PhysCot: A Benchmark for Multi-Image Physical Reasoning

Anonymous ACL submission

Abstract

Given the temporal nature of physical processes, evaluating scientific LLMs requires assessing whether they understand state transitions by establishing connections across multiple image states. To bridge this gap, we introduce PhysCot, a benchmark of 400 physics-grounded visual sequence completion problems probing MLLMs' ability to reason across multiple images. Models observe four temporally ordered simulation frames spanning 8 subfields, from Kinematics to Electromagnetic Induction, predicting the correct next state from four visually similar but physically distinct candidates. To locate failures, PhysCot provides three disentangled protocols: the full setting tests end-to-end reasoning; a single-frame variant isolates visual state perception; and a text-augmented variant isolates physical world modeling. We observe a substantial performance gap: no model surpasses 50% accuracy despite a 25% random baseline, whereas all human experts exceed 90%, underscoring the considerable distance between current MLLMs and human-level physical reasoning. Our results reveal current MLLMs are fundamentally deficient at multi-image physical reasoning, with fine-grained visual state perception and physical world modeling as the underlying causes.

1 Introduction

A defining characteristic of scientific phenomena, and of physical processes in particular, is that they unfold over time under the governance of objective natural laws. A collision is resolved through successive momentum transfers; a thermodynamic process advances through an ordered sequence of state transitions; the trajectory of a projectile, the diffusion of a solute, or the evolution of an electromagnetic field can only be understood by tracking how the system's variables co-evolve across multiple instants. Faithful scientific reasoning therefore requires more than recognizing the content of an isolated observation: it demands integrating temporally ordered visual evidence, inferring the causal transitions that connect successive physical states, and predicting subsequent states from candidates that are visually similar but physically inconsistent. For Multimodal Large Language Models aspiring to scientific competence, the ability to reason over *sequential, multi-image* visual inputs grounded in physical law is thus not an auxiliary skill but a foundational one, and any rigorous evaluation of scientific MLLMs must place it at the center.

Despite this need, current scientific benchmarks systematically under-test multi-image, physically grounded reasoning. Early efforts such as MATH (Hendrycks et al., 2021b) target mathematical problem solving, while broader suites including SciEval (Sun et al., 2023), MMLU (Hendrycks et al., 2021a), CMMLU (Li et al., 2023), and MMLU-Pro (Yue et al., 2025) extend coverage across physics, chemistry, biology, and astronomy, but evaluate models exclusively through text and therefore leave the visual reasoning capabilities of frontier MLLMs largely unexamined. Multimodal benchmarks such as ScienceQA (Lu et al., 2022) and ChemVLM (Li et al., 2025b) introduce scientific imagery, and MME-CoT (Jiang et al., 2025) and RxnBench (Li et al., 2025a) further probe structured reasoning via chain-of-thought evaluation, yet all remain anchored to a single-image paradigm in which each question is answered from one isolated visual input. VisChainBench (Wenbo Lyu, 2025) advances multi-image reasoning evaluation but remains confined to general-purpose visual chains whose inter-image dependencies are not grounded in domain-specific scientific laws; consequently, it does not measure whether models can reason about the empirical and theoretical principles that causally link scientific observations.

To address this gap, we introduce **PhysCot**, a benchmark of 400 physics-grounded visual sequence completion problems designed to evaluate

the multi-image sequential reasoning capacities of MLLMs.As shown in Fig. 1, each problem presents four temporally ordered context frames depicting a physical process, such as a ball traversing a semi-circular track, charges interacting in an electrostatic field, or a gas undergoing thermodynamic state transitions, with the fifth and final frame deliberately withheld. The model must select the correct missing final state from four visually similar candidates, demanding joint understanding of the underlying physical dynamics, temporal coherence of the depicted process, and fine-grained visual discrimination among near-identical alternatives. By grounding both context frames and candidate options in procedurally generated physical simulations, **PhysCot** demands that models not only perceive fine-grained visual evidence but also internalize the objective physical laws governing sequential processes, two capabilities that, as our experiments reveal, remain fundamentally underdeveloped in current MLLMs. Our contributions are as follows:

- **A Sequential Multi-Image Evaluation Framework for Physical Reasoning.** Existing benchmarks evaluate scientific images as independent collections, overlooking a fundamental property of physical processes: frames are temporally and causally connected, governed by underlying natural principles. Faithful evaluation of multi-image physical reasoning must therefore respect this sequential structure. We identify this as a critical and underexplored gap in current MLLM benchmarking and propose a multi-image sequential evaluation framework to assess the model capabilities of multi-image reasoning.

- **PhysCot: A Benchmark with Three Disentangled Protocols.** We instantiate the above framework as **PhysCot**, a dataset of 400 procedurally generated visual sequence completion problems across 8 physics subfields, in which distractors are generated via physically meaningful perturbations. To enable precise failure attribution, we further introduce three orthogonal evaluation protocols: **PhysCot** jointly evaluates perception and physical reasoning; **PhysCot**^{T} isolates *physical world modeling* by augmenting candidates with textual state descriptions; **PhysCot**^{S} isolates *fine-grained visual perception* via a single context frame.

Table 1: Scientific benchmark comparison analysis. This table compares the key attributes of existing benchmarks to highlight **the scarcity of multi-image benchmarks and visual QA**, which is precisely the motivation for our work. **Legend:** *Subjects*: Math, Physics, Chemistry, etc. *Eval. Methods*: rule- / model- / answer-level based. *Difficulty*: Knowledge Recall, Concept Application, Cognitive Reasoning. *#Multi-Modal Data Construction*: Manual Manual / Auto Automatic annotation.

Benchmark	Subjects	Multi-modal	Multi-images	Visual Options	Chain of Thought	Eval.	Difficulty	#Data.
C-Eval (Huang et al., 2023)		×	×	×	✓	/		-
MATH (Hendrycks et al., 2021b)		×	×	×	✓	/		-
SciEval (Sun et al., 2023)		×	×	×	✓	/		-
MMLU (Hendrycks et al., 2021a)		×	×	×	×	/		-
CMMLU (Li et al., 2023)		×	×	×	×	/		-
MMLU-Pro (Yue et al., 2025)		×	×	×	×	/		-
EESE (Wang et al., 2025)		×	×	×	✓	/		-
AI2D (Kembhavi et al., 2016)		✓	✓	×	×	/ /	-	Manual
ScienceQA (Lu et al., 2022)		✓	×	×	×	/		Manual
MathVista (Lu et al., 2024)		✓	×	×	✓	/ /		Manual
MMMU (Yue et al., 2024)		✓	×	×	×	/		Manual
Q-Mirror (Junying Wang, 2025)		✓	×	×	×	/	-	-
ChemVLM (Li et al., 2025b)		✓	×	×	✓	/	-	Auto
HLE (Phan et al., 2025)		✓	×	×	✓			Manual
MME (Jiang et al., 2025)		✓	×	×	✓	/	-	Manual
RxnBench (Li et al., 2025a)		✓	×	×	✓		-	Manual
VisChainBench (Wenbo Lyu, 2025)		✓	✓	×	✓		-	Manual
PhysCot(Our Work)		✓	✓	✓	✓			Manual

- **Comprehensive Evaluation and Diagnostic Insights.** Benchmarking diverse MLLMs against human experts reveals a fundamental inability to reason across multiple images. Performance on **PhysCot**^{S} falls drastically below human levels, pinpointing *visual perception* as the core deficiency. While **PhysCot**^{T} yields significant improvements by removing visual ambiguity, models still fail to reach human performance, demonstrating that their *physical world modeling* also remains substantially inferior. These findings expose dual bottlenecks in multi-image perception and physical reasoning, providing clear diagnostic targets for future MLLM development and offering actionable guidance for advancing model capabilities in physically grounded visual reasoning and scientific research.

2 Related Work

Scientific Benchmarks. Progress on scientific LLMs has been driven jointly by higher-quality training corpora and increasingly rigorous evaluation suites. Early efforts focus on text-only assessment: MATH (Hendrycks et al., 2021b), C-Eval (Huang et al., 2023), SciEval (Sun et al., 2023), MMLU (Hendrycks et al., 2021a), CMMLU (Li et al., 2023), and MMLU-Pro (Yue et al., 2025) collectively cover a broad range of scientific disciplines through textual questions and answers. With the emergence of large multimodal models, scientific image-grounded benchmarks such as ChemVLM (Li et al., 2025b), Sci-

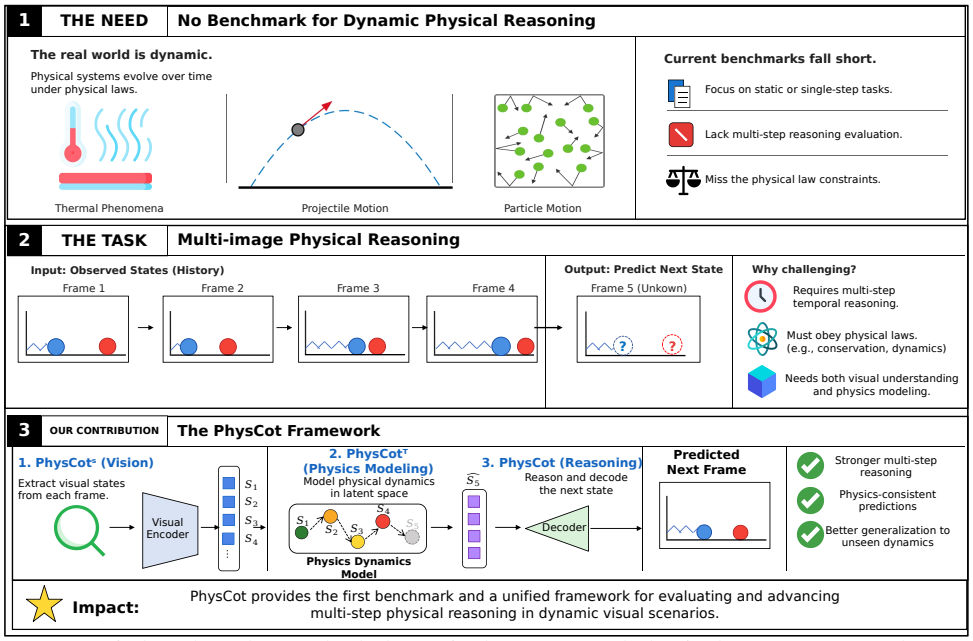


Figure 1: Current static benchmarks overlook dynamic, law-governed physical processes (Panel 1). We formulate the benchmark task as next-state prediction from four sequential frames (Panel 2), and instantiate it via PhysCot's three disentangled protocols, **PhysCot**^{S} (perception), **PhysCot**^{T} (modeling), and **PhysCot** (full) in (Panel 3).

enceQA (Lu et al., 2022), MMMU (Yue et al., 2024), and MathVista (Lu et al., 2024) extend evaluation to single-image scientific reasoning. More recently, VisChainBench (Wenbo Lyu, 2025) introduces multi-image questions, pushing multimodal scientific evaluation toward sequential visual contexts. Despite this progress, as summarized in Table 1, existing benchmarks remain *answer-level*: they score a model by string-matching its textual output against a textual ground truth, while the candidate options and target outcomes themselves are never expressed visually. Scientific reasoning in the real world is inherently visual and sequential and physical phenomena do not unfold as text strings but as observable, continuously evolving states. This omission leaves a fundamental capability untested in the scientific domain, whether a model can genuinely interpret a sequence of images and predict the next-frame visual state, which is precisely the gap our benchmark targets.

