

# VISION LANGUAGE MODELS KNOW LAW OF CONSERVATION WITHOUT UNDERSTANDING MORE-OR-LESS

**Dezhi Luo<sup>1,\*,</sup>, Haiyun Lyu<sup>2,+,</sup>, Qingying Gao<sup>3,</sup>, Haoran Sun<sup>3,</sup>, Yijiang Li<sup>4,</sup>, Hokin Deng<sup>5,\*</sup>**

<sup>1</sup>University of Michigan <sup>2</sup>University of North Carolina at Chapel Hill

<sup>3</sup>Johns Hopkins University <sup>4</sup>University of California, San Diego

<sup>5</sup>Carnegie Mellon University <sup>+</sup>Equal Contribution

\* ihzedoul@umich.edu, hokind@andrew.cmu.edu

## ABSTRACT

Understanding law of conservation is a critical milestone in human cognitive development considered to be supported by the apprehension of quantitative concepts and the reversibility of operations. To assess whether this critical component of human intelligence has emerged in Vision Language Models, we have curated the ConserveBench, a battery of 365 cognitive experiments across four dimensions of physical quantities: volume, solid quantity, length, and number. The former two involve transformational tasks which require reversibility understanding. The latter two involve non-transformational tasks which assess quantity understanding. Surprisingly, we find that while Vision Language Models are generally good at transformational tasks, they tend to fail at non-transformational tasks. There is a dissociation between understanding the reversibility of operations and understanding the concept of quantity, which both are believed to be the cornerstones of understanding law of conservation in humans.

## 1 INTRODUCTION

Conservation is the idea required for understanding that physical quantities do not change in amount when they undergo spatial transformations (Piaget, 1965; Halford, 2011; Piaget & Inhelder, 1974; McGarrigle & Donaldson, 1974; Church & Goldin-Meadow, 1986; Goldin-Meadow & Beilock, 2010; Houdé et al., 2011). First studied by Jean Piaget, it has been considered a landmark ability of human cognitive development for it not only signifies the understanding of quantitative concepts in rudimentary, concrete domains, but also lays the foundation for formal, abstract thinking required in advanced cognitive abilities such as mathematical reasoning (Piaget, 1965; 1971; McGarrigle & Donaldson, 1974). In Piaget’s theory of cognitive development, the acquisition of conservation marks children’s transition from the pre-operational stage to the concrete operational stage (Piaget, 1965; Miller, 2016). The pre-operational stage is characterized by children’s reliance on a single attribute of an object while ignoring others when reasoning about the physical world. Piaget posits that this is because their mental representations supporting physical attributes are still isolated and juxtaposed (Piaget, 1952; 1971; Houdé, 1997). As these representations are stabilized into organized concepts of numerosity, volume, length, and so on, the representations would become supportive of mental operations, which are internalized actions organized by logic structures and could be manipulated in systematic ways (Piaget, 1950; Miller, 2016). Children enter the concrete operational stage, as they are able to perform mental operations which allow them to simulate reversible physical transformations of quantitative objects. In this sense, we say that children have acquired law of conservation.

Given the highly informative nature of the acquisition of conservation in terms of cognitive abilities and their developmental trajectories, we suggest that it can be applied as a benchmark for assessing the cognitive functions possessed by large Vision Language Models (VLMs). We have created the ConserveBench, which contains 365 cognitive experiments designed based on Piaget’s four classic conservation tasks, to investigate the law of conservation in current Vision Language Models. We have aligned 5 models for our analysis (Li et al., 2023; OpenAI; Bai et al., 2023). The

