COGDEVELOP2K: REVERSED COGNITIVE DEVELOPMENT IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Are Multi-modal Large Language Models (MLLMs) stochastic parrots? Do they genuinely understand? This paper aims to explore the core cognitive abilities that human intelligence builds upon to perceive, comprehend, and reason in MLLMs. To this end, we propose **CogDevelop2K**, a comprehensive benchmark that spans 12 sub-concepts from primitive knowledge like object permanence and boundary to more complex abilities like intentionality understanding, structured via the developmental trajectory of a human mind. We evaluate 46 MLLMs on our benchmarks. Surprisingly, we observe a **reversed** cognitive developmental trajectory compared to humans. Comprehensively, we further evaluate the influence of evaluation strategies and prompting techniques.

1 INTRODUCTION

Building on the foundation of advanced large language models (LLMs), multi-modal large language models (MLLMs) have recently demonstrated human-level performance in complex tasks involving high-level reasoning, perception, and cognition (Li et al., 2024a; Liu et al., 2024; Gemini, 2023; Fu et al., 2023; OpenAI, 2023), such as Spatial Reasoning (Chen et al., 2024; Cai et al., 2024), OCR (Mori et al., 1999), Scene Understanding (Cordts et al., 2016; Chen et al., 2017), Action Recognition (Jhuang et al., 2013; Herath et al., 2017) and Prediction (Lan et al., 2014; Kong & Fu, 2022). The progress in MLLMs has reignited hopes for achieving Artificial General Intelligence (AGI). However, recent studies have shown that even state-of-the-art MLLMs face critical limitations as compared to human intelligence. On the one hand, they perform poorly on rudimentary reasoning tasks like counting Paiss et al. (2023), compositional reasoning Yuksekgonul et al. (2022) and spatial reasoning Hoehing et al. (2023) despite their excellence at high-level reasoning tasks on similar domains (Paiss et al., 2023; Rahmanzadehgervi et al., 2024). On the other hand, said excellency often does not appear to translate to more generalized and real-world contexts (Shiffrin & Mitchell, 2023; Zhang et al., 2024). To explore the underlying reason for these limitations, we draw inspirations from human cognitive development.

Past research has shown that humans exhibit a series of rudimentary yet robust abilities in domains such as object, number, space, action, and social cognition at a very young age. Such abilities, often known as ”core” cognition, grounds the set of diverse and complex abilities of human intelligence that develop later (Spelke et al., 1992; 1994; 1995; Spelke & Kinzler, 2007; Baillargeon & Carey, 2012; Mitchell, 2020; 2021). From infancy to early adulthood, human cognition develops along a structured trajectory, with interdependent relations between early, simple abilities and late, complex abilities. For instance, the ability to imagine the perspectives of others typically develops between the ages of 3 and 6 (Piaget & Inhelder, 1969), while the capacity to fully comprehend others’ intentions matures around age 12 (Wimmer & Perner, 1983; Wellman et al., 2001; Liu et al., 2008). At the same time, the ability to understand other people’s intentions largely depends on the the ability to understand other people’s perspectives (Iacoboni, 2009; De Waal & Preston, 2017; Liu et al., 2017; Caviola et al., 2021; Ninomiya et al., 2020). An influential account of human learning has suggested that cognitive development is fundamentally driven by the increase of computational/representational power of the system, which allows for more complex mental operations to be performed on external data (Fodor, 1975; Pylyshyn, 1980; Halford et al., 1998; Fodor, 2008). However, while high-level abilities emerge directly due to enhanced operational resources, these operations are critically guided by the ”core” cognition system that has enabled the system to possess a rudimentary understanding of each cognitive domain. This early-stage grounding not only empowers humans to achieve a reliable performance at basic yet widely-applicable tasks starting from very young ages, but is also precisely what supports high-level abilities to robustly direct task-relevant behaviors despite the nuanced signals exist in the environment (Mitchell, 2021).

Given this line of evidence, we suggest that the absence of a core cognition system may provide a joint account of the two limitations faced by current MLLMs: poor performance on basic reasoning and the lack of robustness with high-level reasoning. In order to assess this hypothesis, We draw on theoretical and empirical approaches from developmental science to create benchmarks that evaluate simple and complex cognitive abilities in large vision-language models that are interrelated along the developmental trajectory. On a high level, we follow Jean Piaget's theory of cognitive development, which identifies four stages in children: sensorimotor, preoperational, concrete operational, and formal operational (Piaget, 1950; Piaget & Inhelder, 1969; 1974). During the sensorimotor stage, infants acquire knowledge through sensory experiences and actions, developing an understanding of basic object properties, such as permanence, continuity, and boundaries. In the preoperational stage, symbolic representation emerges, along with a grasp of basic physical properties. The concrete operational stage is characterized by the development of logical thinking and an understanding of intuitive physics. Finally, the formal operational stage introduces more advanced cognitive abilities, including abstraction, hypothetical reasoning, counterfactual thinking, and tool use. The interdependence and developmental trajectories of these abilities can be mapped in terms of a tree-like structure (as illustrated in Fig. 1).

To evaluate the performance of MLLMs on the core cognitive abilities, we curate the first-ever vision cognitive development benchmark, termed as CogDeveop2K, which consists of a total of 2519 questions with 2517 images and 455 videos. Then, we evaluate 46 MLLM models on our benchmark that spans all four stages of cognitive development. We introduce a novel multi-frame question format to evaluate models' co-reference, cognitive reasoning and temporal understanding capability simultaneously. Forty-seven models are compared against a human baseline under zero-shot conditions using 11 different prompts (including no prompt). Surprisingly, while prompts can boost model performance by 8.1%, models still demonstrate reversed trends in cognitive development against those observed in children.