Multimodal Reasoning. Reasoning lies at the core of large language models (LLMs), as it ultimately determines whether a model can solve complex, multi-step problems. Seminal works such as Chain-of-Thought prompting (Jason Wei, 2022) and zero-shot "Let's think step by step" reasoning (Takeshi Kojima, 2022) first established protocols for eliciting and evaluating step-wise reasoning in text-only LLMs. As multimodal capabilities have become a central axis of LLM development, this line of work has been extended to the visual domain: Multimodal-CoT (Zhuosheng Zhang, 2024) and ScienceQA (Lu et al., 2022) couple textual rationales with image inputs, while MME-CoT (Jiang et al., 2025) provides a systematic benchmark for chain-of-thought reasoning over visual content. Yet existing efforts overwhelmingly target symbolic, mathematical, or commonsense reasoning, whereas reasoning grounded in the *natural sciences*, constructing an internal model of the physical world and inferring future states from underlying physical principles, remains largely unexplored. Our work directly addresses this gap by introducing a multi-image reasoning benchmark that requires a model to perceive a sequence of frames depicting a physical process, internally model its governing dynamics, and predict the target physical frame.

3 Benchmark Construction

3.1 Key Principles

Multi-Image Reasoning. Evaluating a model's capacity to reason over sequential visual inputs represents a fundamental challenge in assessing scientific understanding. **PhysCot** targets **Multi-Image Reasoning** as its primary evaluation objective, requiring models to compare temporally ordered frames, track the evolution of physical states across time, and predict the missing final state. As illustrated in Fig. 2(3), models must jointly analyze multiple frames captured at different time steps and

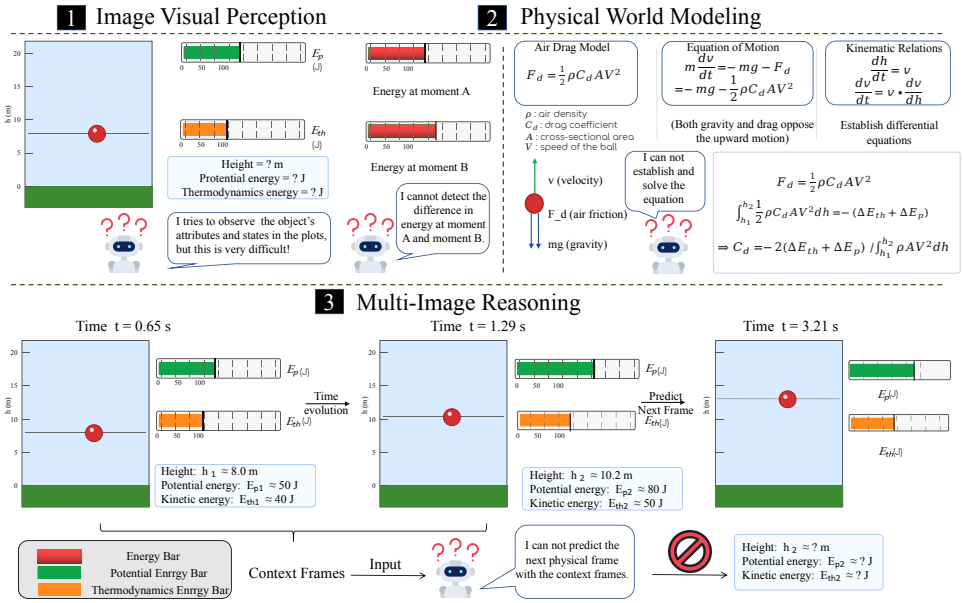


Figure 2: **Overview of key principles.** PhysCot is established on three core principles: (1) evaluating MLLMs' visual perception of physical states from images; (2) probing MLLMs' capacity to model physical world; (3) assessing MLLMs' ability to reason across multiple temporally ordered images.

infer the underlying physical process connecting them, rather than interpreting each image independently. This overarching ability inherently depends on two indispensable sub-capabilities: the ability to accurately *perceive physical state information from visual evidence*, and the ability to *model the physical world* by applying governing laws to predict process evolution. **PhysCot** is designed to evaluate both sub-capabilities as described below.

Image Visual Perception. A necessary prerequisite for multi-image reasoning is the ability to extract and interpret precise physical state information directly from visual inputs, such as reading object positions, energy levels, or gauge readings from rendered scene images. As shown in Fig. 2(1), this requires models to accurately perceive fine-grained visual attributes (e.g. height, velocities, energy values) from plots and scene visualizations, a task that is highly challenging due to the subtle visual differences between physically distinct states. To isolate and evaluate **Image Visual Perception**, **PhysCot**^{S} presents models with only a single context image paired with visually similar candidates lacking explicit textual state descriptions, directly probing whether models can ground fine-grained perceptual features onto physically meaningful quantities without frames context.

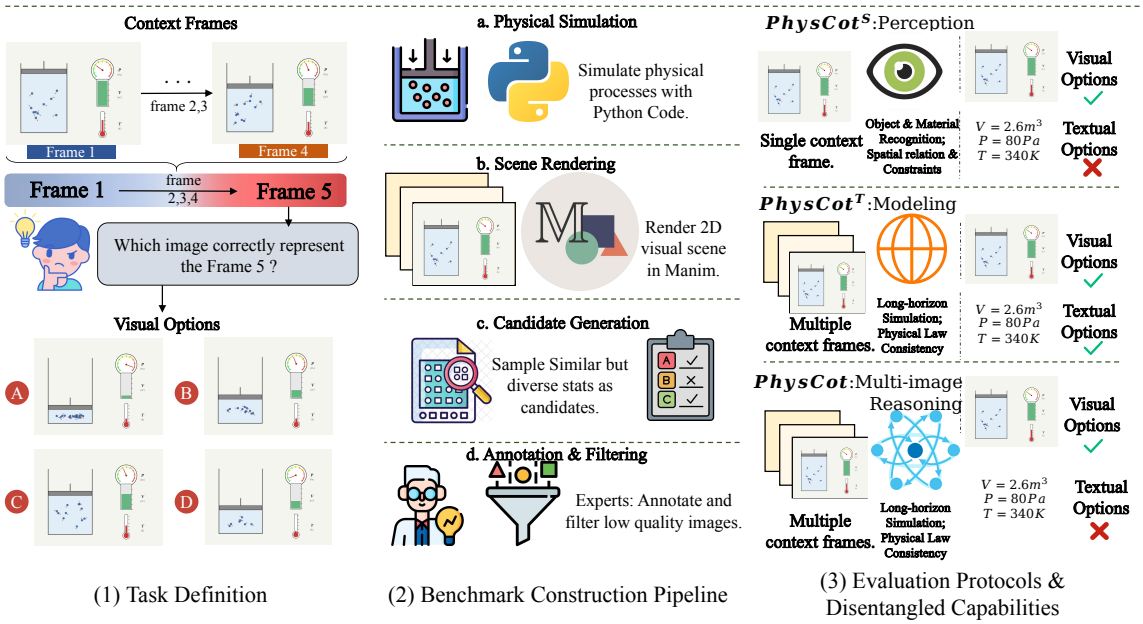
Physical World Modeling. Beyond perception, accurate prediction of physical process evolution requires models to internalize and apply the governing physical laws of the depicted system, constructing an internal physical model, formulating the relevant equations of motion, and performing principled deductions to determine the correct future state. As illustrated in Fig. 2(2), this involves reasoning through dynamic models (e.g.energy conservation) and distinguishing physically consistent outcomes from plausible distractors. To isolate **Physical World Modeling**, **PhysCot**^{T} provides models with explicit textual descriptions of physical state quantities in the candidate options, removing the visual perception bottleneck and directly assessing whether models can apply objective physical principles to reconstruct and predict process evolution from symbolic state information alone.

3.2 Task Definition

As illustrated in Fig. 3, given a sequence of observed context frames (Frame 1 through Frame 4), models are required to predict the correct next state (Frame 5) by selecting from four visually similar but physically distinct candidate options. **PhysCot** provides a principled evaluation of perception and modeling capabilities that together enable accurate multi-image physical reasoning. The benchmark encompasses 8 distinct physical subfields and comprises 400 questions stratified across three difficulty levels, easy, medium, and hard. The distribution of question categories is shown in Fig. 4.

3.3 Benchmark Construction Pipeline

All problems are procedurally generated through a four-stage pipeline to ensure physical accuracy



(1) Task Definition (2) Benchmark Construction Pipeline (3) Evaluation Protocols & Disentangled Capabilities

Figure 3: **PhysCot Framework Overview.** (1) Multi-image physical reasoning task formulation; (2) benchmark construction pipeline; (3) evaluation protocols with disentangled capability assessment.