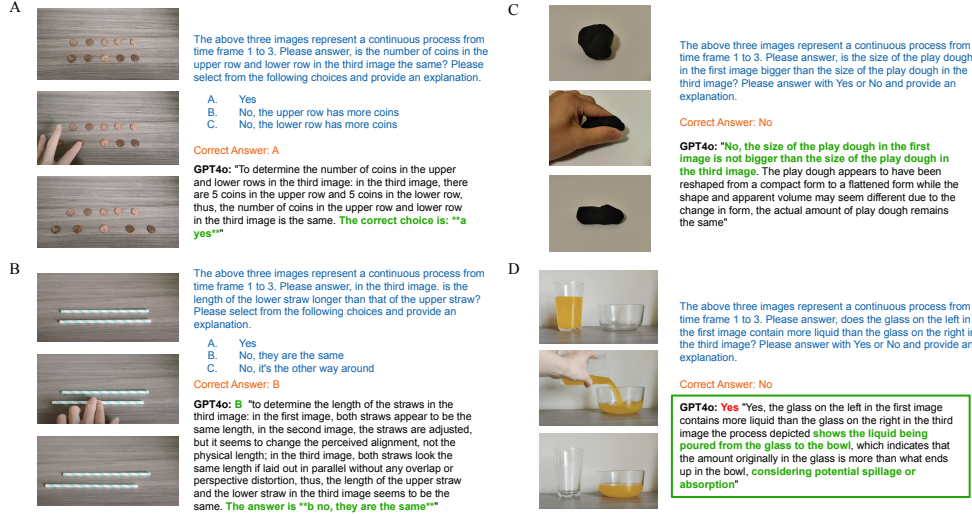


Figure 1: VLMs behaviors on Number, Length, Solid Quantity, and Liquid Volume experiments on ConserveBench. Transformational Tasks.

tasks are composed of transformational and non-transformational tasks, which the former tasks assess understanding of reversibility, and the latter tasks assess understanding of quantity. Incredibly, we find that VLMs are able to perform well on transformational tasks and nevertheless fail dramatically on quantity understanding tasks, suggesting that they understand law of conservation without knowing what’s more-or-less.

## 2 METHODS

### 2.1 DATASET

We have curated the ConserveBench to investigate the law of conservation in Vision Language Models. ConserveBench contains 39 multi-image and 326 single-image cognitive experiments, and therefore in total 365 cognitive experiments.

### 2.2 COGNITIVE EXPERIMENTS

#### 2.2.1 TRANSFORMATIONAL TASKS

Following classic Piagetian design (Piaget, 1965; Halford, 2011; McGarrigle & Donaldson, 1974; Church & Goldin-Meadow, 1986; Lozada & Carro, 2016), our cognitive experiments are separated into four groups, each probing one dimension of physical quantity: number, length, solid quantity, and liquid volume, as shown in Figure 1. In real-life, the conservation tasks consist of the experimenter showing the child the process of physical transformation by hands-on manipulating the objects in front of them. Given that VLMs process visual information on a discrete, frame-by-frame basis, such demonstration of physical transformation is operationalized into three phase: the *Initial Phase*, the *Manipulation Phase*, and the *End Phase*, represented by three images that are consecutively fed to the models. The prompt of the question provides the information that the series of images depicts a continued process, which is mandatory in order to prevent VLMs to directly cross-compare the quantity across images without acknowledging the transformation. Below introduced the tasks for different dimensions separately in details.

1. **Number:** *Initial Phase* depicts two parallel lines of objects aligned perfectly by their positions on the lines; *Manipulation Phase* depicts the experimenter’s fingers moving one line of objects; *End Phase* depicts the line of the objects moved being more spread out than the other, whilst the number of coins remains the same. Experiments in virtual setting are also tested.

2. **Length:** *Initial Phase* depicts two linear objects placed parallel to each other and aligned perfectly; *Manipulation Phase* depicts the experimenter’s fingers moving one of the linear objects; *End Phase* depicts the linear object moved misaligned with the other straw. Both virtual and reality settings are tested.
3. **Solid Quantity:** *Initial Phase* depicts a round-shaped piece of play dough; *Manipulation Phase* depicts the experimenter’s hand rubbing the play dough; *End Phase* depicts the play dough appearing notably extended.
4. **Liquid Volume:** *Initial Phase* depicts a tall glass partially filled with colored liquid placed next to an empty, shorter glass. *Manipulation Phase* depicts the experimenter’s hand holding the tall glass, pouring the colored water into the short glass. *End Phase* depicts the short glass now partially filled with colored water, while the tall glass next to it is now empty.

### 2.2.2 NON-TRANSFORMATIONAL TASKS

To probe VLMs’ understanding of quantity and its relationship with conservation, we leverage a section of ConserveBench, which consists entirely of single-image tasks featuring Number and Length dimensions (as shown in Figure 4-5), which each are in the format of the *End Phase* of respective conservation tasks as described above. The overall set of cognitive experiments, therefore, consists of what is henceforth labeled Transformational Tasks and Non-Transformational Tasks.

### 2.3 MODEL SELECTION AND EXPERIMENT

For a fair comparison, 5 typical VLMs were tested on our dataset using the same prompt under a zero-shot, open-ended generation task (Figure 2). In order to analyze the reasoning abilities of VLMs, we ask the models to explain their answers after they have given the answers.

