Beyond Accuracy: Diagnosing Large Multimodal Models Reasoning Failures in Multimodal Physics

Anonymous ACL submission

Abstract

Current evaluations of multimodal large language models in physics rely predominantly on final answer accuracy, implicitly equating correct answers with correct reasoning. This assumption overlooks the structured nature of physics problem solving, which requires the accurate perception of visual scenes, the correct interpretation of problem descriptions, and the principled application of physics concepts. This work addresses the mentioned gap by introducing a diagnostic evaluation for multimodal physics reasoning that goes beyond final answer mathcing based accuracy. We propose a fine-grained error taxonomy that disentangles perception, explanation, concept selection, and value interpretation errors, and apply it consistently across physcis reasoning settings and input modalities. Rather than ranking models, our analysis focuses on revealing how and why errors arise. Rather than ranking models, our study provides mechanistic insights into how and why multimodal reasoning breaks down, establishing a foundation for more rigorous and interpretable assessment of physics reasoning systems.

1 Introduction

Reasoning is a fundamental component of intelligent behavior, requiring the integration of observations, background knowledge, and logical inference to derive valid conclusions (Yu et al., 2024; Huang and Chang, 2023; Walton, 1990). In scientific domains such as physics, reasoning additionally demands structured manipulation of symbolic representations, equations, and diagrams, where correctness depends not only on conceptual knowledge but also on faithful interpretation of visual and quantitative information. In multimodal settings, scientific reasoning therefore hinges on the alignment between visual perception and analytical deduction, enabling models to ground physics principles in diagrammatic representations.

multimodal large language models (MLLMs) capabilities remain limited in physics problems that require precise diagram grounding, context-sensitive application of physics laws, and multi-step quantitative reasoning. While advances in training and post-training have significantly improved text-only mathematical and logical reasoning in LLMs (Hendrycks et al., 2021; Lightman et al., 2023), analogous progress in multimodal scientific reasoning has been far more uneven, particularly in physics-intensive settings (Liu et al., 2024).

Chain-of-Thought (CoT) (Wei et al., 2023) prompting has emerged as a widely adopted technique for improving performance on complex reasoning tasks by encouraging models to generate intermediate reasoning steps to further amplify this capability, achieving higher final answer accuracy across a range of benchmarks (Seßler et al., 2024; Xu et al., 2025). However, increased accuracy and more fluent explanations do not necessarily imply faithful or correct reasoning (Zheng et al., 2025b; Tyen et al., 2024). In multimodal tasks (Bi et al., 2025; Huang et al., 2025), errors in visual grounding or early conceptual interpretation can propagate through reasoning chains, producing confident but incorrect output or explanations that are difficult to detect.

We present a diagnostic evaluation of multimodal physics reasoning that disentangles perception and reasoning failures beyond final answer accuracy as discussed in section (§5 and §6). We introduce a unified error taxonomy covering perception and reasoning-level errors in §3 and a corresponding evaluation protocol for step-level analysis in §4.4. we analyze models under both caption-based and direct visual input settings (§4.3 and §5) , and show that improved visual grounding enhances perception but does not reliably resolve conceptual reasoning errors.

Work	Multimodal	Physics-specific	Perception vs Reasoning	Step-level Analysis
(Lightman et al., 2023)	No	No	No	Yes
(Turpin et al., 2023)	No	No	No	Yes
(Liu et al., 2024)	Yes	Partial	No	No
(Xiang et al., 2025a)	Yes	Yes	Partial	No
This work	**Yes**	**Yes**	**Yes**	**Yes**

Table 1: Comparison with prior work along modality, domain specificity, and evaluation granularity. *Partial* indicates limited coverage.