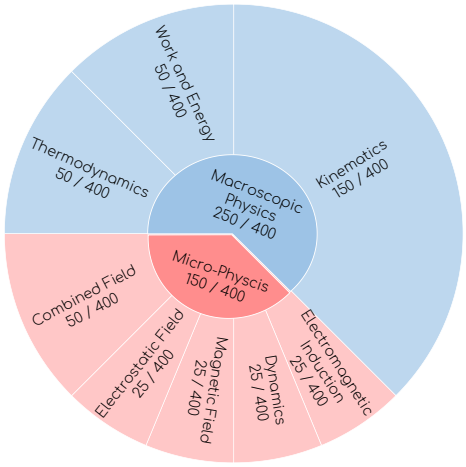


Figure 4: **Physical Subfield Distribution**

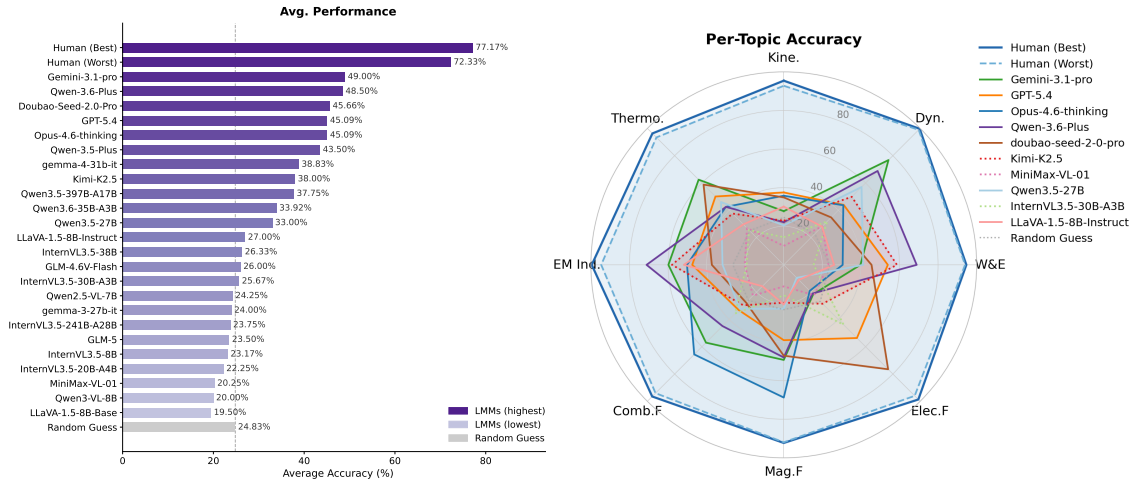
and visual consistency. **(1) Physics Simulation:** Ground-truth physical trajectories are computed via numerical integration using NumPy, governed by Newtonian mechanics, electrostatics, or thermodynamic laws depending on the topic. **(2) Scene Rendering:** Simulated states are rendered into 2D visual scenes with consistent styles using Manim, supplemented by Matplotlib for thermodynamic gauge plots. **(3) Candidate Generation:** Visually similar but physically distinct states are sampled as distractor candidates, ensuring that correct selection requires genuine physical understanding rather than superficial visual matching. **(4) Annotation & Filtering:** Physical states are annotated and low-quality or trivial samples are filtered to ensure benchmark reliability. Full implementation details are provided in Appendix A.1. Question samples are attached in Appendix A.5.

3.4 Evaluation Protocols and Disentangled Capabilities

To systematically disentangle the contributions of visual state perception and physical world modeling to multi-image reasoning, three evaluation conditions are constructed, each targeting a distinct capability as shown in Fig. 3. Evaluation protocols details are shown in Appendix A.3.

PhysCot: Models are presented with four temporally ordered context frames alongside four visually similar candidate images lacking explicit textual state descriptions. This condition evaluates **Multi-Image Reasoning** in its most demanding form, requiring models to extract physical state information directly from images, establish inter-frame relational dependencies, and identify the physically consistent candidate through visual understanding alone, without any symbolic grounding.

PhysCot^T: Models are presented with four context frames, but options are augmented with detailed textual descriptions of the underlying physical states, including numerical values of key physical quantities. Removing the visual perception bottleneck isolates **Physical World Modeling**, probing whether models can apply governing physical



(a) Overall results of **PhysCot**.

(b) Detailed results of **PhysCot**.

Figure 5: **PHYSCOT RESULTS. (a)** Overall accuracy of humans, 24 proprietary and open-source LMMs, and a random-guess baseline. **(b)** Accuracy gap between top LMMs and human experts per subfield across PHYSCOT.

principles to identify the correct outcome.

PhysCot^S: Models are presented with only a single context frame paired with candidates lacking textual descriptions. By withholding intermediate sequential frames, this condition isolates **Image Visual Perception**, assessing object and material recognition, physical state estimation, and the ability to discriminate between visually similar candidates without multi-frame temporal context.

4 Experiment Results

4.1 Benchmark Candidates

To have a comprehensive and updated overview of the evaluation results on the benchmark, state-of-the-art proprietary MLLM and open source MLLM that are widely used in vision tasks are selected for this task. The **Proprietary LMMs** (*closed-source*) include Claude-Opus-4.6 thinking (Anthropic, 2025), GPT-5.4 (OpenAI, 2026), Gemini-3.1-pro (Google DeepMind, 2026a), Qwen3.6-Plus (Qwen Team, 2026e), Qwen3.5-Plus (Qwen Team, 2026c) and Doubao-Seed-2-0-pro-260215 (ByteDance Seed, 2026). The **Open Source LMMs** include Kimi-K2.5 (Moonshot AI, 2026), MiniMax-VL-01 (MiniMax, 2025), Gemma-3-27B-it (Gemma Team, 2025), Gemma-4-31B-it (Google DeepMind, 2026b), Qwen2.5-VL-7B-Instruct (Yang et al., 2024), Qwen3-VL-8B-Instruct (Yang et al., 2025), Qwen3.5-27B (Qwen Team, 2026a), Qwen3.5-397B-A17B (Qwen Team, 2026b), Qwen3.6-35B-A3B (Qwen Team, 2026d), InternVL3_5-8B (Wang et al., 2025), InternVL3_5-GPT-OSS-20B-A4B-Preview (Wang et al., 2025), InternVL3_5-38B (Wang et al., 2025), InternVL3_5-241B-A28B (Wang et al., 2025), LLaVA-OneVision-1.5-8B-Instruct (An et al., 2025) and LLaVA-OneVision-1.5-8B-Base (An et al., 2025).All LMMs are tested with a zero-shot setting and adopted a unified instruction prompt to ensure a consistent evaluation protocol.

4.2 Human Performance

To establish a reliable human baseline for PhysCot, we recruit **10** physics experts, including **3** gold medalists of the Physics Olympiad (PhO), to complete the full 400-problem benchmark through a dedicated evaluation interface. (Details in Appendix A.2) For each problem, the interface displays the four context images together with the textual description of the initial state and the question, and the participant must select, from four candidates, the image that depicts the hidden fifth scene, relying solely on the given conditions and their physics knowledge. To match the conditions under which LMMs are evaluated, we randomize the question order, withhold any information beyond the question and answer options, and provide no feedback during the session. We report the best and worst participant scores as direct human references for comparison with model performance.

4.3 Findings of PhysCot

General Observation: Human experts substantially outperform current state-of-the-art models across all conditions. Fig. 5(a) reveals three consistent patterns. First, under the multi-image text-invisible condition, all LMMs achieve accu-

Categories	Classical Mechanics			Electromagnetism				Thermo.	Avg.↑
LMM *(LLM)*	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (WORST)	92.79%	98.95%	94.00%	96.00%	91.95%	94.03%	94.65%	93.38%	94.47%
HUMAN (BEST)	95.43%	99.70%	94.67%	98.67%	92.29%	96.47%	99.56%	96.31%	96.64%
Proprietary LMMs:									
OPUS-4.6-THINKING	35.81%	43.67%	30.54%	19.06%	**68.76%**	**65.65%**	50.17%	42.50%	44.52%
GPT-5.4	**37.52%**	43.95%	53.95%	53.64%	39.04%	33.09%	47.45%	50.08%	44.84%
GEMINI-3.1-PRO	27.81%	**76.78%**	39.76%	21.58%	49.27%	57.02%	59.82%	**62.42%**	**49.31%**
QWEN-3.6-PLUS	22.11%	68.80%	**68.91%**	21.04%	47.99%	44.89%	**71.14%**	42.45%	48.42%
QWEN-3.5-PLUS	16.46%	67.81%	60.47%	11.05%	36.20%	55.72%	48.15%	50.38%	43.28%
GLM-5	33.72%	16.09%	21.00%	21.58%	39.90%	26.00%	18.16%	14.70%	23.89%
DOUBAO-SEED-2-0-PRO-260215	34.96%	34.76%	45.51%	**76.53%**	47.08%	27.65%	37.28%	58.81%	45.32%
Open-source LMMs:									
KIMI-K2.5	22.87%	49.87%	58.68%	28.50%	19.63%	29.65%	**57.95%**	37.55%	38.09%
MINIMAX-VL-01	9.84%	28.28%	23.15%	22.27%	11.17%	22.58%	20.22%	26.92%	20.55%
GEMMA-3-27B-IT	**33.22%**	33.45%	41.81%	23.51%	16.41%	16.33%	10.03%	18.34%	24.14%
GEMMA-4-31B-IT	25.81%	16.47%	53.39%	**55.10%**	44.77%	31.78%	34.07%	48.86%	**38.78%**
GLM-4.6V-FLASH	24.10%	33.42%	21.34%	36.37%	27.16%	17.66%	20.11%	31.23%	26.42%
Qwen2.5-VL-7B-Instruct	32.51%	23.16%	19.97%	27.57%	15.49%	32.19%	26.49%	16.16%	24.19%
Qwen3-VL-8B-Instruct	9.24%	33.14%	20.72%	19.06%	13.81%	**43.71%**	10.53%	10.33%	20.07%
QWEN3.5-27B	20.47%	57.02%	41.18%	9.48%	22.75%	32.17%	31.74%	46.00%	32.60%
QWEN3.5-397B-A17B	24.98%	**63.57%**	**67.67%**	7.74%	28.65%	42.29%	16.04%	**52.36%**	37.91%
QWEN3.6-35B-A3B	23.05%	51.42%	44.22%	3.72%	29.01%	43.20%	34.84%	43.13%	34.07%
INTERNVL3_5-8B	19.41%	36.57%	26.29%	28.73%	17.37%	15.11%	20.31%	26.16%	23.74%
INTERNVL3_5-GPT-OSS-20B-A4B	30.58%	8.98%	22.15%	33.62%	13.04%	25.24%	23.50%	22.69%	22.47%
INTERNVL3_5-30B-A3B	14.37%	31.87%	15.52%	43.78%	15.80%	35.62%	19.22%	25.49%	25.21%
INTERNVL3_5-38B	25.25%	15.15%	33.74%	35.58%	33.71%	26.24%	24.24%	18.20%	26.51%
INTERNVL3_5-241B-A28B	19.44%	20.44%	36.01%	29.03%	12.88%	30.99%	22.80%	19.90%	23.94%
LLAVA-ONEVISION-1.5-8B-BASE	25.49%	13.85%	16.31%	27.66%	**48.43%**	24.59%	1.13%	2.94%	20.05%
LLAVA-ONEVISION-1.5-8B-INSTRUCT	30.17%	28.04%	26.14%	10.99%	20.53%	15.66%	51.79%	29.65%	26.62%
random guess	24.60%	22.80%	23.80%	26.80%	23.20%	26.00%	26.40%	25.00%	24.82%

