Easy for Children, Hard for AI: The Limits of Multimodal LLMs in Early Childhood Learning

Jingping Liu[1], Xueyan Wu[2], Hanxuan Chen[3], Ziyan Liu[2], Zhangquan Chen[4], Ronghao Chen[5*], Huacan Wang[6*]

[1] School of Software Engineering, Sun Yat-sen University
[2] School of Information Science and Engineering, East China University of Science and Technology
[3] College of Electrical and Information Engineering, Hunan University
[4] Tsinghua Shenzhen International Graduate School, Tsinghua University
[5] School of Environmental and Energy, Peking University
[6] School of Materials Science and Optoelectronic Technology, University of Chinese Academy of Sciences
liujp68@mail.sysu.edu.cn, {y40250690, y30241069}@mail.ecust.edu.cn, chenhanxuan@hnu.edu.cn
czq23@mails.tsinghua.edu.cn, chenronghao@alumni.pku.edu.cn, wanghuacan17@mails.ucas.ac.cn

Abstract

Early childhood is a critical stage for cognitive development, involving core skills such as visual perception and reasoning. While multimodal large language models (MLLMs) have made rapid progress in various general-purpose tasks, their ability to support early education remains largely underexplored. Existing research on child-related AI largely centers on modeling language, emotion, or behavior, with limited focus on evaluating cognitive tasks relevant to early learning. To address this gap, we propose ChildBench, a multimodal benchmark designed to assess models on tasks inspired by early childhood cognitive development. It covers five key domains through ten tasks, including spatial reasoning, visual reasoning, visual discrimination, counting skills, and visual tracking. The benchmark includes 4,890 carefully constructed images and 5,346 manually annotated samples, ensuring both diversity and age-appropriate content. We evaluate a range of state-of-the-art (SoTA) open-source and closed-source MLLMs—including GPT-4o, Gemini, and Qwen2.5-VL—on ChildBench. Despite strong performance on other benchmarks, the best 7B-parameter model with LoRA tuning achieves only 52.01% accuracy, far below the 96% achieved by 5-year-old children. These results reveal critical limitations in fine-grained perception and reasoning. We further analyze failure cases and discuss directions for future model development.

Datasets — https://github.com/Jderder/ChildBench

Introduction

Recently, multimodal large language models have advanced rapidly, achieving impressive results in complex tasks such as multimodal dialogue (Li and Tajbakhsh 2023) and action planning (Yang et al. 2023). However, despite their performance on high-level tasks, they still struggle with basic abilities like object localization (Fu et al. 2024) and spatial reasoning (Liu et al. 2025). This gap highlights the importance of understanding the limits of their fundamental capabilities.

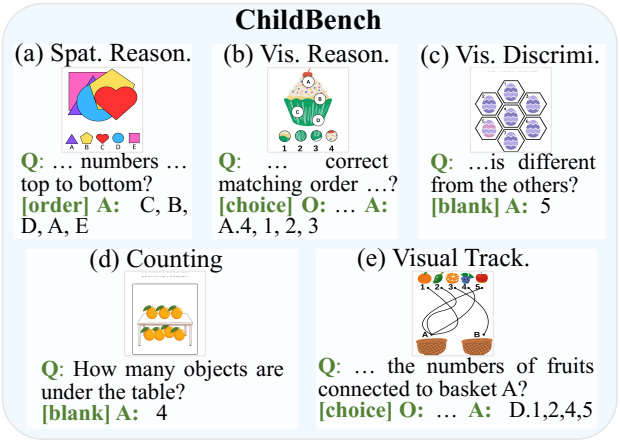

Figure 1: Five cognitive abilities assessed in ChildBench.

Early education is the foundation of human cognitive development and an important way to evaluate the abilities of MLLMs. These tasks mirror the initial stages of human intelligence and effectively assess a model's performance in fundamental visual perception and reasoning. This includes identifying differences between objects (e.g., shape, color, and size), recognizing detailed features (e.g., texture, quantity, and orientation), and understanding spatial relationships (e.g., relative position and occlusion) within images—often in combination with textual instructions to make appropriate judgments. For instance, the task illustrated in Figure 1(a) requires the model to determine the order of occlusion among shapes and output their sequence from top to bottom. A model can only be considered to exhibit human-like cognitive abilities when it consistently performs such tasks with high accuracy.

To the best of our knowledge, there remains a notable lack of benchmarks specifically tailored to early childhood education, which hinders the systematic evaluation of MLLMs on cognitive tasks at this early stage. Existing relevant benchmarks can be roughly divided into two cate-

*Corresponding Author.

gories. The first category focuses on children's language development and basic behaviors. Typical examples include CBT (Hill et al. 2016), EmoReact (Nojavanasghari et al. 2016), SAYCam (Sullivan et al. 2021), PInSoRo (Lemaignan et al. 2018), LIRIS-CSE (Khan et al. 2019), ChildMandarin (Zhou et al. 2024a), and ChildACT (Sandygulova et al. 2025). These benchmarks mainly focus on language learning, emotional expression, and social interaction. While they offer useful insights into child development, they lack data for evaluating key multimodal cognitive abilities such as spatial reasoning and visual tracking. The second category emphasizes multimodal educational tasks, including benchmarks like CLEVR (Johnson et al. 2017), MATH-V (Wang et al. 2024), MathVista (Lu et al. 2023), CMM-Math (Liu et al. 2024), MathScape (Zhou et al. 2024b), MME-Reasoning (Yuan et al. 2025), EXAMS-V (Das et al. 2024), and MMGeoLM (Sun et al. 2025). These primarily test mathematical reasoning and cross-disciplinary multimodal capabilities, making them more suitable for older children or adults. However, they do not cover cognitive tasks central to early childhood development.