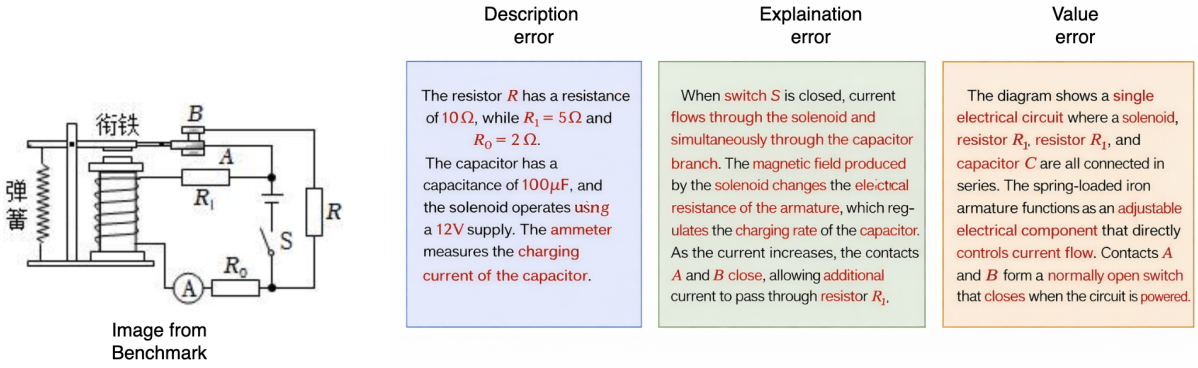


Figure 1: The figure highlights how incorrect scene interpretation as shown in red leads to different perception reasoning failures from left to right. (a) Description error: The model incorrectly collapses the control and working circuits into a single series circuit, misreading the relay structure. (b) Explanation error: Mechanical armature motion is falsely treated as an electrical resistance that regulates current and capacitor charging. (c) Value error: Numerical values (e.g., resistances, capacitance, voltage) are hallucinated despite not being specified in the diagram.

2 Multimodal Physics Benchmarks

2.1 MDK12

MDK12-Bench (Zhou et al., 2025) is a large-scale multimodal benchmark derived from real-world K–12 educational assessments. It spans multiple academic disciplines; in this work, we use the physics subset, which contains 6,827 labeled knowledge points with difficulty annotations, detailed solution explanations, and year-based partitions, enabling fine-grained evaluation of physics reasoning.

2.2 PHYSICS

PHYSICS (Zheng et al., 2025a) is a large-scale dataset for structured physics problem solving, consisting of 16,568 expert-authored and reviewed problems across five domains: mechanics, electromagnetism, thermodynamics, optics, and modern physics. Problems span four difficulty levels (basic, intermediate, advanced, and Olympiad) and are designed to evaluate symbolic reasoning, equation selection, and multi-step derivations.

2.3 PhysReason

PhysReason (Zhang et al., 2025) is a multimodal benchmark for evaluating step-wise physics reasoning. It consists of 1,200 curated problems drawn from international physics competitions and college entrance examinations, spanning four difficulty levels (Knowledge, Easy, Medium, Hard). Each problem includes multimodal inputs and step-by-step reference solutions, enabling evaluation of both final answer accuracy and intermediate reasoning quality.

2.4 PhysUniBench

PhysUniBench (Wang et al., 2025) is a multimodal benchmark for evaluating undergraduate-level physics reasoning in vision–language models. It contains 3,304 human-verified problems in both Multiple-Choice (MCQ) and Open-Ended (OE) formats. The English subset, where each problem is paired with a corresponding diagram, requiring joint visual and textual reasoning. The dataset spans multiple undergraduate-level physics domains, supporting systematic assessment of conceptual understanding, analytical reasoning, and visual grounding.

2.5 SeePhys

SeePhys (Xiang et al., 2025b) is a multimodal benchmark designed to evaluate vision-based physics reasoning in language models. It contains 2,000 curated problems spanning difficulty levels from middle-school concepts to PhD qualifying derivations. Each problem integrates visual diagrams and questions, and includes both multiple-choice and open-ended formats, providing a challenging testbed for analyzing limitations in multimodal scientific reasoning.

3 Physics Error Taxonomy

We propose a unified taxonomy of recurring failure modes in multimodal physics reasoning, derived from a systematic analysis of models outputs across benchmarks. The taxonomy organizes errors into two tightly coupled categories: *perception errors*, which arise from incorrect interpretation of visual entities, structures, or relations in an image; and *reasoning errors*, which arise from failures in problem formulation, physics concept selection, or mathematical execution within a solution. This taxonomy provides a structured framework for diagnosing multimodal reasoning failures beyond final answer accuracy.

3.1 Perception Errors

Perception errors arise when models fail to construct a faithful interpretation of visual inputs in multimodal physics problems. These failures include misidentification of diagram elements, incorrect interpretation of physics or spatial relationships, and erroneous extraction or hallucination of quantitative information associated with visual components. Such errors lead to flawed internal representations of the physics scene, which can subsequently propagate into downstream reasoning failures. Figure 1 illustrates the proposed perception error taxonomy.

