PROBING MECHANICAL REASONING IN LARGE VISION LANGUAGE MODELS

Haoran Sun[1], Yijiang Li[2], Qingying Gao[1], Haiyun Lyu[3],
Dezhi Luo[4,5,*], Hokin Deng[6,*]
[1]JHU, [2]UCSD, [3]UNC Chapel Hill, [4]UMich, [5]UCL, [6]CMU
[*] ihzedoul@umich.edu, hokind@andrew.cmu.edu

ABSTRACT

Mechanical reasoning is a hallmark of human intelligence, defined by its ubiquitous yet irreplaceable role in human activities ranging from routine tasks to civil engineering. Embedding machines with mechanical reasoning is therefore an important step towards building human-level artificial intelligence. Here, we leveraged 155 cognitive experiments to test the understanding of system stability, gears and pulley systems, leverage principle, inertia and motion, and fluid mechanics in 26 Vision Language Models (VLMs). Results indicate that VLMs consistently perform worse than humans on all domains, while demonstrate significant difficulty in reasoning about gear systems and fluid mechanics. Notably, their performance on these tasks do not improve as number of parameters increase, suggesting that current attention-based architecture may fail to grasp certain underlying mechanisms required for mechanical reasoning, particularly those pertaining to mental simulations.

Keywords: mechanical reasoning; vision language models; model-based reasoning; intuitive physics; cognitive AI

1 INTRODUCTION

Humans are uniquely capable of working with complex mechanical systems, ranging from routine tasks, such as assembling furniture, to large-scale civil endeavors, such as designing architectural structures and developing advanced technologies (Harari, 2014). These capabilities are underpinned by the cognitive ability to reason about the relationships and interactions of physical objects, an ability known as mechanical reasoning (Clark, 2010; Harman, 2011; Vaesen, 2012). While there is evidence that some animal species demonstrate limited mechanical reasoning, such as understanding basic tool use or object interactions (Shumaker et al., 2011), the human capacity for mechanical reasoning is unparalleled in its flexibility, sophistication, and creative potential. This distinction allows humans to innovate, solve complex problems, and adapt tools and systems to a wide range of environments and challenges (Allen et al., 2020; Allen, 2021). As such, mechanical reasoning is a cornerstone of human intelligence, driving technological and cultural progress throughout history. Implementing mechanical reasoning is therefore a vital step toward developing artificial intelligence systems capable of achieving human-level performance in real-life scenarios. Given the rapid advancements in large language models (LLMs), particularly their variants that support visual operations, i.e. vision language models (VLMs), mechanical reasoning appears to be a critical area of assessment to evaluate their current reasoning capabilities and identify potential limitations.

Although mechanical reasoning is a high-level cognitive ability that does not emerge until late childhood, it has been found to rely heavily on more foundational cognitive strategies (Hegarty et al., 1988; Kim & Spelke, 1999; Allen, 2021). In particular, decades of research in cognitive science suggest that mental simulation, the process of constructing and operating mental models of the world to guide reasoning, is critical for many aspects of mechanical reasoning (Hegarty, 2004). For example, studies employing methods such as interviews and eye-tracking have revealed that individuals dynamically construct spatial representations of gear and pulley systems to infer their properties as they mentally simulate stages of motion (Hegarty, 1992; Lehrer & Schauble, 1998; Kubricht et al., 2017; Rozenblit et al., 2002). On the other hand, while it remains an ongoing de-

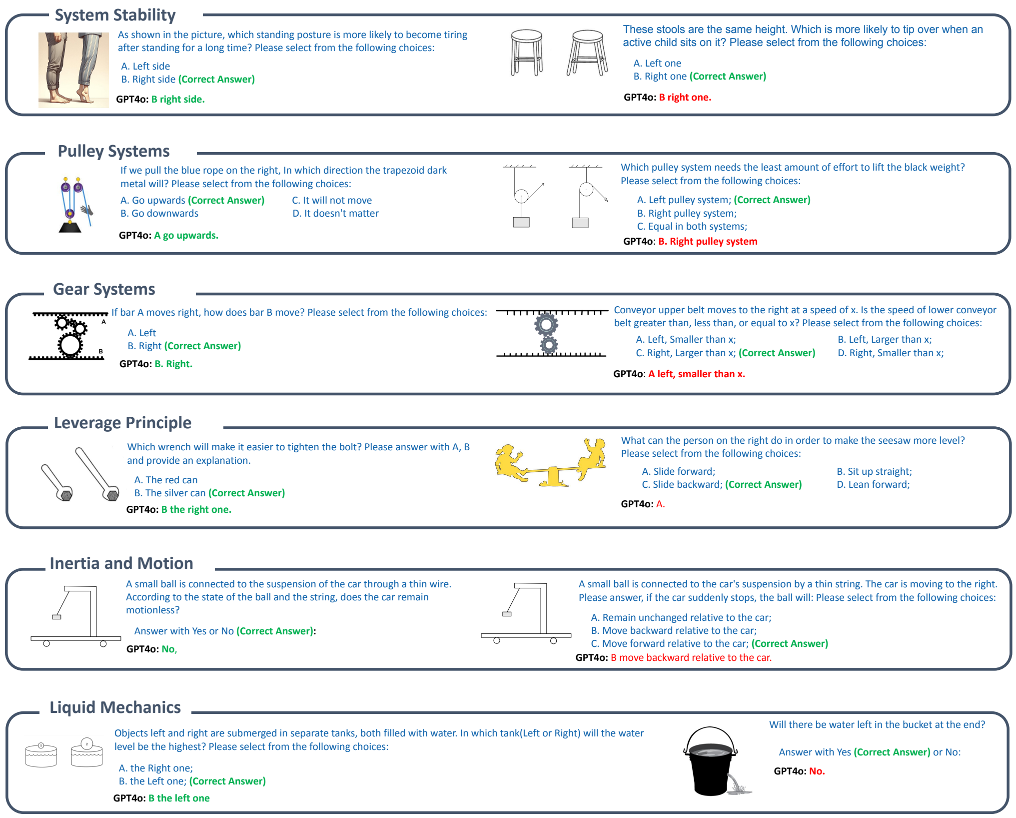

Figure 1: **Sample Tasks on the Six Domains of MechBench.** For each domain, GPT4o answered correctly to the task on the left and failed the task on the right

bate, several studies have indicated that the ability to construct mental models of the world to guide reasoning, particularly in the context of spatial dimensions, remains significantly underdeveloped even in state-of-the-art LLMs (Hao et al., 2023; Mitchell & Krakauer, 2023; Goddu et al., 2024; Gao et al., 2024). Since mental simulations necessitate this model-based reasoning capacity on the visuospatial domain, a thorough evaluation of the mechanical reasoning capabilities of VLMs could therefore provide valuable insights into this debate given the critical role of mental simulation in solving mechanical reasoning tasks.