Hence, in this study, we introduce ChildBench, a multi-task, multi-format evaluation benchmark tailored for early childhood education. It aims to systematically assess the basic cognitive abilities of MLLMs, including spatial reasoning, visual reasoning, visual discrimination, counting skills, and visual tracking, as illustrated in Figure 1. In ChildBench, the images are primarily created manually using the Canvas[1] platform, with a smaller portion sourced from freely licensed early education websites such as print-kids.net, as well as from open-source datasets like CMM-Math (Liu et al. 2024) and MathVista (Lu et al. 2023). In total, the benchmark contains 4,890 high-quality images, covering a diverse range of scenes and topics carefully selected to reflect the cognitive characteristics of early learners. To ensure data quality, we establish a detailed and standardized annotation pipeline. A professional team—consisting of three annotators, two checkers, and one reviewers—carried out a three-stage quality control process with clearly defined roles. This rigorous process resultes in 5,346 high-quality annotated samples, covering five task types across ten tasks, enabling a comprehensive and fine-grained evaluation of MLLMs' core perception and reasoning capabilities. Based on this benchmark, we conduct a comprehensive evaluation of several representative MLLMs, including closed-source models (i.e., GPT-4o (Achiam et al. 2023) and Gemini 2.5 Pro (Team et al. 2023)) and open-source models (e.g., LLaMA-3.2-11B (Grattafiori et al. 2024), LLaVA (Liu et al. 2023), and Qwen-VL (Bai et al. 2025)). The results reveal both the strengths and limitations of current models in handling early-education-related cognitive tasks and offer valuable insights for future model development.

In brief, our contributions are summarized as follows:

- We propose ChildBench, a high-quality benchmark comprising manually constructed images and annotated samples, designed to systematically evaluate the core capabilities of MLLMs in early childhood education scenarios.
- ChildBench covers 10 tasks across 5 categories of children's cognitive assessments, including spatial reasoning, visual reasoning, visual discrimination, counting skills, and visual tracking.
- We test several open-source and closed-source MLLMs on ChildBench. The results show that even the best-performing models—Gemini-2.5-Pro and fine-tuned Qwen2.5vl-7B—only achieved 44.60% and 52.01% accuracy, far below the 96% performance of 5-year-old children. We also suggest future research directions to bridge this gap.

[1]https://www.canva.cn/

Related Work

Child-Oriented benchmarks. To support AI research on child development tasks, existing benchmarks can be roughly divided into two categories: (1) language and cognitive development and (2) perception and behavior. The first category focuses on language understanding, generation, and reasoning. For instance, CBT (Hill et al. 2016) is built from children's books and includes four answer types: verbs, pronouns, named entities, and common nouns. GSM8K (Verschaffel et al. 2020) includes 8.5K elementary math-related problems that require multi-step reasoning. FairytaleQA (Xu et al. 2022) contains 27 fairy tales and 10,580 QA pairs for evaluating story understanding. ChildMandarin (Zhou et al. 2024a) is a Mandarin speech dataset for children aged 3–5. The second category focuses on how children see and act in the world, including visual recognition, motion analysis, and emotion understanding. EmoReact (Nojavanasghari et al. 2016) includes 1,102 video clips from children aged 4–14, labeled with 17 emotions. PInSoRo (Lemaignan et al. 2018) records 45+ hours of child–child and child–robot interaction. LIRIS-CSE (Khan et al. 2019) has facial expression videos from 12 children aged 6–12, covering six basic emotions. SAYCam (Sullivan et al. 2021) provides 415+ hours of video from three children aged 6 months to 2.5 years. ChildACT (Sandygulova et al. 2025) includes multi-view videos of 200 children performing seven actions. While these datasets are useful for studying language, behavior, and emotion, there is still a lack of data that targets key cognitive abilities in early education, such as spatial reasoning, visual discrimination, and visual tracking.