Description Errors : Description errors as shown in the Fig 1(a) arise from failures in constructing an accurate representation of the physics scene, including misidentification of objects, motion, or spatial structure, as well as omission or mischaracterization of relevant entities, relationships, or constraints.

Explanation Errors : Explanation errors refer to failures in producing a scientifically valid physics account of the system, where the model's interpretation as shown in the Fig 1(b) does not correctly adhere to governing physics laws, fundamental principles, or causal relationships required to explain the observed configuration.

Value Errors : Value errors as shown in the Fig 1(c) arise from incorrect identification, interpretation, or use of quantitative or symbolic information, leading to invalid or hallucinated parameterization of physics variables, constants, or conditions within the reasoning process.

3.2 Reasoning Errors

Reasoning errors arise when models fail to correctly formulate, organize, or apply physics knowledge during problem solving. These errors operate over the model's internal representation of the problem and therefore depend on both the quality of upstream perception and the model's ability to perform context-sensitive inference. As a result, inaccurate visual interpretation can bias subsequent reasoning, while correct perception alone does not ensure correct problem understanding or inference. Figure 2 summarizes the proposed taxonomy of reasoning errors analyzed in this work.

Problem Miscomprehension : As shown in Fig 2(a) problem miscomprehension occurs when a model constructs an incorrect formulation of the task itself, arising from misinterpretation of the problem objective, overlooked constraints, or misunderstanding of the relevant physics concepts. Because this error appears at the earliest stage of reasoning, it fundamentally shapes all subsequent inference: once the problem is incorrectly framed, later reasoning steps may remain internally consistent yet target the wrong physics question. This failure mode is particularly common in problems that require coordinating multiple physics concepts, interpreting interdependent statements, or performing sequential inference where intermediate results condition later reasoning.

Conceptual Misapplication : As shown in Fig 2(b) Conceptual misapplication refers to failures in selecting or applying appropriate physics principles, laws, or formulas to the problem at hand. These errors arise when models invoke incorrect concepts, combine incompatible principles, or apply valid laws outside their applicable regimes. As a result, the reasoning process may appear structured yet fails to satisfy physics correctness., reflecting limitations in context-sensitive concept se-

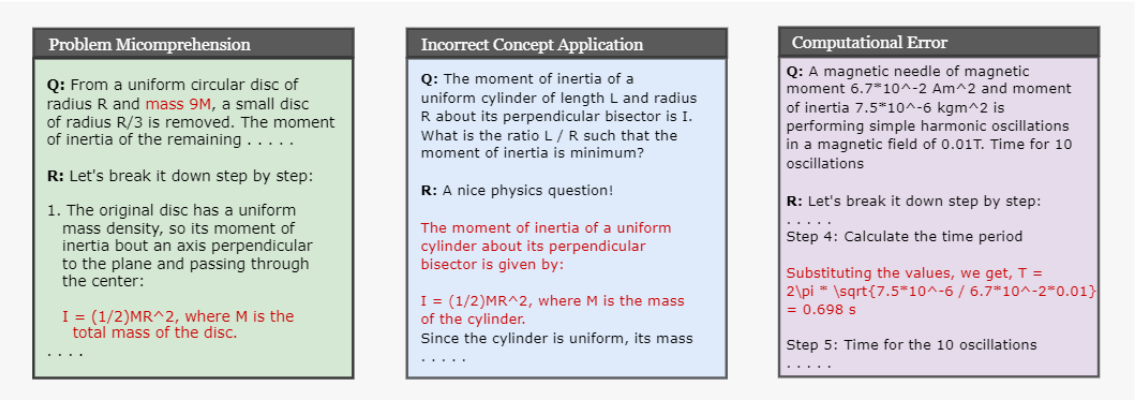


Figure 2: The illustration of three key error as red color observes in the CoT solution of model for physics problems. From left to right : (a) Problem miscomprehension, where the model response uses the incorrect value of variables given in the question here, M instead of 9M, (b) Incorrect concept application in the LLM response, here incorrect moment of inertia formula for uniform cylinder, (c) Computational error within LLM response here, incorrect calculation of time period.

lection rather than a lack of factual knowledge.

