PHYSCOT: A BENCHMARK FOR MULTI-IMAGE PHYSICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical reasoning requires models to track ordered states, infer dynamics, and predict physically valid futures. We introduce PhysCot, a benchmark of 400 physics-grounded visual sequence-completion problems for multimodal large language models. Each problem provides four simulation frames from one of eight subfields and asks the model to select the fifth from four visually similar candidates generated by physical-state perturbations. PhysCot combines coherent sequences, image candidates, traceable states, and controlled distractors. Because candidates differ in rendered physical state rather than answer text, the task is designed to evaluate state extraction and consistency checking under a shared explicit physics scaffold. Its three protocols evaluate full visual reasoning, focus on visual state perception with reduced temporal context (PhysCot^S), and probe physical modeling with textual candidate states (PhysCot^T). These protocols are diagnostic rather than causal interventions and provide a common basis for comparing model failure patterns across physical domains. Within the fixed synthetic simulator–renderer distribution, no model exceeds 50% on the full protocol, while the strongest model reaches 78.56% in the mixed text-augmented condition. Temporal-context effects are heterogeneous across models and input configurations; these results characterize in-distribution behavior and do not establish robustness to held-out parameters, renderings, or real imagery outside the benchmark's controlled evaluation setting.

1 INTRODUCTION

A defining characteristic of physical phenomena is that they unfold through temporally ordered state transitions governed by laws. Understanding a collision, thermodynamic process, or charged-particle trajectory requires tracking how state variables evolve, not merely recognizing a frame. Visual physical reasoning therefore combines sequential evidence, process inference, and physically consistent future-state prediction. For multimodal large language models (MLLMs) used in scientific reasoning, models must use scientific images as evidence for law-governed state evolution.

Scientific and multimodal benchmarks test knowledge, perception, and reasoning, but partially cover this setting. Text-only suites such as MATH (Hendrycks et al., 2021), SciEval (Sun et al., 2024), MMLU (Hendrycks et al., 2020), CMMLU (Li et al., 2024), and MMLU-Pro (Wang et al., 2024) lack visual evidence. ScienceQA (Lu et al., 2022), ChemVLM (Li et al., 2025b), MME-CoT (Jiang et al., 2025), CoMT (Cheng et al., 2025a), ViC-Bench (Wu et al., 2025), and VERIFY (Bi et al., 2025) extend evaluation to scientific images and multimodal reasoning. ReMI (Kazemi et al., 2024), MMRB (Cheng et al., 2025b), and OMIBench (Chen et al., 2026) study multi-image evidence, while MVPBench (Dong et al., 2025), PhysUniBench (Wang et al., 2025c), PhysicsArena (Dai et al., 2025), ContPhy (Zheng et al., 2024), and Physion-Eval (Zhang et al., 2026) address physical reasoning and realism. These works typically use textual answers, general multi-image reasoning, videos, diagrams, or explicit variables. PhysCot instead asks models to infer a missing state from coherent visual observations and select among physically perturbed image candidates.

To study this capability, we introduce **PhysCot**, a benchmark of 400 physics-grounded visual sequence-completion problems. Each problem presents four ordered frames from one of eight physics subfields, with the fifth withheld; the model selects the next state from four visually similar candidates. Distractors are generated by controlled perturbations to physical quantities such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

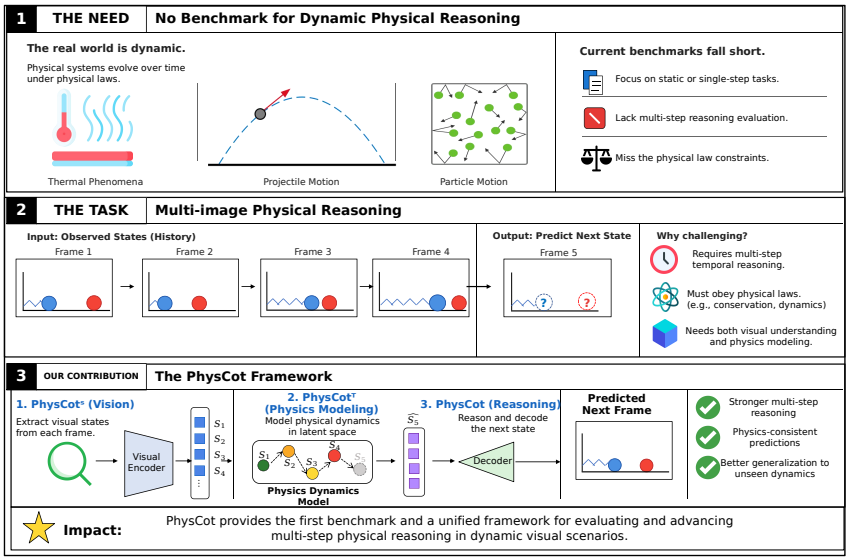


Figure 1: Static benchmarks often overlook dynamic, law-governed physical processes (Panel 1). We formulate next-state prediction from four sequential frames (Panel 2) and instantiate three diagnostic protocols: **PhysCot**^{S} (perception-focused), **PhysCot**^{T} (text-augmented modeling), and full **PhysCot** (Panel 3).

as position, momentum, energy, or thermodynamic variables. Traceable simulation states support reproducibility and analyses beyond aggregate accuracy. All reported items come from the same simulator–renderer families, so subfield coverage should not be interpreted as evidence of out-of-distribution or real-image generalization.

PhysCot combines temporally coherent sequences, image-based answer candidates, controlled distractors, traceable states, and diagnostic protocols. The full **PhysCot** setting evaluates end-to-end visual physical reasoning; **PhysCot**^{S} focuses on visual state perception with one context frame; and **PhysCot**^{T} probes physical modeling with textual candidate states. Together, they provide evidence about visual, physical, and temporal bottlenecks without claiming a causal decomposition. Our contributions are as follows:

- **A Sequential Multi-Image Evaluation Framework for Physical Reasoning.** We formulate visual physical reasoning as next-state prediction from temporally ordered observations, reflecting that consecutive frames are linked states generated by physical laws.
- **PhysCot: A Benchmark with Three Controlled Protocols.** We instantiate this framework as **PhysCot**, a dataset of 400 procedurally generated problems across 8 physics subfields with image candidates, controlled distractors, and traceable states. It provides three protocols: full **PhysCot**, perception-focused **PhysCot**^{S}, and text-augmented **PhysCot**^{T}.
- **Comprehensive Evaluation and Diagnostic Insights.** On the evaluated synthetic distribution, the best model remains below 50% accuracy in the full setting, while the strongest model reaches 78.56% in the mixed text-augmented condition. Context effects vary across models and input configurations, indicating that the three protocols expose different model-dependent failure patterns.