In the present work, to systematically evaluate VLMs' mechanical reasoning abilities, we constructed the MechBench, which contains around 155 cognitive experiments. The experiments investigate six generalized domains of mechanical reasoning: **system ability**, **pulley systems**, **gear systems**, **leverage principle**, **inertia and motion**, and **fluid mechanics**.

2 Methods

2.1 Cognitive Experiments

Mechanical reasoning has been widely explored in cognitive psychology and educational research. This study uses single-image adaptations of classic cognitive tasks from the literature to assess mechanical reasoning, focusing on six key aspects: system stability, pulley systems, gear systems, leverage principle, inertia and motion, and fluid mechanics. Example cases on each domains are presented in Figure 1. This categorization is intentionally designed to encompass the full range of mechanical reasoning, ensuring each category represents a distinct but fundamental dimension of physical understanding. Below, we provide explanations for each domain.

System Stability The understanding of stability is essential for reasoning about states of equilibrium within physical systems, such as predicting whether a stack of blocks will remain upright or collapse (McCloskey, 1983a). In our experiments, models are presented with images of objects like stools with varying base widths or angles of inclination. The task involves selecting the most stable configuration. Stability reasoning encompasses factors such as the center of gravity, base area, and force distribution, making it a critical baseline for evaluating mechanical reasoning.

Pulley Systems Pulley systems are widely used in cognitive psychology and physics education to study how individuals reason about force and motion (Hegarty & Sims, 1994), requiring an understanding of force distribution and machinery functions. For example, a simple task might involve determining which pulley system requires less effort to lift a weight. By testing VLMs' ability to distinguish between fixed and movable pulleys and to predict object movement, we evaluate their capacity to infer about dynamic relationships within real-life scenarios.

Gear Systems Gear systems are deterministic mechanical setups governed by well-defined rules, such as adjacent gears rotating in opposite directions and gear ratios determining relative speeds. These properties make gears an ideal domain for testing logical and causal reasoning (Hegarty et al., 1988). Tasks in this category involve analyzing diagrams of connected gears to predict their rotational direction and speed.

Leverage Principle The leverage principle illustrates the relationship between force, distance, and torque. Balance-scale experiments in cognitive psychology have shown how humans progressively develop an understanding of leverage through iterative learning and application (Peirce, 1992). Tasks in this category include determining how shifting weights on a seesaw or applying force to a lever affects balance.

Inertia and Motion Inertia and motion are dynamic aspects of mechanical reasoning that require understanding how forces influence the movement of objects over time. These concepts are central to Newtonian mechanics and intuitive physics (McCloskey, 1983b). Human cognition integrates spatial and temporal information to make predictions about motion and forces, as seen in studies of tool use and physical reasoning (Allen, 2021). Our experiments include scenarios such as predicting the trajectory of an object on a moving cart or identifying the kinetic energy distribution of a pendulum. These tasks thereby probe VLMs to integrate information about multiple aspects of the physical world.

Fluid Mechanics Fluid mechanics involves understanding the behavior of liquids under various conditions, such as flow, external pressures, and volume changes. Although gounded in the intuitive understanding of fluid dynamics emerged very early in humans' cognitive development (Hespos et al., 2016), reasoning about liquid behaviors in mechanical systems require simultaneous consideration of geometry, force, and dynamics. These tasks represent a highly important dimension of mechanical reasoning concerning specifically about liquid motions as opposed to solid objects alone.

2.2 Model Selection and Experiment

We evaluated the mechanical reasoning abilities of three categories of VLMs. To ensure a fair comparison, all VLMs are evaluated on their ability to reason over images and texts under a zero-shot generation task. A complete list of models is reported in the results section as shown in Figure 2. Model size data are curated at the same time. The models are categorized as follows:

1. **Open-source VLMs with Multi-Image Reasoning**: Includes models with different sizes and other variants such as `CogVLM` Series(Hong et al., 2024), `Qwen` series(Qwen-VL (Bai et al., 2023), Qwen-2 (Wang et al., 2024)), and `Blip2` (Li et al., 2023), LLaVA-Next (Liu et al., 2024) , which are capable of reasoning over interleaved multiple images and texts.
2. **Closed-source VLMs with Multi-Image Reasoning**: Includes proprietary models such as GPT series (OpenAI) (`GPT-4v`, `GPT-4-turbo`, `GPT-4o-mini`), Gemini Series (Gemini), and Claude Series (claude). These models also support reasoning across interleaved images and texts,

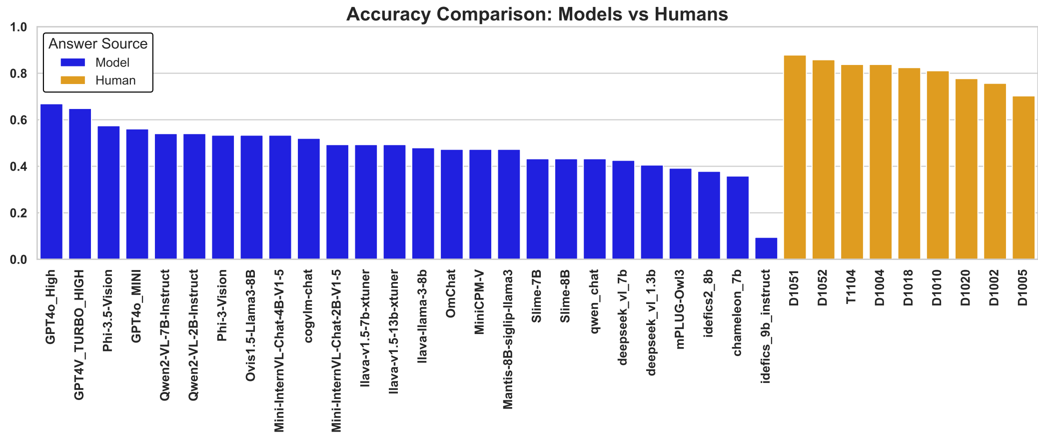

Figure 2: **Model Performance on MechBench As Compared to Human Performance**

3. **Open-source VLMs with single-Image Reasoning**: Includes models designed to process a single image alongside continuous text. InstructBlip Series (Dai et al., 2023), LLaVA Series (Liu et al., 2023a) (Liu et al., 2023b)

In total, we aligned 26 models for evaluation. In order to analyze the reasoning abilities of VLMs, we asked the models to explain their answers after they have given the answers by adding "please provide an explanation" in the prompt.