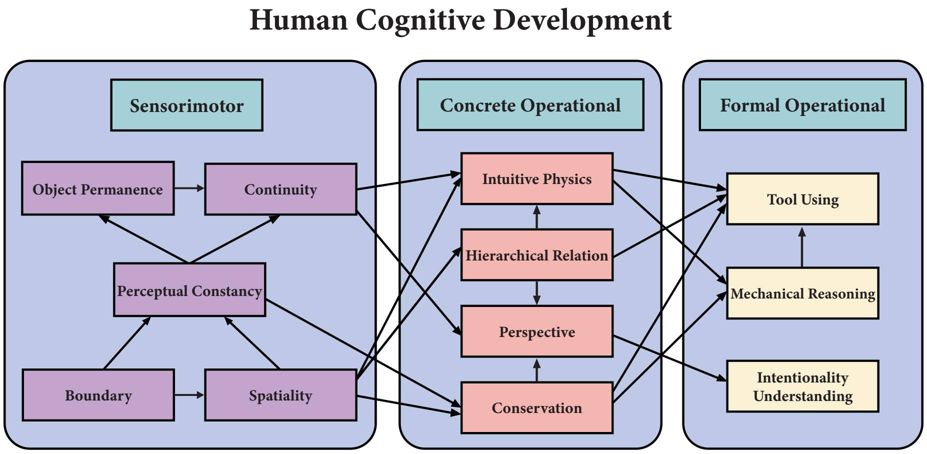


Figure 1: Map of core cognitive concepts during human developmental stages

2 CogDevelop2K

2.1 Multi-modal Large Language Models

The Vision Language Model (VLM) has a long history from Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) (Karpathy & Fei-Fei, 2014; Vinyals et al., 2015) to unified modeling of visual and text modality with transformers (Li et al., 2019; Xu et al., 2023; Tan & Bansal, 2019; Alayrac et al., 2022; Radford et al., 2021). With the advancement of Large Language

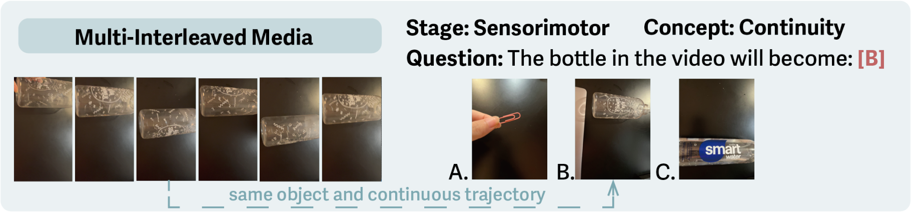

Multi-Interleaved Media
Stage: Sensorimotor
Concept: Continuity
Question: The bottle in the video will become: [B]
A.
B.
C.
smart
same object and continuous trajectory

Figure 2: A video-image interleaved example of multi-frame questions. To correctly infer the answer, model needs to understand the question by mapping each image (co-reference) to its option letter, to understand correlation between frames (temporal understanding) and to infer the possible trajectory of the bottle (reasoning).

Models (LLMs), existing state-of-the-art MLLMs (Liu et al., 2024; Li et al., 2023) adopt open-sourced Large Language Models such as Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023). Instruction tuning is also introduced to further improve the task generalization ability of MLLMs (Liu et al., 2024; Dai et al., 2023). To acquire open-ended conversation abilities, LLaVA (Liu et al., 2024) proposes to distill the conversational abilities of ChatGPT to MLLMs, boosting performance by a large margin, which becomes a defacto procedure in the area (Wang et al., 2023; Bai et al., 2023; Gemini, 2023; Team, 2024; Sun et al., 2023; Li et al., 2022).

2.2 Human Cognitive Development

The sensorimotor stage is the first stage of cognitive development proposed by Jean Piaget (Piaget, 1952; Piaget & Inhelder, 1974). Spanning from birth to approximately 2 years of age, this stage is characterized by infants' understanding of the world through their sensory experiences and motor actions. Several prominent features of human intelligence develop during this period. First, infants develop object permanence, that they realize objects and people continue to exist even when not in direct sight, or being heard or touched (Baillargeon et al., 1985). They start to understand that there is a sense of continuity for the ways that objects exist, and the inductive bias of continuity is essential, e.g., for recognizing objects when occluded or for continuously tracking objects (Spelke et al., 1995; Le Poidevin, 2000). Infants also develop the sense of boundary during this stage, namely, the ability to recognize where one object ends and another begins (Kestenbaum et al., 1987; Jackendoff, 1991). Lastly, infants develop spatial and perceptual constancy by the end of sensorimotor stage. Spatiality refers to the ability to perceive the position and distance of objects relative to oneself and each other, and recognize the spatial invariance between them when presented by various sensory experiences (Hermer & Spelke, 1996; Bell & Adams, 1999).

Preoperational and concrete operational stage are the second and third stage of Piaget's cognitive development. Typically spanning over 2 to 7 years of age, preoperational stage is the transitional stage to concrete operational stage, which children enter around 7 years of age. During this period, children begin to develop internalized mental actions supported by organized structures that can be manipulated and reversed in systematic ways, known as mental operations (Janet, 1905; Kirkpatrick, 1908; Piaget, 1950; Piaget & Inhelder, 2014; Miller, 2016). Through mental operations, children are then able to rigidly perform tasks that are previously unreachable, such as thinking from other people's perspectives, understanding hierarchical relations of objects, and reasoning about physical events in the world. These tasks require not only rudimentary understandings of physical concepts, which gradually became in place during preoperational stage, but also relational and transformational reasoning that can only be done through mental operations (Piaget & Inhelder, 1974; Church & Goldin-Meadow, 1986; Houdé, 1997). Since preoperational stage is mostly meaningful as the transitional period preceding concrete operational stage, we do not have evaluation dimensions specifically targeting the stage. However, tasks targeting concrete operational stage could assess the existence of knowledge associated with preoperational stage, such as law of conservation (Piaget, 1952; Halford, 2011; Houdé, 1997).