## 3 RESULTS

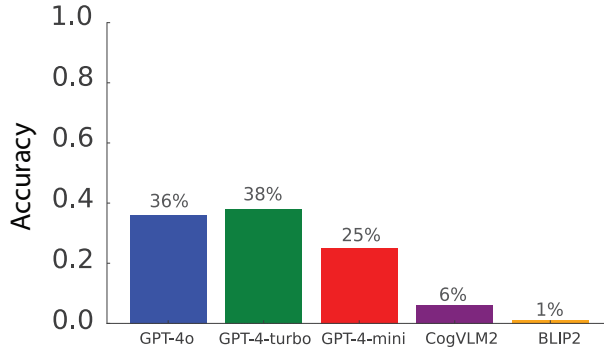


Figure 2: VLMs Performance on ConserveBench

First of all, we have tested our ConserveBench on 5 typical Vision Language Models. Interestingly, the models achieve a mediocre performance on our ConserveBench (Figure 2). We further look into their performances in terms of Transformational Tasks and Non-Transformational Tasks (Figure 3). Interestingly, VLMs achieve good performances across Transformational Tasks. This indicates that VLMs could recognize the reversibility of physical operations. However, intriguingly, in Non-Transformational Tasks for number and length dimensions, VLMs perform significantly poorer, exhibiting consistent errors comparable to pre-operational children with extremely limited understanding of quantity. This indicates that VLMs have difficulties in what requires a rudimentary conceptual understanding of quantity. Our results agree with recent studies reporting that major language models and as well as their corresponding vision models fail at simple counting tasks (Rane et al., 2024; Rahmanzadehgervi et al., 2024). Notably, however, a particularly surprising finding of the present study is that among the tasks that ask more-or-less questions (as seen in Figure 4, 5, and

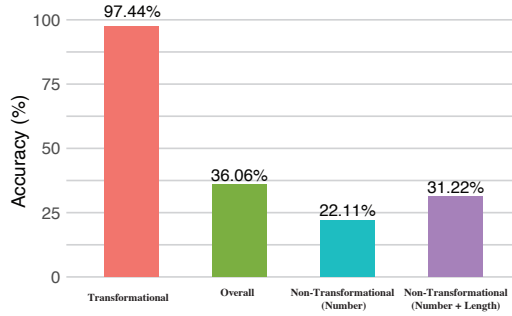


Figure 3: GPT-4o performance on ConserveBench. We observe that GPT-4o achieve very high performance on Transformational Tasks whereas fail badly on quantity understanding, Number and Length, Non-Transformational tasks.

6), VLMs consistently give answers that are opposite to the typical human fallacy. There are 95 tasks from ConserveBench that specifically probe what is known as the length-equals-number fallacy, which is understood to be produced by the heuristics that visually longer or bigger stuff tend to have more stuff among them (Houdé, 1997; Viarouge et al., 2019). Said heuristics remain presented in older children and adults as a common System-1 strategy used in daily life but can be suppressed with a reasonable amount of effort when the task at hand requires (Harnishfeger & Bjorklund, 1990; Poirel et al., 2012). Notably, we observe that **every** quantity understanding, non-transformational task among the said 95 tasks that GPT-4o fails is by choosing the choice opposite to what demonstrates the length-equals-number fallacy. The consistency of this phenomenon reveals a significant divergence between human and VLM reasoning with respect to quantity understanding.

We further calculate the accuracy scores on all tasks in ConserveBench of GPT-4o (Figure 3). GPT-4o is able to achieve 97.44% on Transformation tasks. Nevertheless, it only has achieved 31.22% on non-transformational single-image tasks in general, and only 22.11% in non-transformational single-image number tasks. The results really show that VLMs understand conversation yet do not understand more-or-less.

## 4 DISCUSSIONS

It is surprising that VLMs employ a misleading strategy in number-dimension, non-transformational task that is entirely opposite to human intuition. This indicates that the numerical cognition of VLMs may differ from humans at the respective level (i.e., pre-operational children who fail at the same task) in certain foundational ways. Contrary to the length-equals-number strategy, VLMs’ failure to achieve a rudimentary understanding of quantity seems to be supplemented by the exploitation of a dense-equals-more strategy, as shown by their tendencies to report that lines that are more packed have more objects among them.

Further research is required to explain how these VLM strategies emerge, why they diverge from human intuition, and whether they generalize to other contexts. Importantly, we need to conduct human experiments on the same ConserveBench to validate these observations and refine our understanding of both AI and human cognition at this foundational level.