2.3 Human Baseline

We recruited a total of 9 participants, all of whom were college students proficient in English. Participants were instructed to skip any question that was ambiguously phrased or too complex to answer within 90 seconds. A question was marked as failed if the participant did not provide an answer. For each question, at least 80% of participants needed to answer correctly; otherwise, we modified the question, and new annotators completed the revised version. The human baseline result for each question was normalized based on the number of participants who provided an answer.

In addition to answering the experimental questions, participants completed a test of visuospatial fluid intelligence using Raven's Advanced Progressive Matrices (APM), a widely used non-verbal assessment tool for evaluating fluid intelligence (Raven et al., 2000). The APM results were collected to explore correlations between participants' general visuospatial reasoning abilities and their performance on the experimental tasks.

3 Results

3.1 General Results

Our study reveals a significant disparity between human and model performance across multiple evaluation dimensions. As shown in Figure 2 and 3, humans consistently outperform models both in overall accuracy and in task-specific dimensions. These results highlight the limitations of current VLMs in replicating human-like reasoning in mechanical and intuitive physics tasks. Among the evaluating dimensions, Pulley Systems exhibits the largest performance gap, with human accuracy nearing 90%, compared to model accuracy averaging around 50%.

3.2 Human vs. Model Performance Across Each Dimensions

The ANOVA test ($F = 11.8111, p < 0.0001$) revealed significant differences in human performance across the six categories. Tukey HSD post-hoc analysis (visualized in the heatmap) shows that the performance of humans handling the pulley system is significantly different from the other five

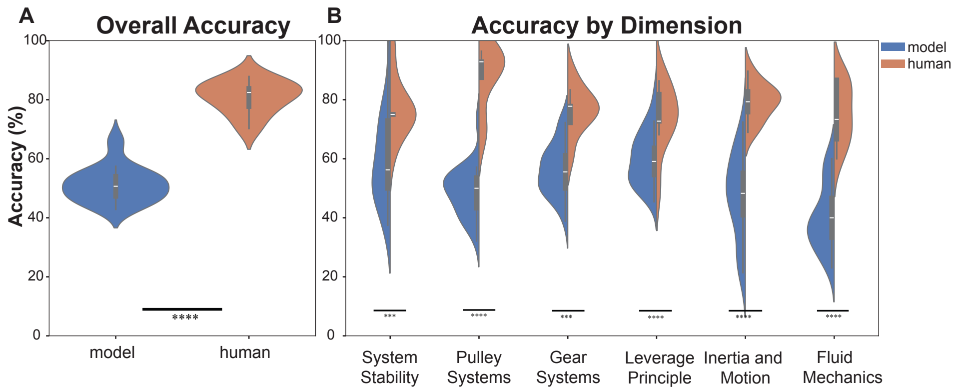

Figure 3: **Overall and Dimension-Wise Accuracy: Humans vs. Models**. A. For overall accuracy across tasks, human participants outperform models significantly (p < 0.0001). B. Human participants consistently outperform models in each dimension (all categories p < 0.001).

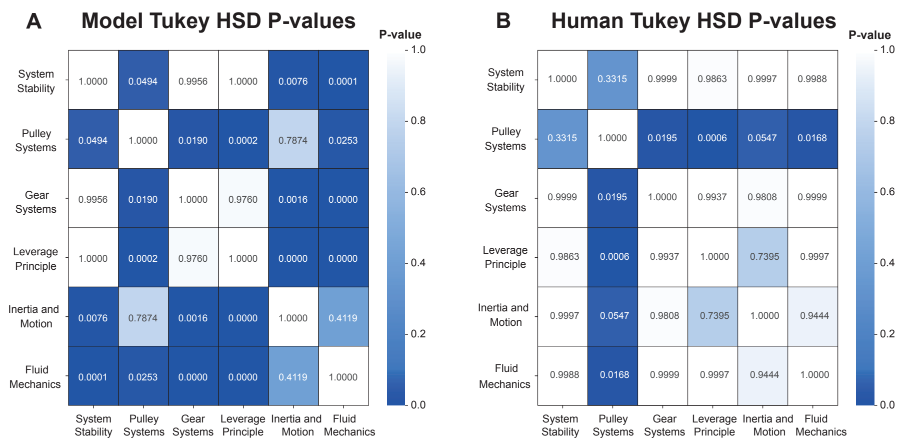

Figure 4: **Comparison of Tukey HSD P-values for Model and Human Performance Across Different Dimensions.** A. The ANOVA results for model performance reveal a significant difference across the six dimensions (F-statistic = 11.8111, p ¡ 0.0001), indicating that the models exhibit varying accuracy depending on the task. Tukey HSD tests show that significant differences exist between several pairs of dimensions, with System Stability, Inertia and Motion, and Fluid Systems exhibiting large disparities in accuracy. Leverage Principle and Gear Systems show less pronounced differences. B. Human performance across the six dimensions shows no significant difference (F-statistic = 1.6809, p = 0.1430). Tukey HSD tests for human performance indicate that none of the comparisons between different dimensions reach statistical significance, suggesting a more uniform performance across dimensions.

categories. As can be seen from the heat map, the performance of humans handling the pulley system is significantly better than the other five categories. In contrast, ANOVA results for model performance ($F = 1.6809, p = 0.1430$) indicated no significant differences across task categories. Tukey HSD analysis further reveals that models exhibited similar accuracy across all categories with the exception of Fluid System, which deviates significantly with other dimensions. Together, as shown in Figure 3, the distribution of human performance is narrower, indicating more consistent accuracy, whereas the models display a wider range of variability.

