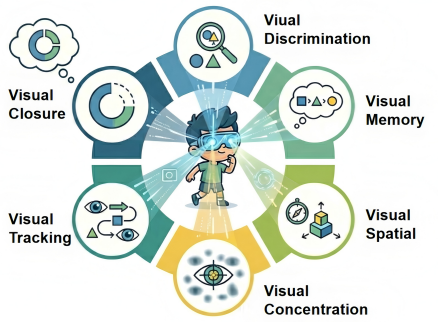



Figure 1: The scheme of KidVis benchmark for comprehensively assessing the visual perceptual capabilities.

Our contributions are: (1) We introduce KidVis, a child-level, low-semantic, and motor-free benchmark that maps six foundational visual primitives to ten controlled tasks for evaluating MLLMs beyond high-level semantic reasoning. (2) Across 22 mainstream MLLMs, KidVis reveals a substantial gap between current models and a human reference baseline, especially on fine-grained, low-semantic visual tasks. (3) Our rationale analysis shows that visual grounding failures can derail otherwise coherent reasoning chains, indicating that strong language reasoning does not guarantee stable visual perception. (4) Several model families show non-monotonic or saturated performance trends, suggesting that parameter scaling alone may be insufficient for foundational visual perception.

## 2 Related Works

### 2.1 Multimodal Large Language Models

Recent breakthroughs in Large Language Models (LLMs) have driven the emergence of MLLMs, extending the reasoning prowess of LLMs into the visual modality. Pioneering architectures like Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023b) introduced Perceiver and Q-Former modules to align frozen visual encoders with language models via trainable projection layers, establishing the dominant paradigm of visual-to-text feature alignment. Subsequent open-source projects have focused on Visual Instruction Tuning. Models such as LLaVA (Liu et al., 2023) and MiniGPT-4 (Zhu et al., 2023) leverage high-quality multimodal instruction data synthesized by GPT-4, significantly enhancing instruction-following capabilities and conversational interactivity. On the proprietary MLLMs front, models like GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2025a) demonstrate exceptional cross-modal generalization through large-scale end-to-end pre-training.

### 2.2 Benchmarks for MLLMs

The evolution of MLLMs has required a parallel advancement in evaluation benchmarks. Traditional benchmarks like VQA v2.0 (Goyal et al., 2017) and COCO Caption (Chen et al., 2015), which target specific tasks, have become largely saturated by modern models. Consequently, comprehensive benchmarks such as MME (Fu et al., 2025), MMBench (Liu et al., 2024), and MMMU (Yue et al., 2024) have emerged to assess a broader spectrum of capabilities, including commonsense reasoning, knowledge retrieval, and OCR. Furthermore, addressing the reliability of automated evaluation, AIBench (Zhang et al., 2025a;b) introduces a trustworthy evaluation framework under the 45° Law. More recently, targeted evaluation frameworks have uncovered critical deficiencies in large-scale models. POPE (Li et al., 2023c) and HallusionBench (Guan et al., 2024) expose severe Object Hallucination—the generation of plausible yet non-existent details. Meanwhile, MathVista (Lu et al., 2024) highlights limitations in visual mathematical reasoning. However, these benchmarks often

Table 1: Comparison of KidVis with existing MLLM evaluation benchmarks.

| Benchmark | Primary evaluation regime | Primitive-perception evaluation features | | | |
|---|---|---|---|---|---|
| | | Isolated perception | Primitive decomposition | Child level | Low semantic |
| MME | General multimodal evaluation | × | × | × | × |
| MMBench | General multimodal evaluation | × | × | × | × |
| MMMU | Expert-level reasoning | × | × | × | × |
| MathVista | Visual math reasoning | × | × | × | × |
| HallusionBench | Diagnostic benchmark | × | × | × | × |
| **KidVis (ours)** | **Basic visual perception** | ✓ | ✓ | ✓ | ✓ |

confuse perception with higher-order reasoning. In contrast, our work isolates core visual primitives, serving as a complementary diagnostic tool to existing semantic-heavy evaluations. Table 1 positions KidVis relative to representative MLLM evaluation benchmarks.

### 2.3 Theoretical foundations of visual perception

The systematic study of visual perception possesses a deep theoretical foundation within psychology and cognitive science. In seminal cognitive factor analysis, Carroll (Carroll, 1993) deconstructed human visual behavior into distinct factors, including Visualization, Spatial Relations, Closure Speed, and Perceptual Speed. Based on this theoretical framework, standardized assessment systems for visual development have been widely adopted. These classical paradigms typically employ multidimensional subtests. For instance, DTVP-2 (Hammill et al., 1993) assesses dimensions such as Eye-Hand Coordination and Spatial Relations, while TVPS-4 (Martin, 2017) and MVPT-3 (Colarusso & Hammill, 1972) focus on motor-free components like Visual Discrimination, Memory, and Closure. These mature classification systems, originally designed for the human visual system, provide the theoretical grounding and methodological basis for our proposed evaluation of visual primitives in MLLMs.

## 3 KidVis benchmark

### 3.1 Classification of visual capabilities

To strictly assess the foundational visual perceptual capabilities of MLLMs, we move beyond high-level semantic reasoning tasks. Drawing on Carroll's Three-Stratum Theory of Cognitive Capabilities and MVPT-3, we establish a classification method for six atomic visual primitives, grounded in developmental psychology.

**Visual Concentration (VC):** The capabilities to sustain visual attention on specific stimuli while actively suppressing irrelevant objects from complex backgrounds (Warren, 1993). This serves as the prerequisite for valid visual processing.

**Visual Tracking (VT):** The oculomotor ability to execute smooth pursuit along a path or trajectory, essential for maintaining topological continuity in reading and motion perception (Warren, 1993).

**Visual Discrimination (VD):** The ability to distinguish fine-grained features, such as subtle variations in color, shape, position or texture, to categorize objects distinctively (Brown, 2011).

**Visual Memory (VM):** Includes Visual Short-term Memory for buffering transient visual details (sequence, location) to support sequential reasoning (Newton & McGrew, 2010), and Long-term Memory for retrieving encoded visual prototypes (Brady et al., 2008; Standing, 1973).

**Visual Spatial (VS):** Entails perceiving spatial relationships (e.g., orientation, relative position) and performing mental manipulation (e.g., rotation) within a defined coordinate system (Colarusso & Hammill, 1972; Newton & McGrew, 2010).

**Visual Closure (VCl):** The ability to identify an unknown visual object when only presented with a visual stimulus that is obscure, disconnected, incomplete, or vague (Newton & McGrew, 2010).

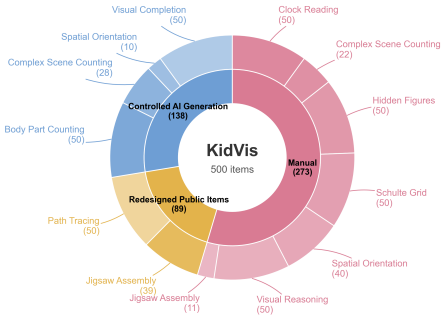



Figure 2: Overview of the data construction routes in KidVis.

Table 2: Mapping of KidVis tasks to paired target visual capabilities.

| **Task name** | **Cap. 1** | **Cap. 2** |
|---|---|---|
| Body Part Counting | VD | VM |
| Clock Reading | VS | VD |
| Complex Scene Counting | VC | VD |
| Hidden Figures | VT | VCl |
| Schulte Grid | VC | VT |
| Spatial Orientation | VS | VD |
| Visual Completion | VM | VCl |
| Visual Reasoning | VD | VM |
| Jigsaw Assembly | VCl | VS |
| Path Tracing | VT | VC |

## 3.2 Task instantiation and dataset construction

### 3.2.1 Evaluation construct.

The capability taxonomy and task paradigms are grounded in established cognitive and motor-free visual-perception frameworks and were defined before model evaluation, rather than selected based on model performance. Each task is assigned to the two capabilities most directly required for successful completion (Table 2). This dual-factor mapping represents principal perceptual requirements rather than process-pure isolation. The tasks are low-semantic and motor-free to reduce semantic and motor confounds.

### 3.2.2 Data construction and quality assurance.

KidVis items are constructed from task prototypes grounded in child visual ability exercises and standardized visual assessment paradigms. All question images are rendered at a resolution of at least 2K to preserve fine-grained perceptual details. As shown in Figure 2, we instantiate the final items through three routes: manual authoring for geometric and logic-based tasks requiring precise spatial control, redesigning publicly available visual cognitive materials permitted for academic use, and controlled AI-assisted generation to increase appearance diversity while preserving the intended perceptual structure. For each item, the gold answer is determined by a deterministic construction rule and independently verified through manual review. Distractors are designed to be visually plausible but perceptually incorrect, so that solving the item requires the targeted visual mechanism rather than semantic guessing. Three human reviewers cross-verify each item, and samples with ambiguous answers, multiple plausible interpretations, unintended cues, or recognizable generation artifacts are rejected and regenerated.

### 3.2.3 Task instantiation.

The task instances are described as follows and illustrated in Figure 3.

**Task 1: Body Part Counting.** This task fuses abnormal biological features (e.g., polydactyly) with naturalistic textures. Participants must **separate** specific features from background noise and **contrast** them against biological prototypes retrieved from long-term memory.

**Task 2: Clock Reading.** Participants must **distinguish** the hour and minute hands based on subtle length and width differences, and **map** their precise angular positions to the clock face reference frame to acquire temporal information.

**Task 3: Complex Scene Counting.** This task requires **Figure-Background Separation** under challenging conditions, such as occlusion or noise. Participants must sustain high visual concentration to **filter out** interference while employing fine-grained discrimination to identify targets based on local cues.