Multimodal education benchmarks. Existing benchmarks in multimodal education can be broadly categorized into two types: those for math reasoning and those for general reasoning. Math-related benchmarks focus on solving math problems that include both text and images. MathVista (Lu et al. 2023) encompasses a wide range of math tasks in visual settings. CMM-Math (Liu et al. 2024) is a Chinese benchmark with questions from elementary to high school, some accompanied by images. MATH-V (Wang et al. 2024) is based on math competition questions and evaluates visual reasoning ability. MathScape (Zhou et al. 2024b) also covers school-level math and focuses on multi-step reasoning in visual contexts. MMGeoLM (Sun et al. 2025) leverages real geometry exam

questions to test models' geometric understanding. General reasoning benchmarks cover broader logic and multimodal tasks. CLEVR (Johnson et al. 2017) is a foundational benchmark for visual reasoning tasks such as counting and comparison. EXAMS-V (Das et al. 2024) spans multiple subjects and languages with image-text inputs, requiring complex, cross-lingual reasoning. MME-Reasoning (Yuan et al. 2025) tests three types of reasoning: inductive, deductive, and abductive. Although these benchmarks support research on multimodal education, they mostly focus on math or general reasoning, and do not adequately capture the unique cognitive challenges present in early learning tasks.

Problem Formulation

To comprehensively evaluate the performance of MLLMs on early learning tasks, ChildBench includes three types of questions: multiple-choice, fill-in-the-blank, and sequential QA. All question types consist of a query Q paired with an image I as input. For multiple-choice questions, additional candidate options are provided, which can be either text-based or image-based. The model need to select the correct answer from a variable number of options (see Figure 1(b) and (e)). In fill-in-the-blank tasks, the model is required to generate a single word, letter, or number as the answer (Figure 1(c) and (d)). For sequential QA questions, the model is expected to produce a sequence as the response (Figure 1(a)).

ChildBench Construction

In this section, we describe the construction of ChildBench in detail, including the image collection and annotation procedures.

Image Collection

The image set in ChildBench is constructed from two primary sources. The first comprises freely accessible, copyright-free early childhood education websites (e.g., print-kids.net), along with existing open-source datasets such as CMM-Math (Liu et al. 2024) and MathVista (Lu et al. 2023). From these sources, we collect 250 images. Early childhood websites contribute fewer images, as many of their materials are designed for hands-on interaction (e.g., circling, coloring), which are not easily adapted for use with MLLMs. Remaking such content for machine-readable formats is also non-trivial. Similarly, while the open-source datasets focus on mathematical reasoning, they contain only a limited number of early-learning-style images. To supplement this, the majority of ChildBench images are manually constructed using the free design platform Canva. Canva provides a rich library of child-friendly image assets and a flexible design interface, making it well-suited for creating age-appropriate content while avoiding copyright concerns.

Based on this platform, the image construction process is as follows: First, we refer to questions from early childhood education resource websites and design different image templates for various task types, as illustrated in Figure 2 (left). In total, we manually create 86 image templates. Next, for each template, we select child-appropriate images from the Canva platform and populate the templates accordingly, as shown in Figure 2 (right). To improve efficiency, we reuse the same template with different image categories—for example, replacing animal-themed images with fruit-themed ones. Following this process, we obtain a total of 4,640 images.

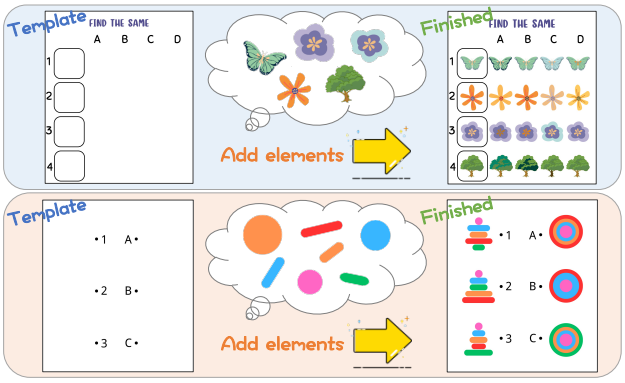

Figure 2: Image creation process in Canva.

Annotation Procedure

In this section, we introduce the tasks in ChildBench and describe the benchmark annotation process in detail.

To fully evaluate the ability of MLLMs in early childhood education tasks, we construct the ChildBench benchmark, which includes a range of representative cognitive tasks. It covers five core task types: spatial reasoning (including perspective-taking reasoning, paper folding reasoning, and image overlapping reasoning tasks), visual reasoning (graphic matching and transformation reasoning), visual discrimination (odd-one-out and identical item detection), counting skill (planar counting and spatial counting), and visual tracking. To support efficient and consistent data creation, we provide multiple question templates for each task to help annotators quickly generate high-quality question–answer pairs. Example templates are shown in Appendix "Question Template" of the Github link.

The annotation process of ChildBench consists of three stages: data annotation, data checking, and data review. The entire process is completed by a team of six college students, including three annotators, two checkers, and one reviewer. Before starting the formal annotation, all team members first annotate some sample data, which are then reviewed and discussed by the core authors to ensure consistency in standards and accuracy in execution.

Stage 1: data annotation. We evenly distribute the 4,890 collected images to three annotators. For each image, they refer to task-specific question templates to write 1–3 questions and annotate the correct answers. Question types include multiple-choice, fill-in-the-blank, and sequential QA. For multiple-choice questions, distractors are first generated by a script using simple heuristic rules (e.g., adding or subtracting up to 2 from numbers in the correct answer). Annotators then select the most appropriate distractors from the generated options. For example, if the correct answer is "(A)

5, 8, 4, 4", the script creates variations like "(B) 6, 8, 4, 3" and the annotators pick the ones that seem most confusing.