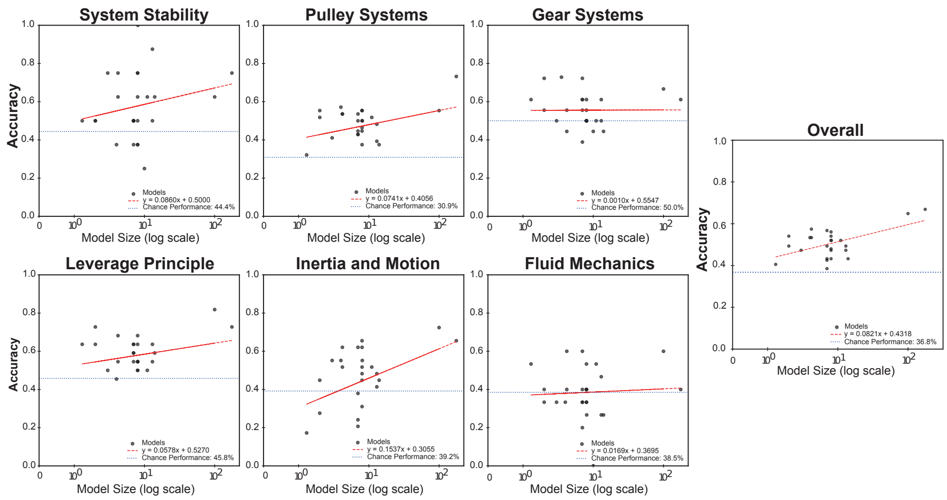

Figure 5: **The Relationship Between Model Performance and Model Size** The six plots on the left illustrate how model size, as measured by numbers of parameter in the models' neural networks (log scale), affects accuracy across six evaluation dimensions. The plot on the right shows the overall accuracy across tasks.

3.3 Relationship Between Model Performance and Model Size

A widely held belief in the machine learning community is that an increase in a model's scale, measured by the number of parameters, leads to systematic improvements in its reasoning abilities (Sutton, 2019; Kaplan et al., 2020), a concept known as the scaling law. However, this assumption is an empirical observation without theoretical proof. To evaluate whether the scaling law applies to mechanical reasoning, we further examined the relationship between model performance on mechanical reasoning tasks and model size, as measured by the number of parameters (Figure 5). For overall accuracy across task, regression analysis yielded a formula $y = 0.0821x + 0.4318$ (p = 0.0053, R^2 = 0.2917). This trend was particularly evident in models with parameter counts exceeding 10 billion, such as GPT-4o High and GPT4V-TURBO-HIGH, which consistently outperformed smaller models. In most of the dimensions (System Stability, Pulley Systems, Leverage Principle, Inertia and Motion, Fluid Mechanics), accuracy increases with larger model sizes. However, Gear Systems and Fluid Systems show very weak relationships, with slopes and R-squared values close to zero (Gear Systems: $y = 0.0010x + 0.5547$, p = 0.981628, R^2 = 0.000023; Fluid Systems: $y = 0.0169x + 0.3695$, p = 0.787137, R^2 = 0.003097). These findings suggest that while model size often correlates with improved performance in many mechanical reasoning domains, some domains may not benefit significantly from increased model scale. This indicates a potential distinction between the underlying mechanisms across different domains of mechanical reasoning.

Additionally, in conjunction with participants' scores on the Raven's Advanced Progressive Matrices (APM), we found that that humans' performance correlates significantly with fluid intelligence (Pearson correlation coefficient $R^2 = 0.24$). This holds for all evaluating dimensions with the exception of System Stability, which exhibited a moderately negative correlation with fluid intelligence (Figure 6). This irregularity likely indicates that tasks in this category rely more on intuitive judgment than on advanced logical reasoning. This static relationship between System Stability and fluid intelligence, however, is not replicated between model performance on said domain and model size. This result potentially indicated a divergence between the underlying process of models and humans in solving System Stability tasks, where models rely more on scalable abilities as opposed to intuitive judgment.

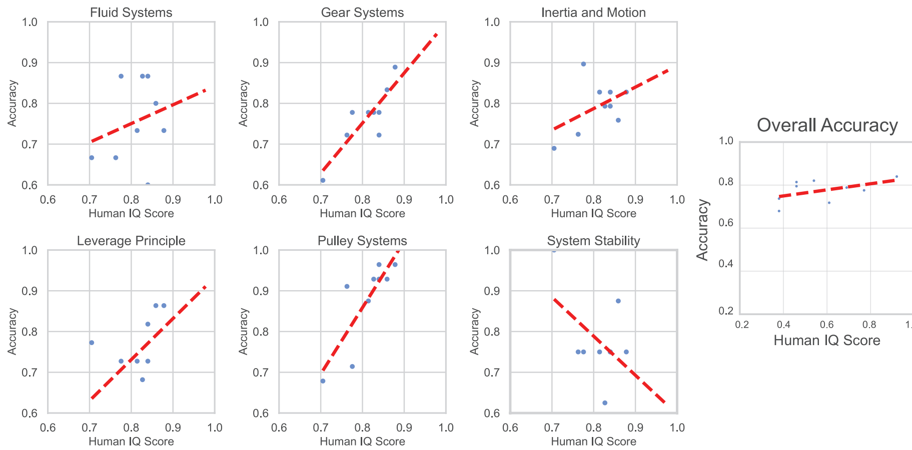

Figure 6: **The Relationship Between Human Performance on MechBench and Human Fluid Intelligence.** We further validate our MechBench using correlations with human fluid intelligence, measured via Raven's Advanced Progressive Matrices (Raven et al., 2000). Interestingly, System Ability seems to be an innate cognitive ability that's orthogonal to human fluid intelligence.

4 Discussions

Here, we constructed MechBench, an assay of 155 cognitive experiments that include six aspects of mechanical reasoning. We investigated VLMs' ability in reasoning about System Stability, Gear Systems, Pulley Systems, Leverage Principle, Inertia and Motion, and Fluid Mechanics. Despite falling behind humans on all domains, in general they achieved significantly above chance performance, indicating certain levels of mechanical reasoning abilities.

Notably, VLMs' performance on two dimensions—Gear Systems and Fluid Mechanics—remains close to chance and does not improve with increased model size, serving as a counterexample to the scaling law hypothesis in machine learning (Sutton, 2019; Kaplan et al., 2020). Such a trend indicate that the underlying architecture of these models is yet to be able to support certain mechanisms that are required for a system to have mechanical reasoning abilities regarding Gear System and Fluid Mechanics. Similar trends have previously been reported in VLMs' performance on level-2 perspective-taking tasks, in which the models are asked to tell how another agent would see the same spatial arrangement in a perspective different from their own (Gao et al., 2024; Piaget & Inhelder, 1969; Moll & Meltzoff, 2011). Interestingly, like mechanical reasoning, said ability is also known to be supported by mental simulation in humans (Hegarty, 2004; Zhao et al., 2016). Together, these findings suggest that VLMs may lack the ability to perform model-based reasoning, possibly highlighting a fundamental limitation in the architecture of the foundational models used in current VLMs.