Calculation Errors : Calculation errors occur when models violate mathematical consistency as shown in Fig 2(c) during the execution of an otherwise plausible physics solution. These failures include incorrect mathematical operations, breakdowns in algebraic or analytical rules, or hallucinated intermediate computations. Although less frequent than conceptual errors, calculation errors can propagate through the reasoning chain and invalidate the final outcome, often amplifying earlier miscomprehension or conceptual mistakes.

4 Experiments

4.1 Models

We evaluate a diverse set of multimodal large language models spanning a wide range of scales and architectural designs, including both open-source and proprietary systems. The open-source models include the Gemma-3 family (4B, 12B, and 27B) (Team et al., 2025), LLaMA-3.2 Vision models (11B and 90B) (Grattafiori et al., 2024), Phi-4 Multimodal (Abdin et al., 2024), Qwen2.5-VL series (7B, 32B, and 72B) (Bai et al., 2025). In addition, we evaluate leading closed-source multimodal models, Gemini-2.5 Flash (Comanici and other authors, 2025) to contextualize open-source performance against state-of-the-art proprietary systems.

4.2 Setup

Answer-Only : For multiple-choice benchmarks, models are required to select a single or multiple answer option. Evaluation is therefore conducted in an Answer-Only setting, where only the final prediction is available for analysis. This setting supports assessment of final answer accuracy but limits direct inspection of intermediate reasoning. In this work, Answer-Only evaluation is performed under both Caption + Question and Image + Question input formulations, enabling analysis of how different input representations influence final answer selection. This approach is applied to MDK12 single-choice (SCQ) and multiple-choice (MCQ) subsets across difficulty levels MDK12 (SCQ) easy (54), medium (73), and hard (110), MDK12 (MCQ) easy (131), medium (123), and hard (122) as well as PhysUni-MCQ (393).

Chain-of-Thought : For open-ended benchmarks, models are required to generate free-form solutions. Evaluation in this setting leverages Chain-of-Thought outputs, allowing explicit inspection of intermediate reasoning steps. This enables fine-grained analysis of reasoning trajectories, explanation faithfulness, and error propagation beyond final answer correctness. Similar to Answer-Only evaluation, Chain-of-Thought analysis is conducted under both Caption + Question and Image + Question input formulations to disentangle the effects of visual abstraction and direct visual grounding on reasoning behavior. This approach is applied to MDK12 easy (59), medium (54), and hard (58), PhysUni-OE (629), PHYSICS (298), PhysReason (941), and SeePhys (1818).

4.3 Input Modalities

Caption + Question : In the Caption + Question setting, models internally generate a structured tex-

You are an expert physicist. Your task is to provide a comprehensive analysis of the provided image of a physics problem. Your response should be structured into three distinct sections:

Description: Provide a detailed and concise description of the scene in the image. Identify all relevant objects, their physical states, and their arrangement. Use precise physics terminology where applicable (e.g., "a block of mass m is at rest on an inclined plane," or "a positive point charge Q is centered within a spherical Gaussian surface").

Step-by-Step Breakdown (Explanation): Explain the physical principles and processes at play. Break down the scenario into a logical, sequential analysis. For a mechanics problem, this might involve identifying forces, drawing a free-body diagram, and setting up the relevant equations. For an optics problem, it might involve tracing the path of light rays. This section should clearly explain the "why" and "how" of the physics involved.

Values and Quantities: List all the numerical values and variables explicitly shown or implied in the diagram. This includes masses, lengths, angles, velocities, charges, and any other relevant quantities. Clearly state the units for each value. If a value is unknown but represented by a variable (e.g., "m," "L," "θ"), list the variable.

Figure 3: Expert-physicist protocol used to generate structured captions from physics diagrams during inference in the Caption + Question setting.

tual description of the diagram at inference time following a fixed expert-physicist protocol (Figure 3). The generated caption encodes the physics scene, governing principles, and relevant quantities, and is provided alongside the question in place of the original image. Because this formulation requires the model to first interpret visual content and then abstract it into text, errors introduced during this perceptual abstraction stage can influence subsequent reasoning. This setting therefore allows controlled analysis of how perception-driven abstraction affects problem understanding and downstream reasoning behavior.

Image + Question : In the Image + Question setting, models are provided with the original diagram together with the corresponding question, requiring direct visual interpretation in addition to physics reasoning. This formulation preserves full visual grounding and avoids intermediate textual abstraction, serving as a reference point for reasoning under direct perception. Comparing performance between the Image + Question and Caption + Question settings enables us to distinguish errors arising from perceptual abstraction from those caused by reasoning or knowledge application.