The formal operational stage is the fourth and final stage in Piaget's theory of cognitive development, typically emerging around 11 or 12 years of age and continuing into adulthood (Inhelder & Piaget, 1958). Starting in this stage, one is able to systematic and flexibly apply mental operations to not only concrete, physical domains but also abstract, formal domains (Kuhn & Angelev, 1976; Shayer, 1979; Huitt & Hummel, 2003). Foremost, this stage is characterized by the development of complex thinking and reasoning abilities, such as abstraction, pattern recognition, the employment of logic, and hypothetical and counterfactual reasoning (Piaget, 1950; Inhelder & Piaget, 1958). These cognitive advancements pave the way for more sophisticated abilities to interact with the physical world, marked by mechanical reasoning and tool use (O'Brien & Shapiro, 1968). Together, there is the advancement in social cognition, characterized by a deeper understanding of intentions, actions, and the reasoning behind them (Meltzoff, 1999).

2.3 Evaluation Dimension

Boundary Boundary refers to the cognitive understanding of where one object ends and another begins, an essential aspect of perceiving and understanding the physical world (Kestenbaum et al., 1987). Without understanding boundary, it seems very hard to construct a concept of object (Berkeley, 1709; Jackendoff, 1991).

Spatiality Spatiality, particularly demonstrated through the A-not-B task, involves a child's understanding of the location of objects in relation to their environment (Bell & Adams, 1999). In a classic A-not-B task, an object is hidden at location A (such as under a cup) and the child successfully finds it several times. Then, the object is visibly moved to a different location B (under a different cup), in full view of the child. Younger infants often make the error of searching for the object at the original location A, indicating a developmental stage where their understanding of object spatiality is still forming.

Perceptual Constancy Perceptual constancy is the cognitive ability to perceive objects as being constant in their properties, such as size, shape, and color, despite changes in perspective, distance, or lighting (Rutherford & Brainard, 2002; Khang & Zaidi, 2004; Green, 2023). For instance, consider a red ball being thrown in a park. To an observer, the ball appears smaller as it moves farther away, yet the observer understands it remains the same size throughout its trajectory.

Object Permanence Permanence, or specifically object permanence, is the idea that objects continue to exist even when they are not visible (Baillargeon, 1986; Spelke et al., 1992). Imagine a simple scene: a small child playing peek-a-boo. In the beginning, when the caregiver covers their face with their hands, the child might seem surprised or even distressed, thinking the person has disappeared. However, as children's understanding of permanence develops, they begin to realize that just because they can't see the person's face, it doesn't mean the person is gone.

Continuity Continuity is the cognitive prior in humans that in our world, objects usually exist in a consistent and continuous manner, even moving out of sight (Spelke et al., 1995; Le Poidevin, 2000; Spelke et al., 1994; Yantis, 1995; Yi et al., 2008; Bertenthal et al., 2013). Picture a train moving through a tunnel: as it enters one end, yet we naturally expect it to emerge from the other end, if the train is long enough. This expectation demonstrates our understanding of object continuity. Even though the train is not visible while it's inside the tunnel, we know it continues to exist.

Conservation Conservation refers to the ability to understand that certain properties of physical entities are conserved after an object undergoes physical transformation (Piaget & Inhelder, 1974). This is instantiated in their ability to tell that quantities of physical entities across different domains, such as number, length, solid quantity and liquid volume, will remain the same despite adjustments of their arrangement, positioning, shapes, and containers (Halford, 2011; Craig et al., 1973; Piaget & Inhelder, 1974; Houdé et al., 2011; Poirel et al., 2012; Marwaha et al., 2017; Viarouge et al., 2019). For example, when a child watches water being poured from a tall, narrow glass into a short, wide one, a grasp of liquid conservation would lead them to understand that the amount of water remains the same even though its appearance has changed.

Perspective-taking Perspective-taking is the ability to view things from another's perspective. This ability has seminal importance both to the understanding of the physical world as well as to the competence in social interactions (Wimmer & Perner, 1983; Wellman, 1992; Liu et al., 2008; Barnes-Holmes et al., 2004). The Three Mountain Task first invented by Jean Piaget is widely used