There is however an important concern with this interpretation: if the limitations of VLMs in mechanical reasoning are indeed due to their inability to perform model-based reasoning, why do they not also under-perform in domains like Pulley Systems, which are also known to rely on such reasoning? Two possible explanations could account for this discrepancy. Foremost, the role of mental simulations in mechanical reasoning is known to be mediated by low-level cognitive abilities, such as intuitions about object relations and motions—often referred to as intuitive physics (McCloskey, 1983a; Kubricht et al., 2017). Intuitive physics involves understanding how objects and substances in the environment are subject to physical laws without relying on formal, abstract knowledge. While intuitive physics may draw on domain-general abilities like model-based reasoning, it also depends on the differential application of key physical principles, such as motion, gravity, and states of matter (Kaiser et al., 1986; Kubricht et al., 2017). Different domains of mechanical reasoning are supported by an intuitive understanding of these physical principles. For instance, reasoning about fluid mechanics heavily relies on understanding liquid properties, which differentiates from that of solid objects (Kawabe et al., 2015). It is possible that models can perform model-based reasoning at a lower level but fail to apply it effectively to certain physical domains, explaining their varied performance across mechanical reasoning tasks.

On the other hand, it may just be that models have learned shortcuts that allow them to bypass model-based reasoning when tackling certain domains. Pulley Systems, for example, are heavily represented in online materials related to physics education. Given that Pulley Systems often involve a small number of components (e.g., fixed and moving pulleys), with limited variety in their arrangements, models may exploit spurious correlations between benchmark questions and similar examples already present in their training data. In other words, the performance of models on certain domains of mechanical reasoning may not reflect genuine, step-by-step reasoning but rather statistical correlations (Bender et al., 2021; Bleeker et al., 2024). This is supported by the observation that very large models, such as GPT-4o, perform particularly well on Pulley System tasks, while smaller models tend to perform much worse.

5 Conclusion

In this paper, we presented the first comprehensive assessment of VLMs' mechanical reasoning abilities. Drawing on experimental paradigms from cognitive science literature, our findings reveal that current VLMs fail to fully acquire the necessary competencies for mechanical reasoning across various domains. Furthermore, the observation that performance on certain tasks does not improve with increased model size underscores potential fundamental limitations in existing model architectures. In line with previous research, we specifically highlighted model-based reasoning—the ability to mentally simulate dynamic relations within physical systems. However, given that mechanical reasoning is a high-level cognitive ability reliant on a range of basic cognitive processes, the precise mechanisms underlying these limitations remain unclear. Further exploration of these mechanisms could provide crucial insights into the constraints of current AI models and inform the development of more capable architectures for mechanical reasoning tasks.

REFERENCES

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.

Kelsey Rebecca Allen. *Learning to act with objects, relations and physics*. PhD thesis, Massachusetts Institute of Technology, 2021.

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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Maurits Bleeker, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. Demonstrating and reducing shortcuts in vision-language representation learning. *arXiv preprint arXiv:2402.17510*, 2024.

Andy Clark. *Supersizing the mind: Embodiment, action, and cognitive extension*. oxford university Press, 2010.

claude. Claude models - anthropic. `https://docs.anthropic.com/en/docs/welcome#models`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Qingying Gao, Yijiang Li, Haiyun Lyu, Haoran Sun, Dezhi Luo, and Hokin Deng. Vision language models see what you want but not what you see. *arXiv preprint arXiv:2410.00324*, 2024.

Gemini. Gemini models— gemini api— google ai for developers. `https://ai.google.dev/gemini-api/docs/models/gemini`.

Mariel K Goddu, Alva Noë, and Evan Thompson. Llms don't know anything: reply to yildirim and paul. *Trends in Cognitive Sciences*, 2024.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014.

Graham Harman. *Tool-being: Heidegger and the metaphysics of objects*. Open Court, 2011.

Mary Hegarty. Mental animation: Inferring motion from static displays of mechanical systems. *Journal of experimental psychology: learning, memory, and cognition*, 18(5):1084, 1992.

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

Mary Hegarty and Valerie K. Sims. Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, 22(4):411–430, 1994. doi: 10.3758/BF03200867.

Mary Hegarty, Marcel Adam Just, and Ian R. Morrison. Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. *Cognitive Psychology*, 20(2): 191–236, 1988. doi: 10.1016/0010-0285(88)90019-9.

Susan J. Hespos, Alycia L. Ferry, Emily M. Anderson, Erin N. Hollenbeck, and Leonard J. Rips. Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science*, 27(2):244–256, 2016. doi: 10.1177/0956797615617897.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Mary Kister Kaiser, John Jonides, and Joanne Alexander. Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, 14:308–312, 1986.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Takahiro Kawabe, Kazushi Maruya, Roland W Fleming, and Shin'ya Nishida. Seeing liquids from visual motion. *Vision research*, 109:125–138, 2015.

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.

James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

Richard Lehrer and Leona Schauble. Reasoning about structure and function: Children's conceptions of gears. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 35(1):3–25, 1998.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Michael McCloskey. Intuitive physics. *Scientific american*, 248(4):122–131, 1983a.

Michael McCloskey. Naive theories of motion. *In Mental models (pp. 299-324)*, 1983b.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Henrike Moll and Andrew N Meltzoff. How does it look? level 2 perspective-taking at 36 months of age. *Child Development*, 82(2):661–673, 2011.

OpenAI. Models - openai api. `https://platform.openai.com/docs/models/gpt-4o`.

Charles S. Peirce. *The Essential Peirce - Selected Philosophical Writings (1867-1893), Vol. 1*. Indiana University Press, Bloomington, IN, 1992.

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

John C Raven, John E Raven, and Jacqueline H Court. *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford Psychologists Press, 2000.

Leon Rozenblit, Michael Spivey, and Julie Wojslawowicz. Mechanical reasoning about gear-and-belt diagrams: Do eye-movements predict performance? In *Diagrammatic representation and reasoning*, pp. 223–240. Springer, 2002.