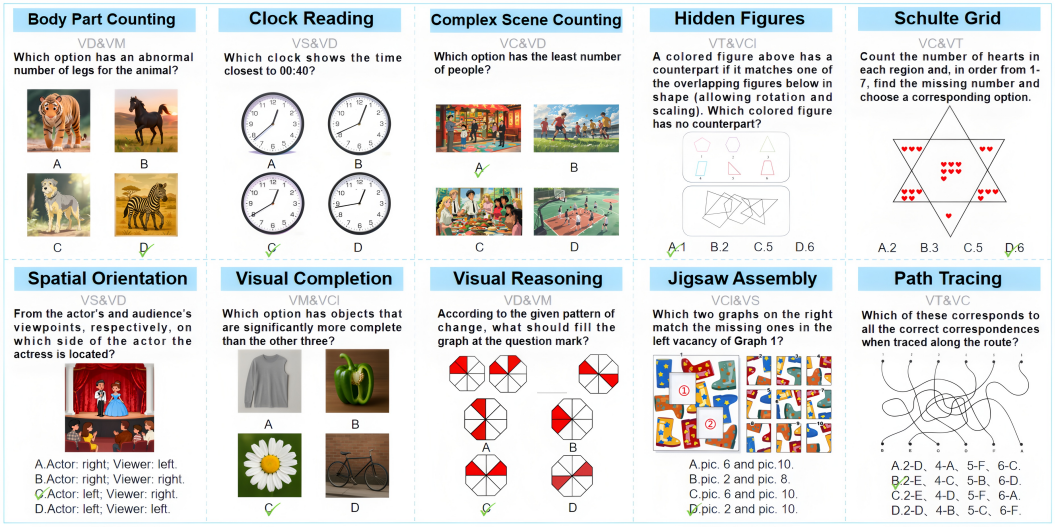



Figure 3: Representative task instances in KidVis.

**Task 4: Hidden Figures.** Participants need **track** a specific geometric shape embedded within a complex network of intersecting lines. The **Visual Closure mechanism** is invoked to mentally complete the truncated segments, allowing participants to perceive the shape as a coherent Gestalt.

**Task 5: Schulte Grid.** Participants are required to **search for** a strictly ascending numerical sequence (e.g., 1-16) within a scrambled grid. Participants must utilize Working Memory to maintain the search target while performing **rigorous saccadic eye movements** to locate the next target, inhibiting interference from neighboring digits.

**Task 6: Spatial Orientation.** Sub-tasks involve: (1) Determining relative object positions (e.g., coverage, rotation) within a reference frame; (2) 3D Structural Completion (calculating missing blocks); (3) Viewpoint transformation (e.g., identifying top-down views). Participants must construct 2D/3D coordinate systems to facilitate **mental rotation** and **perspective taking**.

**Task 7: Visual Completion.** Presented with objects containing structural defects, participants must **retrieve the standard prototype** from memory. The cognitive system detects interrupted contours and performs **modal completion** to instinctively fill in missing parts, thereby identifying the complete or defective object.

**Task 8: Visual Reasoning.** Sub-tasks include pattern completion and Visual Sudoku. Participants are required to utilize **Visual Discrimination** to isolate features and detect state changes. Crucially, **Visual Memory** acts as a Hypothesis Buffer, retaining previous states to infer the underlying rules of transformation.

**Task 9: Jigsaw Assembly.** Sub-tasks involve selecting fragments based on content compatibility, geometric decomposition, and boundary topology. **Visual Closure** acts as the generative prior, envisioning the complete state. **Visual Spatial** serves as the execution engine, manipulating fragments in a mental buffer to verify geometric fit and texture consistency.

**Task 10: Path Tracing.** Sub-tasks involve distinguishing entwined lines and following spiral mazes. Participants must execute smooth pursuit along thin paths, requiring **high concentration** to suppress interference and **maintain continuity** across long-range dependencies.

# 4 Experiments

## 4.1 Experimental setup

**Multimodal large language models.** Based on KidVis benchmark, we evaluate 22 state-of-the-art MLLMs, spanning both proprietary and open-source models. The suite includes leading proprietary

MLLMs (e.g., GPT-5, Gemini-3.1-Pro) and comprehensive open-source MLLMs (e.g., Qwen3-VL, InternVL3.5, LLaVA-Next), covering parameter scales from 2B to 110B. All models are evaluated in a zero-shot setting using default configurations to assess their visual perceptual capabilities without task-specific tuning.

**Human reference baseline.** We recruited ten children aged 6–7 years as an age-specific human reference; none had specialized KidVis-related training. We interpret results under standardized conditions as a reference, not a physiological ceiling or a developmentally matched MLLM comparison; details are in Appendix F.

**Evaluation metrics.** We employ a two-tiered metric system to quantify performance. **Task Score (0-100):** This is calculated strictly based on the correctness of the final option. For each of the 50 questions per task, participants award 2 points for a correct response and 0 points for an incorrect one, resulting in a normalized 100-point scale. **Capability Score (0-100):** This is obtained as the weighted average of tasks probing the same visual perceptual capabilities. For instance, the capability score for Visual Tracking is computed as the weighted average of results from Task 4, Task 5, and Task 10.

### 4.2 Main results

The quantitative results on our KidVis benchmark are shown in Table 3 and Figure 4, which reveal a profound separation between semantic reasoning and visual perception.

**Human performance advantage.** The human group demonstrates near-saturation performance, achieving a composite score of 95.31 across all capabilities. The relatively narrow participant-level dispersion (SD = 1.64; observed range: 93.16–98.18) indicates that the high aggregate performance was consistent across participants rather than being driven by a small subset.

**Proprietary MLLMs gap.** In contrast, GPT-5 attains 67.33, about 28 points below the human baseline. Performance is non-monotonic even within proprietary MLLMs: GPT-5.4 scores 50.31, suggesting that improvements in general-purpose multimodal models do not automatically translate into stronger foundational visual perception.

**Open-Source MLLMs lag substantially.** Leading series such as Qwen3-VL and InternVL3.5 fall behind the human baseline substantially, with average scores typically below 40. In tasks requiring precise spatial manipulation such as Visual Tracking, many open-source models remain close to chance-level performance.

## 5 Analysis

### 5.1 The visual scaling law paradox

A common principle in modern Artificial Intelligence is the Scaling Law, which dictates that performance improves consistently with model size. However, our evaluation across three major open-source families (Qwen3-VL, InternVL3.5, LLaVA-Next) reveals a surprising exception: visual perception does not scale linearly with parameter count, as shown in Figure 5.

We observe that the Qwen3-VL series reaches its best overall performance at 8B, followed by a decline at 30B. We hypothesize that this non-monotonic trend may reflect semantic interference: as model capacity increases, the language decoder may over-interpret simple perceptual inputs, producing semantically plausible but visually ungrounded answers. In contrast, InternVL3.5 and LLaVA-Next show performance saturation, where larger parameter counts yield only marginal gains in visual perception. These results suggest that, although many reasoning abilities benefit from scaling, foundational visual perception may require architectural or training-level innovations beyond parameter expansion alone.

### 5.2 Perception-Reasoning misalignment in MLLMs

To validate KidVis's diagnostic utility beyond aggregate accuracy, we conduct a fine-grained manual analysis of free-form rationales generated by three representative open-source MLLMs. We sample 100 balanced instances across all sub-tasks. To preserve consistency with the main evaluation, we

Table 3: Results of mainstream MLLMs on visual perceptual capabilities. Human performance is shaded in green, the best proprietary MLLM results are shaded in orange, and the best open-source MLLM results are shaded in blue.