2 Related Work

Scientific and Multimodal Benchmarks. Scientific evaluation has progressed from text-only knowledge and problem-solving suites to visual and multimodal benchmarks. Text-only benchmarks such as MATH (Hendrycks et al., 2021), C-Eval (Huang et al., 2023), SciEval (Sun et al., 2024), MMLU (Hendrycks et al., 2020), CMMLU (Li et al., 2024), and MMLU-Pro (Wang et al., 2024) cover broad scientific and mathematical domains, but evaluate reasoning through text. Multimodal benchmarks such as AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), MathVista (Lu et al., 2023), MMMU (Yue et al., 2024), ChemVLM (Li et al., 2025b), RxnBench (Li et al., 2025a),

Table 1: Scientific benchmark comparison analysis. This table compares representative scientific and multimodal benchmarks along dimensions relevant to PhysCot, including modality, multi-image input, visual answer options, reasoning format, evaluation style, difficulty coverage, and data construction. **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., 2021)		×	×	×	✓	/		-
SciEval (Sun et al., 2024)		×	×	×	✓	/		-
MMLU (Hendrycks et al., 2020)		×	×	×	×	/		-
CMMLU (Li et al., 2024)		×	×	×	×	/		-
MMLU-Pro (Wang et al., 2024)		×	×	×	×	/		-
EESE (Wang et al., 2025a)		×	×	×	✓	/		-
AI2D (Kembhavi et al., 2016)		×	✓	×	×	/ /	-	Manual
ScienceQA (Lu et al., 2022)		✓	×	×	×	/		Manual
MathVista (Lu et al., 2023)		✓	×	×	✓	/ /		Manual
MMMU (Yue et al., 2024)		✓	×	×	×	/		Manual
Q-Mirror (Wang et al., 2025b)		✓	×	×	×	/	-	-
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 (Lyu et al., 2025)		✓	✓	×	✓		-	Manual
PhysCot		✓	✓	✓	✓			Auto

and MME-CoT (Jiang et al., 2025) extend evaluation to diagrams, scientific images, and image-grounded rationales. These datasets are important, but most use single images or textual answer formats. PhysCot instead focuses on visual state completion: the model must infer a valid missing state and select it from image candidates, keeping both evidence and answers visual.

Multi-Image Reasoning. Recent work has begun to evaluate reasoning across multiple images. ReMI (Kazemi et al., 2024) studies reasoning from image combinations, MMRB (Cheng et al., 2025b) evaluates multimodal multi-image reasoning, OMIBench (Chen et al., 2026) targets olympiad-level multi-image problems, and VisChainBench (Lyu et al., 2025) introduces multi-turn, multi-image visual reasoning beyond language priors. These benchmarks show that integrating distributed visual evidence is difficult. However, their primary goal is general multi-image reasoning, visual pattern inference, or multimodal question answering. PhysCot is complementary: its context images are temporally ordered states generated by one physical process, requiring models to track evolution and select a dynamically consistent future frame from visual options.

Physical Reasoning Benchmarks. Physical reasoning has also been studied through benchmarks for intuitive physics, physics problem solving, and physical realism. MVPBench (Dong et al., 2025) evaluates graph-based visual physical Chain-of-Thought reasoning over multi-path physical reasoning; PhysUniBench (Wang et al., 2025c) and PhysicsArena (Dai et al., 2025) test undergraduate or structured physics reasoning; ContPhy (Zheng et al., 2024) studies continuum physical concept learning from videos; and Physion-Eval (Zhang et al., 2026) evaluates physical realism in generated videos through human reasoning. SeePhys (Xiang et al., 2025) is particularly relevant to the perception–reasoning interface: it evaluates physics questions grounded in heterogeneous diagrams and contrasts vision-essential, text-plus-vision, and text-only conditions. Its primary unit is diagram-grounded problem solving rather than temporal completion of a simulated state sequence. MV-Physics (Anonymous, 2026) further studies physical visual scenes with a hierarchical evaluation spanning single-image and multi-image settings. Compared with these efforts, PhysCot focuses on temporally ordered simulation states, image-based next-state candidates, and controlled distractors whose differences are tied to physical variables. Thus, PhysCot complements prior visual-physics benchmarks by targeting temporal state prediction rather than static diagram interpretation or general multi-image question answering.

Multimodal Chain-of-Thought and Diagnostic Evaluation. Chain-of-Thought prompting (Wei et al., 2022) and zero-shot step-by-step reasoning (Kojima et al., 2022) established a practical way to elicit intermediate reasoning in text-only LLMs. Multimodal-CoT (Zhang et al., 2024), MME-

Table 2: Comparison with closely related multi-image and physical reasoning benchmarks. We summarize design dimensions relevant to visual physical state prediction: temporally coherent physical sequences, image-based answer candidates, physically meaningful distractors, traceable physical states, and controlled diagnostic protocols. A checkmark indicates that the property is explicitly supported by the published methodology; a cross indicates false.