Table 2: Benchmark results on the **PhysCot** subset. The best performance is marked in **bold** and the second performance is underlined for both proprietary and open-source LMMs respectively.

racy below 50%, and the weakest proprietary model falls below the random baseline (24.82%), highlighting a fundamental deficiency in multi-image physical reasoning. Among open-source models, fewer than half exceed chance performance. Second, the radar chart in Fig. 5(b) shows that top LMMs exhibit pronounced sub-category imbalances, whereas human evaluators maintain consistently high accuracy across all topics. Third, even the weakest human expert (94.47%) exceeds the strongest LMM, Gemini-3.1-pro (49.31%), by over 45 percentage points, with no model surpassing 80% on any sub-topic. These results demonstrate that current LMMs lag far behind human-level competency in physical sequential process reasoning.

Findings of PhysCot^T: Explicit textual grounding substantially improves accuracy, revealing that visual state perception, not physical reasoning, is the primary bottleneck. Table 3 yields four key findings. (1) Providing physical state information in text yields accuracy gains exceeding 15% for all models; MiniMax-VL-01 improves most dramatically (+29.78%), indicating heavy reliance on symbolic rather than visual cues across all architectures. (2) Under the text-visible condition, all proprietary models exceed 59% and all open-source models exceed 40%, yet the same models fail to extract equivalent information directly from images, confirming that their reasoning is grounded in textual rather than visual evidence. (3) Even under full textual grounding, the best model, Gemini-3.1-pro (78.40%), remains well below the weakest human expert (94.47%), showing that current LMMs cannot reconstruct complete physical processes from visual frames alone. (4) Collectively, these results indicate that current models possess latent physical reasoning capability that is fundamentally gated by insufficient weak visual state perception, and that advances in fine-grained visual grounding and physics-aware representation learning represent the critical path toward closing the human–LMM gap.

Findings of PhysCot^S: Removing context images produces negligible performance change, indicating that current LMMs do not genuinely exploit multi-image sequential information. Table 4 yields three consistent findings. (1) Withholding context images leaves LMM accuracy largely unchanged, demonstrating that models neither leverage the additional information provided by multiple frames nor establish meaningful cross-frame connections to infer the physical principles governing the depicted process. (2) The

Table 3: Benchmark results on **PhysCot** under the *Multi-Image + Text-Visible* condition. The best performance is marked in **bold** and the second in underline for both proprietary and open-source LMMs respectively. Avg. is the arithmetic mean of the eight topic columns.

Categories	Classical Mechanics			Electromagnetism				Thermo.	Avg.↑
LMM *(LLM)*	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (WORST)	92.79%	98.95%	94.00%	96.00%	91.95%	94.03%	94.65%	93.38%	94.47%
HUMAN (BEST)	95.43%	99.70%	94.67%	98.67%	92.29%	96.47%	99.56%	96.31%	96.64%
Proprietary LMMs:									
OPUS-4.6-THINKING	**79.11%**	99.27%	76.54%	**99.14%**	**64.76%**	53.65%	86.17%	56.50%	76.89%
GPT-5.4	65.49%	95.95%	73.95%	69.64%	51.04%	**55.09%**	51.45%	58.08%	65.09%
GEMINI-3.1-PRO	73.81%	**99.49%**	**93.76%**	97.58%	57.27%	47.02%	91.82%	66.42%	**78.40%**
QWEN-3.6-PLUS	53.44%	84.80%	76.91%	57.04%	27.99%	34.89%	75.14%	**66.45%**	59.58%
DOUBAO-SEED-2-0-PRO-260215	51.13%	63.81%	48.47%	79.05%	52.20%	43.72%	**96.15%**	48.38%	60.36%
Open-source LMMs:									
KIMI-K2.5	**65.72%**	68.09%	63.00%	**93.58%**	7.90%	10.00%	**94.16%**	**66.70%**	58.64%
MINIMAX-VL-01	62.99%	**90.76%**	55.51%	40.53%	19.08%	29.65%	57.28%	46.81%	50.33%
GEMMA-3-27B-IT	64.87%	81.87%	60.68%	28.50%	51.63%	21.65%	17.95%	23.55%	43.84%
GEMMA-4-31B-IT	54.51%	88.28%	**73.15%**	50.27%	**63.17%**	**38.58%**	52.22%	58.92%	**59.89%**
QWEN3.6-35B-A3B	53.89%	53.45%	65.81%	43.51%	32.41%	28.33%	90.03%	50.34%	52.22%
random guess	24.60%	22.80%	23.80%	26.80%	23.20%	26.00%	26.40%	25.00%	24.82%

Table 4: Benchmark results on **PhysCot** under the *Single-Image + Text-Invisible* condition. The best performance is marked in **bold** and the second in underline for both proprietary and open-source LMMs. Avg. is the arithmetic mean of the eight topic columns.

Categories	Classical Mechanics			Electromagnetism				Thermo.	Avg.↑
LMM *(LLM)*	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (WORST)	92.79%	98.95%	94.00%	96.00%	91.95%	94.03%	94.65%	93.38%	94.47%
HUMAN (BEST)	95.43%	99.70%	94.67%	98.67%	92.29%	96.47%	99.56%	96.31%	96.64%
Proprietary LMMs:									
OPUS-4.6-THINKING	**36.97%**	33.81%	33.61%	9.96%	72.06%	45.23%	34.61%	44.09%	38.79%
GPT-5.4	36.09%	62.07%	**63.03%**	**58.87%**	39.64%	39.35%	51.26%	**49.04%**	49.92%
GEMINI-3.1-PRO	23.11%	**75.67%**	56.54%	7.06%	**80.76%**	**73.65%**	**54.17%**	42.50%	**51.68%**
DOUBAO-SEED-2.0-PRO	36.82%	47.95%	27.95%	49.64%	55.04%	45.09%	27.45%	24.08%	39.25%
QWEN3.6-PLUS	26.48%	60.78%	41.76%	21.58%	17.27%	37.02%	39.82%	20.42%	33.14%
Open-source LMMs:									
KIMI-K2.5	22.78%	16.80%	32.91%	29.04%	27.99%	10.89%	16.14%	18.45%	21.75%
MINIMAX-VL-01	11.79%	15.81%	16.47%	11.05%	12.20%	21.72%	**44.15%**	22.38%	19.45%
GEMMA-3-27B-IT	32.39%	20.09%	35.00%	25.58%	27.90%	14.00%	10.16%	18.70%	22.98%
GEMMA-4-31B-IT	**39.66%**	34.76%	**47.51%**	**56.53%**	**51.08%**	**35.65%**	13.28%	**34.81%**	**39.16%**
QWEN3.6-35B-A3B	26.20%	**57.87%**	42.68%	8.50%	39.63%	**35.65%**	29.95%	21.55%	32.75%
random guess	24.60%	22.80%	23.80%	26.80%	23.20%	26.00%	26.40%	25.00%	24.82%

negligible sensitivity to context removal confirms that models treat visual inputs as auxiliary rather than substantive evidence, anchoring primarily to textual cues in the problem statement. (3) Notably, Qwen3.6-Plus and Kimi-K2.5 show counter-intuitive accuracy *improvements* upon context removal—an effect particularly pronounced for Kimi-K2.5 across most sub-topics—suggesting that context images not only provide no useful physical signal but actively interfere with reasoning, functioning as distracting noise. Collectively, these findings indicate that current LMMs have yet to develop genuine multi-image physical reasoning: they extract no meaningful information from sequential visual context, exploit no temporal dependencies across frames, and rely on superficial pattern matching rather than physically grounded inference.

5 Conclusion

Current MLLMs exhibit fundamental limitations when reasoning over physically grounded and temporally ordered image sequences, despite strong performance on general visual tasks. Through our three systematic diagnostic protocols, we show that this deficiency stems from two compounding shortcomings: the inability to perceive physical states from visual inputs, and the failure to apply governing physical principles even under explicit state supervision. Progress in multi-image physical reasoning will therefore require fundamentally new paradigms for grounding visual representations in physical dynamics, beyond mere scaling or broader pretraining. We hope PHYSCOT offers a rigorous and reproducible foundation for future investigation into this persistent critical capability gap.

Limitations

We acknowledge five limitations. **Topic scope.** PhysCot covers eight topics spanning classical mechanics, electromagnetism, and thermodynamics, but omits optics, fluid dynamics, quantum mechanics, and relativity; broader coverage would yield a more complete picture of physical reasoning. **Synthetic visuals.** All frames are procedurally rendered with Manim and NumPy, producing clean 2D diagrams. Whether performance transfers to photographs, noisy sensor data, or 3D scenes remains open. **Fixed task format.** Every item predicts the fifth frame from four context frames as four-way multiple choice; alternative formats such as intermediate-frame prediction, free-form generation, or variable context length could probe complementary skills. **Distractor calibration.** Distractors are produced by perturbing the correct frame, but perturbation magnitude is not systematically varied; a controlled difficulty gradient would enable finer-grained capability assessment.