| Method | Mean | VC | VT | VD | VM | VS | VCl |
|---|---|---|---|---|---|---|---|
| Human children (6–7 years old) | 95.31 | 94.93 | 96.07 | 94.84 | 95.73 | 93.93 | 96.33 |
| *Proprietary multimodal large language models* | | | | | | | |
| GPT-4o (Hurst et al., 2024) | 40.05 | 34.67 | 28.67 | 41.60 | 52.67 | 35.33 | 47.33 |
| GPT-5 (Singh et al., 2025) | **67.33** | 58.00 | 48.00 | **78.00** | **82.00** | **67.33** | **70.67** |
| GPT-5.4 (OpenAI, 2026) | 50.31 | 50.00 | 39.33 | 55.20 | 60.00 | 44.67 | 52.67 |
| Gemini-2.5-Flash (Comanici et al., 2025) | 56.71 | 49.33 | 39.33 | 63.60 | 75.33 | 52.67 | 60.00 |
| Gemini-2.5-Pro (Comanici et al., 2025) | 65.51 | **64.00** | 50.67 | 74.40 | 81.33 | 60.00 | 62.67 |
| Gemini-3.1-Pro (Google DeepMind, 2026) | 66.89 | 63.33 | **56.67** | 74.00 | 78.67 | 61.33 | 67.33 |
| *Open-source multimodal large language models* | | | | | | | |
| Qwen3-VL-4B-Instruct (Team, 2025) | 34.31 | 35.33 | 24.00 | 39.20 | 40.67 | 32.67 | 34.00 |
| Qwen3-VL-8B-Instruct (Team, 2025) | **38.58** | **43.33** | **32.67** | **40.80** | **42.00** | 34.00 | 38.67 |
| Qwen3-VL-30B-Instruct (Team, 2025) | 36.27 | 41.33 | 30.00 | 39.60 | 38.67 | 33.33 | 34.67 |
| InternVL3_5-4B (Wang et al., 2025) | 31.69 | 35.33 | 29.33 | 30.80 | 29.33 | 33.33 | 32.00 |
| InternVL3_5-8B (Wang et al., 2025) | 30.27 | 31.33 | 26.67 | 29.60 | 30.67 | 30.00 | 33.33 |
| InternVL3_5-14B (Wang et al., 2025) | 34.13 | 40.00 | 28.00 | 38.80 | 31.33 | **36.00** | 30.67 |
| InternVL3_5-38B (Wang et al., 2025) | 33.82 | 40.67 | 30.00 | 29.60 | 30.67 | 32.00 | **40.00** |
| llava-next-8b-hf (Li et al., 2024a) | 23.78 | 17.33 | 18.00 | 28.00 | 29.33 | 24.67 | 25.33 |
| llava-next-72b-hf (Li et al., 2024a) | 30.09 | 28.00 | 21.33 | 33.20 | 41.33 | 24.00 | 32.67 |
| llava-next-110b-hf (Li et al., 2024a) | 29.78 | 30.67 | 26.67 | 30.00 | 32.67 | 26.67 | 32.00 |
| llava-onevision-qwen2-7b-ov-hf (Li et al., 2024b) | 29.42 | 34.00 | 24.67 | 31.20 | 32.67 | 24.00 | 30.00 |
| llava-onevision-qwen2-72b-ov-hf (Li et al., 2024b) | 33.64 | 35.33 | 27.33 | 33.20 | 36.00 | 30.00 | **40.00** |
| gemma-3-27b-it (Team et al., 2025b) | 33.07 | 38.00 | 28.00 | 34.40 | 32.67 | 30.67 | 34.67 |
| Kimi-VL-A3B-Instruct (Team et al., 2025c) | 34.89 | 38.67 | 25.33 | 38.00 | 39.33 | 33.33 | 34.67 |
| mPLUG-owl3-2B (Ye et al., 2024) | 24.84 | 24.00 | 26.00 | 24.40 | 24.00 | 27.33 | 23.33 |
| mPLUG-owl3-7B (Ye et al., 2024) | 25.42 | 23.33 | 20.00 | 27.20 | 30.00 | 25.33 | 26.67 |

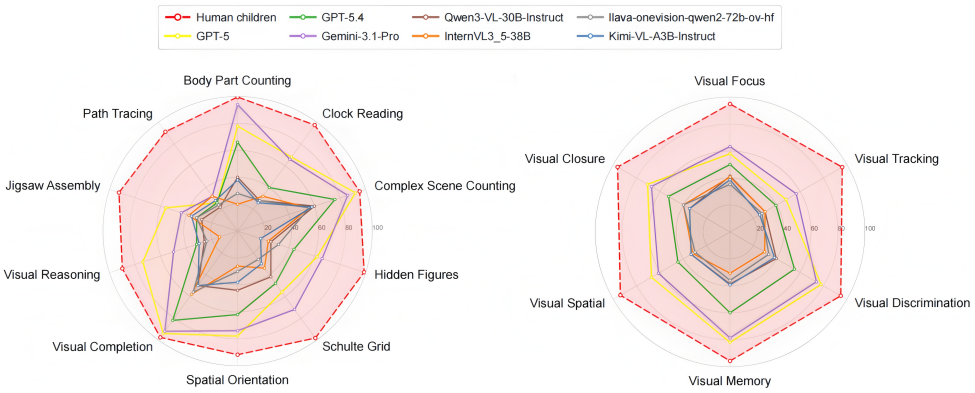



Figure 4: Comparison of representative MLLMs on KidVis benchmark. Left: Performance on 10 tasks within KidVis benchmark. Right: Performance on six visual perception capabilities.

use the same base prompt as in the benchmark setting, except that we remove the answer-only constraint so that the models can produce natural free-form rationales before giving the final option.

We categorize each response according to the position of the first decisive visual grounding failure. A response is labeled as **EVF** (Early Visual Failure) if the model makes a critical visual error at the initial stage, causing the subsequent reasoning chain to be built on an incorrect perceptual premise. It is labeled as **LVF** (Late Visual Failure) if the initial interpretation and reasoning are correct, but a later visual error derails the remaining reasoning and leads to an incorrect final answer. Finally, a response is labeled as **GCR** (Grounded and Coherent Response) if the visual interpretation remains

correctly grounded and the reasoning process stays coherent throughout, resulting in the correct answer. Three annotators independently labeled the 300 traces before discussion (88.00% mean pairwise agreement; 82.33% unanimous agreement; Fleiss' $\kappa = 0.809$); 53 cases required consensus resolution.

The results show that visual failures (both EVF and LVF) overwhelmingly dominate the incorrect responses across all evaluated models. The prevalence of EVF exposes flawed initial global encoding, as models override genuine observation with semantic priors. More critically, that LVFs constitute the leading failure mode exposes MLLMs' inability to sustain continuous visual attention. Even equipped with correct logical frameworks, models fail to re-attend to fine-grained details during multi-step reasoning, ultimately derailing execution. These patterns suggest that, for the analyzed traces, unstable visual grounding can remain a bottleneck even when the surrounding reasoning is otherwise coherent. More details are in Appendix C

### 5.3 Inter-Capability Correlation Analysis

We further analyze the relationships among the six capability scores across the 22 evaluated MLLMs using Pearson correlation, as shown in Figure 6. The six capabilities exhibit strong positive correlations, with most pairwise coefficients above 0.8. This suggests that current MLLMs tend to improve or fail jointly across perceptual dimensions. In particular, Visual Discrimination is highly correlated with Visual Memory and Visual Spatial, while Visual Spatial is also strongly correlated with Visual Tracking. These patterns are consistent with the hypothesis that fine-grained visual encoding acts as a shared constraint for downstream spatial processing, tracking, and memory-dependent visual inference.

### 5.4 Path-Tracing Controls

We conducted two additional controls on Path Tracing. First, we proportionally downsampled each image along its longest dimension by $2\times$ and $4\times$, preserving aspect ratio without cropping. Although overall performance decreased at lower resolutions, Path Tracing did not exhibit a monotonic trend: GPT-5 scored 26/20/28 and Qwen3-VL-8B-Instruct scored 32/18/28 at the original, $2\times$, and $4\times$ resolutions, respectively. These results indicate that resolution affects performance, but resolution scaling alone is insufficient to explain the low Path Tracing accuracy.

Second, we constructed a color-change control while preserving the original path geometry, intersections, endpoints, questions, and answer options. Depending on the item, one to three target paths were assigned distinct red, yellow, or blue colors, providing stable visual identity cues while leaving the required endpoint correspondence unchanged. Under this condition, GPT-5 improved from 26 to 66, Gemini-3.1-Pro from 32 to 70, and Qwen3-VL-8B-Instruct from 32 to 34. The substantial gains for GPT-5 and Gemini-3.1-Pro suggest that, under the original same-color condition, insufficiently salient path identity is an important source of error. Stable color cues help the models maintain path identity across intersections, thereby improving endpoint correspondence and answer selection.

### 5.5 Analysis of atomic capabilities

**Visual concentration and attentional inhibition.** Humans employ hard attention (fixation) to physically exclude surrounding stimuli. In contrast, MLLMs rely on soft attention, assigning non-zero weights to all visual tokens. In high-interference tasks like the Schulte Grid, this global receptivity causes noise accumulation and attentional drift, as the models lack an inherent inhibition mechanism to actively suppress background distractors.

**Visual tracking in discrete architectures.** Visual Tracking is the weakest capability in our evaluation. The Path-Tracing controls above suggest that its difficulty cannot be explained by input resolution alone. In contrast, providing salient color cues substantially improves performance for GPT-5 and Gemini-3.1-Pro, indicating that maintaining path identity across intersections is an important source of error. ViT patch discretization may contribute to this difficulty by fragmenting continuous paths into discrete visual tokens (Darcet et al., 2023; Naseer et al., 2021; Gu et al., 2022).

**Visual discrimination under semantic prior bias.** While proprietary MLLMs like GPT-5 demonstrate robust capabilities (scoring 78.00), open-source models significantly fall behind. We found

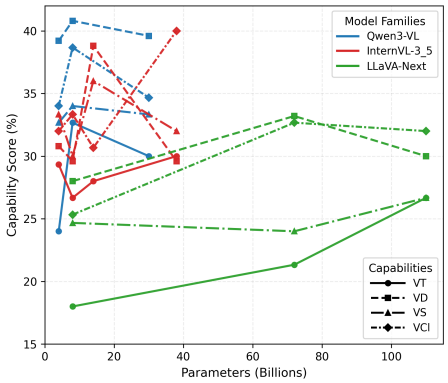



Figure 5: Scaling trends of three open-source MLLM families on four capabilities.

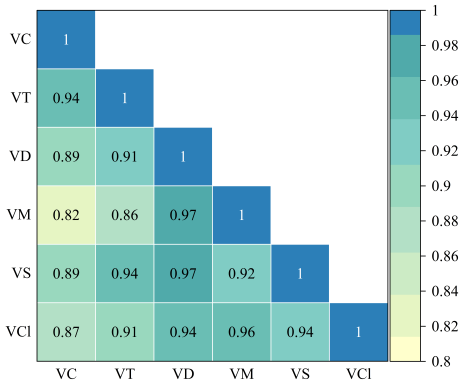



Figure 6: Correlation heatmap of the six visual capability scores across MLLMs.

two distinct failure mechanisms. First, MLLMs exhibit a severe Semantic Prior Bias. In tasks like Fine-grained Body Part Counting, when visual evidence (e.g., a five-legged cat) conflicts with pre-trained knowledge, models frequently hallucinate biologically correct but visually false answers, confirming a "Concept > Percept" hierarchy. Second, models suffer from fundamental Perceptual Imprecision. This is starkly illustrated in Task 3 (e.g., identifying pentagons). Although models correctly ground their search in the semantic criteria (e.g., 5-sided closed polygons), they lack the visual acuity to strictly differentiate these specific contours from visually similar distractions.