4.4 Evaluation

All examples were independently annotated by two annotators with formal physics training using shared written guidelines. Inter-annotator agreement was assessed on a randomly sampled subset, showing high consistency for perception labels and lower agreement for reasoning judgments due to inherent interpretive ambiguity. Disagreements were resolved through discussion and, when necessary, adjudication by a third annotator to produce final labels for analysis

Perception Evaluation. Perception is evaluated by comparing a model's interpretation of a visual input with the reference across three dimensions: scene description, scientific interpretation, and value extraction. Each dimension is labeled as correct or incorrect, with brief justifications for errors to identify the primary source of perceptual failure.

Reasoning Evaluation. Reasoning is evaluated by assessing whether intermediate steps are logically valid and consistent with the reference solution, independent of final answer correctness. The evaluation emphasizes internal coherence, appropriate principle application, and logical progression between steps, rather than outcome accuracy.

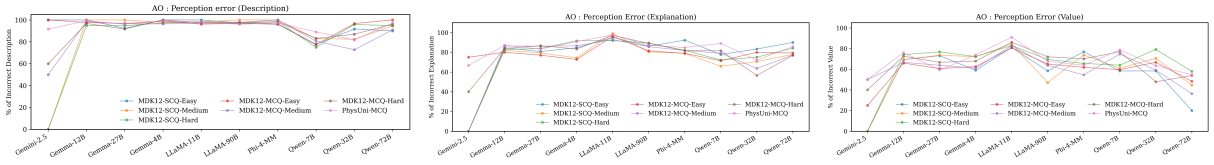
5 Results

5.1 Caption + Question

Fig 6a and 7a summarize final answer accuracy under the Caption + Question setting for Answer-Only and Chain-of-Thought evaluation. Across all benchmarks, performance remains constrained when models rely solely on internally generated textual abstractions of diagrams. Accuracy is highest on simple problems and degrades sharply with increasing complexity, with particularly weak results on multiple-choice benchmarks. Chain-of-Thought prompting consistently improves on open-ended datasets such as MDK12-OE and PhysReason. However, these gains do not alter the overall performance profile: accuracy remains highly variable across benchmarks and model families, and no model achieves robust performance under caption-only inputs. This indicates that while explicit reasoning helps models better utilize available information, it cannot compensate for limitations introduced by caption-based abstraction in the absence of direct visual input.

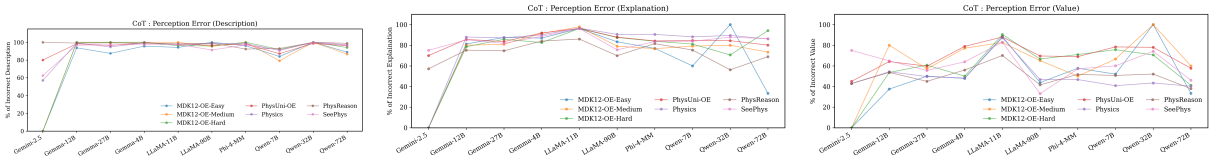
5.2 Image + Question

Fig 6b and 7b summarize final answer accuracy under the Image + Question setting for Answer-Only and Chain-of-Thought evaluation, respectively. Providing the original diagram consistently improves performance relative to caption-based in-



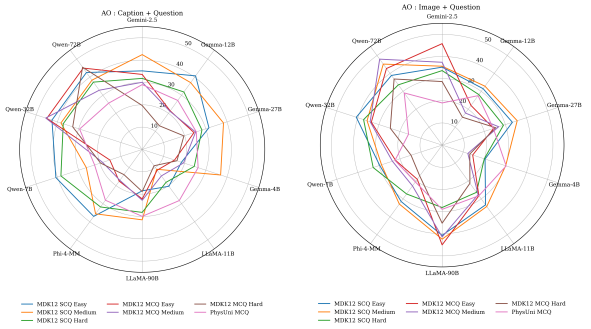
(a) Description (b) Explanation (c) Value

Figure 4: Perception error analysis under answer-only (AO) prompting. We report the percentage of incorrect (left) descriptions, (middle) explanations, and (right) numerical values across models and benchmarks.