Enabling safer embodied and scientific systems. Robots, autonomous vehicles, and laboratory assistants all act under physical laws; their safety depends on predicting trajectories, forces, and conservation constraints from visual evidence. By exposing where today's MLLMs fail at these predictions, PhysCot directs the field toward models that are safer to deploy in settings where mistaken physical reasoning has real-world consequences.

Expanding who can do physics-aware AI research. We release the full simulation and rendering pipeline, every scene state, and every evaluation protocol. This lowers the cost of building physics-grounded benchmarks for new domains and lets researchers without access to proprietary data or large annotation budgets participate in a research agenda that has historically required both.

Ethical Considerations

This work introduces a benchmark for evaluating the physical reasoning capabilities of Multimodal Large Language Models. All data used in PHYSCOT are synthetically generated from physics simulations and do not involve human subjects, personal data, or sensitive information. We foresee no direct ethical risks associated with the benchmark itself. However, we acknowledge that advances in scientific reasoning for MLLMs may have broader societal implications, and we encourage responsible use of such systems in high-stakes scientific domains.

References

2025 An et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv perprint arXiv:2509.23661*.

Anthropic. 2025. Introducing Claude Opus 4.6. Accessed: 2026-04-21.

ByteDance Seed. 2026. Doubao-Seed-2.0: Official launch. Accessed: 2026-04-21.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2026a. Gemini 3.1 Pro: A smarter model for your most complex tasks. Accessed: 2026-04-21.

Google DeepMind. 2026b. Gemma 4: Byte for byte, the most capable open models. Accessed: 2026-04-21.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Dale Schuurmans Maarten Bosma Brian Ichter Fei Xia Ed Chi Quoc Le Denny Zhou Jason Wei, Xuezhi Wang. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, and 1 others. 2025. Mmecot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*.

Ye Shen Yalun Wu Yingji Liang Yijin Guo Farong Wen Wenzhe Li Xuezhi Zhao Qi Jia Guangtao Zhai Junying Wang, Zicheng Zhang. 2025. Q-mirror: Unlocking the multi-modal potential of scientific text-only qa pairs. *arXiv preprint arXiv:2509.24297*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Hanzheng Li, Xi Fang, Yixuan Li, Chaozheng Huang, Junjie Wang, Xi Wang, Hongzhe Bai, Bojun Hao, Shenyu Lin, Huiqi Liang, Linfeng Zhang, and Guolin Ke. 2025a. Rxnbench: A multimodal benchmark for evaluating large language models on chemical reaction understanding from scientific literature. *arXiv preprint arXiv:2512.23565*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025b. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

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

MiniMax. 2025. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Moonshot AI. 2026. Kimi K2.5. *arXiv preprint arXiv:2602.02276*.

OpenAI. 2026. Introducing GPT-5.4. Accessed: 2026-04-21.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Qwen Team. 2026a. Qwen3.5-27B. Accessed: 2026-04-22.

Qwen Team. 2026b. Qwen3.5-397B-A17B. Accessed: 2026-04-22.

Qwen Team. 2026c. Qwen3.5: Towards native multimodal agents. Accessed: 2026-04-21.

Qwen Team. 2026d. Qwen3.6-35B-A3B. Accessed: 2026-04-22.

Qwen Team. 2026e. Qwen3.6-Plus: Towards real world agents. Accessed: 2026-04-21.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. Scieval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.

Machel Reid Yutaka Matsuo Yusuke Iwasawa Takeshi Kojima, Shixiang Shane Gu. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Junying Wang, Zicheng Zhang, Yijin Guo, Farong Wen, Ye Shen, Yingji Liang, Yalun Wu, Wenzhe Li, Chunyi Li, Zijian Chen, Qi Jia, and Guangtao Zhai. 2025. The ever-evolving science exam. *arXiv preprint arXiv:2507.16514*.

2025 Wang et al. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv perprint arXiv:2508.18265*.

Jinglin Zhao Xianton Zhen Ling Shao Wenbo Lyu, Yingjun Du. 2025. Vischainbench: A benchmark for multi-turn, multi-image visual reasoning beyond language priors. *arXiv preprint arXiv:2512.06759*.

2024 Yang et al. 2024. Qwen2 technical report. *arXIv perprint arXiv:2407.10671*.

2025 Yang et al. 2025. Qwen3 technical report. *arXiv perprint arXiv:2505.09388*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of ACL*.

Mu Li Hai Zhao George Karypis Alex Smola Zhuosheng Zhang, Aston Zhang. 2024. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

A Technical appendices and supplementary material

A.1 Physical Images Collection

All PHYSCOT images are produced by a deterministic Python pipeline that couples a custom `NumPy` physics simulator with the `Manim` Community edition animation engine for 2D scene rendering. This appendix reports the actual code used, abridged to its physically meaningful core and grouped by pipeline stage.

Directory layout. Each PHYSCOT subcategory is organized as one top-level folder (`Motion`, `Shoot`, `Collision`, `Spring`, `Gravity`, `Synthesis`, `new`) containing one numbered sub-directory per instance, each with four context frames, four candidate images, and a `JSON` file recording the simulated physical parameters.

Stage 1: Physics Simulation

The simulator integrates the governing equations of motion with a fixed time step $\Delta t = 5\times10^{-3}$ s, records per-step state, and detects discrete events (collisions, track transitions). The listing below is reproduced from `generate_and_render_51_100.py` for the two-body collision subcategory (spring-panel plus vertical semicircular track).

```python
import numpy as np

g = 9.8
dt = 0.005

def simulate_collision(k, natural_length, m_blue, m_red,
                       track_radius, panel_length,
                       x_blue_init, v_blue_init,
                       red_angle_init, red_omega_init,
                       mu_panel=0.15, max_time=3.0):
    """Integrate spring-panel + semicircular track two-body system."""
    x_blue, v_blue = x_blue_init, v_blue_init
    red_on_track = True
    red_angle, red_omega = red_angle_init, red_omega_init
    x_red_panel, v_red_panel = None, 0.0

    r_blue = 0.20 * np.sqrt(m_blue)
    r_red  = 0.20 * np.sqrt(m_red)

    trajectory, t, last_collision_time = [], 0.0, -999.0
    while t < max_time:
        # --- Blue ball: spring + friction ---
        spring_force  = -k * (x_blue - natural_length)
        friction_blue = -mu_panel * m_blue * g * np.sign(v_blue) \
                        if abs(v_blue) > 1e-2 else 0.0
        a_blue = (spring_force + friction_blue) / m_blue
        v_blue += a_blue * dt
        x_blue += v_blue * dt

        # --- Red ball: rolling on vertical semicircular track ---
        if red_on_track:
            v_tangent    = red_omega * track_radius
            normal_force = m_red * (v_tangent**2 / track_radius) \
                           - m_red * g * np.sin(red_angle)
            if normal_force < 0 and red_angle > 0:
                red_on_track = False
                x_red_panel  = panel_length - 1e-2
                v_red_panel  = -abs(v_tangent)
            else:
                # Rolling sphere: a_t = -(5/7) g cos(theta)
                alpha = -(5.0/7.0) * g * np.cos(red_angle) / track_radius
                red_omega += alpha * dt
                red_angle += red_omega * dt
                if red_angle <= -np.pi/2:
                    red_on_track = False
                    v_red_panel  = red_omega * track_radius
                    x_red_panel  = panel_length - 1e-2
        else:
            # --- Red ball: on panel with friction ---
            friction_red = -mu_panel * m_red * g * np.sign(v_red_panel) \
                           if abs(v_red_panel) > 1e-2 else 0.0
            v_red_panel += (friction_red / m_red) * dt
            x_red_panel += v_red_panel * dt

            # --- Elastic collision ---
            if (abs(x_red_panel - x_blue) < (r_blue + r_red)
                and t - last_collision_time > 0.1):
                last_collision_time = t
                v_blue_new = ((m_blue - m_red) * v_blue
                              + 2 * m_red * v_red_panel) / (m_blue + m_red)
                v_red_new  = ((m_red - m_blue) * v_red_panel
                              + 2 * m_blue * v_blue) / (m_blue + m_red)
                v_blue, v_red_panel = v_blue_new, v_red_new

        trajectory.append({
            'time': t, 'x_blue': x_blue, 'v_blue': v_blue,
            'red_on_track': red_on_track,
            'red_angle': red_angle, 'red_omega': red_omega,
            'x_red': x_red_panel, 'v_red': v_red_panel
        })
        t += dt
    return trajectory
```

Listing 1: Two-body collision simulator (abridged).

Per-step state is serialized to the instance `JSON` (example from `Collision/1/1.json`):

```
{ "scene_id": 1, "m_blue": 2.78, "m_red": 1.24,
  "x_blue_init": -3.03, "x_red_init": 2.68, "y_init": 4.53,
  "vx_blue_init": 1.71, "vx_red_init": -0.54,
  "vx_blue_final": 0.32194, "vx_red_final": 2.57194,
  "t_land": 0.961014, "t_collision": 2.166993 }
```

Listing 2: Ground-truth parameters recorded per instance.