**Visual memory versus dynamic simulation.** Visual memory stands as the strongest capability for proprietary models, with GPT-5 achieving a high score of 82.00. This confirms that top-tier MLLMs possess robust static contextual storage. However, open-source models fall behind significantly, with the majority scoring below 40. This stark gap suggests that the bottleneck lies not in raw storage capacity, but in operational fluency. Unlike humans, who utilize visual memory as a dynamic read-write workspace for mental simulation (e.g., rotating objects in Visual Reasoning), MLLMs may treat visual history as a static read-only buffer, failing to execute the active state transitions required for reliable, step-by-step sequential inference.

**Visual spatial processing and reference frames.** Analysis of Spatial Orientation reveals a critical failure in processing depth cues like perspective, where models struggle to infer coherent 3D scenes from 2D images. This points to a fundamental absence of an internal reference frame. Without a grounded coordinate system, models cannot execute viewpoint transformation, leading to an inability to distinguish between inherent object features and viewpoint-dependent states (e.g., the relative relationship of left and right). Instead, they rely on coarse semantic approximation , as seen in Clock Reading (e.g., pointing near 7) , lacking the precise metric computation required for strict spatial reasoning.

**Visual closure and texture-geometry separation.** Our analysis of Jigsaw Assembly exposes a fundamental texture-geometry separation. While humans use the Gestalt Law of Continuity to mentally reconstruct the spatial geometry of missing parts by matching them via characteristic edge outlines or internal details, MLLMs prioritize superficial texture consistency such as similar colors or patterns. Consequently, models struggle to visually trace connecting lines, failing to strictly verify whether the boundary and interior textures are truly compatible.

## 6 Conclusion

By establishing the KidVis benchmark, this study systematically quantifies MLLM performance across six foundational visual capabilities, shifting evaluation from expert-level reasoning to child-level perception. Across the evaluated models, substantial deficits remain on low-semantic visual tasks relative to the age-specific child reference, and several model families exhibit non-monotonic or saturated scaling trends. These findings suggest that foundational visual perception remains a distinct bottleneck and may require training or architectural innovations beyond parameter scaling alone.

### AI USE STATEMENT

In this work, generative AI tools were used in the controlled AI-assisted data-generation route illustrated in Figure 2 to increase appearance diversity while preserving the intended perceptual structure. Generative AI tools were also used to polish the manuscript for language, clarity, and readability. The authors take responsibility for the final content of this work, including AI-assisted artifacts.

### ETHICS STATEMENT

The human baseline involved 10 children aged 6–7. Written informed consent was obtained from parents or legal guardians, and age-appropriate assent was obtained from the children. The study protocol was reviewed and approved by the institutional ethics committee. To protect participant privacy, only aggregate human performance statistics are reported and no raw child-level data are released. The human baseline is used only as an age-specific reference for the benchmark and is not presented as a clinical assessment of visual development.

### REPRODUCIBILITY STATEMENT

KidVis construction, task definitions, evaluation metrics, and the zero-shot evaluation setting are described in the main paper. The appendix provides confidence-interval and sensitivity analyses, diagnostic controls, detailed task instantiations, the human-baseline protocol, per-task model results, and representative failure cases. These materials are intended to make the benchmark design and reported comparisons auditable and reproducible under the stated protocol.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105:14325–14329, 2008.

Ted Brown. Construct validity of the three motor-reduced subscales of the developmental test of visual perception—adolescent and adult (dtvp-a): A rasch analysis model evaluation. *British Journal of Occupational Therapy*, 2011.

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

John Bissell Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge university press, 1993.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Ronald P Colarusso and Donald D Hammill. *Motor-free visual perception test*. Academic Therapy Pub., 1972.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, et al. Palm-e: An embodied multimodal language model. 2023.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Google DeepMind. Gemini 3.1 pro - model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

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.

Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pp. 404–421. Springer, 2022.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

D. D. Hammill, N. A. Pearson, and J. K. Voress. *Developmental Test of Visual Perception: Second Edition (DTVP-2)*. PRO-ED, 2 edition, 1993.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, et al. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024a. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, et al. Llava-onevision: Easy visual task transfer, 2024b. URL https://arxiv.org/abs/2408.03326.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023c. URL https://openreview.net/forum?id=xozJw0kZXF.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, et al. Core knowledge deficits in multi-modal language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EIK6xxIoCB.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Nancy Martin. *Test of Visual Perceptual Skills-4th Edition*. Academic Therapy Publications Novato, CA, 2017.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

Jocelyn H Newton and Kevin S McGrew. Introduction to the special issue: Current research in cattell–horn–carroll–based assessment. *Psychology in the Schools*, 47, 2010.

OpenAI. GPT-4V(ision) System Card. OpenAI, September 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf. System Card.

OpenAI. Introducing gpt-5.4, March 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Caitlin R Siu and Kathryn M Murphy. The development of human visual cortex and clinical implications. *Eye and brain*, pp. 25–36, 2018.

Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25: 207–222, 1973.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. Gemini: A family of highly capable multimodal models, 2025a. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025b.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, et al. Kimi-VL technical report, 2025c. URL https://arxiv.org/abs/2504.07491.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Mary Warren. A hierarchical model for evaluation and treatment of visual perceptual dysfunction in adult acquired brain injury, part 1. *The American Journal of Occupational Therapy*, 47(1):42–54, 1993.

Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Shi, et al. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*, 2023.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, et al. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1877–1893, 2025. doi: 10.1109/TPAMI.2024.3507000.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, et al. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, 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 (CVPR)*, pp. 9556–9567, June 2024.

Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, et al. Aibench: Towards trustworthy evaluation under the 45° law. *Displays*, pp. 103255, 2025a. ISSN 0141-9382. doi: 10.1016/j.displa.2025.103255.

Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, et al. Large multimodal models evaluation: A survey. *SCIENCE CHINA Information Sciences*, 68(12): 221301–221369, 2025b. doi: https://doi.org/10.1007/s11432-025-4676-4.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

# A Limitations and Future Work

KidVis is designed as a controlled diagnostic benchmark rather than a standalone training or fine-tuning corpus. Its current scale therefore limits its suitability for training, but does not undermine its primary purpose of held-out evaluation under zero-shot settings. The task-to-capability mapping identifies shared primary perceptual demands; assigning two tasks to the same perceptual primitive does not imply that they share the same visual statistics in an MLLM's input space. Human results serve as an age-specific reference baseline rather than a physiological ceiling. Future work will further diversify the task types in KidVis and rely more extensively on programmatic generation to improve scalability, controllability, and coverage.

# B Confidence Intervals and Sensitivity Analyses

## B.1 Confidence intervals

For each model's overall score and six capability scores, we used a task-stratified item bootstrap with 10,000 replicates. Complete 95% confidence intervals were calculated for all evaluated MLLMs. For all 220 model–task combinations, we calculated Wilson 95% binomial confidence intervals based on the 50 items in each task; for example, InternVL3.5-4B scores 26.00 on Task 8 with a 95% CI of [15.87, 39.55]. For the child baseline, we jointly resampled the ten participants and the items within each task. Table 4 reports representative overall-score confidence intervals.

Table 4: Representative 95% confidence intervals for overall KidVis scores.

| **Quantity** | **Estimate** | **95% CI** |
|---|---|---|
| Child overall | 95.31 | [93.72, 96.79] |
| GPT-5 overall | 67.33 | [63.56, 70.98] |
| Gemini-3.1-Pro overall | 66.89 | [63.02, 70.62] |

## B.2 Sensitivity to evaluation settings

We conducted controlled sensitivity experiments on GPT-5 and Qwen3-VL-8B-Instruct by varying the answer constraint, prompt wording, and image resolution while keeping all other evaluation conditions fixed. For answer-format sensitivity, we replaced the original answer-only instruction with: "You may briefly explain your reasoning. End your response with 'Answer: X,' where X is A, B, C, or D." For prompt-wording sensitivity, only the question sentence was modified and the answer-only constraint was retained. Two semantically equivalent candidates were generated for each item and manually screened so that the target object, comparison relation, visual-attribute definitions, and required solution steps remained unchanged. For image-resolution sensitivity, we

Table 5: Sensitivity of capability scores to answer format, prompt wording, and image resolution.

| Evaluation setting | VC | VT | VD | VM | VS | VCl | Mean |
|---|---|---|---|---|---|---|---|
| **GPT-5** | | | | | | | |
| Original | 58.00 | 48.00 | 78.00 | 82.00 | 67.33 | 70.67 | 67.33 |
| Explanation allowed | 53.33 | 40.00 | 73.20 | 75.33 | 62.00 | 64.67 | 61.42 |
| Semantically equivalent prompt | 54.00 | 42.00 | 77.60 | 82.67 | 64.67 | 68.67 | 64.93 |
| Resolution at 50% | 50.00 | 42.00 | 70.80 | 74.67 | 60.00 | 66.00 | 60.58 |
| Resolution at 25% | 42.00 | 35.33 | 50.80 | 61.33 | 41.33 | 55.33 | 47.69 |
| **Qwen3-VL-8B-Instruct** | | | | | | | |
| **Evaluation setting** | **VC** | **VT** | **VD** | **VM** | **VS** | **VCl** | **Mean** |
| Original | 43.33 | 32.67 | 40.80 | 42.00 | 34.00 | 38.67 | 38.58 |
| Explanation allowed | 45.33 | 34.00 | 43.60 | 41.33 | 37.33 | 40.00 | 40.27 |
| Semantically equivalent prompt | 41.33 | 32.00 | 41.60 | 44.67 | 36.00 | 39.33 | 39.16 |
| Resolution at 50% | 37.33 | 30.67 | 37.60 | 42.67 | 30.00 | 38.00 | 36.04 |
| Resolution at 25% | 39.33 | 32.00 | 37.20 | 39.33 | 30.00 | 36.00 | 35.64 |

proportionally downsampled each image along its longest dimension by factors of $2\times$ and $4\times$, preserving aspect ratio and applying no cropping. All experiments used deterministic decoding.