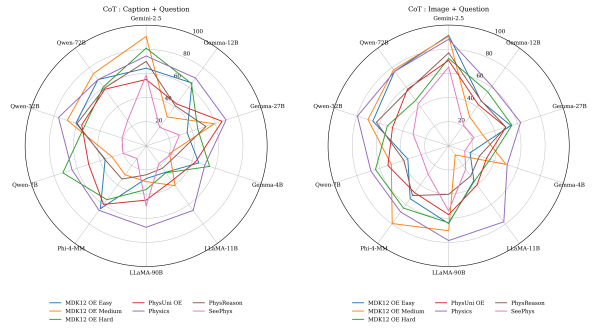
(a) Description (b) Explanation (c) Value

Figure 5: Perception error analysis under chain-of-thought (CoT) prompting. We report the percentage of incorrect descriptions, explanations, and numerical values across models and benchmarks.



(a) AO: Caption + Question (b) AO: Image + Question

Figure 6: Performance comparison across models under different answer-only (AO) input settings. Report accuracy across MDK12 and PhysUni benchmarks with varying difficulty levels.



(a) CoT: Caption + Question (b) CoT: Image + Question

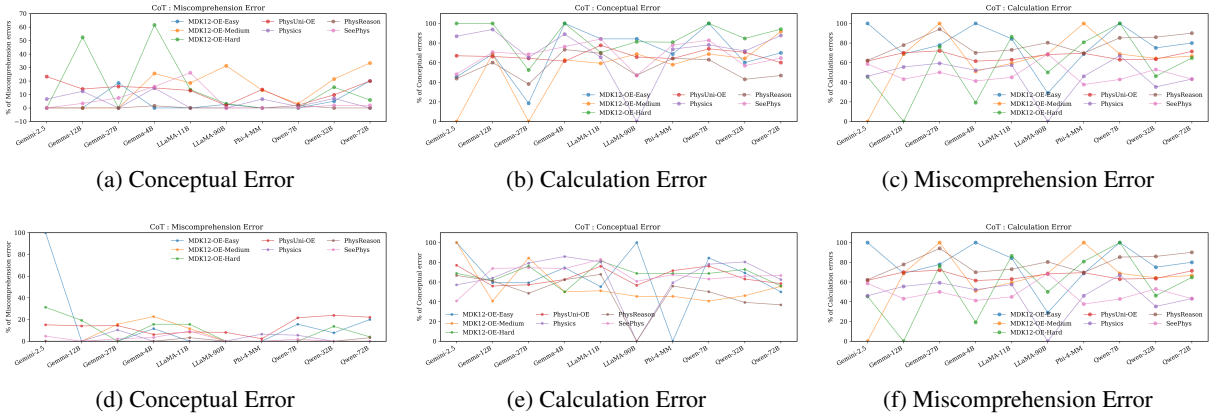
Figure 7: Model performance on open-ended physics benchmarks with direct visual grounding. Radar plots report accuracy across difficulty levels.

puts across most benchmarks and difficulty levels. Under Answer-Only evaluation, gains are most evident on MDK12 multiple-choice subsets, particularly on easy and medium splits, while improvements diminish on harder problems. Chain-of-Thought prompting further boosts accuracy, yielding the highest overall performance across benchmarks, with especially strong gains on visually intensive datasets such as PhysUni-OE, Physics, and SeePhys. However, despite the combined benefits of visual input and explicit reasoning, accuracy still declines on harder subsets and remains inconsistent across model families. These results indicate that direct visual grounding substantially improves problem interpretation and answer selection, but does not fully resolve challenges in complex multimodal physics reasoning.

6 Error Analysis

6.1 Perception Errors

Description Errors : Description errors constitute the dominant perception bottleneck in multimodal physics reasoning under both Answer-Only and Chain-of-Thought evaluation. Across benchmarks, they account for the majority of incorrect predictions (Fig 4a and 5a), indicating persistent difficulty in forming an accurate internal representation of the physics scene. While increasing model scale reduces their frequency, such errors remain prevalent, showing that perceptual abstraction especially when mediated through model-generated captions is inherently brittle. Under Chain-of-Thought prompting, these errors are not corrected but instead explicitly verbalized and propagated through subsequent reasoning, resulting in confi-



(a) Conceptual Error (b) Calculation Error (c) Miscomprehension Error

(d) Conceptual Error (e) Calculation Error (f) Miscomprehension Error

Figure 8: Reasoning Error-type breakdown under chain-of-thought (CoT) prompting across conceptual, calculation, and miscomprehension categories. Caption + Images (a,b,c) and Image + Question (d,e,f) on Benchmarks

dent yet misgrounded explanations. This identifies description-level perception as a structural limitation that neither scaling nor explicit reasoning reliably resolves.