Stage 2: Scene Rendering

Once a trajectory is computed, five temporally ordered frames are rendered through `Manim`. The renderer is invoked by generating a scene-specific Python script on the fly and calling the `manim` executable. The listing below is reproduced from `render_slope_scene.py` / `render_spring_frames.py`.

```python
import subprocess, tempfile, shutil
from pathlib import Path

MANIM_BIN = "/path/to/manim-env/bin/manim"

def render_frame(frame_data, scene_num, frame_idx, output_path):
    """Render a single PNG frame at 1920x1080 using Manim."""
    script = f'''
from manim import *
import numpy as np

config.pixel_height = 1080
config.pixel_width  = 1920
config.frame_rate   = 60

class PhysFrame(Scene):
    def construct(self):
        self.camera.background_color = WHITE

        # --- load simulated state ---
        t   = "{frame_data['time']}"
        k   = float("{frame_data['k'].split()[0]}")
        L0  = float("{frame_data['length'].split()[0]}")
        m1  = float("{frame_data['m_blue'].split()[0]}")
        x1  = float("{frame_data['x_blue'].split()[0]}")
        m2  = float("{frame_data['m_red'].split()[0]}")
        x2  = float("{frame_data['x_red'].split()[0]}")
        E   = float("{frame_data['E_total'].split()[0]}")

        # --- geometry ---
        surface_y = -2.8
        flat = Line([-6, surface_y, 0], [2, surface_y, 0], color=GRAY)
        wall = Rectangle(width=0.2, height=1.2, color=GRAY,
                         fill_opacity=0.5).move_to([-5.9, surface_y+0.6, 0])

        # --- spring (zigzag) ---
        def spring(x_left, x_right, N=8, amp=0.15):
            pts = []
            for i in range(N+1):
                x = x_left + (x_right - x_left) * i / N
                y = surface_y + (amp if i % 2 else -amp)
                pts.append([x, y, 0])
            return VMobject().set_points_as_corners(pts).set_color(BLACK)

        # --- objects with mass-scaled radii ---
        r1 = 0.20 * m1**0.5
        r2 = 0.20 * m2**0.5
        ball1 = Circle(radius=r1, color=BLUE, fill_opacity=0.8)\\
                .move_to([-5 + x1*0.6, surface_y + r1, 0])
        ball2 = Circle(radius=r2, color=RED, fill_opacity=0.8)\\
                .move_to([-5 + x2*0.6, surface_y + r2, 0])

        # --- overlays: time, energy, velocity arrows ---
        time_lbl = Text(f"t = {{t}} s", font_size=28, color=BLACK)\\
                   .to_edge(UP)
        energy_lbl = Text(f"E = {{E:.2f}} J", font_size=24, color=BLACK)\\
                     .next_to(time_lbl, DOWN)

        self.add(flat, wall,
                 spring(-5.9, -5 + x1*0.6 - r1),
                 ball1, ball2, time_lbl, energy_lbl)
'''
    tmp = tempfile.mkdtemp()
    script_path = Path(tmp) / "frame.py"
    script_path.write_text(script)
    subprocess.run([MANIM_BIN, "-ql", "-s", "-o", output_path.name,
                    "--media_dir", tmp, str(script_path), "PhysFrame"],
                   check=True)
    shutil.copy(Path(tmp) / "images/frame" / f"{output_path.stem}.png",
                output_path)
    shutil.rmtree(tmp)
```

Listing 3: Manim-based frame rendering (abridged).

For thermodynamic scenes , the rendering pipeline additionally overlays gauge readings for pressure P, temperature T, and volume V using `matplotlib` sub-plots composited into the `Manim` frame.

Stage 3: Candidate Generation

Three distractors are created by perturbing the ground-truth terminal state along physically meaningful axes and re-rendering through the same pipeline.

```python
def generate_distractors(final_state, perturbation_axes, n=3):
    """Produce physically inconsistent but visually similar candidates."""
    distractors = []
    for axis in perturbation_axes[:n]:
        state = dict(final_state)
        if axis == "momentum_violation":
            state['v_blue'] *= -1.0              # wrong collision direction
        elif axis == "energy_violation":
            state['E_total'] *= np.random.uniform(1.15, 1.3)
        elif axis == "position_offset":
            state['x_red'] += np.random.uniform(0.5, 1.0) \
                              * np.random.choice([-1, 1])
        elif axis == "thermo_perturbation":
            state['T'] *= np.random.uniform(0.85, 1.15)
            state['V'] *= np.random.uniform(0.85, 1.15)
            state['P']  = (state['n'] * 8.314 * state['T']) / state['V']
        distractors.append(state)
    return distractors

# Correct candidate + 3 distractors -> rendered as choice_A..D.png
def save_candidates(true_state, distractors, out_dir):
    all_states = [true_state] + distractors
    order = np.random.permutation(len(all_states))
    answer_idx = int(np.where(order == 0)[0])
    for i, idx in enumerate(order):
        tag = "ABCD"[i]
```

```
        render_frame(all_states[idx], scene_num=out_dir.name,
                     frame_idx=tag,
                     output_path=out_dir / f"choice_{tag}.png")
    return answer_idx  # index of the correct option in A..D
```

Listing 4: Physically meaningful distractor generation (abridged).

Stage 4: Batch Orchestration and Filtering

Scenes are generated in batches by per-subfield orchestrators (`batch_generate_scenes_*.py`). Three filters are then applied: (i) an automatic physical-consistency check (conservation laws, integrator drift), (ii) a visual-readability filter (clipped gauges, occluded objects), and (iii) independent review by two graduate-level physics annotators.

A.2 Human Evaluation Protocol

To establish a reliable human performance reference for PhysCot, we conducted a structured expert evaluation on the full 400-problem benchmark. This appendix details the participant pool, the evaluation interface, and the procedural safeguards used to ensure that the reported human scores accurately reflect the upper bound of human expert capability on our task.

Participant pool. We recruited 10 physics experts who had no prior exposure to PhysCot content. Among them are 3 gold medalists of the Physics Olympiad (PhO), 4 PhD students in physics or applied physics with an active publication record, and 3 senior undergraduates majoring in physics who had previously ranked in the top 5% of national physics competitions. All participants self-reported at least university-level proficiency across the eight PhysCot subcategories: Kinematics, Dynamics, Work & Energy, Electric Fields, Magnetic Fields, Combined Fields, Electromagnetic Induction, and Thermodynamics. Each participant provided informed consent prior to evaluation, and participation was compensated at a rate consistent with local IRB guidelines.

Evaluation interface. Participants completed the benchmark through a dedicated web-based evaluation interface, shown in Fig. 6. The interface presents each problem on a single scrollable page and enforces the same input format given to the evaluated MLLMs: (i) the four temporally ordered context frames rendered at full resolution, (ii) the textual question statement, and (iii) the four candidate next-state images labeled A–D. To prevent information leakage across problems, the interface shuffles problem order independently for each participant and disables browser back-navigation between items. Response selection is recorded together with per-item response time, and no external tools, calculators, or reference materials are permitted.

Question-answering interface. Within each trial, the question-answering panel (Fig. 7) displays the four context frames in a fixed left-to-right temporal order along the top row, with the four candidate images shown as a 2×2 grid directly beneath. A single-click radio selector assigns the participant's answer to one of {A, B, C, D}, and an explicit "Confirm and Next" button must be pressed before the next problem is revealed. Participants may revise their choice at any time before confirmation but cannot return to a previously confirmed problem. This irrevocability mirrors the single-shot evaluation protocol applied to MLLMs and removes the possibility of iterative refinement.

Procedure. Each participant completed the full 400-problem benchmark in one or more sessions, with no individual session exceeding two hours to mitigate fatigue effects. Between sessions, participants were free to rest but were prohibited from discussing specific problems with one another or consulting external sources. A short five-problem calibration block, disjoint from the released benchmark, was administered before the main evaluation to familiarize participants with the interface and response format; calibration responses are excluded from the reported scores.

Aggregation of human scores. We report two human reference levels in our main tables. *Human (Best)* denotes the highest per-subcategory accuracy achieved by any single participant, reflecting the ceiling of human expert performance on the released problems. *Human (Worst)* denotes the lowest per-subcategory accuracy among the 10 participants, providing a conservative lower bound. The gap between these two bounds remains narrow across all eight subcategories (within $\approx$ 3–5 percentage points), indicating consistent expert performance and confirming that the benchmark admits stable, uniquely identifiable ground-truth answers.

Quality-control checks. To verify the integrity of the collected responses we applied three auto-

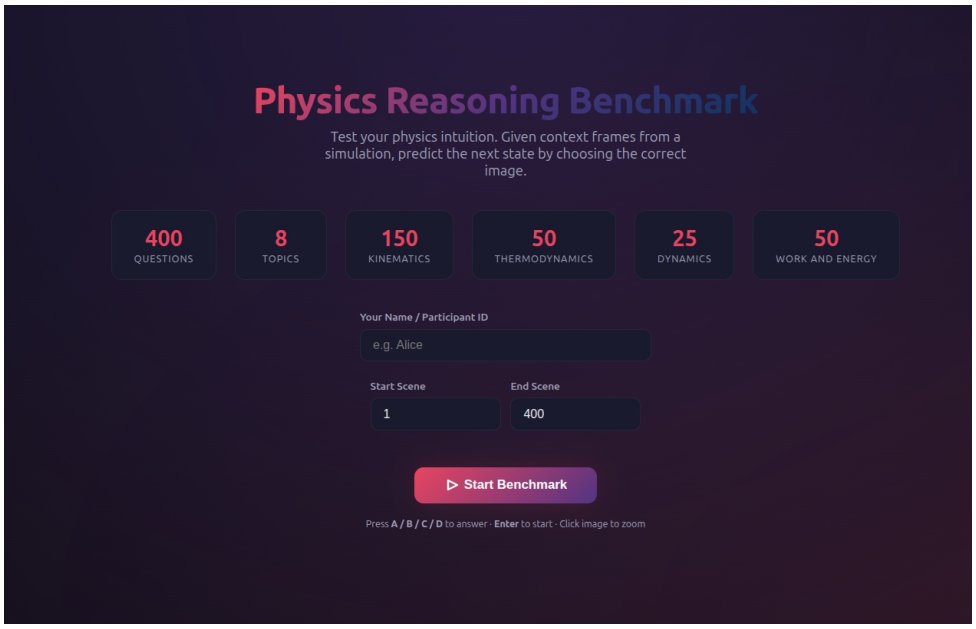


Figure 6: Illustration of the interface for the human test.

mated checks. First, response-time outliers below 5 s or above 20 min per problem were flagged for manual review. Second, per-participant accuracy was compared against the group mean on control problems with known difficulty to detect disengagement. Third, we measured inter-annotator agreement across all 400 problems and observed agreement ≥ 0.92 (Fleiss' κ), confirming that human expert selections are consistent and that residual disagreements cluster on the small subset of problems intentionally designed at the highest difficulty level.

A.3 Benchmark Evaluation Protocols

PHYSCOT is administered to all evaluated models through three disentangled evaluation protocols, PHYSCOT, PHYSCOT^S, and PHYSCOT^T, that share a common textual question but differ in (i) the number of context frames supplied to the model and (ii) whether candidate next-state images are accompanied by textual descriptions of the underlying physical state. This appendix documents the precise input format of each protocol and the Python code used to construct model queries.