Benchmark	Temporal Physical Sequence	Image Answer Candidates	Meaningful Distractors	Traceable States	Controlled Protocols
MVPBench (Dong et al., 2025)	✓	×	×	×	×
ReMI (Kazemi et al., 2024)	×	×	×	×	×
PhysUniBench (Wang et al., 2025c)	×	×	×	×	×
PhysicsArena (Dai et al., 2025)	×	×	×	×	×
ContPhy (Zheng et al., 2024)	✓	×	×	✓	×
Physion-Eval (Zhang et al., 2026)	✓	×	×	×	×
MMRB (Cheng et al., 2025b)	×	×	×	×	×
OMIBench (Chen et al., 2026)	×	×	×	×	×
SeePhys (Xiang et al., 2025)	×	×	×	×	✓
MV-Physics (Anonymous, 2026)	×	×	×	×	✓
PhysCot	✓	✓	✓	✓	✓

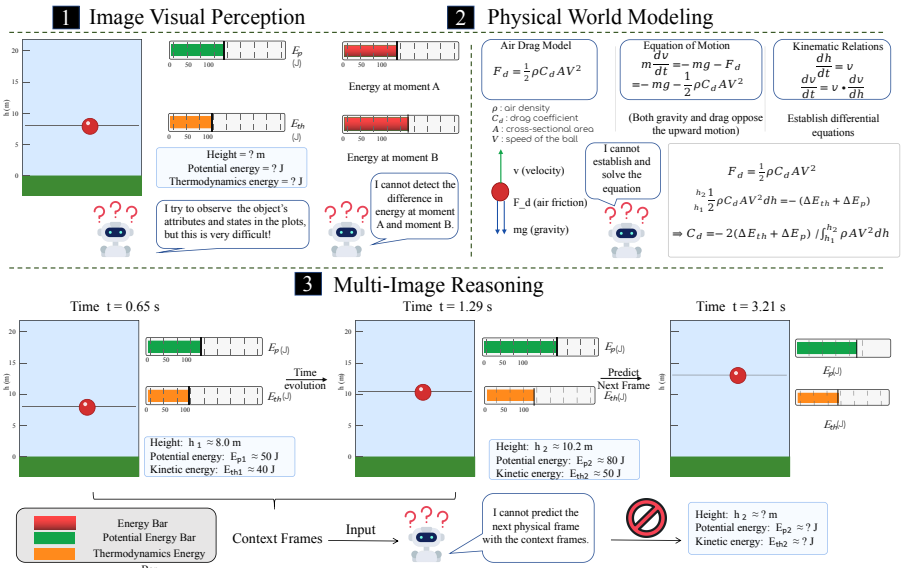


Figure 2: **Overview of PhysCot's evaluation principles.** PhysCot is designed to diagnose three capabilities required for visual physical reasoning: perceiving physical state information from images, modeling the underlying physical world, and integrating temporally ordered observations to infer a missing future state.

CoT (Jiang et al., 2025), CoMT (Cheng et al., 2025a), ViC-Bench (Wu et al., 2025), and VERIFY (Bi et al., 2025) extend this direction to visual inputs, multimodal reasoning traces, and reasoning fidelity. PhysCot follows this diagnostic spirit, but does not rely only on generated rationales. Instead, it evaluates whether models recover the correct physical state under controlled variants: PhysCot^S emphasizes fine-grained visual state perception under reduced temporal context, PhysCot^T probes physical world modeling with textual state information, and the full PhysCot setting evaluates their integration over ordered images. This structure provides evidence about visual, physical, and temporal errors without treating them as perfectly separable causes.

3 Benchmark Construction

3.1 Key Principles

Multi-Image Physical Reasoning. Physical processes unfold as ordered state transitions rather than independent snapshots. PhysCot evaluates whether a model can compare temporally ordered frames, track evolving physical quantities, and infer the missing final state. As illustrated

in Fig. 2(3), this requires connecting multiple observations through an underlying process instead of interpreting each image in isolation. This capability depends on extracting physical state information from visual evidence and applying the supplied physical principles across time in each sequence.

Visual State Perception. A necessary prerequisite for multi-image physical reasoning is reading fine-grained physical state information from rendered scientific images, such as object positions, heights, velocities, energy levels, field directions, or thermodynamic gauge values. As shown in Fig. 2(1), these cues may differ only subtly between physically distinct states. PhysCot^S provides a perception-focused condition with a single context frame, visually similar candidates, and no explicit candidate-state descriptions. It tests whether models can use perceptual features to identify physical quantities when temporal evidence is limited, while the future-state target means that the condition is not a pure measure of perception independent of physical prediction.

Physical World Modeling. Beyond perception, accurate prediction requires using a model of the depicted physical system. Models must apply the supplied physical principles to state variables and distinguish consistent outcomes from plausible distractors. As shown in Fig. 2(2), this includes dynamics, conservation laws, fields, and thermodynamic processes. PhysCot^T probes physical world modeling by adding textual state descriptions to candidate options, reducing candidate-side visual ambiguity and testing whether models can apply the provided guidance to the observed process.

3.2 Task Definition

Given four observed context frames, models select the hidden fifth frame from four visually similar but physically distinct image candidates. Each instance includes a short setting description, four context images, and four answer options. The correct option comes from the underlying simulation, while distractors perturb physical quantities, forcing prediction from temporally coherent visual evidence rather than textual answer matching.

Formally, states evolve as $s_{t+1} = F^c_{\Delta t}(s_t; \theta)$ and render as $x_t = R_c(s_t)$. Distractors are either admissible but transition-inconsistent or inadmissible under subfield constraints; Appendix A.1 gives the full definitions. Thus, an incorrect option need not be an impossible state, only an incorrect continuation.

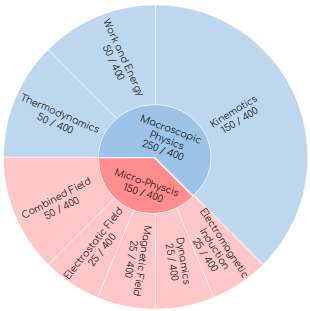


Figure 3: **Subfield distribution.** 400 questions across 8 physics subfields.

As shown in Fig. 3, PhysCot contains 400 questions spanning 8 subfields: kinematics, dynamics, work and energy, thermodynamics, electrostatic fields, magnetic fields, combined fields, and electromagnetic induction. Difficulty is stratified by distractor magnitude: larger perturbations are easier, while smaller ones demand finer perception and physical modeling.