## 5 CONCLUSION

Law of conservation has been recognized as a core cognitive ability. Here, we have created the Conservebench to investigate such abilities in Vision Language Models. Remarkably, although Vision Language Models generally appear able to perform transformational tasks, that require understanding of reversibility, they often struggle at non-transformational tasks, which require understanding of quantity, that are commonly taken as the foundations for genuinely understanding law of conservation.

## REFERENCES

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- R. Breckinridge Church and Susan Goldin-Meadow. The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23:43–71, 1986.
- Susan Goldin-Meadow and Sian Beilock. Action’s influence on thought: the case of gesture. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 5(6): 664–674, 2010.
- 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.
- Katherine Kipp Harnishfeger and David F Bjorklund. The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental review*, 10(1):48–71, 1990.
- Olivier Houdé. Numerical development: From the infant to the child. *Cognitive Development*, 12(3):373–391, 1997.
- Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poiré, Guy Perchey, Céline Lanoë, Amélie Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lambert, 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.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- Mariana Lozada and Natalia Carro. Embodied action improves cognition in children: Evidence from a study based on piagetian conservation tasks. *Frontiers in psychology*, 7(393), 2016.
- James McGarrigle and Margaret Donaldson. Conservation accidents. *Cognition*, 3(4):341–330, 1974.
- Patricia H Miller. *Theories of developmental psychology (6th ed.)*. Macmillan Higher Education, 2016.
- OpenAI. Models - openai api. <https://platform.openai.com/docs/models/gpt-4o>.
- Jean Piaget. *The Psychology of Intelligence*. Harcourt, Brace, 1950.
- Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, 1952.
- Jean Piaget. *The Child’s Conception of Number*. W.W. Norton and Company, 1965.
- Jean Piaget. *Biology and knowledge*. University of Chicago Press, 1971.
- Jean Piaget and Bärbel Inhelder. *The Child’s Construction of Quantities: Conservation and Atomism*. Psychology Press, 1974.
- Nicolas Poiré, 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.
- Pooyan Rahmzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint*, 2024.
- Sunayana Rane, Alexander Ku, Jason Baldridge, Ian Tenney, Tom Griffiths, and Been Kim. Can generative multimodal models count to ten? *Proceeding of the Annual Meeting of the Cognitive Science Society*, 46:1235–1241, 2024.

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.

## Appendix

### A MORE EXAMPLES

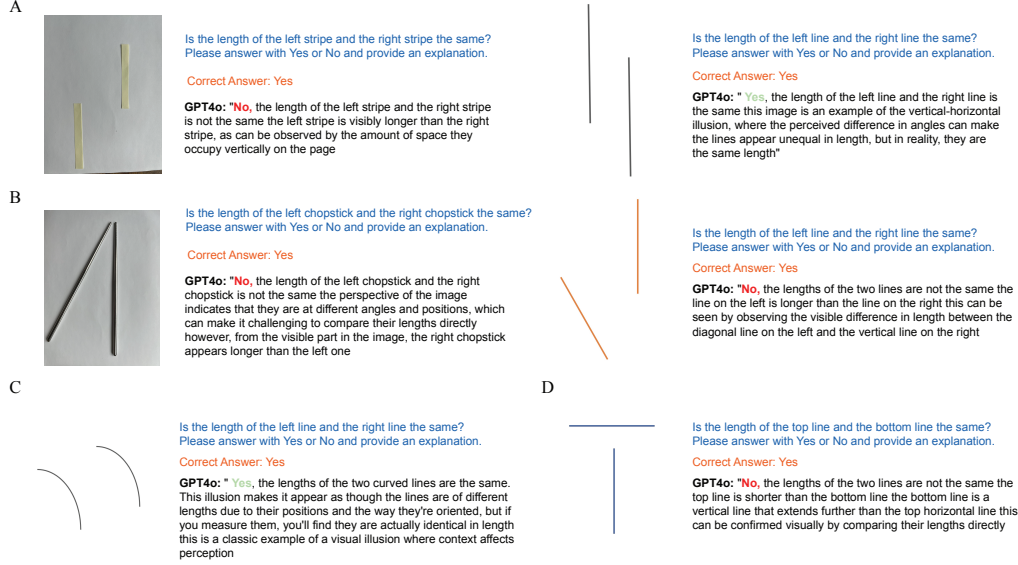


Figure 4: VLMs performance on Length experiments in ConserveBench. Non-Transformational Tasks.