Robert W Shumaker, Kristina R Walkup, and Benjamin B Beck. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Krist Vaesen. The cognitive bases of human tool use. *Behavioral and brain sciences*, 35(4):203–218, 2012.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xuan Zhao, Bertram F Malle, and Hyowon Gweon. Is it a nine, or a six? prosocial and selective perspective taking in four-year-olds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38, pp. 924–929, 2016.

Appendix

A Detailed Analysis of Model Performance on Each Domains of Mechanical Reasoning

A.1 System Stability

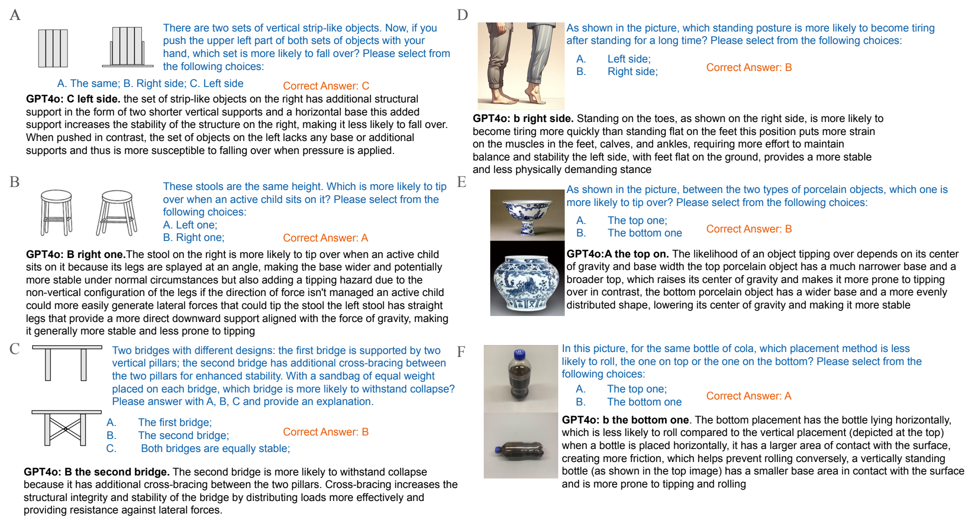

Figure 7: VLMs performance on reasoning about mechanical system stability in MechBench
.

VLMs performance in intuitively evaluating system stability is not ideal, however here we observe some interesting phenomena. Models excel at identifying objects in images, and they not only understand what the objects are, the models also successfully recognize their mechanical states as well. Also, the models could effectively connect mechanical descriptions with the corresponding mechanical scenarios (Figure 7). In Experiments 7B and 7F, the models could recognize the two chairs and the two bottles, and their mechanical situations; however, the models still failed to provide the correct answers in the experiments. For instance, in Experiment 7B, the model explains, ”the stool on the right is more likely to tip over when an active child sits on it because its legs are splayed at an angle, making the base wider and potentially more stable under normal circumstances but also adding a tipping hazard due to the non-vertical configuration of the legs”. The model correctly notices that it’s the angle of the leg that matters for the stability of the system. However, it reasons completely the opposite way to correct answers. When the leg’s angle wider, it’s actually more stable. It’s a very intuitive physical problem for humans but the models fail, even though they still demonstrate step-by-step reasoning abilities in this case. In Experiment 7F, the model is correct that the bottle on the bottom ”has a larger area of contact with the surface, creating more friction, which helps prevent rolling”. However, the model fails to realize one bottle is standing, and one bottle is rolling, and sliding friction and rolling friction are completely different. In contrast, humans can easily solve these problems intuitively.

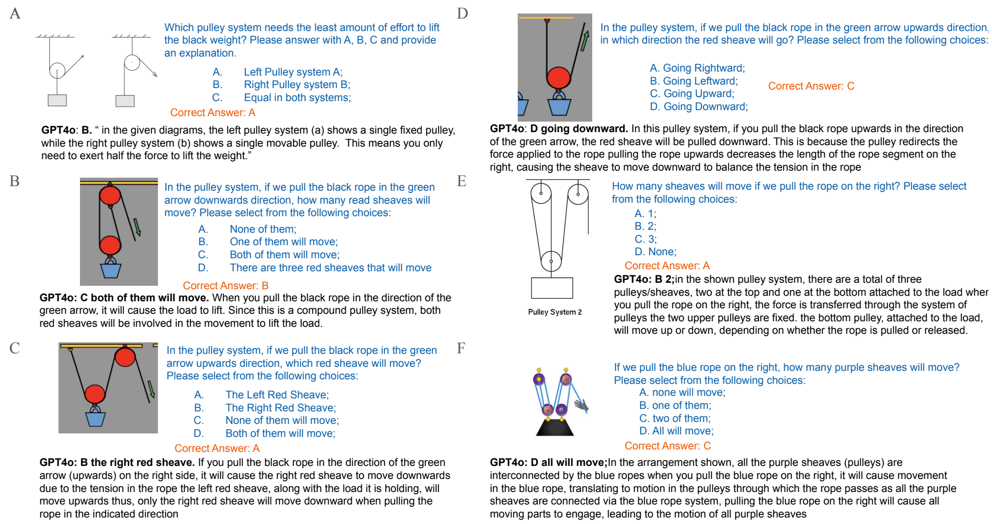

Figure 8: VLMs performance on reasoning about pulley systems in MechBench

.

A.2 Pulley Systems

We find that current VLMs struggle to handle pulley systems (Figure 8). We observe that there are generally three failures in VLMs reasoning about pulley systems: first, VLMs are not able to identify which are the movable pulleys in the system, and second, VLMs exhibit relatively low accuracy in determining whether an object is rising or falling through pulley systems.

VLMs perform poorly in recognizing movable pulley systems. In Experiment 8A, the image includes a standard single movable pulley system and a standard single fixed pulley system. The question "Which system requires less effort?" is essentially asking whether the model can correctly select the movable pulley. Clearly, the model failed in its selection, as it straightforwardly provided an incorrect answer in its explanation. VLMs also struggle in predicting whether a suspended weight is being lifted or lowered through a pulley system. Experiments 8B-F either directly or indirectly reflect this issue, with Experiment 8D being the most direct and concise. In Experiment D, the weight is directly attached to the movable pulley, and by pulling the other end of the rope, the pulley and the weight are lifted. However, the model's response was the exact opposite of the correct answer. In its explanation, the model seemed to imply that the pulley was not fixed (though it did not explicitly state that it was a movable pulley), and the physics it provided was entirely incorrect. Therefore, we can hypothesize that the model's poor performance in predicting the weight's movement may be due to its limited ability to recognize movable pulleys. However, the specific reasons require further experiments to be analyzed in detail.