3.3 Benchmark Construction Pipeline

As summarized in Fig. 4, all problems are procedurally generated through a four-stage pipeline that preserves the correctness of the reference trajectory, visual consistency, and traceability. First, NumPy simulations compute ground-truth trajectories and state variables for mechanics, electromagnetism, and thermodynamics. Second, Manim renders physical states as 2D scientific visualizations, with Matplotlib used for thermodynamic plots and gauges. Third, answer candidates pair the correct next frame with transition-inconsistent or inadmissible distractors produced by controlled perturbations to physical quantities such as position, velocity, momentum, energy, charge-to-mass ratio, field direction, or thermodynamic variables. Finally, simulation-state metadata is attached to each image, and trivial or ambiguous samples are filtered (see Appendix A.1). The current evaluation uses these fixed simulator–renderer families throughout and includes no designated generator-shift, rendering-shift, corruption, compositional-system, or real-image test set.

3.4 Evaluation Protocols and Diagnostic Capabilities

PhysCot. The full protocol presents four temporally ordered context frames and four image candidates, without explicit textual candidate-state descriptions. Under the shared physics scaffold,

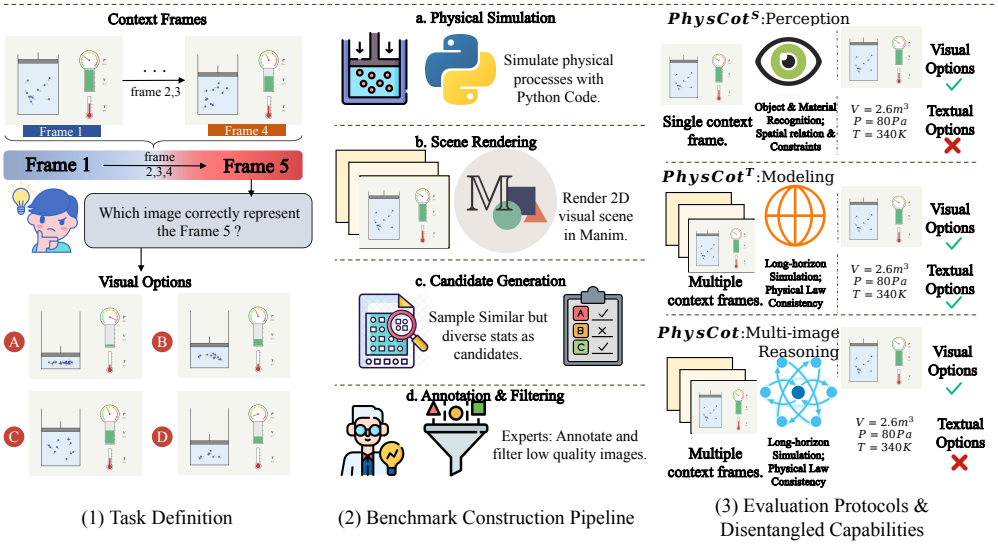


Figure 4: **PhysCot framework overview.** (1) The task is formulated as next-state prediction from temporally ordered visual observations. (2) Each instance is generated through simulation, rendering, candidate generation, and filtering. (3) Three diagnostic protocols compare full multi-image physical reasoning with perception-focused and text-augmented conditions.

models must recover physical state from images, connect the ordered context using the supplied principles, and select the consistent rendered future state.

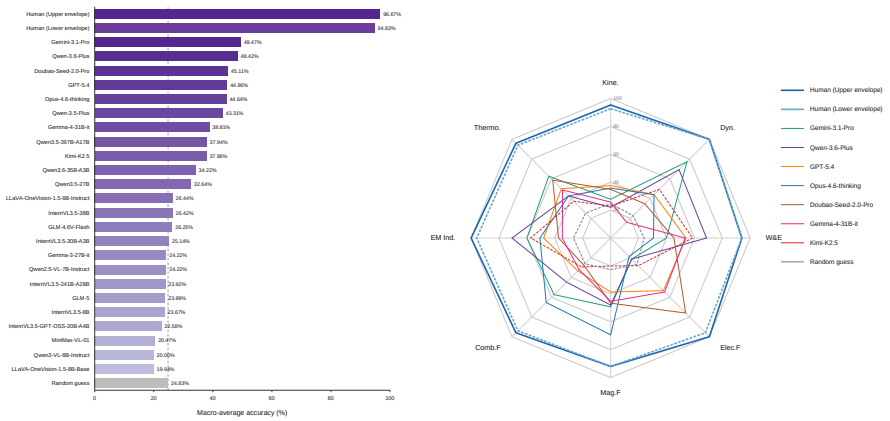
PhysCot^S. The perception-focused protocol presents a single context frame paired with visually similar candidate images and no candidate state descriptions. By removing intermediate context frames, it emphasizes fine-grained visual state perception: models must identify physical quantities from the image and distinguish subtle but physically meaningful candidate differences. Because the target remains a future state, this condition also retains a physical-prediction component.

PhysCot^T. The modeling protocol preserves the four-frame context but augments candidate options with textual descriptions of their underlying physical states. This reduces the need to identify state variables from candidate images and probes whether models can apply physical principles to select the correct future state. Comparing PhysCot, PhysCot^S, and PhysCot^T provides diagnostic evidence about perception, physical world modeling, and temporal integration, but does not by itself establish a causal decomposition of these capabilities (see Appendices A.3–A.4).

4 Experiment Results

4.1 Evaluation Setup

We evaluate proprietary and open-source MLLMs used for multimodal reasoning and vision-language tasks. Proprietary models include Claude-Opus-4.6-thinking (Anthropic, 2025), GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro (Google DeepMind, 2026a), Qwen-3.6-Plus (Qwen Team, 2026d), Qwen-3.5-Plus (Qwen Team, 2026b), GLM-5 (Z.ai, 2026), and Doubao-Seed-2.0-Pro (ByteDance Seed, 2026). Open-source models include Kimi-K2.5 (Moonshot AI, 2026), MiniMax-VL-01 (MiniMax, 2025), Gemma-series models (Gemma Team, 2025; Google DeepMind, 2026b), Qwen-VL-series models (Yang et al., 2024; 2025; Qwen Team, 2026a;c), InternVL3.5-series models (Wang et al., 2025), GLM-4.6V-Flash (Z.ai, 2025), and LLaVA-OneVision-1.5 models (An et al., 2025). All use the same zero-shot physics scaffold; results therefore do not test autonomous law discovery. Each model and protocol receives three fixed evaluation passes. For a subfield with n questions, we report aggregated accuracy $100k/(3n)$; Avg. is the



(a) Overall results on **PhysCot**. **(b)** Per-subfield results on **PhysCot**.