Explanation Errors : Explanation errors increase with problem difficulty across both Answer-Only (AO) and Chain-of-Thought (CoT) settings (Fig 4b and 5b). Under AO evaluation, these errors often remain latent, as models may select correct answers despite incorrect or incomplete physics reasoning. Under CoT prompting, the same failures become explicit: models generate well-structured but invalid physics explanations, exposing a gap between reasoning fluency and scientific validity. This pattern indicates that while models can recognize scene elements, they frequently fail to apply physics principles correctly to govern their interactions.

Value Errors : Value errors occur infrequently across both Answer-Only (AO) and Chain-of-Thought (CoT) evaluations. As shown in (Fig 4c and 5c), incorrect predictions rarely stem from misinterpretation of numerical quantities, symbolic variables, or units, even on medium and hard problem subsets. This pattern holds under both implicit and explicit reasoning, indicating that numerical extraction and basic symbol handling are not primary bottlenecks in multimodal physics reasoning.

6.2 Reasoning Errors in CoT

Problem Miscomprehension : Problem miscomprehension is a dominant failure mode under both Caption + Question (CQ) and Image + Question (IQ) settings, with distinct patterns across benchmarks (Fig 8c and 8f). Under CQ, miscomprehension is especially prevalent on structurally complex datasets such as PhysUni-OE, PhysReason, and SeePhys, where model-generated captions often omit critical constraints or oversimplify relationships, leading to incorrect problem formulations. Providing direct visual input (IQ) reduces miscomprehension on simpler, well-structured problems by preserving spatial and relational information, but does not consistently resolve errors in longer problems that require identifying implicit objectives, interdependent quantities, or non-obvious physics concepts. Larger models benefit more from IQ than CQ, while smaller models exhibit similar miscomprehension rates under both settings. Overall, these results indicate that caption-based abstraction and direct visual grounding each alleviate different aspects of problem understanding, but neither reliably ensures correct inference of problem intent in complex physics tasks.

Conceptual Misapplication : Conceptual misapplication is the primary source of reasoning failure across all benchmarks under CoT evaluation. As shown in the (Fig 8a and 8d)., errors are most concentrated on PhysReason and SeePhys, where problems require adapting physics principles to non-canonical configurations rather than direct formula application. Comparing input modalities reveals a consistent pattern: Caption + Question (CQ) exhibits higher conceptual error rates than Image + Question (IQ), indicating that internally generated captions often abstract or distort critical conditions, leading models to apply generic but incorrect solution strategies. While direct visual input (IQ) reduces such abstraction-induced errors on simpler datasets, it does not resolve conceptual failures on

complex benchmarks. Even with correct visual grounding, models frequently select inappropriate laws or assumptions when multiple constraints interact. These findings show that visual access improves perception but does not reliably improve concept selection, establishing conceptual misapplication as a core limitation of multimodal physics reasoning.

Calculation Errors : Across benchmarks, calculation errors are consistently less frequent than problem miscomprehension and conceptual misapplication under both Caption + Question (CQ) and Image + Question (IQ) settings (Fig 8b and 8e)., indicating that arithmetic execution is not the primary bottleneck in multimodal physics reasoning. However, their persistence under Chain-of-Thought prompting shows that explicit reasoning does not reliably enforce mathematical correctness. Notably, providing direct visual input does not significantly reduce calculation errors compared to caption-based inputs, suggesting that visual grounding primarily affects problem interpretation and concept selection rather than downstream mathematical stability. Overall, calculation errors appear largely as downstream consequences of earlier reasoning failures rather than as isolated numerical deficiencies.