The above two issues confirm that VLMs still have limitations in recognizing pulley systems, whereas for individuals with some mechanical experience, identifying simple pulley systems through basic diagrams is not difficult (especially in the case of Experiments 8A and 8D).

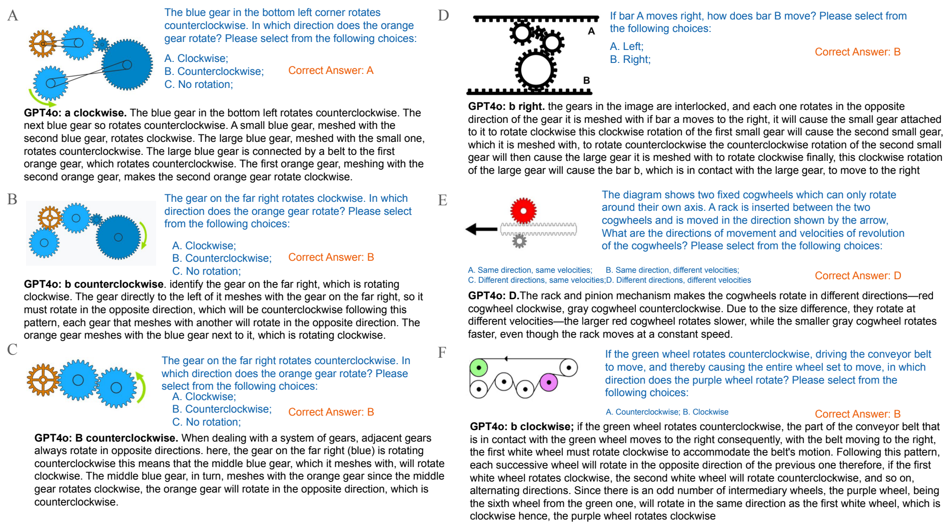

Figure 9: VLMs performance on reasoning about gear systems in MechBench

.

A.3 Gear Systems

On gear and conveyor belt problems, VLMs are able to give highly accurate judgments (Figure 9). In particular, VLMs can reliably determine the rotation of one pulley based on the direction of another.

Compared to pulley systems, we speculate several reasons why VLMs perform better in gear systems. First, gear problems typically involve simple mechanical setups with fixed scenarios, meaning the layout and interactions between gears are often straightforward. This reduces the complexity for the model when analyzing the system, as it does not need to account for too many variables or unpredictable changes in the structure. In other words, a gear system could be reduced into a logical diagram, and solved correctly using that diagram.

Secondly, gear problems rely on a few key physical principles. Specifically, there are two important rules: adjacent gears always rotate in opposite directions, and smaller gears rotate faster than larger ones. Unlike problems that involve complex mathematical calculations, these rules are straightforward for the model to understand and apply consistently. As a result, the model can accurately analyze and predict gear behavior. Similarly, these physical rules can be reduced to logical rules, allowing the model to solve the problem using fixed logical principles.

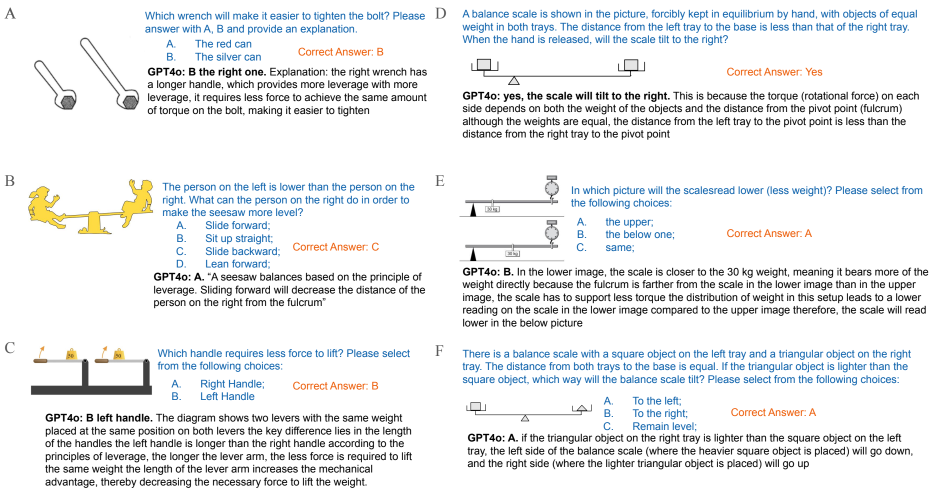

Figure 10: VLMs performance on reasoning about seesaw-like systems and leverage principle in MechBench

.

A.4 Seesaw-like systems and Leverage Principle

We observe diverse behaviors in VLMs on solving seesaw-like mechanical systems and applying leverage principle (Figure 10). Experiments 10A, 10C, 10D, and 10F involve relatively simple scene-based question answering and basic extrapolation. Experiments 10A and 10D are straightforward question-answer tasks. As long as the model identifies the key information in the images—such as the length of the wrench in Experiment 10A or the unequal torques in Experiment 10D—it can easily solve the problems. Experiments 10C and 10F involve simple extrapolations based on scene information, like understanding how changes in the length of the resistance arm affect force in Experiment 10C or predicting the outcome when the weights on a balance scale are unequal in Experiment 10F. It is evident that the model performed exceptionally well in these experiments, providing both correct answers and detailed explanations. However, in Experiments 10B and 10E, VLMs failed to give the correct answers. We suspect that VLMs struggle when the mechanical reasoning processes, particularly applying the leverage principle and solving seesaw-like systems, require multiple steps. For example, in Experiment 10B, while VLMs recognize it is a seesaw-like system, they fail to understand how to operate within this system: sliding backward would give more leverage. This reasoning process may require backtracking through several steps: first, identifying the seesaw-like system; second, understanding how to apply leverage; third, realizing that sliding backward increases leverage; fourth, recognizing that more leverage is needed to level the system; and finally, concluding that sliding backward is necessary. The same issue arises in Experiment 5E.