Figure 5: **PHYSCOT results.** (a) Overall accuracy of proprietary and open-source MLLMs, a random-guess baseline, and two legacy aggregate human envelopes whose participant-level source records are unavailable. (b) Raw per-subfield accuracies for selected MLLMs, the random baseline, and the same legacy human envelopes.

unweighted mean over eight subfields. Missing responses are incorrect, and repeats are consistency checks rather than independent trials. The supplementary artifact retains selected first-pass records for six models under the three formal protocols; these files are not repeat averages and do not reconstruct every three-pass main-table result (Appendix A.3). The displayed random baseline is an empirical 24.83% realization of a theoretical 25% chance rate; shortcut and external-validity tests remain unmeasured (see Appendices A.3–A.7).

4.2 MAIN RESULTS ON PHYSCOT

Table 3 and Fig. 5 report full-protocol results. All evaluated models remain below 50% macro-average accuracy: Gemini-3.1-Pro reaches 49.47%, and the strongest open model, Gemma-4-31B-it, reaches 38.81%. The table also reproduces two legacy aggregate human envelopes for descriptive scale only; the current artifact does not retain participant-level records needed to verify their provenance or support statistical comparison (Appendix A.2). The model results therefore provide the primary empirical evidence. The end-to-end protocol is difficult within this synthetic distribution, although current baselines do not isolate physical reasoning from shortcuts.

4.3 DIAGNOSTIC ANALYSES

Diagnostic error analysis. We report legacy aggregate counts for overlapping perception, reasoning, and temporal-integration labels on one designated GPT-5.4 pass (204 errors among 400 questions; 49.0% micro accuracy). Item-level labels, intersections, annotator metadata, and agreement records are unavailable in the current artifact, so these rates are provisional marginals rather than a reproducible, mutually exclusive, or causal decomposition. They are distinct from the 44.86% GPT-5.4 macro-average in Table 3, which aggregates three passes within each subfield and then weights subfields equally. Appendix A.5 specifies a prospective reannotation protocol.

The higher perception and temporal-integration marginals identify these as the dominant observed failure factors for GPT-5.4 in the designated pass and provide evidence that visual state extraction and context integration are major practical bottlenecks under the full PhysCot protocol. Because the labels overlap and item-level assignments and reliability records are unavailable, the reported percentages should not be interpreted as mutually exclusive shares, population prevalence estimates, or a causal decomposition.

Distractor type and difficulty. The legacy panel-specific aggregates in Table 4(b) decrease from 40.5% on Easy questions to 36.8% on Medium and 27.9% on Hard questions. Position-offset dis-

Table 3: Benchmark results on the full **PhysCot** protocol. Within each model group, bold marks every tied best score and underlining marks every tied second-highest distinct score. Human rows reproduce legacy per-subfield extrema and are descriptive only because participant-level source records are unavailable. Avg. is the arithmetic mean over eight subfields.

Categories	Classical Mechanics			Electromagnetism				Thermo.	Avg.↑
LMM	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (LOWER ENVELOPE)	92.67%	100.00%	94.00%	96.00%	92.00%	94.00%	96.00%	94.00%	94.83%
HUMAN (UPPER ENVELOPE)	95.33%	100.00%	94.00%	100.00%	92.00%	96.00%	100.00%	96.00%	96.67%
Proprietary MLLMs:									
OPUS-4.6-THINKING	35.78%	44.00%	30.67%	18.67%	**69.33%**	**65.33%**	50.67%	42.67%	44.64%
GPT-5.4	**37.56%**	44.00%	54.00%	53.33%	38.67%	33.33%	48.00%	50.00%	44.86%
GEMINI-3.1-PRO	27.78%	**77.33%**	40.00%	21.33%	49.33%	57.33%	60.00%	**62.67%**	**49.47%**
QWEN-3.6-PLUS	22.00%	69.33%	**68.67%**	21.33%	48.00%	44.67%	**70.67%**	42.67%	48.42%
QWEN-3.5-PLUS	16.44%	68.00%	60.67%	10.67%	36.00%	56.00%	48.00%	50.67%	43.31%
GLM-5	33.78%	16.00%	20.67%	21.33%	40.00%	26.00%	18.67%	14.67%	23.89%
DOUBAO-SEED-2.0-PRO	34.89%	34.67%	45.33%	**76.00%**	46.67%	27.33%	37.33%	58.67%	45.11%
Open-source MLLMs:									
KIMI-K2.5	22.89%	49.33%	58.67%	28.00%	20.00%	29.33%	**57.33%**	37.33%	37.86%
MINIMAX-VL-01	9.78%	28.00%	23.33%	22.67%	10.67%	22.67%	20.00%	26.67%	20.47%
GEMMA-3-27B-IT	**33.11%**	33.33%	42.00%	24.00%	16.00%	16.00%	10.67%	18.67%	24.22%
GEMMA-4-31B-IT	25.78%	16.00%	53.33%	**54.67%**	45.33%	32.00%	34.67%	48.67%	**38.81%**
GLM-4.6V-FLASH	24.00%	33.33%	21.33%	36.00%	26.67%	17.33%	20.00%	31.33%	26.25%
Qwen2.5-VL-7B-Instruct	32.44%	22.67%	20.00%	28.00%	16.00%	32.00%	26.67%	16.00%	24.22%
Qwen3-VL-8B-Instruct	9.33%	33.33%	20.67%	18.67%	13.33%	**44.00%**	10.67%	10.00%	20.00%
QWEN3.5-27B	20.44%	57.33%	41.33%	9.33%	22.67%	32.00%	32.00%	46.00%	32.64%
QWEN3.5-397B-A17B	24.89%	**64.00%**	**68.00%**	8.00%	28.00%	42.00%	16.00%	**52.67%**	37.94%
QWEN3.6-35B-A3B	23.11%	52.00%	44.00%	4.00%	29.33%	43.33%	34.67%	43.33%	34.22%
INTERNVL3.5-8B	19.33%	36.00%	26.00%	29.33%	17.33%	15.33%	20.00%	26.00%	23.67%
INTERNVL3.5-GPT-OSS-20B-A4B	30.67%	9.33%	22.00%	33.33%	13.33%	25.33%	24.00%	22.67%	22.58%
INTERNVL3.5-30B-A3B	14.44%	32.00%	15.33%	44.00%	16.00%	35.33%	18.67%	25.33%	25.14%
INTERNVL3.5-38B	25.33%	14.67%	34.00%	36.00%	33.33%	26.00%	24.00%	18.00%	26.42%
INTERNVL3.5-241B-A28B	19.33%	20.00%	36.00%	29.33%	13.33%	30.67%	22.67%	20.00%	23.92%
LLAVA-ONEVISION-1.5-8B-BASE	25.56%	13.33%	16.00%	28.00%	**48.00%**	24.67%	1.33%	2.67%	19.94%
LLAVA-ONEVISION-1.5-8B-INSTRUCT	30.22%	28.00%	26.00%	10.67%	20.00%	15.33%	52.00%	29.33%	26.44%
Random guess	24.67%	22.67%	24.00%	26.67%	22.67%	26.00%	26.67%	25.33%	24.83%