Shared input schema. Every PHYSCOT instance exposes the following fields:

- `question["context_images"]`: a list of 4 temporally ordered context-frame paths.
- `question["choices"]`: a list of four candidate entries, each with a `choice_id` (A–D) and an `image_path`.
- `question["candidates"]`: a dictionary mapping `choice_id` to a textual description of that candidate's physical state (e.g., P, T, V for a thermodynamic scenario; position, velocity, energy for a mechanical scenario).

Protocol	Context	Candidate	Capability Isolated
PHYSCOT	4 frames	image only	Multi-image reasoning
PHYSCOT^S	1 frame	image only	Visual perception
PHYSCOT^T	4 frames	image + text	Physical world modeling

Table 5: The three PHYSCOT evaluation protocols.

Images are base64-encoded and submitted at `detail="high"` resolution so that fine-grained state attributes (object positions, gauge readings, energy bars) remain legible to the model.

Protocol summary. Table 5 summarizes the three protocols. The dispatcher `build_query(question, protocol)` selects the appropriate builder at evaluation time.

Protocol 1: PHYSCOT (Full). The full setting supplies all four temporally ordered context frames and presents candidates as images only, without textual state descriptions. This jointly exercises visual perception, temporal integration, and physical reasoning.

```
def build_query_physcot(self, question):
    """PhysCot (full): multi-image context + image
        -only candidates."""
    content = []
    ctx_paths = question.get("context_images", [])
```

Figure 7: Illustration of the interface for Question-Answering.

```
    content.append({"type": "text",
                    "text": f"\n**Context Frames ({len(ctx_paths)} images):**"})
    for idx, img_path in enumerate(ctx_paths):
        image_data = self.encode_image(img_path)
        content.append({"type": "text", "text": f"\nFrame {idx}:"})
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})

    content.append({"type": "text",
                    "text": "\n**Choice Images (which is the correct "
                            "hidden frame?):**"})
    for choice in question["choices"]:
        cid = choice["choice_id"]
        content.append({"type": "text", "text": f"\nChoice {cid}:"})
        image_data = self.encode_image(choice["image_path"])
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})
    return content
```

Listing 5: Query builder for PHYSCOT (full setting).

Protocol 2: PHYSCOT^S (Perception). This protocol supplies only the first context frame, removing temporal information and forcing the model to ground its prediction entirely in fine-grained visual features of a single image. Candidates remain image-only, so neither symbolic state descriptions nor multi-frame reasoning are available. Any performance gap between PHYSCOT^S and the full setting can therefore be attributed to the contribution of visual perception in isolation.

```
def build_query_physcot_s(self, question):
    """PhysCot^S: single context frame + image-only candidates.
    Isolates visual perception by removing multi-frame temporal context.
    """
    content = []
    ctx_paths = question.get("context_images", [])
    ctx_paths = [ctx_paths[0]]                    # <-- single-image context
    content.append({"type": "text",
                    "text": f"\n**Context Frames ({len(ctx_paths)} images):**"})
    for idx, img_path in enumerate(ctx_paths):
        image_data = self.encode_image(img_path)
        content.append({"type": "text", "text": f"\nFrame {idx}:"})
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})

    content.append({"type": "text",
                    "text": "\n**Choice Images (which is the correct "
                            "hidden frame?):**"})
    for choice in question["choices"]:
        cid = choice["choice_id"]
        content.append({"type": "text", "text": f"\nChoice {cid}:"})
        image_data = self.encode_image(choice["image_path"])
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})
    return content
```

Listing 6: Query builder for PHYSCOT^S (perception protocol).

Protocol 3: PHYSCOT^T (Physical World Modeling). This protocol supplies all four context frames and additionally augments each candidate with an explicit textual description of the underlying physical state, for example, P=80 Pa, T=340 K, V=2.6 m^3 for a thermodynamic scenario, or position, velocity, and energy values

for a mechanical one. By exposing the candidate states symbolically, this protocol removes the visual-perception bottleneck and isolates whether the model can apply governing physical laws (conservation, dynamics, long-horizon simulation) to select the physically consistent outcome.

```python
def build_query_physcot_t(self, question):
    """PhysCot^T: multi-image context + text-augmented candidates.
    Removes the visual perception bottleneck by exposing per-candidate
    physical state descriptions; isolates physical world modeling.
    """
    content = []
    ctx_paths = question.get("context_images", [])
    content.append({"type": "text",
                    "text": f"\n**Context Frames ({len(ctx_paths)} images):**"})
    for idx, img_path in enumerate(ctx_paths):
        image_data = self.encode_image(img_path)
        content.append({"type": "text", "text": f"\nFrame {idx}:"})
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})

    content.append({"type": "text",
                    "text": "\n**Choice Images (which is the correct "
                            "hidden frame?):**"})
    for choice in question["choices"]:
        cid = choice["choice_id"]
        content.append{
                                                    # <-- text-augmented candidate
            "type": "text",
            "text": f"\nChoice {cid}:{question['candidates'][cid]}"})
        image_data = self.encode_image(choice["image_path"])
        content.append({"type": "image_url",
            "image_url": {"url": f"{image_data}",
                          "detail": "high"}})
    return content
```

Listing 7: Query builder for PHYSCOT^T (physical modeling protocol).

Unified dispatcher. At evaluation time, a single dispatcher routes each question to the appropriate builder, keeping the rest of the evaluation harness (decoding, answer extraction, scoring) identical across protocols.

```python
def build_query(self, question, protocol="full"):
    """
    protocol:
      "full" -> PhysCot       (multi-image context, image-only choices)
      "s"    -> PhysCot^S     (single context frame, image-only choices)
      "t"    -> PhysCot^T     (multi-image context, text-augmented choices)
    """
    if   protocol == "full": return self.build_query_physcot(question)
    elif protocol == "s":    return self.build_query_physcot_s(question)
    elif protocol == "t":    return self.build_query_physcot_t(question)
    raise ValueError(f"Unknown protocol: {protocol!r}")
```

Listing 8: Protocol dispatcher.

Answer extraction and scoring. All protocols share an identical scoring procedure. Each model is prompted with the same system instruction that requires a single letter from $\{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$ at the end of its response. Answers are parsed by a strict regular expression matching a trailing final-answer tag; responses that fail to produce a parseable letter are scored as incorrect. Accuracy is reported per subcategory and as the arithmetic mean over the eight subcategories, alongside the random-guess baseline (25%).

A.4 Scene Initial State Prompt

For each PhysCot item, the evaluation harness concatenates an initial-state description to the question text before querying the model. The `_describe_initial_state(topic, state)` dispatcher inspects the keys of the scene state and emits one of eight subfield-specific templates. Each template exposes (i) the governing quantities the model must read, (ii) the conserved or evolving quantities it must track, and (iii) a "key physics" checklist that scaffolds the reasoning chain without revealing the answer.

Table 6: Mapping from state-key signatures to subfield prompts.

Subfield	Trigger Keys	Prompt
Kinematics	A, omega, ball_positions	KIN-SHM
Dynamics	m_blue, m_red, vx_blue_init	DYN-COLL
Work & Energy	k, m_blue, v_blue	WE-SPRING
	KE_J, PE_J	WE-GRAV
Electric Fields	q_sign, point_charge	EF-POINT
Magnetic Fields	vx, vy, arc_len	MF-TRAJ
Combined Fields	E_field_N_C, B_field_T	CF-LORENTZ
EM Induction	m_bullet, v_bullet_0	EMI-IMPULSE
Thermodynamics	P_Pa or P_kPa	THERMO-GAS

Kinematics (Kin-SHM).

```
Initial state: simple harmonic motion with amplitude A = {A} m and angular
frequency omega = {omega} rad/s. Ball positions at t=0: {ball_positions}.
Key physics to apply:
 - x(t) = A cos(omega t + phi); v(t) = -A omega sin(omega t + phi).
 - Period T = 2 pi / omega; energy is conserved across the cycle.
 - Identify phase from the provided context frames before predicting the target.
```

Dynamics (Dyn-Coll).

```
Initial state: two-ball collision on a frictionless track. Blue ball m1 = {m_blue}
kg, initial velocity v1 = {vx_blue_init} m/s. Red ball m2 = {m_red} kg at rest
(or with stated v2).
Key physics to apply:
 - Linear momentum is conserved: m1 v1 + m2 v2 = m1 v1' + m2 v2'.
 - For elastic collision, kinetic energy is also conserved.
 - Post-collision velocities determine subsequent positions; infer elasticity
   from the context frames, do not assume it.
```

Work & Energy – spring (WE-Spring).

```
Initial state: horizontal spring-mass system. Stiffness k = {k} N/m, natural
length L0 = {length} m, attached mass m = {m_blue} kg with initial velocity
v0 = {v_blue} m/s.
Key physics to apply:
 - Total mechanical energy E = (1/2) m v^2 + (1/2) k (x - L0)^2 is conserved.
 - Turning points occur where v = 0; maximum speed at the natural length.
 - Match energy and phase to the rendered frames before selecting the target.
```

Electric Fields (EF-Point).

```
Initial state: a test charge in the field of a point charge with sign
q_sign = {q_sign} located at {point_charge}. Test-charge initial position and
velocity are shown in the context frames.
Key physics to apply:
 - Coulomb force F = k q Q / r^2 along the radial direction; sign governs
   attraction vs. repulsion.
 - Trajectory is a conic section for inverse-square central forces.
 - Use the frame sequence to infer whether the orbit is bound or unbound.
```

Magnetic Fields (MF-Traj).

```
Initial state: charged particle entering a uniform magnetic field with initial
velocity (vx, vy) = ({vx}, {vy}) m/s; observed arc length so far = {arc_len} m.
Key physics to apply:
 - Magnetic force F = q v x B is always perpendicular to v; speed is constant.
 - Radius r = m v / (|q| B); direction of curvature is set by the sign of q.
 - Extrapolate along the circular arc; do not assume straight-line motion.
```

Combined Fields (CF-Lorentz).

```
Initial state: charged particle in combined fields with E = {E_field_N_C} N/C
and B = {B_field_T} T. Directions of E, B and the initial velocity are visible
in the context frames.
Key physics to apply:
 - Lorentz force F = q (E + v x B); decompose motion parallel and perpendicular
   to B.
 - E || B gives helical motion with accelerating pitch; E perp B gives a
   cycloid-family trajectory with drift velocity v_d = (E x B) / B^2.
 - Match curvature and drift direction in the context frames before choosing
   the target continuation.
```

EM Induction (EMI-Impulse).

```
Initial state: bullet-block-plank system on a frictionless rail inside a
magnetic region. Bullet mass m_b = {m_bullet} kg with initial speed
v_b0 = {v_bullet_0} m/s strikes the block embedded in the conducting plank.
Key physics to apply:
 - Embedding collision: momentum conservation gives post-embedding velocity.
 - Moving conductor induces EMF = B L v; induced current exerts a retarding
   magnetic force on the plank.
 - The system evolves toward a common drift velocity; identify which frame
   corresponds to steady-state.
```

Thermodynamics (Thermo-Gas).

```
Initial state: ideal gas in a cylinder with pressure P = {P_Pa} Pa
(or {P_kPa} kPa), volume V0, temperature T0, and piston constraints shown in
the context frames.
Key physics to apply:
 - Ideal gas law P V = n R T relates the three state variables.
 - Identify the process type from the frames: isobaric (P const), isochoric
   (V const), isothermal (T const), or adiabatic (P V^gamma const).
 - Use the correct work expression W = integral P dV for the inferred process
   before predicting the target frame's state.
```

In PhysCot^T this prompt is appended to the textual candidates so that physical world modeling is tested in isolation from perception; in PhysCot and PhysCot^S it is placed in the system prompt, supplying consistent physics scaffolding without leaking the target frame's state variables.