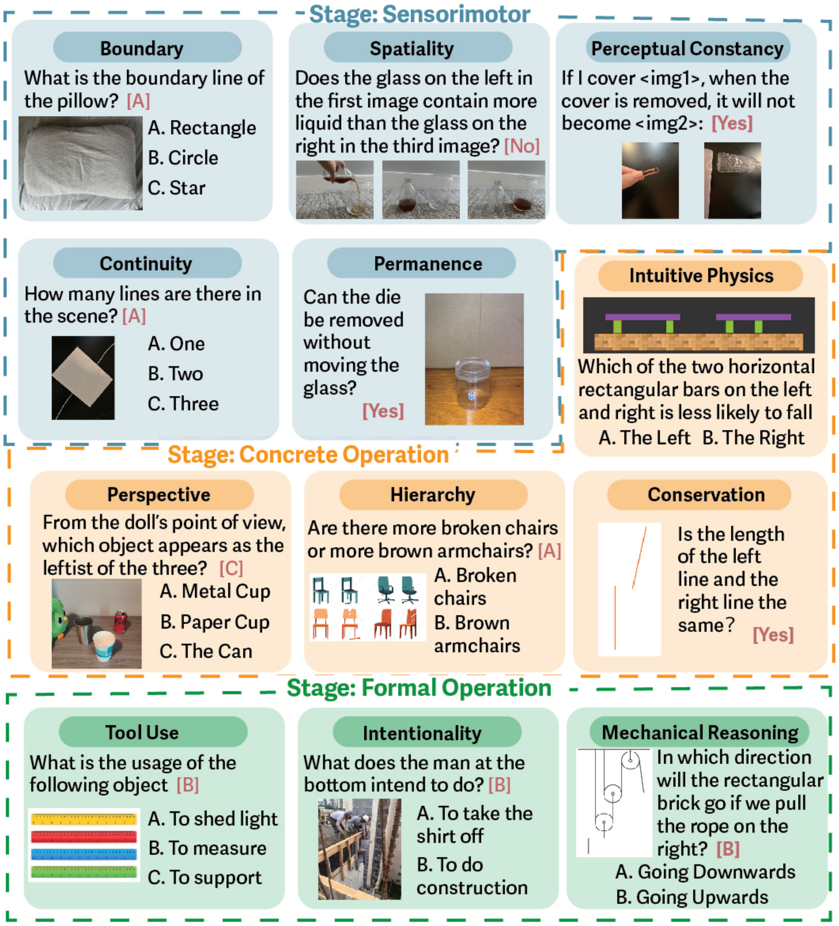

Stage: Sensorimotor
Boundary
What is the boundary line of the pillow? [A]
A. Rectangle
B. Circle
C. Star
Spatiality
Does the glass on the left in the first image contain more liquid than the glass on the right in the third image? [No]
Perceptual Constancy
If I cover <img1>, when the cover is removed, it will not become <img2>: [Yes]
Continuity
How many lines are there in the scene? [A]
A. One
B. Two
C. Three
Permanence
Can the die be removed without moving the glass?
[Yes]
Intuitive Physics
Which of the two horizontal rectangular bars on the left and right is less likely to fall
A. The Left B. The Right
Stage: Concrete Operation
Perspective
From the doll's point of view, which object appears as the leftist of the three? [C]
A. Metal Cup
B. Paper Cup
C. The Can
Hierarchy
Are there more broken chairs or more brown armchairs? [A]
A. Broken chairs
B. Brown armchairs
Conservation
Is the length of the left line and the right line the same? [Yes]
Stage: Formal Operation
Tool Use
What is the usage of the following object [B]
A. To shed light
B. To measure
C. To support
Intentionality
What does the man at the bottom intend to do? [B]
A. To take the shirt off
B. To do construction
Mechanical Reasoning
In which direction will the rectangular brick go if we pull the rope on the right? [B]
A. Going Downwards
B. Going Upwards

Figure 3: We demonstrate examples of different sub-concepts from the three stages.

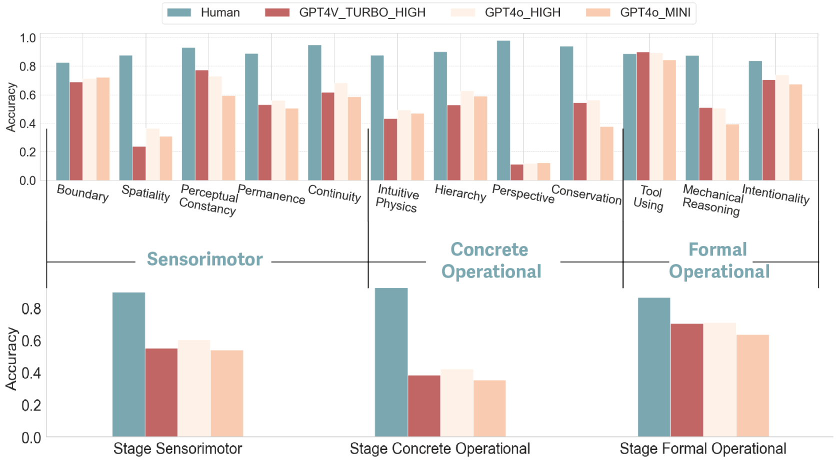


Figure 4: Reversed Cognitive Development in Advanced Models

in developmental psychology laboratories as the gold standard for testing perspective-taking abilities in children (Piaget & Inhelder, 1969)

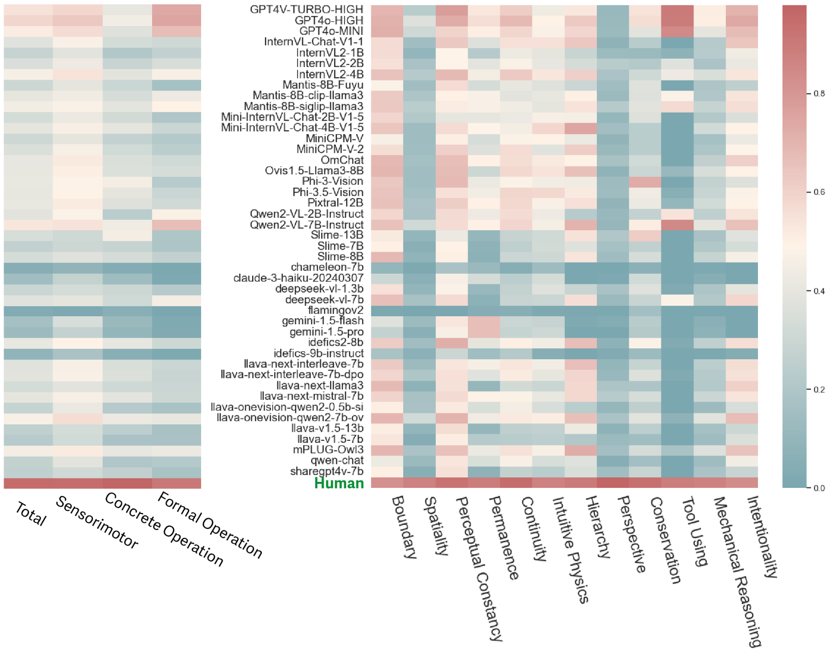


Figure 5: Multimodal Large Language Models and Human Performances

Hierarchical Relation Hierarchical relation refers to the ability to organize objects or concepts into structured categories and subcategories, which are supported by the development of mental operations marked by class inclusion and transitivity (Shipley, 1979; Winer, 1980; Chapman & McBride, 1992). Class inclusion refers to the ability to recognize that some classes or groups of objects are subsets of a larger class. For example, a child in the concrete operational stage is able to understand that all roses are flowers, but not all flowers are roses (Borst et al., 2013; Politzer, 2016). This concept is essential for one's systematic and logical organizations of conceptual knowledge. Transitivity refers to the ability to understand logical sequences and relationships between objects (Andrews & Halford, 1998; Wright & Smailes, 2015). For instance, if a child knows that Stick A is longer than Stick B, and Stick B is longer than Stick C, they can deduce that Stick A is longer than Stick C.

Intuitive Physics Intuitive physics refers to the ability of humans to predict, interact with, and make assumptions about the physical behavior of objects in their world (Michotte, 1963). As children grow, they transition from simplistic understandings, such as expecting unsupported objects to fall, to more complex theories, such as grasping the principles of inertia (Spelke et al., 1994; Kim & Spelke, 1999) and gravity (Vasta & Liben, 1996; Kim & Spelke, 1999; Li et al., 1999).