Semantically equivalent prompt rewriting produces only modest overall changes: $-2.40$ points for GPT-5 and $+0.58$ points for Qwen3-VL-8B-Instruct. Allowing explanations has a larger but model-dependent effect, decreasing GPT-5 by 5.91 points while increasing Qwen3-VL-8B-Instruct by 1.69 points. Nevertheless, the overall capability profiles remain broadly consistent, with Visual Tracking remaining the weakest capability for both models. Resolution has a stronger and more model-dependent effect; the additional reduction from 50% to 25% of the original resolution affects GPT-5 most strongly on Clock Reading, Complex Scene Counting, Hidden Figures, and Visual Reasoning.

## C Reasoning Trace Classification

For the failure-mode analysis, we sampled 100 balanced KidVis instances and collected one response from each of the three evaluated models, yielding 300 model-response traces. Three annotators independently labeled each trace as EVF, LVF, or GCR before discussion. The annotations achieved a mean pairwise agreement of 88.00%, a unanimous agreement rate of 82.33%, and a Fleiss' $\kappa$ of 0.809; 53 of 300 cases (17.67%) required subsequent resolution. Disagreements were then discussed to determine the consensus labels reported in Table 6.

Table 6: Fine-grained classification of reasoning traces on 100 sampled KidVis instances.

| Model | EVF | LVF | GCR |
|---|---|---|---|
| Qwen3-VL-8B-Instruct | 13 | 56 | 31 |
| Kimi-VL-A3B-Instruct | 31 | 47 | 22 |
| llava-onevision-7b | 22 | 50 | 28 |

## D Task Installation

To demonstrate the quality and diversity of the KidVis benchmark, we provide a detailed breakdown of the sub-tasks for each task category.

### D.1 Task 1: Body Part Counting

This task fuses abnormal biological features with naturalistic textures, as shown in Figure 7. Participants must separate specific features from background noise and contrast them against biological prototypes retrieved from long-term memory. It consists of two sub-categories. Finger Count: This sub-task presents images of human hands. The target option features polydactyly , while distractors display normal anatomy. Animal Leg Count: This sub-task depicts various common animal species in naturalistic environments. The target image contains a morphological violation specifically involving the legs, contrasting with biologically correct distractors.

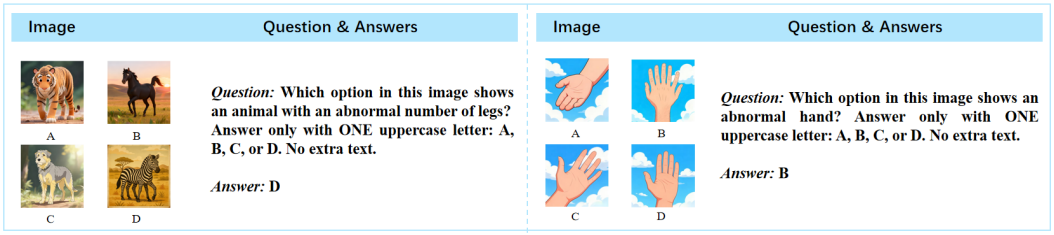



Figure 7: Sub-tasks of Body Part Counting.

### D.2 Task 2: Clock Reading

Participants must distinguish the hour and minute hands based on subtle length and width differences, and map their precise angular positions to the clock face reference frame to acquire temporal information, as shown in Figure 8.

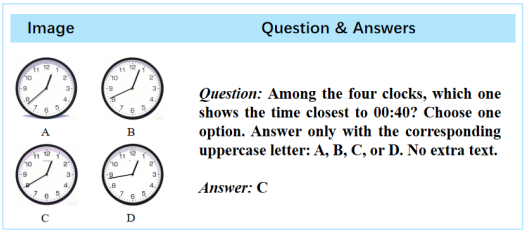



Figure 8: Sub-tasks of Clock Reading.

### D.3 Task 3: Complex Scene Counting

This task requires Figure-Background Separation under challenging conditions, such as occlusion or noise. Participants must sustain high visual concentration to filter out interference while employing fine-grained discrimination to identify targets based on local cues, as shown in Figure 9 . It consists of three sub-categories. Geometric Attribute Counting: Participants are presented with groups of visually similar geometric items and must count a specific subset based on fine-grained attributes. This involves distinguishing specific shapes or identifying polygons based on their number of sides. Structured Scene Counting: This task focuses on counting items within organized or demarcated spatial structures. It includes identifying objects in compartmentalized containers like pears in boxes or products on shelves, and determining the number of vehicles within the grid-like constraints of parking lots. Unstructured Scene Counting: This task requires identifying and counting the number of people in complex, non-structured backgrounds where targets often feature heavy occlusions and overlapping boundaries.

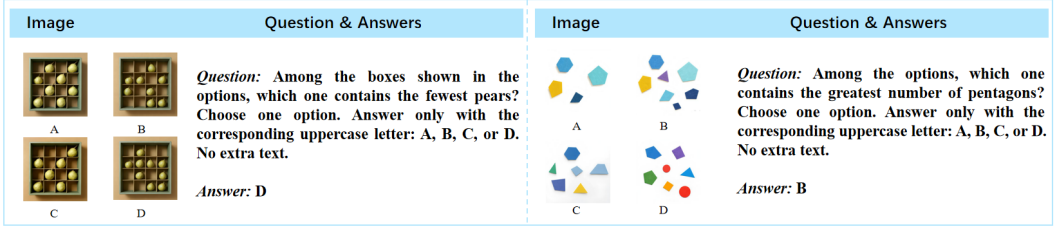



Figure 9: Sub-tasks of Complex Scene Counting.

### D.4 TASK 4: HIDDEN FIGURES

Participants must track a specific geometric shape embedded within a complex network of intersecting lines, as shown in Figure 10 . The Visual Closure mechanism is invoked to mentally complete the truncated segments, allowing participants to perceive the shape as a coherent Gestalt.

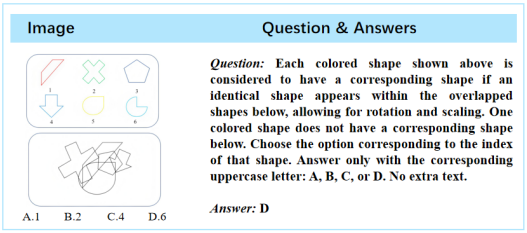



Figure 10: Sub-tasks of Hidden Figures.

### D.5 TASK 5: SCHULTE GRID

This task evaluates the model's ability to locate and track a strictly ascending numerical sequence within a visually scrambled environment, as shown in Figure 11. To assess the robustness of visual search mechanisms, the task employs diverse visual instantiations beyond standard digit-grid pairs. Symbolic Representation: The sequence elements are not limited to explicit digits but also include countable symbols like hearts or geometric shapes. Spatial Layout: The search space varies from standard grids to irregular geometric deformations.

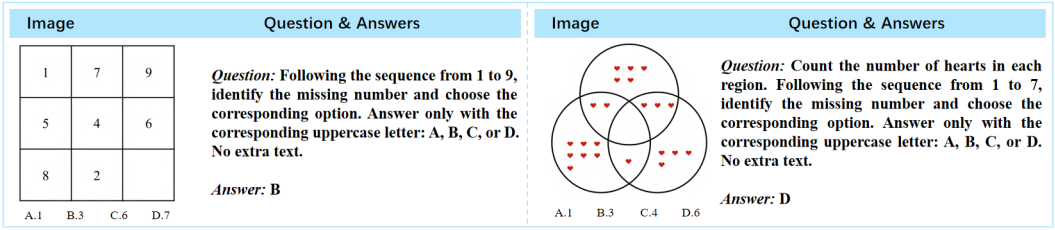



Figure 11: Sub-tasks of Schulte Grid.

### D.6 TASK 6: SPATIAL ORIENTATION

Sub-tasks involve: (1) Determining relative object positions (e.g., coverage, rotation) within a reference frame; (2) 3D Structural Completion (calculating missing blocks); (3) Viewpoint transformation (e.g., identifying top-down views). Participants must construct 2D/3D coordinate systems to facilitate mental rotation and perspective taking. The examples of sub-task are shown in Figure 12.