Table 4: **Compact GPT-5.4 diagnostic analyses.** (a) Exploratory overlapping error-label marginals from one designated pass; (b) legacy panel-specific accuracy (%) by distractor type and difficulty on randomly sampled diagnostic questions; (c) context-input comparison, with Δ measured relative to ordered four-frame input.

(a) Error Factors

Factor	GPT-5.4 pass
Perception	152/204 (74.5%)
Reasoning	90/204 (44.1%)
Temporal	124/204 (60.8%)

(b) Distractor Type and Difficulty

Distractor	Easy	Med.	Hard	Avg.
Position offset	44.6	36.6	24.5	34.1
Momentum violation	55.6	50.0	35.8	38.7
Energy violation	35.1	35.1	–	35.1
Thermo. perturbation	36.0	–	–	36.0
Overall	40.5	36.8	27.9	35.0

(c) Context-Input Comparison

Setting	Acc.	Δ
4 frames, ordered	44.9	–
4 frames, shuffled	41.5	-3.4
Last frame only	37.8	-7.1
First frame only	**50.0**	+5.1

tractors show the largest decline, while momentum violations have the highest reported average. Because the questions were randomly sampled and the retained artifact does not include the original per-cell sampling record, these values support a descriptive difficulty trend rather than a population-level estimate.

Context-input comparison. Table 4(c) shows that shuffled four-frame and last-frame-only inputs underperform ordered four-frame input, whereas the confirmed first-frame-only PhysCot^S result reaches 50.0%. This pattern indicates sensitivity to how visual context is presented, but it does not isolate temporal-order effects because the retained aggregate summaries do not share fully recoverable accounting details.

4.4 DIAGNOSTIC PROTOCOL RESULTS

PhysCot^T: text-visible candidate states. PhysCot^T preserves the four-frame context and adds candidate-state text. An artifact audit finds four distinct descriptions for 350 items, while 50 Kinematics items contain identical text for all four candidates and thus receive no candidate-side textual disambiguation. PhysCot^T is therefore a mixed text-augmented condition rather than a uniform state-text oracle (Appendix A.1).

Table 5: Benchmark results under **PhysCot**^{T}, the multi-image text-augmented condition. Candidate descriptions are distinct for 350 items and identical within each of 50 Kinematics items. Bold marks all tied best scores and underlining all tied second-highest distinct scores within each model group.

Categories	**Classical Mechanics**			**Electromagnetism**				**Thermo.**	*Avg.*↑
LMM	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
Proprietary MLLMs:									
OPUS-4.6-THINKING	**79.11%**	98.67%	76.67%	**98.67%**	**65.33%**	53.33%	86.67%	56.67%	76.89%
GPT-5.4	65.56%	96.00%	74.00%	69.33%	50.67%	**55.33%**	52.00%	58.00%	65.11%
GEMINI-3.1-PRO	73.78%	**100.00%**	**94.00%**	97.33%	57.33%	47.33%	92.00%	**66.67%**	**78.56%**
QWEN-3.6-PLUS	53.33%	85.33%	76.67%	57.33%	28.00%	34.67%	74.67%	**66.67%**	59.58%
DOUBAO-SEED-2.0-PRO	51.11%	64.00%	48.67%	78.67%	52.00%	44.00%	**96.00%**	48.67%	60.39%
Open-source MLLMs:									
KIMI-K2.5	**65.78%**	68.00%	62.67%	**93.33%**	8.00%	10.00%	**94.67%**	**66.67%**	58.64%
MINIMAX-VL-01	62.89%	**90.67%**	55.33%	40.00%	18.67%	29.33%	57.33%	46.67%	50.11%
GEMMA-3-27B-IT	64.89%	81.33%	60.67%	28.00%	52.00%	21.33%	17.33%	23.33%	43.61%
GEMMA-4-31B-IT	54.44%	88.00%	**73.33%**	50.67%	**62.67%**	**38.67%**	52.00%	58.67%	**59.81%**
QWEN3.6-35B-A3B	54.00%	53.33%	66.00%	44.00%	32.00%	28.00%	90.67%	50.67%	52.33%
Random guess	24.67%	22.67%	24.00%	26.67%	22.67%	26.00%	26.67%	25.33%	24.83%

Table 6: Benchmark results under **PhysCot**^{S}, the single-image text-invisible condition that retains the first context frame. This is the sole tabular report of the first-frame-only experiment; candidate answers remain image-only. Bold marks all tied best scores and underlining all tied second-highest distinct scores within each model group.