6.3 Reasoning Errors in Correct Final Answers

Even when models produce the correct final answer under Chain-of-Thought (CoT) prompting, their intermediate reasoning can still contain substantive errors. Across both Caption + Question (CQ) and Image + Question (IQ) settings, such false-positive cases where the final answer is correct despite flawed reasoning are relatively rare, typically accounting for under 10% of correct predictions depending on the benchmark. These patterns are consistent across benchmarks and model families, demonstrating that final answer correctness does not imply faithful or robust reasoning. Notably, access to visual input does not eliminate such erroneous traces: while images can help constrain scene interpretation, they do not reliably enforce correct concept selection or mathematical consistency throughout the reasoning process. This reveals a fundamental limitation of CoT based evaluation models may arrive at correct answers despite internally incoherent or scientifically invalid reasoning underscoring the need for evaluation protocols that assess reasoning validity beyond final-answer accuracy.

7 Limitations

A primary limitation of this study is the absence of reliable and scalable verification frameworks for assessing the correctness of intermediate physics reasoning. Existing approaches, including using large language models as judges, are sensitive to response length and formulation, resulting in inconsistent error attribution. Dedicated physics verifiers capable of robustly identifying perceptual, conceptual, and mathematical errors are not yet available, while rule-based methods fail to generalize across diverse multimodal reasoning patterns. Consequently, our analysis relies on targeted human inspection, which, while accurate, is labor-intensive and limits exhaustive annotation across all models and benchmarks.

8 Future Work

Our analysis is that current multimodal language models treat visual information as a static input rather than an active component of reasoning. Images or generated captions are incorporated only at the beginning of inference and are not revisited during multi-step physics reasoning. As a result, early perceptual inaccuracies propagate through subsequent reasoning steps, contributing to conceptual errors. This suggests that improving multimodal physics reasoning will require architectures that support iterative interaction between perception and reasoning, enabling models to dynamically refine visual understanding as inference progresses.

9 Conclusion

This work presents a diagnostic evaluation of multimodal physics reasoning by systematically analyzing perception and reasoning-level errors across existing benchmarks, moving beyond final answer accuracy. Our analysis shows that many failures and even correct predictions stem from flawed scene interpretation, misapplied physics principles, or inconsistent intermediate reasoning that standard metrics fail to capture. Our findings indicate that multimodal models treat visual inputs as static context, allowing early perceptual errors to propagate unchecked through reasoning. Future progress will require iterative perception–reasoning mechanisms that dynamically refine visual understanding during inference.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C.T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 9 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jing Bi, Guangyu Sun, Ali Vosoughi, Chen Chen, and Chenliang Xu. 2025. Diagnosing visual reasoning: Challenges, insights, and a path forward. *Preprint*, arXiv:2510.20696.

Gheorghe Comanici and 3297 other authors. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, pages 1049–1065.

Kai Huang, Jian Zhang, Xiaofei Xie, and Chunyang Chen. 2025. Seeing is fixing: Cross-modal reasoning with multimodal llms for visual software issue fixing. *Preprint*, arXiv:2506.16136.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Kathrin Seßler, Yao Rong, Emek Gözlüklü, and Enkelejda Kasneci. 2024. Benchmarking large language models for math reasoning tasks. *Preprint*, arXiv:2408.10839.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 173 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2024. Llms cannot find reasoning errors, but can correct them given the error location. *Preprint*, arXiv:2311.08516.

Douglas N Walton. 1990. What is reasoning? what is an argument? *The journal of Philosophy*, 87(8):399–419.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Mingyu Ding, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. 2025. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2502.12054*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. 2025a. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. *Preprint*, arXiv:2505.19099.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. 2025b. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*.

Shenghe Zheng, Qianjia Cheng, Junchi Yao, Mengsong Wu, Haonan He, Ning Ding, Yu Cheng, Shuyue Hu, Lei Bai, Dongzhan Zhou, Ganqu Cui, and Peng Ye. 2025a. Scaling physical reasoning with the physics dataset. *arXiv preprint arXiv:2506.00022*.

Xinyi Zheng, Ningke Li, Xiaokun Luan, Kailong Wang, Ling Shi, Meng Sun, and Haoyu Wang. 2025b. Beyond correctness: Exposing llm-generated logical flaws in reasoning via multi-step automated theorem proving. *Preprint*, arXiv:2512.23511.

Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, Jianwen Sun, Haoquan Zhang, Zizhen Li, Xiaofeng Mao, Wangbo Zhao, Kai Wang, Xiaojun Chang, Wenqi Shao, Yang You, and Kaipeng Zhang. 2025. Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. *arXiv preprint arXiv:2504.05782*.