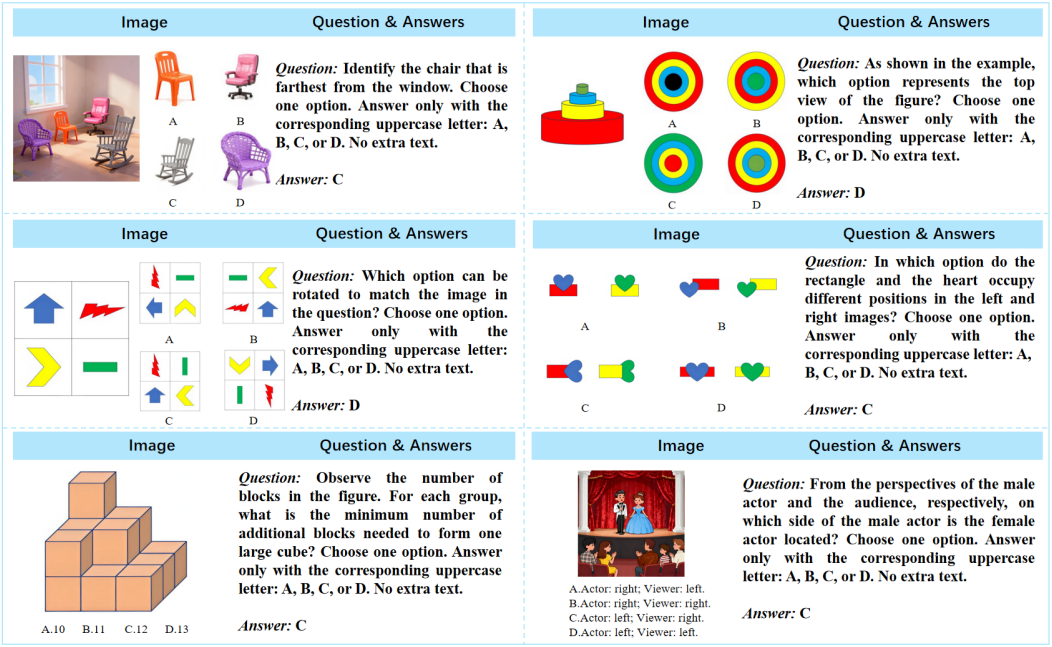



Figure 12: Sub-tasks of Spatial Orientation.

### D.7 Task 7: Visual Completion

Presented with objects containing structural defects, participants must retrieve the canonical prototype from memory, as shown in Figure 13. All objects used in this task are common everyday items that are highly familiar to humans, thereby minimizing additional semantic ambiguity. The cognitive system detects interrupted contours and performs Amodal completion to instinctively fill in the missing parts, thereby identifying the complete or defective object.

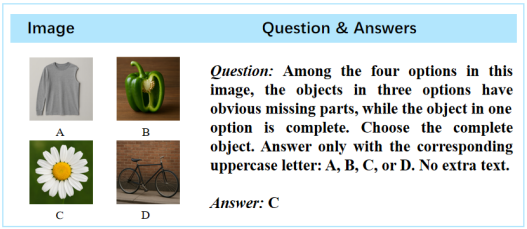



Figure 13: Sub-tasks of Visual Completion.

### D.8 Task 8: Visual Reasoning

This task evaluates the ability to figure out abstract rules from visual sequences and is composed of two sub-tasks: Pattern Completion and Visual Sudoku, as shown in Figure 14. Task Mechanism: Participants are required to identify logical regularities by analyzing state changes across a sequence. These transitions involve diverse visual attributes, including number, color, shape, spatial arrangement, and relative position. Cognitive Process: To solve these puzzles, participants must utilize visual memory as a Hypothesis Buffer, retaining previous states to infer the underlying transformation rules and correctly predict the missing element.

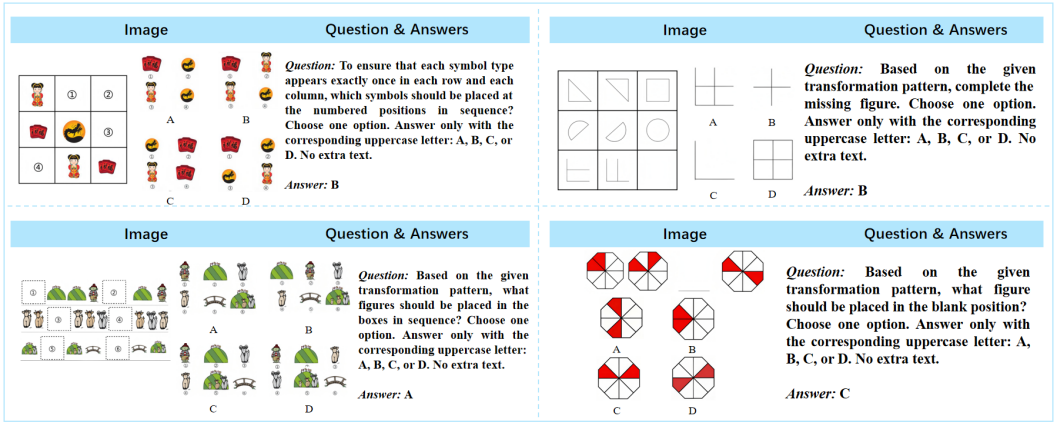



Figure 14: Sub-tasks of Visual Reasoning.

### D.9 Task 9: Jigsaw Assembly

Sub-tasks involve selecting fragments based on content compatibility, geometric decomposition, and boundary topology, as shown in Figure 15. Visual Closure acts as the generative prior, envisioning the complete state. Visual Spatial serves as the execution engine, manipulating fragments in a mental buffer to verify geometric fit and texture consistency.

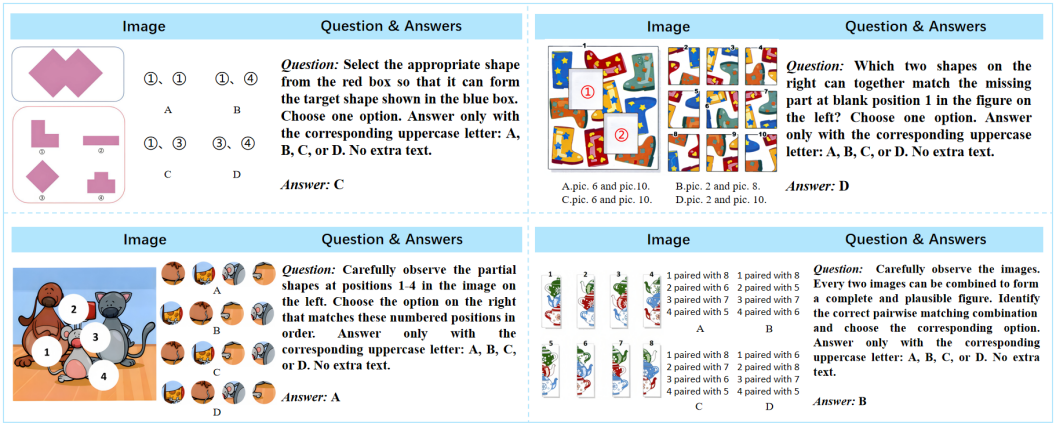



Figure 15: Sub-tasks of Jigsaw Assembly.

### D.10 Task 10: Path Tracing

Sub-tasks involve: (1) distinguishing entwined lines and identifying the correct endpoint correspondences; (2) following spiral or number mazes by tracing a valid route under sequential constraints; and (3) counting black dots distributed along concentric circular paths while maintaining accurate track identity, as shown in Figure 16. Participants must execute ”smooth pursuit” along thin paths, requiring high concentration to suppress interference and maintain continuity across long-range dependencies.

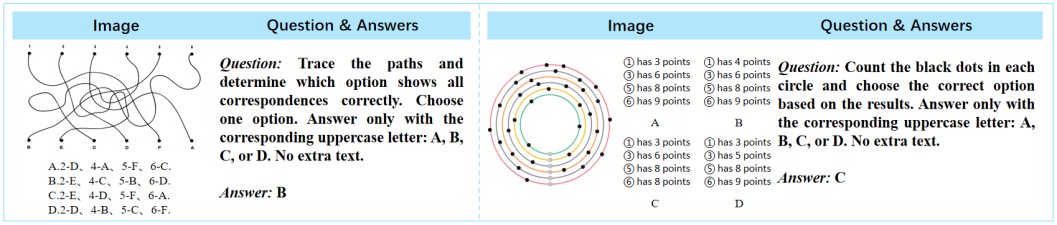



Figure 16: Sub-tasks of Path Tracing.

## E Answer Distribution Analysis

To examine whether KidVis contains positional answer bias, we analyze the distribution of correct options at both the global and task levels. As shown in Table 7, the overall answer distribution is close to balanced across the four options, and a similar pattern is observed within individual tasks.

Table 7: Answer-option distribution in KidVis. The near-balanced option counts reduce position bias in choice evaluation.

| Split | Total | A | B | C | D |
|---|---|---|---|---|---|
| All questions | 500 | 130 | 122 | 122 | 126 |
| Task 1 | 50 | 14 | 11 | 11 | 14 |
| Task 2 | 50 | 12 | 14 | 14 | 10 |
| Task 3 | 50 | 16 | 10 | 9 | 15 |
| Task 4 | 50 | 11 | 14 | 13 | 12 |
| Task 5 | 50 | 12 | 15 | 12 | 11 |
| Task 6 | 50 | 13 | 12 | 13 | 12 |
| Task 7 | 50 | 13 | 11 | 11 | 15 |
| Task 8 | 50 | 13 | 11 | 14 | 12 |
| Task 9 | 50 | 13 | 12 | 12 | 13 |
| Task 10 | 50 | 13 | 12 | 13 | 12 |

## F Human Baseline

### F.1 Participants and Background

The cohort comprised ten children aged 6–7 years, including five boys and five girls, recruited in mainland China and enrolled either in kindergarten or Grade 1. Before participation, we conducted individual pre-study screening interviews and consulted the children's legal guardians and, where applicable, their teachers regarding educational experience and prior exposure to related activities. Based on this screening, none of the participants had received systematic or specialized training specifically targeting KidVis or closely related task formats through preschool enrichment programs, professional tutoring, extracurricular courses, or structured in-home instruction. This screening was intended to exclude task-specific training rather than ordinary age-appropriate experiences, such as counting, clock reading, or puzzle activities.