Intention Understanding Intention understanding involves recognizing and interpreting the actions of others (Searle, 1979; Rosenthal, 1991). This process is not just about observing a behavior but also about understanding the goal behind it (Baker et al., 2009; Gandhi et al., 2021). For example, seeing someone reaching for a cup is not just about recognizing the physical action but understanding the intention behind it (e.g., they want to drink).

Mechanical Reasoning Mechanical reasoning refers to the ability to understand and apply mechanical concepts and logical principles to solve problems (Allen et al., 2020). This cognitive concept first involves the ability to interpret and predict the behaviors of complex physical systems and understand how different mechanisms of the systems work. Second, mechanical reasoning requires the ability to apply logic rules, such as induction, abduction, syllogism (O'Brien & Shapiro, 1968; Cesana-Arlotti et al., 2018), and reasoning forms, such as hypotheticals and counterfactual (Byrne, 2016), to figure out how to manipulate these systems to achieve a desired outcome (Hegarty, 2004).

Table 1: Main statistics in CogDevelop2k. All the questions are image/video-text interleaved.

Statistic	Number
single-frame	1677
multi-frame	842
* multiple images	200
* single video	401
* multiple videos	124
* video-image-text	117
total	2519

Tool Using Tool using refers to the ability to manipulate tool objects in its environment as aids in achieving a specific goal, such as obtaining food or modifying the surroundings. A lot of cognitive components involved in tool using ability, such as affordances, referring to computing the action possibilities offered to the agent by the tool with reference to the agent's sensorimotor capabilities (Gibson, 1979). For example, a door handle affords pulling or pushing, as how the door should be operated by a human agent.

2.4 Data Source

CogDevelop2K comprises 2517 images and 445 videos with multimodal options and questions, crawled primarily from networks as well as self-recorded content. Our targeted platforms include Wikipedia, Reddit, Twitter, Quora, and TieBa. Some of the captions were adapted from user comments to ensure content diversity and relevance. Videos for intuitive physics were either self-recorded or produced using Physion[1].

All concept questions were annotated by four researchers with cognitive science and computer science background, then reviewed by two independent researchers. For a question to pass the screening stage, a minimum correctness rate of 95% was required from both reviewers.

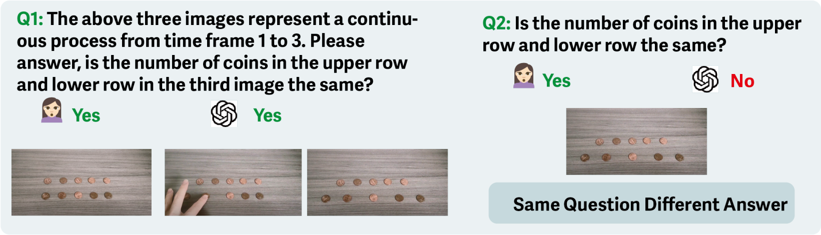
2.5 Dataset Design

Existing datasets typically support only one question-answer format or single modality type, which hinders the assessment of reasoning capabilities across different modalities within the same domain. For instance, current interleaved image understanding and video understanding models cannot be effectively compared on the same question. To address this limitation, we include multiple Q&A formats (e.g., multiple-choice, true/false, and numeric question-answer) and complicate question-answering by incorporating a new image-video-text interleave format as shown in Fig. 2. To further explore the cognitive development capabilities of models across these modalities, we optimize our formality as follows:

Addressing Weak Image-Text Correlation and Imbalance In existing interleaved image-text datasets, the correspondence between images and text is often loose, and text provides marginal information for image modeling. This imbalance can cause models to over-rely on textual information, especially when text segments are lengthy (Lin et al., 2023) . To address this issue and focus on the image understanding abilities of the model, we eliminate sentences that describe the image. This ensures that the textual information is highly relevant to but does not overlap with the image content.

Testing Co-Reference, Reasoning, and Temporal Understanding with novel Multi-Frame Questions Multi-frame questions can simultaneously evaluate a model's three inference ability: *Co-Reference*, *Reasoning*, and *Temporal Understanding* (Jiang et al., 2024). Co-reference involves linking natural language descriptions with specific image inputs (e.g., ”the first image” or ”A.<img 1>”). Reasoning requires models to make decisions based on cognitive knowledge, such as describing spatial relationships. Temporal Understanding, on the other hand, tests the model's capability to

[1] https://physion.net/


Q1: The above three images represent a continuous process from time frame 1 to 3. Please answer, is the number of coins in the upper row and lower row in the third image the same?

Q2: Is the number of coins in the upper row and lower row the same?

Figure 6: MLLMs' Dissociation Between Law of Conservation and Rudimentary Quantity Understanding as Exemplified by GPT-4o

comprehend sequences of frames in terms of temporal order (multi-frame) and correlation (multi-view) (Li et al., 2024b). Existing interleaved multi-image datasets can not adequately test all three properties simultaneously. For example, video datasets with temporal information often include only a single video, while multi-image datasets that require co-reference lack temporal dependencies. To address this, we introduce multi-video interleaving and video-image interleave formats (multi-frame) to evaluate all three properties concurrently. The statistics of the dataset are presented in Table 1.

2.6 Evaluation Strategy

We comprehensively evaluate models' capability of cognitive reasoning using ***46 multi-image interleave MLLMs*** with 11 different promopts. The two evaluation baselines are outlined as follows:

Human baseline We recruit 22 participants, all of whom are college students proficient in English. Participants are instructed to skip a question if the question is phrased ambiguously or is too complicated to answer in 90 seconds. This question is marked as failed if the human participant does not provide an answer.

Zero-Shot-448^2-Circular Baseline Similar to previous studies (Lu et al., 2022), the zero-shot setup follows the format of $Q(M)T \rightarrow A$, where the input includes the question text (Q), task description (T), and multiple options (M) concatenated as tokens, with the output being the predicted answer (A). Given that model predictions can exhibit bias in multiple-choice settings, we implemented circular evaluation as baseline. In circular evaluation, all answer options are shifted one position at a time, ensuring that the correct answer appears in each option slot. Only when the model correctly predicts all shifted answers is it considered accurate (Liu et al., 2023). All images and videos were resized to 488^2.