A.5 Benchmark Samples

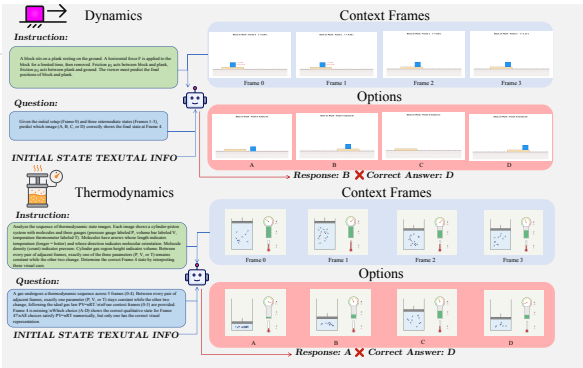


Figure 8: Sample Questions from Dynamics and Thermodynamics

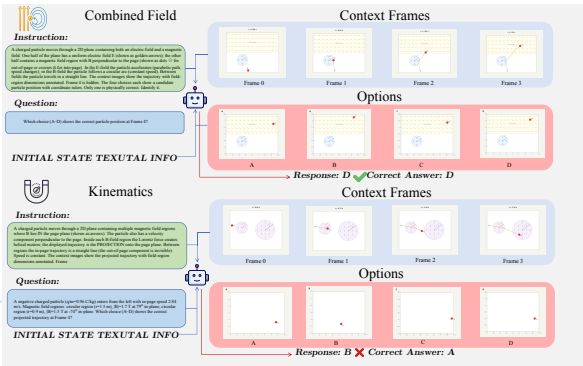


Figure 9: Sample Questions from Combined Field and Kinematics

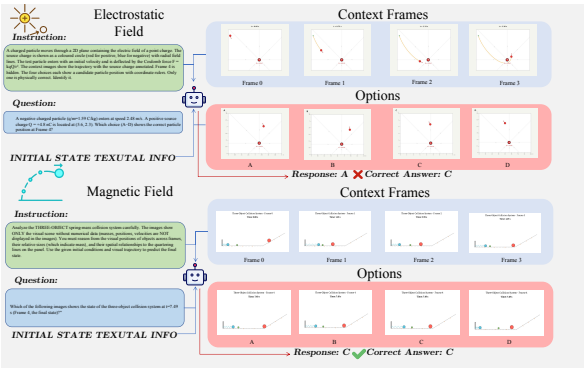


Figure 10: Sample Questions from Electrostatic Field and Magnetic Field

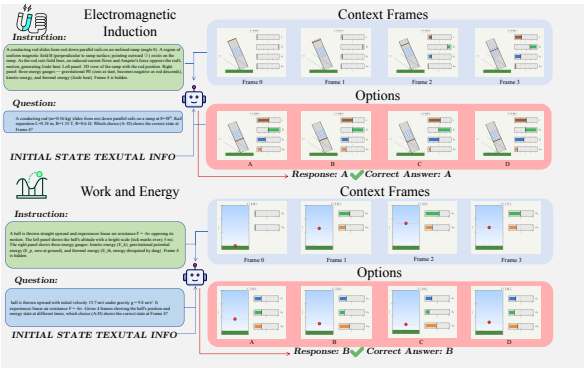


Figure 11: Sample Questions from Electromagnetic Induction and Work & Energy

Categories	Classical Mechanics			Electromagnetism				Thermo.	Avg.↑
LMM (LLM)	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (WORST)	92.79%	98.95%	94.00%	96.00%	91.95%	94.03%	94.65%	93.38%	94.47%
HUMAN (BEST)	95.43%	99.70%	94.67%	98.67%	92.29%	96.47%	99.56%	96.31%	96.64%
Proprietary LMMs:									
OPUS-4.6-THINKING	35.81%	43.67%	30.54%	19.06%	**68.76%**	**65.65%**	50.17%	42.50%	44.52%
GPT-5.4	**37.52%**	43.95%	53.95%	53.64%	39.04%	33.09%	47.45%	50.08%	44.84%
GEMINI-3.1-PRO	27.81%	**76.78%**	39.76%	21.58%	49.27%	57.02%	59.82%	**62.42%**	**49.31%**
QWEN-3.6-PLUS	22.11%	68.80%	**68.91%**	21.04%	47.99%	44.89%	**71.14%**	42.45%	48.42%
QWEN-3.5-PLUS	16.46%	67.81%	60.47%	11.05%	36.20%	55.72%	48.15%	50.38%	43.28%
GLM-5	33.72%	16.09%	21.00%	21.58%	39.90%	26.00%	18.16%	14.70%	23.89%
DOUBAO-SEED-2-0-PRO-260215	34.96%	34.76%	45.51%	**76.53%**	47.08%	27.65%	37.28%	58.81%	45.32%
Open-source LMMs:									
KIMI-K2.5	22.87%	49.87%	58.68%	28.50%	19.63%	29.65%	**57.95%**	37.55%	38.09%
MINIMAX-VL-01	9.84%	28.28%	23.15%	22.27%	11.17%	22.58%	20.22%	26.92%	20.55%
GEMMA-3-27B-IT	**33.22%**	33.45%	41.81%	23.51%	16.41%	16.33%	10.03%	18.34%	24.14%
GEMMA-4-31B-IT	25.81%	16.47%	53.39%	**55.10%**	44.77%	31.78%	34.07%	48.86%	**38.78%**
GLM-4.6V-FLASH	24.10%	33.42%	21.34%	36.37%	27.16%	17.66%	20.11%	31.23%	26.42%
Qwen2.5-VL-7B-Instruct	32.51%	23.16%	19.97%	27.57%	15.49%	32.19%	26.49%	16.16%	24.19%
Qwen3-VL-8B-Instruct	9.24%	33.14%	20.72%	19.06%	13.81%	**43.71%**	10.53%	10.33%	20.07%
QWEN3.5-27B	20.47%	57.02%	41.18%	9.48%	22.75%	32.17%	31.74%	46.00%	32.60%
QWEN3.5-397B-A17B	24.98%	**63.57%**	**67.67%**	7.74%	28.65%	42.29%	16.04%	**52.36%**	37.91%
QWEN3.6-35B-A3B	23.05%	51.42%	44.22%	3.72%	29.01%	43.20%	34.84%	43.13%	34.07%
INTERNVL3_5-8B	19.41%	36.57%	26.29%	28.73%	17.37%	15.11%	20.31%	26.16%	23.74%
INTERNVL3_5-GPT-OSS-20B-A4B-PREVIEW	30.58%	8.98%	22.15%	33.62%	13.04%	25.24%	23.50%	22.69%	22.47%
INTERNVL3_5-30B-A3B	14.37%	31.87%	15.52%	43.78%	15.80%	35.62%	19.22%	25.49%	25.21%
INTERNVL3_5-38B	25.25%	15.15%	33.74%	35.58%	33.71%	26.24%	24.24%	18.20%	26.51%
INTERNVL3_5-241B-A28B	19.44%	20.44%	36.01%	29.03%	12.88%	30.99%	22.80%	19.90%	23.94%
LLAVA-ONEVISION-1.5-8B-BASE	25.49%	13.85%	16.31%	27.66%	**48.43%**	24.59%	1.13%	2.94%	20.05%
LLAVA-ONEVISION-1.5-8B-INSTRUCT	30.17%	28.04%	26.14%	10.99%	20.53%	15.66%	51.79%	29.65%	26.62%
random guess	24.60%	22.80%	23.80%	26.80%	23.20%	26.00%	26.40%	25.00%	24.82%

A.6 LMM Experiment Details

The LMMs are evaluated in a zero-shot setting. Proprietary LMMs are tested via their official APIs. All open-source LMMs are run locally on NVIDIA H200 GPUs with 141 GB HBM3e memory, except for KIMI-K2.5, MINIMAX-VL-01, GEMMA-3-27B-IT, GEMMA-4-31B-IT, GLM-4.6V-FLASH, QWEN3.5-397B-A17B, and QWEN3.6-35B-A3B, which are evaluated through their official APIs due to their large size or access restrictions. All models are tested using their default parameters to ensure the reproducibility of the **PhysCot** results.

A.7 Data Statement

The **PhysCot** dataset is released under the **CC BY 4.0** license. This includes all associated questions, answer candidates and correct answers. We believe this precaution will ensure that **PhysCot** retains its long-term value as a benchmark for multi-images evaluation.

A.8 Social Impact

Advancing science and education. A rigorous, disentangled measure of physical reasoning lets researchers diagnose which capability, visual perception or physical world modeling, currently bottlenecks a model. This diagnostic resolution is what turns raw scale into targeted progress, and targeted progress is what enables reliable AI tutors for physics, automated feedback for students in under-resourced classrooms, and AI assistants that help scientists reason over experimental imagery at a scale no single researcher can match.