Categories	**Classical Mechanics**			**Electromagnetism**				**Thermo.**	*Avg.*↑
LMM	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
Proprietary MLLMs:									
OPUS-4.6-THINKING	**36.89%**	33.33%	33.33%	9.33%	72.00%	45.33%	34.67%	44.00%	38.61%
GPT-5.4	36.00%	62.67%	**63.33%**	**58.67%**	40.00%	39.33%	50.67%	**49.33%**	50.00%
GEMINI-3.1-PRO	23.11%	**76.00%**	56.67%	6.67%	**81.33%**	**73.33%**	**54.67%**	42.67%	**51.81%**
DOUBAO-SEED-2.0-PRO	**36.89%**	48.00%	28.00%	49.33%	54.67%	45.33%	28.00%	24.00%	39.28%
QWEN-3.6-PLUS	26.44%	61.33%	42.00%	21.33%	17.33%	37.33%	40.00%	20.67%	33.31%
Open-source MLLMs:									
KIMI-K2.5	22.89%	17.33%	32.67%	29.33%	28.00%	10.67%	16.00%	18.67%	21.94%
MINIMAX-VL-01	11.78%	16.00%	16.67%	10.67%	12.00%	22.00%	**44.00%**	22.67%	19.47%
GEMMA-3-27B-IT	32.44%	20.00%	34.67%	25.33%	28.00%	14.00%	10.67%	18.67%	22.97%
GEMMA-4-31B-IT	**39.56%**	34.67%	**47.33%**	**56.00%**	**50.67%**	**35.33%**	13.33%	**34.67%**	**38.94%**
QWEN3.6-35B-A3B	26.22%	**57.33%**	42.67%	8.00%	40.00%	**35.33%**	29.33%	21.33%	32.53%
Random guess	24.67%	22.67%	24.00%	26.67%	22.67%	26.00%	26.67%	25.33%	24.83%

Table 5 shows higher aggregate performance for all evaluated models in this mixed text-augmented condition. Proprietary models exceed 59% and open-source models exceed 40%, with Gemini-3.1-Pro reaching 78.56%. The gains show that models can benefit when discriminative state text is available, but they do not estimate a uniform text effect because 50 items receive identical candidate descriptions.

PhysCot^S: single-frame image-only input. PhysCot^S is the first-frame-only, image-only condition referenced above, not a separate experiment. It emphasizes fine-grained visual state cues but still requires future-state prediction from limited evidence.

Across models, removing context has non-uniform effects; because it changes both frame number and content, Full-vs-S measures additional-context utility rather than temporal order alone.

5 CONCLUSION

We introduced PHYSCOT, a controlled benchmark for next-state prediction across eight physics subfields. On the fixed synthetic distribution, no model exceeds 50% full-protocol accuracy, while mixed text augmentation reaches 78.56%. Together with the GPT-5.4 error-label marginals, this gap identifies visual state extraction and context integration as major practical bottlenecks: current MLLMs use explicit physical information more effectively than they recover it from images. These conclusions are diagnostic rather than causal because labels overlap and supporting audits remain incomplete; fully logged reruns and held-out synthetic and real-image tests are still necessary.

AI USE STATEMENT

In this work, we used generative AI tools to support method implementation, protocol and experimental-design review, dataset cleaning and reformatting, numerical auditing, software code creation or editing, and preparation or revision of tables and figures. We also used generative AI tools to draft and translate parts of the manuscript, brainstorm research framing, identify gaps, search for information and relevant literature, and edit the paper for readability.

We did not use generative AI tools to execute model or human evaluations, determine ground-truth answers, or autonomously accept scientific claims or citations. Other disclosure categories, including proving mathematical claims, assisting in the writing of proofs, supporting qualitative or thematic data analysis, and transcribing research materials, are not applicable to this work.

All AI-assisted outputs were reviewed by the authors. Code and implementation-related suggestions were manually inspected for consistency with the stated protocols; no new model or human evaluation was executed as part of this manuscript revision. Numerical and protocol audits were checked against the retained data, logs, scripts, and displayed tables. AI-assisted writing and translation were revised for factual accuracy, clarity, and consistency with the experimental results, and literature-related suggestions were checked against the cited papers before inclusion. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICAL CONSIDERATIONS

PhysCot is an evaluation benchmark designed strictly for academic research on multi-image physical reasoning. The dataset is synthetically generated from physics simulations and contains no personal data, offensive content, or sensitive information, posing no direct privacy or safety risks. We foresee no clear path to malicious use, as the benchmark assesses model capabilities rather than enabling harmful applications. However, we acknowledge two indirect risks: (1) overreliance on benchmark scores as a proxy for genuine scientific competence, which may lead to misleading conclusions about model readiness for deployment in high-stakes scientific contexts; and (2) the possibility that future models optimized specifically for PhysCot may overfit to its simulation-based distribution, limiting generalizability to real-world physical reasoning. We encourage the community to treat PhysCot as one diagnostic tool among many, rather than a definitive measure of scientific reasoning capability.

All data in PHYSCOT are synthetically generated from physics simulations and do not involve collection of real-world text, images, or user-generated content. The released fields and rendered scenes depict abstract physical simulations, such as projectile trajectories, collision diagrams, and electromagnetic field configurations; they contain no intended personal, demographic, biometric, offensive, or harmful content. No participant data are part of the benchmark itself, and no anonymization procedure is applied to the synthetic benchmark. The benchmark is intended solely for academic evaluation of multi-image physical reasoning capabilities.

The manuscript reproduces two legacy aggregate human-reference envelopes for descriptive scale only. The current artifact contains no participant-level answers, response-time logs, recruitment records, consent documentation, compensation records, or IRB approval/exemption determination. We therefore make no claims about participant recruitment, consent, compensation, review status, response-time filtering, or inter-annotator agreement, and we do not interpret the aggregate envelopes as individual scores or a population estimate. No participant-level data or direct identifiers are included in the manuscript or supplementary artifact. Appendix A.2 documents the available aggregate values and their reporting limits.

REPRODUCIBILITY STATEMENT

We provide multiple resources to support reproducibility of PHYSCOT and the reported model evaluations. The main paper describes the benchmark construction procedure, task formulation, evaluation metrics, model set, and prompting protocols. Appendix A.1 provides implementation-level details for the synthetic physics-image generation pipeline, Appendix A.3 documents the model evaluation protocols and prompt formats, and Appendix A.7 reports the experimental environment and execution settings. The supplementary artifact releases the complete set of materials that re-

main available: the 400-item benchmark JSON, rendered images, three protocol-aligned evaluation scripts, and 16 retained single-pass JSON logs across six model folders. Appendix A.3 documents their exact coverage and the two excluded malformed or mislabeled files. These logs provide item-level evidence for the covered single passes but do not reconstruct all three passes behind every main-table entry. Because the remaining original pass records were not retained, exact pass-level reconstruction and uncertainty estimates are unavailable; the paper limits its claims accordingly and reports this coverage explicitly.