Prompts Strategically crafted prompts can enhance model performance, regardless of whether fine-tuning is applied (Bsharat et al., 2023; Yang et al., 2023). To mitigate this, we use image-independent contexts, such as relevant concept introductions and character assignments, which encourage models to reason beyond the provided textual information. The prompts we used can be categorized into leading words, deeper thing, role assignment, reward or penalty, and explanation.

3 Results

We systematically evaluate 48 Multi-modal Large Language Models on the CogDevelop2K benchmark, which spans 12 cognitive concepts designed to assess a broad range of the developmental trajectory of Multi-modal Large Language Models. These abilities substantiate core cognition ranging from object permanence and boundary to mechanical reasoning and intentionality understanding. The models were tested across multiple question formats and ten prompt variations, yielding a comprehensive assessment of their core cognition. For example, in the sensorimotor stage, GPT families show moderate performance, with accuracy scores between 0.4 and 0.6. In the concrete operational stage, GPT families show lower performance, with accuracy scores between 0.2 and 0.4. Nevertheless, in the formal operational stage, GPT families show stronger performance, with accuracy scores

Table 2: Evaluation of different prompting techniques. The best result is achieved when the concept explanation is provided to the model. We highlight the improvement over empty string in red.

Category	Prompt	GPT4V Turbo High	GPT4o High	GPT4o Mini
	Empty String	0.519	0.555	0.487
Leading Word	1. Let's think step by step.	0.531	0.577	0.489
	2. Take a deep breath and answer this question carefully.	0.522	0.562	0.489
Deeper Thinking	3. Please answer the question and provide an explanation.	0.518	0.562	0.499
	4.Please answer the question and explain to me in simple terms.	0.476	0.569	0.501
	5. Please answer the question and ensure that your answer is unbiased and doesn't rely on stereotypes.	0.522	0.575	0.478
Role Assignment	6. (Assign assistance's role) You are an expert on cognitive science and are familiar with [Concept name]	0.565	0.617	0.545
	7. (Assign audiance's role) Please answer the question and explain it to me like I am 11 years old.	0.538	0.564	0.496
Reward & Penalty	8. Please answer the question carefully. I'm going to tip you 200 dollars for a better solution.	0.528	0.563	0.487
	9. Please answer the question carefully. You will be penalized if your answer is incorrect.	0.522	0.566	0.491
Explanation	10. Please read the concept explanation and then answer the related question. Concept: [concept description].	**0.586** (+ 0.067)	**0.636** (+ 0.081)	**0.547** (+ 0.06)

between 0.6 and 0.8. Surprisingly, we find an inverse cognitive developmental trajectory compared to humans in more advanced models, which are typically regarded as state-of-the-art (Fig. 3 and Fig. 4).

Influence of Prompts. We investigate the influence of different prompting techniques on the performance of MLLMs on our benchmark. As illustrated in Table 2, we explore 10 different prompting techniques (divided into 5 categories). We observe that most prompts are useful on our benchmark, increasing the averaged performance by at least 1%. Concept explanation, which offers a clearer context of the question to the MLLMs, surpasses all the other prompts by at least 6%.

4 Discussion

Our results demonstrate that MLLMs exhibit an intriguing reverse cognitive development pattern. Namely, they are systematically proficient at complex tasks that are typically understood to require abilities underlying simple tasks that they perform poorly. This finding supports the hypothesis that MLLMs lack a ”core” cognition system, which, as we suggested, could appear as challenge to the current foundational architecture of MLLMs as a long-term solution to achieve human-like general intelligence (Summerfield, 2022). Specifically, the inability to implement core knowledge in artificial intelligence models prevent them from achieving human-level robustness in performances, even if such models seem to excel at certain complex cognitive reasoning tasks (Mitchell, 2020; 2021; Shiffrin & Mitchell, 2023; Palmarini & Mitchell, 2024). MLLMs' poor performances on foundational concepts like spatiality, permanence, continuity, and perspective-taking, which directly reflect upon grasps of core knowledge, while achieve proficiency in complex concepts like tool using and intention understanding exactly exemplifies this concern.Specifically, the developmental trajectory of human cognition is marked by complex cognitive abilities being grounded upon extremely robust understandings of a series of foundational concepts, namely core knowledge (Spelke, 2000; Spelke & Kinzler, 2007). Through early stages of development, children exhibit rudimentary yet stable understandings of objects, actions, number, space, and social partners, each dimension laying the foundations for the acquisitions of complex abilities in later life.

While a straightforward explanation of this intriguing reverse cognitive development pattern is the absence of ”core” cognitive abilities, further research is needed in order to arrive at a comprehensive explanation for the mechanistic details of why such abilities fails to emerge as opposed to the case in humans. At the same time, existing theories may provide some preliminary insights into this question. In particular, it is likely that scaling of parameters and training data mainly enhanced the operational resources available to MLLMs but not necessarily allow the grasp of core knowledge (Fodor & Pylyshyn, 1988; Kello et al., 2010; von der Malsburg, 2024). This is because neural networks do not themselves contain built-in domain-specific information but are only capable of representing such knowledge through the topological structures of the connection units and the values of their corresponding weights. Therefore, while MLLMs may have gathered all the information needed for constructing a system of core knowledge from the training data , it is likely represented in an extremely distributed fashion that prevents reliable tracing and retrieval (Hinton et al., 1986; Fodor & Pylyshyn, 1988; Yang et al., 2022). If the task condition do not offer enough cues to trigger the required aspect of the system, then MLLMs may fail at the task even if they ”know” the concept that is needed for providing the correct answer.

An initial validation of this explanation is that the best model performance on our benchmarks are achieved when the prompts contain an explanation of the concept tested by the experiment (see Table 2). Such an effect is notable given that unlike regular in-context learning prompts, concept explanations contain mainly high-level, abstract information that has no direct relevance to the context of the task. Opposingly, it may be seen as a rather immediate attempt of conceptual pre-training that facilitates the retrieval and synthesis of ”core” conceptual resources distributed throughout the networks. The effectiveness of such an attempt calls for the attention to training measures for MLLMs aiming at bettering the implementation and use of core cognition, which appears to be a promising approach toward more robust and comprehensive foundational models.