The benchmark instances were instantiated from established visual-assessment paradigms through controlled adaptation and redesign while preserving their intended perceptual demands. The participants had not seen these specific instances before evaluation. They received only standardized procedural instructions and five non-benchmark examples for interface familiarization, without task-solving strategies, correctness feedback, or coaching during the formal evaluation. Thus, neither the children nor the evaluated MLLMs were exposed to the benchmark items or received task-solving training before testing. We therefore interpret the child results as an age-specific human reference under standardized evaluation conditions, rather than as a controlled comparison between children and MLLMs with identical developmental or training histories.

Written informed consent was obtained from the legal guardian of each participating child, and age-appropriate assent was obtained from the children before participation. The study protocol was reviewed and approved by the institutional ethics committee. To protect participant privacy, we report aggregate statistics rather than individual-level records and do not release raw child-level data.

### F.2 Administration protocol

Before the formal evaluation, each child received the same standardized instructions explaining the single-choice format, how to select and submit an answer, and that a submitted response could not be revisited or modified. Each child then completed five interface-familiarization examples, which were separate from the benchmark, to become familiar with the task format and interface. During the formal evaluation, the experimenter provided no task-specific hints, correctness feedback, or assistance in solving individual items.

The children completed the full 500-item KidVis evaluation without a per-item time limit. The total testing time was approximately 600 minutes per participant and was distributed across multiple sessions, with breaks provided between task blocks to reduce fatigue. All participants followed the same administration procedure and completed the same benchmark instances and answer options used in the model evaluation.

### F.3 Participant- and task-level dispersion

Macro-averages alone do not fully characterize participant-level variation in the child cohort or performance heterogeneity across the evaluated MLLMs. Table 8 summarizes overall dispersion. The relatively narrow child dispersion indicates that high aggregate performance was consistent across participants rather than being driven by a small subset. For the MLLMs, the reported SD, variance, and range describe performance differences across the 22 evaluated systems rather than repeated-run uncertainty for an individual model.

Table 8: Overall dispersion across the child cohort and the 22 evaluated MLLMs.

| **Evaluation group** | **Mean** | **SD** | **Variance** | **Observed min–max** |
|---|---|---|---|---|
| Human children ($n = 10$) | 95.31 | 1.64 | 2.68 | 93.16–98.18 |
| Evaluated MLLMs ($n = 22$) | 38.67 | 13.61 | 185.10 | 23.78–67.33 |

For the child cohort, Table 9 reports statistics calculated across the ten participant scores on each task. The greatest participant-level dispersion occurs on Spatial Orientation (T6; SD = 6.56) and Path Tracing (T10; SD = 5.35), although performance remains high.

Table 9: Child task-level performance and participant-level dispersion.

| **Task** | **Mean** | **SD** | **Variance** | **Observed min–max** |
|---|---|---|---|---|
| T1: Body Part Counting | 99.6 | 0.84 | 0.71 | 98.00–100.00 |
| T2: Clock Reading | 97.4 | 2.32 | 5.38 | 92.00–100.00 |
| T3: Complex Scene Counting | 95.4 | 3.27 | 10.71 | 88.00–98.00 |
| T4: Hidden Figures | 98.8 | 2.15 | 4.62 | 94.00–100.00 |
| T5: Schulte Grid | 98.2 | 1.99 | 3.96 | 94.00–100.00 |
| T6: Spatial Orientation | 91.8 | 6.56 | 43.07 | 82.00–100.00 |
| T7: Visual Completion | 97.6 | 2.80 | 7.82 | 94.00–100.00 |
| T8: Visual Reasoning | 90.0 | 3.27 | 10.67 | 86.00–96.00 |
| T9: Jigsaw Assembly | 92.6 | 1.35 | 1.82 | 90.00–94.00 |
| T10: Path Tracing | 91.2 | 5.35 | 28.62 | 80.00–98.00 |

For comparison, Table 10 reports the corresponding statistics across the 22 evaluated MLLMs on each task. Model heterogeneity is greatest on Visual Completion (T7; SD = 22.83), Body Part Counting (T1; SD = 20.31), and Complex Scene Counting (T3; SD = 19.60). Low cross-model dispersion does not necessarily indicate strong performance: on Path Tracing (T10), the MLLM mean is only 25.18 with an observed range of 16.00–34.00, while the child group scores 91.20. The highest observed MLLM score on this task is 34.00, yielding the largest task-level gap of 57.20

percentage points. At the capability level, the largest separation occurs for Visual Tracking, where the child score is 96.07 and the highest observed MLLM score is 56.67, corresponding to a gap of 39.40 points.

Table 10: Task-level dispersion across the 22 evaluated MLLMs.

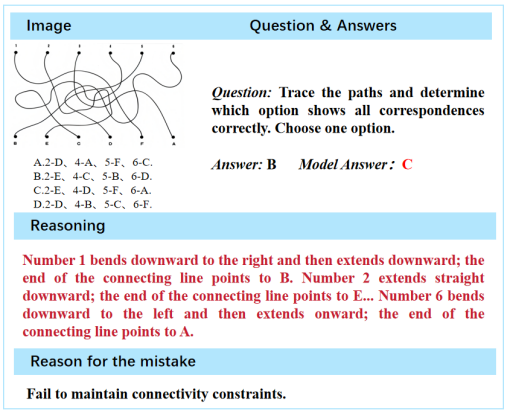
| Task | Mean | SD | Variance | Observed min–max |
|---|---|---|---|---|
| T1: Body Part Counting | 43.00 | 20.31 | 412.48 | 18.00–94.00 |
| T2: Clock Reading | 34.18 | 14.71 | 216.35 | 18.00–68.00 |
| T3: Complex Scene Counting | 55.45 | 19.60 | 384.07 | 16.00–92.00 |
| T4: Hidden Figures | 31.55 | 13.05 | 170.26 | 18.00–66.00 |
| T5: Schulte Grid | 36.09 | 15.66 | 245.13 | 6.00–72.00 |
| T6: Spatial Orientation | 41.55 | 16.77 | 281.12 | 24.00–78.00 |
| T7: Visual Completion | 55.73 | 22.83 | 521.26 | 20.00–94.00 |
| T8: Visual Reasoning | 33.73 | 17.76 | 315.54 | 14.00–76.00 |
| T9: Jigsaw Assembly | 33.18 | 8.83 | 77.97 | 22.00–56.00 |
| T10: Path Tracing | 25.18 | 5.04 | 25.39 | 16.00–34.00 |

## G Experiment Results

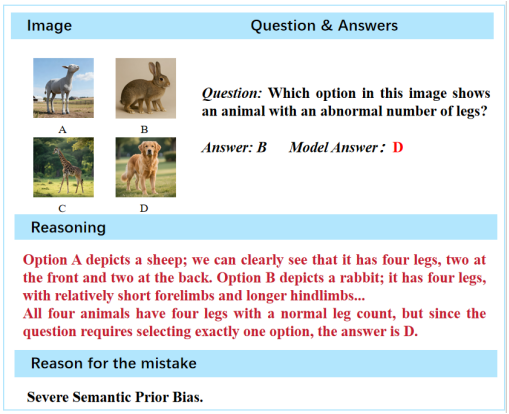
Results of mainstream MLLMs on ten Tasks are illustrated in Table 11.

Table 11: Results of mainstream MLLMs on the ten KidVis tasks. Human performance is shaded in green, the best proprietary MLLM results are shaded in orange, and the best open-source MLLM results are shaded in blue.