Stage 2: data checking. We first run scripts to identify incomplete samples (e.g., missing distractors) and send them back to annotators for correction. After this automatic way, two inspectors manually check all samples in parallel. They verify whether each question matches the image and whether the labeled answer is correct. Samples that do not pass manual inspection, along with the reasons, are returned for revision. This process is repeated until the batch reaches a 95% pass rate.

Stage 3: data review. A reviewer performs a final review on the verified data. First, 20% of the samples from the batch are randomly selected for careful review. If problems are found, the reviewer returns the samples with reasons to the inspectors for correction. If similar issues occur repeatedly, the reviewer discusses them with the whole annotation team to ensure consistency. This process is repeated until the reviewed samples reach a 98% pass rate. In the end, we obtain 5,346 high-quality samples to form ChildBench.

ChildBench Analysis

Benchmark statistics. As reported in Table 1, ChildBench consists of 5,346 samples based on 4,890 images, which are split into training, validation, and test sets in a 7:1:2 ratio. Among all tasks, "planar counting" and "visual tracking" have the largest number of samples, accounting for 14.72% and 13.49% of the total, respectively. In contrast, the "paper folding reasoning" task has a relatively small number of samples (3.54%), but it includes the largest number of unique images, with a total of 845, which accounts for 17.28% of all images in the benchmark.

Question type distribution. ChildBench includes three types of questions: multiple-choice, fill-in-the-blank, and sequential QA, accounting for 50.7%, 43.4%, and 5.9% of the benchmark, respectively. For multiple-choice questions, the number of options ranges from 2 to 6, with an average of 3.85. For fill-in-the-blank questions, the average answer length is 1.42 characters, with a minimum of 1 and a maximum of 10. In sequential QA, answers consist of comma-separated elements, with an average of 5.01 items per response, ranging from 4 to 6. These variations help assess the model's ability to handle questions of differing difficulty levels.

Number of images per question. ChildBench includes both single-image and multi-image input formats. Most samples (4,908) use a single image, while 438 samples require multiple images. Among these, 222 samples include 4 images, 197 include 5 images, and 19 samples contain as many as 7 images. Multi-image samples are fewer because they only appear when the options are images. In these cases, one set of images usually creates only one question, so the total number is limited. Four tasks involve multi-image input: transformation reasoning, perspective-taking reasoning, paper folding reasoning, and graphic matching. Among them, paper folding reasoning uses only multi-image input, with an average of 4.47 images per question.

Task	# Img	Train	Dev	Test	Total	Ratio
Spatial Reasoning						
perspective-taking	623	279	40	80	399	7.46
paper folding	845	132	19	38	189	3.54
image overlapping	312	435	63	124	622	11.63
Visual Reasoning						
graphic matching	624	310	43	89	442	8.27
transformation	498	340	48	97	485	9.07
Visual Discrimination						
odd-one-out	387	415	60	118	593	11.09
ident. item detect.	392	406	58	116	580	10.85
Counting Skills						
planar counting	366	551	79	157	787	14.72
spatial counting	433	370	52	106	528	9.88
Visual Tracking						
-	410	505	72	144	721	13.49
TOTAL	4,890	3,743	534	1,069	5,346	100

Table 1: ChildBench Statistics.

Experiments

In this section, we evaluate the performance of current mainstream MLLMs on our early childhood education benchmark and analyze their limitations in the benchmark.

Baselines

We evaluate two types of baselines: closed-source MLLMs and open-source MLLMs. For closed-source models, we randomly select 500 samples from ChildBench and test them using GPT-4o[2] and Gemini 2.5 Pro.[3] Each model is evaluated under two settings: in the 0-shot setting, the model receives the question, images, and options (if applicable), along with a prompt to guide its response. In the 1-shot setting, we also provide one example from the training set that matches the task and question type of the current sample.

For open-source models, we select eight mainstream MLLMs: LLaMA-3.2-11B (Grattafiori et al. 2024), LLaVA-v1.5-7B (Liu et al. 2023), mPLUG-Owl3-7B (Ye et al. 2024), DeepSeek-VL-7B (Lu et al. 2024), Qwen2.5-VL-7B (Bai et al. 2025), Phi-3.5-Vision-4B (Abdin et al. 2024), InternVL3.0-8B (Chen et al. 2024), and MiniCPM-V-8B (Yao et al. 2024). These models are also evaluated under two settings. In the direct testing setting, the evaluation process is the same as the 0-shot setting used for closed-source models. In the fine-tuning setting, each model is trained on the training set using LoRA (Hu et al. 2021), and then evaluated on the test set.

In addition, we ask two 5-year-old children from the senior kindergarten class, accompanied by their parents, to answer 100 randomly selected questions. A question is counted as correct only if both children answered it correctly.