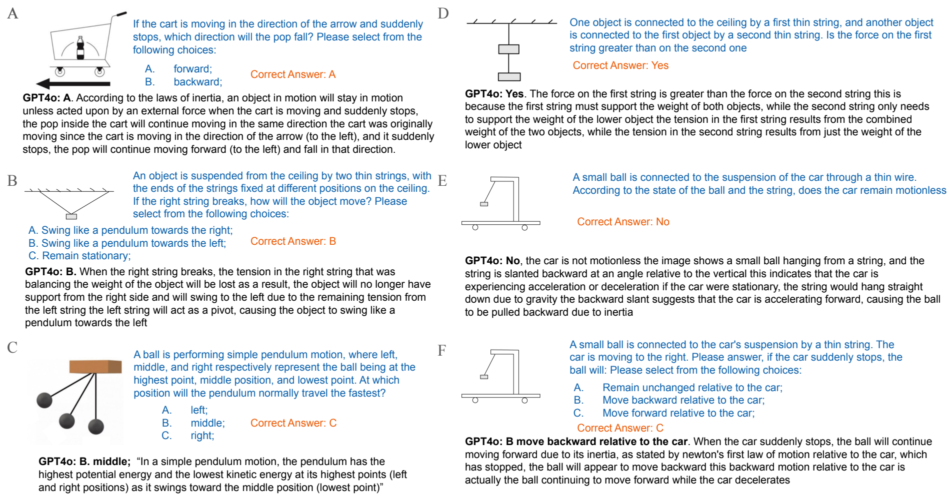

Figure 11: VLMs performance on reasoning about inertia and motion in MechBench

.

A.5 Inertia and Motion

We also observe very diverse behaviors in VLMs' understanding of inertia and motion (Figure 11). Similar to seesaw-like systems, VLMs are able to identify the mechanical situations in the problem settings. However, they lack the ability to effectively predict the next step based on the current scene.

Experiment 11E involves scene-based question. The model not only identified that the cart was not stationary, but it also further analyzed that if the object suspended on the cart were stationary, it would fall vertically due to gravity. However, the depicted scene likely indicates that the cart is accelerating forward. Experiments 11A, 11B, and 11D, on the other hand, are cases of simple extrapolation. They are categorized as "simple extrapolation" because the scenarios are straightforward and involve a single, clear change. For example, in Experiment 11A, the change involves the cart suddenly stopping—a relatively simple scene (involving only two objects, the drink and the cart, with straight-line motion and vertical force equilibrium). Similar patterns apply to Experiments 11B and 11D.

Experiments 11C and 11F, however, involve more complex reasoning related to inertia and motion prediction. Experiment 11F is a derivative question from Experiment 11E, asking what would happen if the cart suddenly stopped. The suspended object would continue moving forward due to inertia. Although the scene in Experiment 11F is similar to Experiment 11E, and the question is similar to that in Experiment 11A, it involves more objects (the pulling rope, the cart, and the pulled object), more physical principles, and a sudden change in force (as the rope loses tension). The model clearly struggled with this problem, providing an incorrect answer and an explanation that did not meet expectations. Notably, Experiment 11C asks a common-sense question based on pendulum motion. The image in Experiment 11C marks three points in the half-arc of the pendulum's swing: the highest point, the quarter-arc point, and the lowest point. This information is also specified in the prompt. However, the model gave an incorrect answer, and its explanation was confused. The model completely misunderstood the designated points in the diagram and in the prompt, mistaking the first and third points as the left and right highest points of the pendulum's motion and the second point as the lowest point. One possible reason for this error is that the model may have mistaken the half-arc pendulum motion in Experiment 11C for a full pendulum motion. A deeper explanation might suggest that the model has issues with perceptual constancy. The image in Experiment 11C might have been misinterpreted as a 3D scene, and this visual misperception could be the root of the error. This hypothesis requires further experimentation to be confirmed.

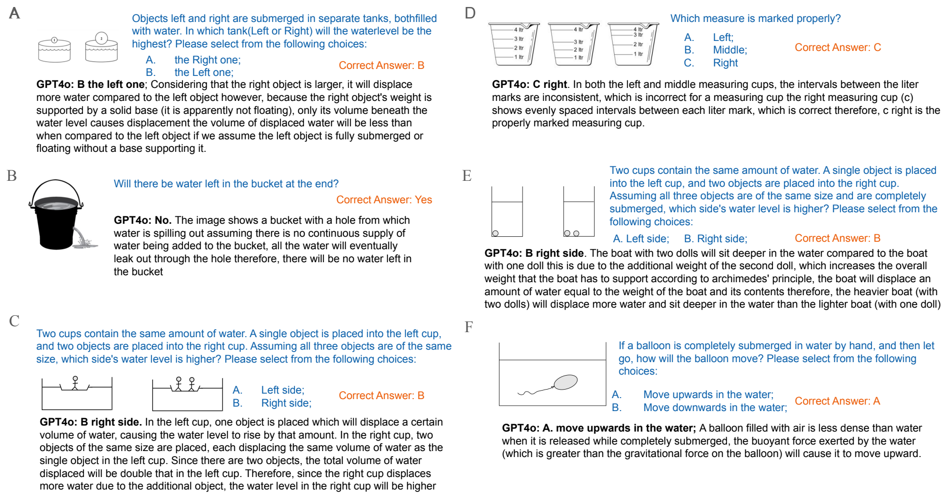

Figure 12: VLMs performance on reasoning about fluid systems in MechBench

.

A.6 Fluid Mechanics

The fluid-related experiments involved properties such as fluid flow, buoyancy, and volume. In fluid-related systems, VLMs still face the aforementioned issues, particularly the challenge of complex inference. However, a notable highlight is that VLMs have demonstrated impressive scene understanding and detail-capturing abilities.

Experiments 12B and 12C highlighted the model's weaker inference abilities. In Experiment 12B, the hole in the bucket was positioned at the middle of the bucket's wall. Clearly, once the water level drops to the level of the hole, water will stop flowing out, meaning some water will remain in the bucket. However, the model failed to capture this crucial detail about the hole's location, leading to an incorrect answer. Similarly, in Experiment 12C, although the model provided the correct answer, its explanation was incorrect.

It is worth noting that the model excelled in image comprehension and detail recognition, especially in Experiments 12D and 12F. In Experiment 12D, the measuring cup had a narrow base and a wider top, meaning the scale markings could not be evenly spaced. The model successfully captured these details, including the design of the measuring cup, the distribution of the scale, and the corresponding numerical values and units, offering a detailed explanation. The same can be said for Experiment 12F. The model accurately identified that the balloon was fully inflated with gas and incorporated the concepts of gas and liquid density to explain the current state and predict the next step of the experiment.