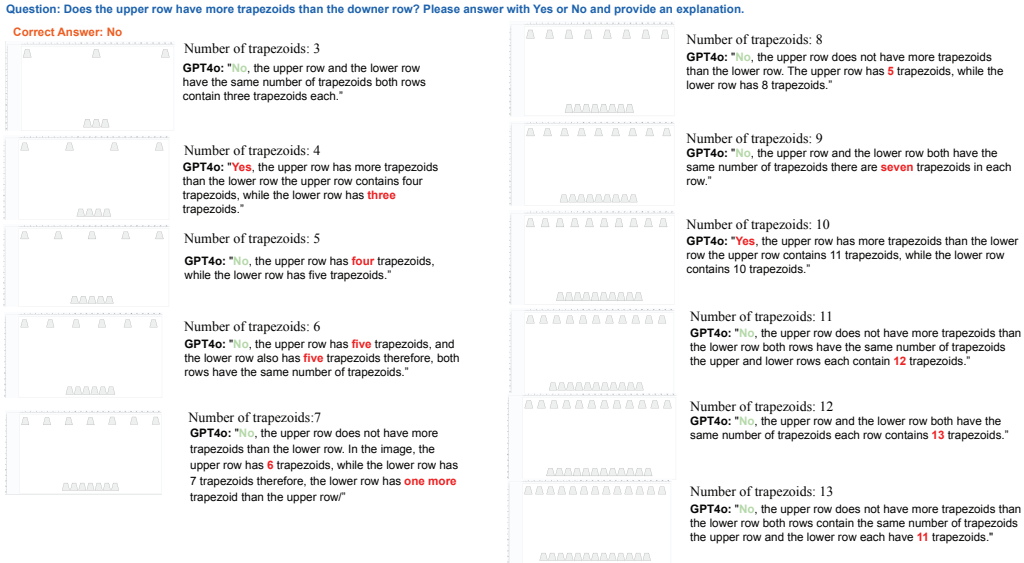


Figure 5: VLMs performance on Number experiments in ConserveBench. Non-Transformational Tasks.

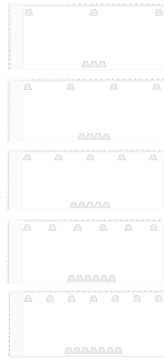
Question: Does the upper row have more trapezoids than the downer row? Please select from the following choices and provide an explanation.

A. Yes

B. No, they are the same

C. No, it's the other way around

Correct Answer: B



Number of trapezoids: 3

**GPT4o:** "C, the upper row has three trapezoids, while the lower row has four trapezoids therefore, the lower row has more trapezoids than the upper row."

Number of trapezoids: 4

**GPT4o:** "C, the upper row has four trapezoids, while the lower row has five trapezoids therefore, the lower row has more trapezoids than the upper row."

Number of trapezoids: 5

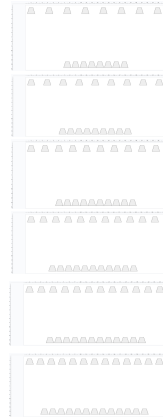
**GPT4o:** "C, the upper row has 4 trapezoids, and the bottom row has 5 trapezoids therefore, the bottom row has more trapezoids than the upper row."

Number of trapezoids: 6

**GPT4o:** "C, explanation: the upper row has 5 trapezoids, and the lower row has 6 trapezoids therefore, the lower row has more trapezoids than the upper row."

Number of trapezoids: 7

**GPT4o:** "C, the upper row has 6 trapezoids, while the lower row has 7 trapezoids therefore, the lower row has more trapezoids than the upper row"



Number of trapezoids: 8

**GPT4o:** "C, the upper row has 6 trapezoids, while the lower row has 7 trapezoids so, the lower row has more trapezoids than the upper row"

Number of trapezoids: 9

**GPT4o:** "C, the upper row has seven trapezoids, while the lower row has eight trapezoids"

Number of trapezoids: 10

**GPT4o:** "C, the upper row has 8 trapezoids, while the lower row has 10 trapezoids therefore, the lower row has more trapezoids than the upper row"

Number of trapezoids: 11

**GPT4o:** "C, the upper row has 12 trapezoids, while the lower row has 14 trapezoids thus, the lower row has more trapezoids than the upper row."

Number of trapezoids: 12

**GPT4o:** "B, both the upper and lower rows contain the same number of trapezoids each row has eight trapezoids."

Number of trapezoids: 13

**GPT4o:** "B, the upper and lower rows each contain the same number of trapezoids counting them in the image, both rows consist of ten trapezoids."

Figure 6: VLMs performance on Number experiments in ConserveBench. Non-Transformational Tasks.