Metrics and Implementations

In our experiments, we report macro-precision (P), macro-recall (R), macro-F1 score, and accuracy (Acc). The task

[2] gpt-4o

[3] gemini-2.5-pro-preview-05-06

Model	Setting	P	R	F1	Acc	Setting	P	R	F1	Acc
					Open-source MLLMs					
LLaMA-3.2-11B	-	31.32	30.89	30.48	23.10	LoRA	26.57	26.80	26.54	25.35
LLaVA-v1.5-7B	-	26.63	25.27	25.70	17.68	LoRA	24.33	25.29	24.04	19.08
mPLUG-owl3-7B	-	34.52	31.97	25.84	25.06	LoRA	40.87	40.87	40.36	39.57
DeepSeek-VL-7B	-	16.59	23.24	17.94	17.50	LoRA	44.32	25.59	21.44	21.23
Qwen2.5-VL-7B	-	34.06	34.23	33.82	32.86	LoRA	53.56	53.64	53.08	**52.01**
Phi-3.5-vision-4B	-	28.00	27.93	27.20	23.15	LoRA	37.94	36.94	36.87	36.11
InternVL3.0-8B	-	27.19	27.24	27.11	26.89	LoRA	41.56	39.08	39.07	41.38
MiniCPM-V-8B	-	23.57	25.07	23.37	23.35	LoRA	41.86	41.63	41.67	40.03
					Closed-source MLLMs					
GPT-4o	0-shot	35.05	34.32	33.07	37.40	1-shot	27.56	26.63	26.84	29.20
Gemini-2.5-Pro	0-shot	59.95	40.59	48.37	**44.60**	1-shot	49.42	49.91	49.01	45.80
					Other Methods					
Human	-	94.95	94.84	94.84	**96.00**	-	-	-	-	-

Table 2: Model comparison (%) of open/closed-source MLLMs. All results are averaged over three runs.

prompts for MLLMs and the hyper-parameter settings are shown in Appendix "Prompts Used in MLLMs" and "Hyperparameter Settings" of the Github link.

Main Results

We evaluate all baseline methods on ChildBench, with the results listed in Table 2. From the table, we observe the following: (1) All MLLMs perform poorly on ChildBench, leaving significant room for improvement compared to senior kindergarten children's performance. For instance, the best-performing open-source and closed-source models, i.e., Qwen2.5-VL-7B and Gemini-2.5-Pro, achieve accuracies of only 52.01% and 44.60%, respectively, while the children accuracy reaches 96.00%. This highlights that, despite strong performance on certain common tasks, MLLMs remain far from achieving truly human-like intelligence. (2) Without fine-tuning, the test performance of open-source models is generally much lower than that of closed-source models. For instance, Qwen2.5-VL-7B achieves only 32.86% accuracy without LoRA tuning—11.74% lower than Gemini-2.5-Pro's 44.60%. However, after LoRA tuning, Qwen2.5-VL-7B's accuracy improves to 52.01%, surpassing Gemini-2.5-Pro by 4.7%. This demonstrates that while large-parameter MLLMs often have an advantage in zero-shot scenarios, models with fewer parameters can outperform them through fine-tuning. (3) Gemini-2.5-Pro shows improved accuracy in the 1-shot setting compared to the 0-shot setting, whereas GPT-4o exhibits the opposite trend, with performance declining in the 1-shot scenario. This is primarily due to the inclusion of multiple images in the input after adding examples. This phenomenon suggests that Gemini-2.5-Pro has stronger multi-image processing capabilities than GPT-4o.

Detailed Analysis

Impact of model parameter size on ChildBench. To evaluate the impact of model size on performance in ChildBench, we test models from the Qwen2.5-VL and InternVL3.0 series. The results are presented in Table 3. Overall, the performance generally improves as the number of model parameters increases. However, the largest model, InternVL3.0-38B, achieves an accuracy of only 63.52%, which remains far below the performance of 5-year-old children. This highlights that even large models still face significant challenges in early education tasks.

Model	Set.	F1	Acc	Set.	F1	Acc
			Qwen-VL Family			
Qwen2.5-VL-7B	-	33.82	32.86	LoRA	53.08	52.01
Qwen2.5-VL-32B	-	36.53	34.42	LoRA	57.05	55.00
			Intern-VL Family			
InternVL3.0-8B	-	27.11	26.89	LoRA	39.07	41.38
InternVL3.0-14B	-	35.68	34.56	LoRA	42.03	40.40
InternVL3.0-38B	-	40.70	40.57	LoRA	63.39	63.52

Table 3: Model performance (%) on ChildBench across different parameter sizes.

Performance analysis across different task types. To evaluate the performance of MLLMs on different types of early childhood education tasks, we calculate their accuracy across various categories, as shown in Table 4. From the table, we notice that: (1) all models perform much worse than 5-year-old children. The best model results are 67.42% from Gemini-2.5-Pro on Counting and 67.09% from Qwen2.5-VL-7B on Visual Discrimination. In comparison, children get over 90% on all tasks, and 100% on three of them. This shows that ChildBench is still very hard for current models. (2) Among open-source MLLMs, most models show significant performance improvements after fine-tuning compared to their original versions. Notably, MiniCPM-V-8B demonstrates substantial gains in the Visual Discrimination and Counting Skills tasks, with accuracy increasing from 25.64% to 52.14% and from 26.24% to 49.05%, respectively—an improvement of 26.5% and 22.81%. For closed-source MLLMs, GPT-4o experiences a decline in performance on most tasks when examples are added, whereas Gemini-2.5-Pro shows improved results under the same conditions. This further suggests that Gemini-2.5-Pro has stronger multi-image processing capabilities. (2) Accuracy in Visual Discrimination and Counting Skills is relatively