| Method | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Human children (6–7 years old) | 99.60 | 97.40 | 95.40 | 98.80 | 98.20 | 91.80 | 97.60 | 90.00 | 92.60 | 91.20 |
| *Proprietary multimodal large language models* | | | | | | | | | | |
| GPT-4o (Hurst et al., 2024) | 50 | 38 | 52 | 34 | 36 | 42 | 82 | 26 | 26 | 16 |
| GPT-5 (Singh et al., 2025) | 78 | **68** | **92** | 62 | 56 | **78** | **94** | 74 | **56** | 26 |
| GPT-5.4 (OpenAI, 2026) | 66 | 40 | 76 | 44 | 48 | 62 | 82 | 32 | 32 | 26 |
| Gemini-2.5-Flash (Comanici et al., 2025) | 64 | 54 | 72 | 42 | 52 | 60 | **94** | 68 | 44 | 24 |
| Gemini-2.5-Pro (Comanici et al., 2025) | 76 | 60 | 88 | 48 | 70 | 72 | 92 | **76** | 48 | **34** |
| Gemini-3.1-Pro (Google DeepMind, 2026) | **94** | 66 | 86 | **66** | **72** | 74 | 92 | 50 | 44 | 32 |
| *Open-source multimodal large language models* | | | | | | | | | | |
| Qwen3-VL-4B-Instruct (Team, 2025) | 34 | 28 | 56 | 22 | 24 | 40 | 50 | **38** | 30 | 26 |
| Qwen3-VL-8B-Instruct (Team, 2025) | **48** | 20 | **62** | 30 | 36 | 40 | 44 | 34 | **42** | **32** |
| Qwen3-VL-30B-Instruct (Team, 2025) | 40 | 28 | 60 | 26 | **42** | **44** | 50 | 26 | 28 | 22 |
| InternVL3_5-4B (Wang et al., 2025) | 22 | 32 | 40 | 22 | 36 | 34 | 40 | 26 | 34 | 30 |
| InternVL3_5-8B (Wang et al., 2025) | 18 | 24 | 40 | 26 | 34 | 36 | 44 | 30 | 30 | 20 |
| InternVL3_5-14B (Wang et al., 2025) | 30 | **36** | 60 | 24 | 34 | 42 | 38 | 26 | 30 | 26 |
| InternVL3_5-38B (Wang et al., 2025) | 20 | 32 | 56 | 24 | 34 | 26 | **58** | 14 | 38 | **32** |
| llava-next-8b-hf (Li et al., 2024a) | 32 | 24 | 26 | 28 | 6 | 28 | 26 | 30 | 22 | 20 |
| llava-next-72b-hf (Li et al., 2024a) | 40 | 18 | 44 | 24 | 24 | 26 | 46 | **38** | 28 | 16 |
| llava-next-110b-hf (Li et al., 2024a) | 32 | 28 | 40 | 28 | 28 | 28 | 44 | 22 | 24 | 24 |
| llava-onevision-qwen2-7b-ov-hf (Li et al., 2024b) | 34 | 18 | 52 | 24 | 30 | 30 | 42 | 22 | 24 | 20 |
| llava-onevision-qwen2-72b-ov-hf (Li et al., 2024b) | 28 | 28 | 56 | **32** | 26 | 30 | 56 | 24 | 32 | 24 |
| gemma-3-27b-it (Team et al., 2025b) | 36 | 28 | 58 | 28 | 28 | 34 | 46 | 16 | 30 | 28 |
| Kimi-VL-A3B-Instruct (Team et al., 2025c) | 38 | 26 | 58 | 18 | 30 | 38 | 50 | 30 | 36 | 28 |
| mPLUG-owl3-2B (Ye et al., 2024) | 30 | 28 | 16 | 22 | 30 | 26 | 20 | 22 | 28 | 26 |
| mPLUG-owl3-7B (Ye et al., 2024) | 36 | 28 | 30 | 20 | 18 | 24 | 36 | 18 | 24 | 22 |




(a) Failure of Path Tracing.




(b) Failure of Body Part Counting.

Figure 17: Representative failure cases from Qwen3-VL-8B-Instruct.

## H Failure Cases

Figures 17–19 illustrate eight representative failure cases from Qwen3-VL-8B-Instruct. Details can be found in Section **5.4 Analysis of atomic capabilities** of the main paper.

## I Evaluation Claims and Intended Use

KidVis is designed as a diagnostic benchmark for foundational visual perception in MLLMs. It evaluates controlled, low-semantic, motor-free visual tasks that target six basic perceptual primitives: visual concentration, tracking, discrimination, memory, spatial processing, and closure. KidVis is intended to complement semantic-heavy multimodal benchmarks, rather than replace them.

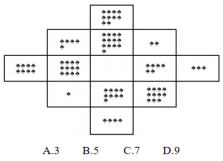
**Supported claims.** KidVis supports comparisons of MLLMs under a shared zero-shot, single-choice protocol, and helps identify task categories and visual primitives where models remain weak. It also enables bounded comparisons between MLLMs and the aggregate performance of 6–7-year-old children on the same benchmark instances. Observed scaling patterns should be interpreted only within the evaluated model families, prompts, and KidVis task distribution.

**Assumptions.** The evaluation assumes that each item has a rule-defined correct answer, that the task-to-capability mapping reflects the principal perceptual requirements involved rather than process-pure isolation, and that single-choice accuracy is a useful proxy for task success.

**Intended uses.** KidVis is intended for diagnostic model evaluation, comparative benchmarking, failure analysis, and research on improving fine-grained visual grounding. It can be used to study how visual perception interacts with language reasoning and to reveal failures that may be obscured in high-level semantic benchmarks.

**Out-of-scope uses and limitations.** KidVis should not be interpreted as a comprehensive measure of general intelligence, human-level vision, embodied perception, or overall multimodal ability. Low performance indicates weakness under the controlled perceptual conditions tested here. The child results serve as a human reference baseline, not a physiological ceiling or clinical assessment of visual development.

**Image** | **Question & Answers**




*Question:* **Count the number of spades in each region. Following the sequence from 1 to 13, identify the missing number and choose the corresponding option.**

***Answer:* B** ***Model Answer :* C**

**Reasoning**

**There are 13 regions in total.**
**The number of spades in the first row is 10. In the second row, the number of spades in each region from left to right is 5, 13, and 2 respectively. In the third row, the number of spades in each region from left to right is. In the fourth row, the number of spades in each region from left to right is 8, 9, 6, and 6 respectively...**

**Reason for the mistake**

**Fail to actively suppress background distractors.**

(a) Failure of Schulte Grid.

**Image** | **Question & Answers**



A.Actor: right; Viewer: left.
B.Actor: right; Viewer: right.
C.Actor: left; Viewer: right.
D.Actor: left; Viewer: left.

*Question:* **From the perspectives of the male actor and the audience, respectively, on which side of the male actor is the female actor located? Choose one option.**

***Answer:* C** ***Model Answer :* B**

**Reasoning**

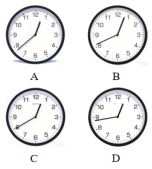
**The male actor stands on the stage, facing the audience. In the stage layout, the male actor is positioned on the left side while the female actor is on the right side. When the male actor faces the audience, his right side corresponds to the right side of the stage; thus, the female actor is on his right...**

**Reason for the mistake**

**Lack a grounded coordinate system.**

(b) Failure of Spatial Orientation.

**Image** | **Question & Answers**




*Question:* **Among the four clocks, which one shows the time closest to 00:40? Choose one option.**

***Answer:* C** ***Model Answer :* B**

**Reasoning**

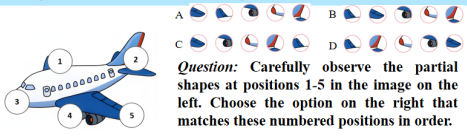
**Determine the time indicated by each clock.**
**Option A: The time is 12:37 (the minute hand lies between the 7 and 8 marks, closer to 7).**
**Option B: The time is 12:40 (the minute hand points directly at 8, corresponding to 40 minutes).**
**Option C: The time is 12:38 (the minute hand lies between the 7 and 8 marks, closer to 8)...**
**Calculate the time difference from the target time (12:40). For Option A, the difference is |40 - 37| = 3 minutes...**

**Reason for the mistake**

**Lacking precise metric computation required for strict spatial reasoning.**

(c) Failure of Clock Reading.

**Image** | **Question & Answers**




*Question:* **Carefully observe the partial shapes at positions 1-5 in the image on the left. Choose the option on the right that matches these numbered positions in order.**

***Answer:* D** ***Model Answer :* B**

**Reasoning**

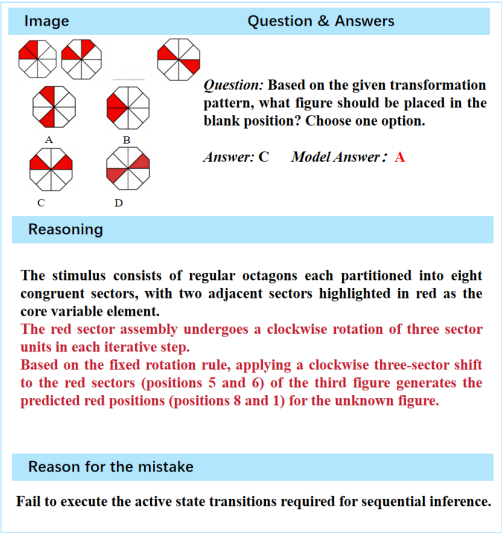
**Key Feature Identification. Item 1 is the left wing (blue wing with white lines); Item 2 is the vertical stabilizer (red-and-blue, vertical tail); Item 3 is the nose cone (white fuselage with blue windows); Item 4 is the landing gear wheel (gray tire with an orange hub); Item 5 is the engine (gray main body with an orange air intake).**
**Option B. Item 1 is a blue wing; Item 2 is a blue wing; Item 3 is a gray-and-orange engine; Item 4 is a red-and-blue vertical stabilizer... The features match with the corresponding serial number positions**

**Reason for the mistake**

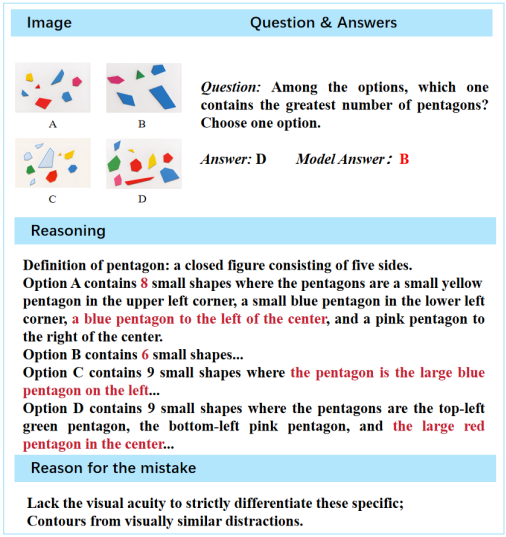
**Prioritize superficial texture consistency;**
**Struggle to visually trace connecting lines.**

(d) Failure of Jigsaw Assembly.

Figure 18: Representative failure cases from Qwen3-VL-8B-Instruct.




(a) Failure of Visual Reasoning.




(b) Failure of Complex Scene Counting.

Figure 19: Additional representative failure cases from Qwen3-VL-8B-Instruct.