5 Conclusion

In this paper, we explored the cognitive capabilities of Multi-modal Large Language Models (MLLMs) through the lens of core cognitive abilities that underpin human intelligence. By introducing CogDevelop2K, a novel benchmark that spans 12 subconcepts across developmental stages, we aimed to assess the fundamental understanding and reasoning capacities of MLLMs. Our evaluation of 46 models revealed intriguing insights, including a reversed cognitive developmental trajectory compared to humans. This finding raises questions about whether MLLMs truly comprehend tasks or simply exhibit performance without genuine understanding. These results underscore the need for further investigation into the cognitive foundations of MLLMs, as well as the influence of evaluation strategies and prompting techniques in shaping their outcomes. Ultimately, this study serves as a step toward unraveling the nature of MLLM intelligence and their potential limitations in mirroring human cognition.

References

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

Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020.

Glenda Andrews and Graeme S Halford. Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development*, 13(4):479–513, 1998.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Renee Baillargeon. Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, 23(1):21–41, 1986.

Renée Baillargeon and Susan Carey. Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*, 1, 2012.

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

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Yvonne Barnes-Holmes, Louise McHugh, and Dermot Barnes-Holmes. Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25, 2004.

Martha Ann Bell and Stephanie E Adams. Comparable performance on looking and reaching versions of the a-not-b task at 8 months of age. *Infant Behavior and Development*, 22(2):221–235, 1999.

George Berkeley. *An Essay Towards A New Theory of Vision*. Dublin, 1709.

Bennett I Bertenthal, Gustaf Gredebäck, and Ty W Boyer. Differential contributions of development and learning to infants' knowledge of object continuity and discontinuity. *Child Development*, 84 (2):413–421, 2013.

Grégoire Borst, Nicolas Poirel, Ariette Pineau, Mathieu Cassotti, and Olivier Houdé. Inhibitory control efficiency in a piaget-like class-inclusion task in school-age children and adults: a developmental negative priming study. *Developmental psychology*, 49(7):1366, 2013.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

Ruth MJ Byrne. Counterfactual thought. *Annual review of psychology*, 67(1):135–157, 2016.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Lucius Caviola, Stefan Schubert, and Joshua D Greene. The psychology of (in) effective altruism. *Trends in Cognitive Sciences*, 25(7):596–607, 2021.

Nicoló Cesana-Arlotti, Ana Martín, Ernő Téglás, Liza Vorobyova, Ryszard Cetnarski, and Luca L Bonatti. Precursors of logical reasoning in preverbal human infants. *Science*, 359(6381):1263–1266, 2018.

Michael Chapman and Michelle L McBride. Beyond competence and performance: Children's class inclusion strategies, superordinate class cues, and verbal justifications. *Developmental Psychology*, 28(2):319, 1992.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

R. Breckinridge Church and Susan Goldin-Meadow. The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23:43–71, 1986.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Grace J Craig, Jean A Love, and Ellis G Olim. An experimental test of piaget's notions concerning the conservation of quantity in children. *Child Development*, 44(2):372–375, 1973.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Frans BM De Waal and Stephanie D Preston. Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, 18(8):498–509, 2017.

Jerry A Fodor. *The Language of Thought*. MIT Press, 1975.

Jerry A Fodor. *LOT 2: The language of thought revisited*. Oxford University Press, 2008.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, 34:9963–9976, 2021.

Gemini. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

James J Gibson. *The Ecological Approach to Visual Perception. (1st ed.)*. Psychology Press, 1979.

EJ Green. Perceptual constancy and perceptual representation. *Analytic Philosophy*, 2023.

G S Halford. An experimental test of piaget's notions concerning the conservation of quantity in children. *Journal of experimental child psychology*, 6(1):33–43, 2011.

Graeme S Halford, William H Wilson, and Steven Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences*, 21(6):803–831, 1998.

Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6): 280–285, 2004.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

Linda Hermer and Elizabeth Spelke. Modularity and development: The case of spatial reorientation. *Cognition*, 61(3):195–232, 1996.

Geoffrey E Hinton et al. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, pp. 12. Amherst, MA, 1986.

Nils Hoehing, Ellen Rushe, and Anthony Ventresque. What's left can't be right–the remaining positional incompetence of contrastive vision-language models. *arXiv preprint arXiv:2311.11477*, 2023.

Olivier Houdé. Numerical development: From the infant to the child. *Cognitive Development*, 12 (3):373–391, 1997.

Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poirel, Guy Perchey, Céline Lanoë, Amélie Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lamberton, Mathieu Vigneau, Gabriel Wisniewski, Jean-René Vicet, and Bernard Mazoyer. Functional magnetic resonance imaging study of piaget's conservation-of-number task in preschool and school-age children: a neo-piagetian approach. *Journal of experimental child psychology*, 110(3):332–346, 2011.

William Huitt and John Hummel. Piaget's theory of cognitive development. *Educational psychology interactive*, 3(2):1–5, 2003.

Marco Iacoboni. Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60(1): 653–670, 2009.

Bärbel Inhelder and Jean Piaget. *The Growth of Logical Thinking from Childhood to Adolescence*. Basic Books, 1958.

Ray Jackendoff. Parts and boundaries. *Cognition*, 41(1-3):9–45, 1991.

Pierre Janet. Mental pathology. *Psychological Review*, 12(2-3):98, 1905.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 3192–3199, 2013.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv: 2310.06825*, 2023.

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

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv: 1412.2306*, 2014.

Christopher T Kello, Gordon DA Brown, Ramon Ferrer-i Cancho, John G Holden, Klaus Linkenkaer-Hansen, Theo Rhodes, and Guy C Van Orden. Scaling laws in cognitive sciences. *Trends in cognitive sciences*, 14(5):223–232, 2010.