Model	Spatial Reasoning		Visual Reasoning		Visual Discriminat.		Counting Skills		Visual Tracking	
					Open-source MLLMs					
	-	LoRA	-	LoRA	-	LoRA	-	LoRA	-	LoRA
LLaMA-3.2-11B	19.01	19.83	28.49	25.27	34.19	32.05	24.33	27.38	18.06	20.14
LLaVA-v1.5-7B	13.22	14.46	26.34	22.04	19.23	25.21	15.21	16.73	24.31	16.67
mPLUG-owl3-7B	23.55	32.23	28.49	28.49	18.38	42.31	30.80	50.95	24.31	40.97
DeepSeek-VL-7B	11.98	14.46	20.43	22.58	19.23	28.21	20.53	21.29	14.58	19.44
Qwen2.5-VL-7B	18.60	**38.84**	33.33	45.70	47.44	**67.09**	41.83	**61.98**	17.36	39.58
Phi-3.5-vision-4B	17.36	30.99	23.66	23.12	27.35	41.03	27.00	46.39	22.22	34.72
InternVL3.0-8B	19.05	26.84	28.26	27.72	35.04	61.11	33.08	58.17	13.19	19.44
MiniCPM-V-8B	21.90	32.23	20.97	27.42	25.64	52.14	26.24	49.05	21.53	33.33
					Closed-source MLLMs					
	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
GPT-4o	**24.14**	18.10	**40.24**	28.04	56.07	37.38	42.42	34.85	15.87	25.40
Gemini-2.5-Pro	10.34	29.31	34.41	**47.56**	**65.42**	47.66	**67.42**	59.09	**38.10**	**42.86**
					Other Methods					
Human	93.33	-	90.00	-	100.00	-	100.00	-	100.00	-

Table 4: Model results (Acc %) on different types of tasks.

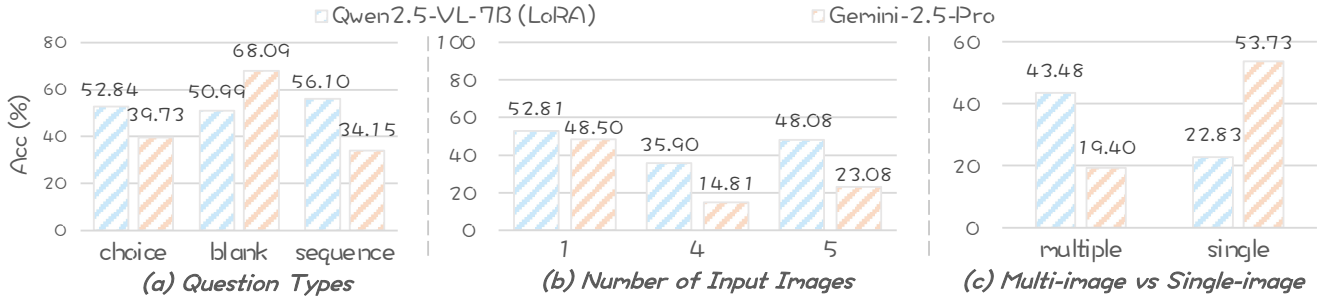

Figure 3: Impact of question type, input image quantity, and multi/single-image input on model performance, respectively.

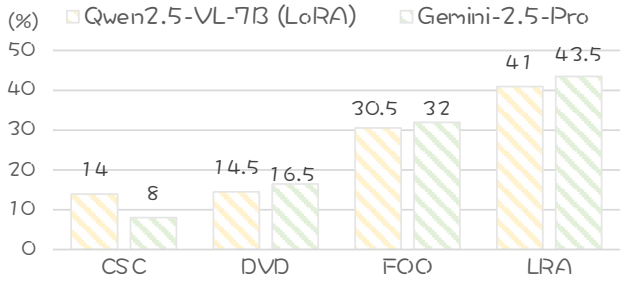

Figure 4: Distribution of error types across the two models.

higher than in other tasks, though still limited. In contrast, performance on Spatial Reasoning and Visual Tracking is the lowest, indicating that these tasks remain particularly challenging. Overall, current models still struggle with early childhood tasks, especially those requiring complex spatial reasoning and visual tracking abilities.

Performance Analysis across different question types. To comprehensively evaluate model performance, ChildBench includes three types of questions: multiple-choice, fill-in-the-blank, and sequential QA. We assess the accuracy of two models, including Qwen2.5-VL-7B and Gemini-2.5-Pro, on each question type, and the results are shown in Figure 3(a). The results reveal clear differences in how each model handles different question types. Qwen2.5-VL-7B demonstrates relatively consistent performance across all three types, with accuracies around 50%, suggesting a stable response to varying formats. In contrast, Gemini-2.5-Pro achieves its highest accuracy on fill-in-the-blank questions, where its performance is approximately 30% higher than on the other two types.

Impact of input image quantity on model performance. We explore how the number of input images affects model performance. Since there are very few samples with seven images in the test set, we exclude this case from statistical analysis. The results for other image quantities are shown in Figure 3(b). The two models show clear differences in their ability to handle varying numbers of input images. Gemini-2.5-Pro performs best with single-image inputs, achieving an accuracy of 48.50%, but its performance drops significantly with more images—down to 14.81% with four images and 23.08% with five images. In contrast, Qwen2.5-VL-7B also performs best with a single image (52.81%), but its accuracy decreases more gradually with additional images—maintaining accuracy at 35.90% (four images) and 48.08% (five images). This indicates that Qwen2.5-VL-7B is more stable across different image input quantities.

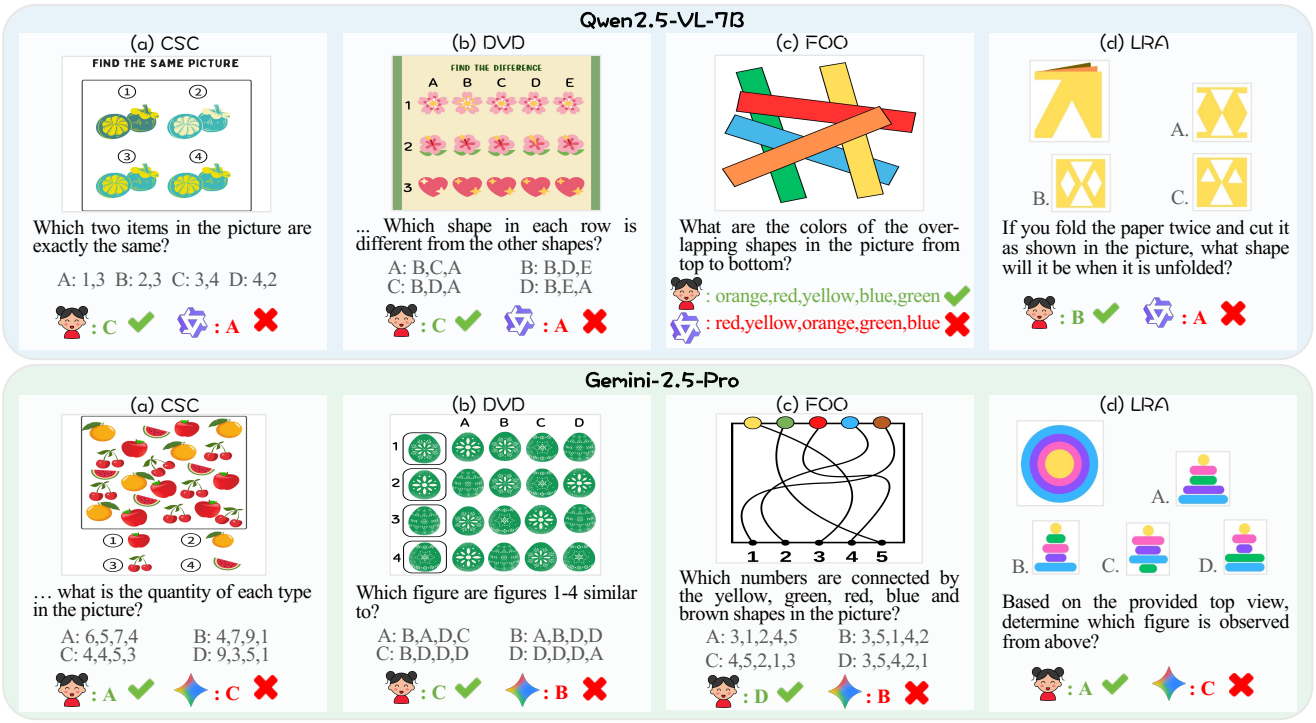

Figure 5: Examples of four error types from two SoTA MLLMs.

Comparison between multi-image and single-image inputs. To further compare the effects of multi-image and single-image inputs, we conduct experiments on the same set of samples. Specifically, we first select all multi-image samples from the ChildBench test set to form dataset C_1. Then, we manually merge each set of images into a single image to construct dataset C_2. We evaluate Qwen2.5-VL-7B (with LoRA) and Gemini-2.5-Pro on both datasets, and the results are presented in Figure 3(c). The results reveal contrasting trends between the two models. Qwen2.5-VL-7B performs significantly better on multi-image inputs (43.48%) than on single-image inputs (22.83%). In contrast, Gemini-2.5-Pro shows the opposite pattern, achieving much higher accuracy with single-image inputs (53.73%) compared to multi-image inputs (19.40%). This suggests that Qwen2.5-VL-7B benefits more from having access to multiple images, while Gemini-2.5-Pro may struggle to effectively process multi-image information.