Roberta Kestenbaum, Nancy Termine, and Elizabeth S Spelke. Perception of objects and object boundaries by 3-month-old infants. *British journal of developmental psychology*, 5(4):367–383, 1987.

Byung-Geun Khang and Qasim Zaidi. Illuminant color perception of spectrally filtered spotlights. *Journal of Vision*, 4(9):2–2, 2004.

In-Kyeong Kim and Elizabeth S Spelke. Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science*, 2(3):339–362, 1999.

EA Kirkpatrick. The part played by consciousness in mental operations. *The Journal of Philosophy, Psychology and Scientific Methods*, 5(16):421–429, 1908.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Deanna Kuhn and John Angelev. An experimental study of the development of formal operational thought. *Child Development*, pp. 697–706, 1976.

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pp. 689–704. Springer, 2014.

Robin Le Poidevin. Continuants and continuity. *The Monist*, 83(3):381–398, 2000.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv: 2205.12005*, 2022.

Chieh Li, Ronald L Nuttall, and Shuwen Zhao. A test of the piagetian water-level task with chinese students. *The Journal of Genetic Psychology*, 160(3):369–380, 1999.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*, 2023.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

David Liu, Henry M Wellman, Twila Tardif, and Mark A Sabbagh. Theory of mind development in chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44(2):523, 2008.

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

Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017.

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? *arXiv preprint arXiv:2307.06281*, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Sugandha Marwaha, Mousumi Goswami, and Binny Vashist. Prevalence of principles of piaget's theory among 4-7-year-old children and their correlation with iq. *Journal of clinical and diagnostic research: JCDR*, 11(8):ZC111, 2017.

Andrew N Meltzoff. Origins of theory of mind, cognition and communication. *Journal of communication disorders*, 32(4):251–269, 1999.

A Michotte. *The perception of causality.* Basic Books, 1963.

Patricia H Miller. *Theories of developmental psychology (6th ed.).* Macmillan Higher Education, 2016.

Melanie Mitchell. On crashing the barrier of meaning in artificial intelligence. *AI magazine*, 41(2): 86–92, 2020.

Melanie Mitchell. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021.

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

Taihei Ninomiya, Atsushi Noritake, Kenta Kobayashi, and Masaki Isoda. A causal role for frontal cortico-cortical coordination in social action monitoring. *Nature communications*, 11(1):5233, 2020.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

Thomas C O'Brien and Bernard J Shapiro. The development of logical thinking in children. *American Educational Research Journal*, 5(4):531–542, 1968.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Alessandro B Palmarini and Melanie Mitchell. Abstract understanding of core-knowledge concepts: Humans vs. llms. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

Jean Piaget. *The Psychology of Intelligence*. Harcourt, Brace, 1950.

Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, 1952.

Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, New York, 1969.

Jean Piaget and Bärbel Inhelder. Intellectual operations and their development. In *Experimental Psychology Its Scope and Method: Volume VII (Psychology Revivals)*, pp. 144–205. Psychology Press, 2014.

Jean Piaget and Bärbel Inhelder. *The Child's Construction of Quantities: Conservation and Atomism*. Psychology Press, 1974.

Nicolas Poirel, Grégoire Borst, Grégory Simon, Sandrine Rossi, Mathieu Cassotti, Arlette Pineau, and Olivier Houdé. Number conservation is related to children's prefrontal inhibitory control: an fmri study of a piagetian task. *PloS one*, 7(7):e40802, 2012.

Guy Politzer. The class inclusion question: a case study in applying pragmatics to the experimental study of cognition. *SpringerPlus*, 5(1):1133, 2016.

Zenon W Pylyshyn. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain sciences*, 3(1):111–132, 1980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*, 2021.

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

David M. Rosenthal. *The Nature of Mind*. Oxford University Press, New York, 1991.

MD Rutherford and DH Brainard. Lightness constancy: A direct test of the illumination-estimation hypothesis. *Psychological Science*, 13(2):142–149, 2002.

John R Searle. The intentionality of intention and action. *Inquiry*, 22(1-4):253–280, 1979.

Michael Shayer. Has piaget's construct of formal operational thinking any utility? *British Journal of Educational Psychology*, 49(3):265–276, 1979.

Richard Shiffrin and Melanie Mitchell. Probing the psychology of ai models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120, 2023.

Elizabeth F Shipley. The class-inclusion task: Question form and distributive comparisons. *Journal of Psycholinguistic Research*, 8:301–331, 1979.

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

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early knowledge of object motion: Continuity and inertia. *Cognition*, 51(2):131–176, 1994.

Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995.

Christopher Summerfield. *Natural General Intelligence: How understanding the brain can help us build AI.* Oxford university press, 2022.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52733.2024.01365.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv: 2405.09818*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv: 2302.13971*, 2023.

Ross Vasta and Lynn S Liben. The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, 5(6):171–177, 1996.

Arnaud Viarouge, Olivier Houdé, and Grégoire Borst. The progressive 6-year-old conserver: Numerical saliency and sensitivity as core mechanisms of numerical abstraction in a piaget-like estimation task. *Cognition*, 190:137–142, 2019.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Christoph von der Malsburg. Fodor and pylyshyn's critique of connectionism and the brain as basis of the mind. *Human Arenas*, pp. 1–15, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Henry M. Wellman. *The Child's Theory of Mind.* MIT Press, Cambridge, MA, 1992.

Henry M Wellman, David Cross, and Julanne Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Gerald A Winer. Class-inclusion reasoning in children: A review of the empirical literature. *Child Development*, pp. 309–328, 1980.

Barlow C Wright and Jennifer Smailes. Factors and processes in children's transitive deductions. *Journal of Cognitive Psychology*, 27(8):967–978, 2015.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10637–10647, 2023.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2022.

Steven Yantis. Perceived continuity of occluded visual objects. *Psychological Science*, 6(3):182–186, 1995.

Do-Joon Yi, Nicholas B Turk-Browne, Jonathan I Flombaum, Min-Shik Kim, Brian J Scholl, and Marvin M Chun. Spatiotemporal object continuity in human ventral visual cortex. *Proceedings of the National Academy of Sciences*, 105(26):8840–8845, 2008.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024.