Error Analysis

To guide future research, we conduct an error analysis on two SoTA models: the open-source Qwen2.5-VL-7B and the closed-source Gemini-2.5-Pro. We randomly sample 200 incorrect predictions from each model and manually categorize the errors. These errors can be broadly grouped into four types: those caused by confusion between similar colors (labeled as CSC), difficulty in recognizing visual details (DVD), failure to understand overlapping objects (FOO), and limitations in reasoning ability (LRA). The distribution of error types across the two models is presented in Figure 4, and representative examples of each error type are shown in Figure 5. From the results, we observe that both models exhibit a similar error distribution. Errors related to color confusion and detail recognition are less frequent, whereas errors involving overlapping objects and reasoning limitations occur more often. Overall, these error types highlight key areas for future improvement.

Conclusion and Limitations

In this paper, we introduce ChildBench, a multimodal benchmark designed to assess basic cognitive abilities in early education. It includes 10 task types across 5 cognitive domains, covering diverse tasks that reflect real-world early learning activities. We evaluate several open-source and closed-source MLLMs on ChildBench and conduct detailed experiments. The results show that current models still struggle with these tasks, performing much worse than 5-year-old children. We also suggest future research directions to improve multimodal understanding in early learning.

Although ChildBench is an effective benchmark for evaluating MLLMs on early education tasks, it still has some limitations. First, the dataset is constructed almost entirely through manual creation and annotation. While this ensures high quality, it also significantly increases the labor cost, limiting the overall scale of the dataset. This restricted scale may affect the fine-tuning performance of multimodal models. In addition, there is some imbalance in task distribution. For example, the paper-folding reasoning task contains more images but fewer annotated samples, which may lead to biased evaluation results for that task.

Ethical Statement

In ChildBench, most images are manually created using free licensed materials from the Canvas platform, which allows non-commercial use. Additional images are collected from Print-Kids.net, a free educational resource site, whose license permits non-commercial research use. We also include images from open-source datasets such as CMM-Math and MathVista, which are released under CC-BY 4.0 or similar academic licenses. All sources are properly cited in the paper. Based on these images, we manually design all questions, answer options, and correct answers. Our ChildBench is also released under the CC-BY 4.0 license. To ensure safety, we carefully check all content and remove anything related to violence, bias, or other material not suitable for early education.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (No. 62306112).

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Das, R.; Hristov, S.; Li, H.; Dimitrov, D.; Koychev, I.; and Nakov, P. 2024. EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7768–7791.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Khan, R. A.; Crenn, A.; Meyer, A.; and Bouakaz, S. 2019. A novel database of children's spontaneous facial expressions (LIRIS-CSE). *Image and Vision Computing*, 83: 61–69.

Lemaignan, S.; Edmunds, C. E.; Senft, E.; and Belpaeme, T. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one*, 13(10): e0205999.

Li, S.; and Tajbakhsh, N. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Liu, Z.; Cen, Z.; Zhou, Y.; Zou, Y.; Zhang, W.; Jiang, H.; and Ruan, T. 2025. Can Multimodal Large Language Models Understand Spatial Relations? *arXiv preprint arXiv:2505.19015*.

Liu, W.; Pan, Q.; Zhang, Y.; Liu, Z.; Wu, J.; Zhou, J.; Zhou, A.; Chen, Q.; Jiang, B.; and He, L. 2024. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C. E.; and Morency, L.-P. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, 137–144.

Sandygulova, A.; Yershov, A.; Zhanatkyzy, A.; and Telisheva, Z. 2025. ChildACT: Child Action Recognition Dataset in RGB Data. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1088–1092. IEEE.

Sullivan, J.; Mei, M.; Perfors, A.; Wojcik, E.; and Frank, M. C. 2021. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5: 20–29.

Sun, K.; Bai, Y.; Yang, Z.; Zhang, J.; Qi, J.; Hou, L.; and Li, J. 2025. Hard Negative Contrastive Learning for Fine-Grained Geometric Understanding in Large Multimodal Models. *arXiv preprint arXiv:2505.20152*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Verschaffel, L.; Schukajlow, S.; Star, J.; and Van Dooren, W. 2020. Word Problems in Mathematics Education: A Survey. *ZDM: The International Journal on Mathematics Education*, 52(1): 1–16.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Xu, Y.; Wang, D.; Yu, M.; Ritchie, D.; Yao, B.; Wu, T.; Zhang, Z.; Li, T. J.-J.; Bradford, N.; Sun, B.; et al. 2022. Fantastic Questions and Where to Find Them: FairytaleQA–An Authentic Dataset for Narrative Comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 447–460.

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Yuan, J.; Peng, T.; Jiang, Y.; Lu, Y.; Zhang, R.; Feng, K.; Fu, C.; Chen, T.; Bai, L.; Zhang, B.; et al. 2025. MME-Reasoning: A Comprehensive Benchmark for Logical Reasoning in MLLMs. *arXiv preprint arXiv:2505.21327*.

Zhou, J.; Wang, S.; Zhao, S.; He, J.; Sun, H.; Wang, H.; Liu, C.; Kong, A.; Guo, Y.; Yang, X.; et al. 2024a. Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. *arXiv preprint arXiv:2409.18584*.

Zhou, M.; Liang, H.; Li, T.; Wu, Z.; Lin, M.; Sun, L.; Zhou, Y.; Zhang, Y.; Huang, X.; Chen, Y.; et al. 2024b. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*.