REFERENCES

2025 An et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. URL https://huggingface.co/papers/2509.23661.

Anonymous. Mv-physics: Benchmark for evaluating multimodal large language models in physical visual scenes, 2026. URL https://openreview.net/forum?id=rjMvoex9jb. Anonymous ACL submission.

Anthropic. Introducing Claude Opus 4.6, 2025. URL https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026-04-21.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, and Chenliang Xu. Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. *arXiv preprint arXiv:2503.11557*, 2025. URL https://arxiv.org/abs/2503.11557.

ByteDance Seed. Doubao-Seed-2.0: Official launch, 2026. URL https://seed.bytedance.com/en/blog/seed2-0-%E6%AD%A3%E5%BC%8F%E5%8F%91%E5%B8%83. Accessed: 2026-04-21.

Qiguang Chen, Chengyu Luan, Jiajun Wu, Qiming Yu, Yi Yang, Yizhuo Li, Jingqi Tong, Xiachong Feng, Libo Qin, and Wanxiang Che. Omibench: Benchmarking olympiad-level multi-image reasoning in large vision-language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 45100–45135, 2026. URL https://arxiv.org/abs/2604.20806.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025a. doi: 10.1609/aaai.v39i22.34538. URL https://arxiv.org/abs/2412.12932.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, et al. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*, 2025b. URL https://arxiv.org/abs/2506.04280.

Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and Xuming Hu. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*, 2025. URL https://arxiv.org/abs/2505.15472.

Zhuobai Dong, Junchao Yi, Ziyuan Zheng, Haochen Han, Xiangxi Zheng, Alex Jinpeng Wang, Fangming Liu, and Linjie Li. Seeing is not reasoning: Mvpbench for graph-based evaluation of multi-path visual physical cot. *arXiv preprint arXiv:2505.24182*, 2025. URL https://arxiv.org/abs/2505.24182.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://huggingface.co/papers/2503.19786.

Google DeepMind. Gemini 3.1 Pro: A smarter model for your most complex tasks, 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/. Accessed: 2026-04-21.

Google DeepMind. Gemma 4: Byte for byte, the most capable open models, 2026b. URL https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/. Accessed: 2026-04-21.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36: 62991–63010, 2023.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. URL https://arxiv.org/pdf/2502.09621.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, and Ahmed Qureshi. Remi: A dataset for reasoning with multiple images. *arXiv preprint arXiv:2406.09175*, 2024. URL https://arxiv.org/abs/2406.09175.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016. URL https://arxiv.org/pdf/1603.07396.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. URL https://arxiv.org/pdf/2205.11916.

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. Rxnbench: A multimodal benchmark for evaluating large language models on chemical reaction understanding from scientific literature. *arXiv preprint arXiv:2512.23565*, 2025a.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. URL https://arxiv.org/abs/2306.09212.

Junxian Li et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025b. URL https://arxiv.org/pdf/2408.07246.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. URL https://arxiv.org/abs/2209.09513.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. URL https://arxiv.org/abs/2310.02255.

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

MiniMax. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025. URL https://huggingface.co/papers/2501.08313.

Moonshot AI. Kimi K2.5. *arXiv preprint arXiv:2602.02276*, 2026. URL https://huggingface.co/papers/2602.02276.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/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, et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Qwen Team. Qwen3.5-27B, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-27B. Accessed: 2026-04-22.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026b. URL https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-04-21.

Qwen Team. Qwen3.6-35B-A3B, 2026c. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed: 2026-04-22.

Qwen Team. Qwen3.6-Plus: Towards real world agents, 2026d. URL https://qwen.ai/blog?id=qwen3.6. Accessed: 2026-04-21.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research, 2024. URL https://arxiv.org/abs/2308.13149.

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. The ever-evolving science exam, 2025a. URL https://arxiv.org/abs/2507.16514.

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

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. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*, 2025c. URL https://arxiv.org/abs/2506.17667.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

2025 Wang et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://huggingface.co/papers/2508.18265.

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

Xuecheng Wu, Jiaxing Liu, Danlei Huang, Xiaoyu Li, Yifan Wang, Chen Chen, Liya Ma, Xuezhi Cao, and Junxiao Xue. Vic-bench: Benchmarking visual-interleaved chain-of-thought capability in mllms with free-style intermediate state representations. *arXiv preprint arXiv:2505.14404*, 2025. URL https://arxiv.org/abs/2505.14404.

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. Seephys: Does seeing help thinking? benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*, 2025. URL https://arxiv.org/abs/2505.19099.

2024 Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. URL https://huggingface.co/papers/2407.10671.

2025 Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://huggingface.co/papers/2505.09388.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://arxiv.org/pdf/2311.16502.

Z.ai. GLM-4.6V-Flash, 2025. URL https://docs.z.ai/guides/vlm/glm-4.6v. Accessed: 2026-08-20.

Z.ai. GLM-5: From vibe coding to agentic engineering, 2026. URL https://z.ai/blog/glm-5. Accessed: 2026-08-20.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*, 2026. URL https://arxiv.org/abs/2603.19607.

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

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024. URL https://arxiv.org/abs/2402.06119.

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 subdirectory 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 $\Delta t{=}5{\times}10^{-3}$ 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).

```
import numpy as np

g = 9.8
```

```python
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`.

```
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 controlled physical axes and re-rendering through the same pipeline.

Distractor taxonomy. For subfield c, let $s_t \in \mathcal{S}_c$ be a state, θ fixed parameters, $x_t = R_c(s_t)$ the rendering, and $s_{t+1} = F^c_{\Delta t}(s_t;\theta)$ the simulator transition. The correct candidate renders $s^\star_4 = F^c_{\Delta t}(s_3;\theta)$, and $\mathcal{A}_c \subseteq \mathcal{S}_c$ denotes states satisfying instantaneous constraints. A *transition-inconsistent* distractor lies in $\mathcal{A}_c$ but differs from $s^\star_4$ beyond a subfield-specific tolerance; an *inadmissible* distractor lies outside $\mathcal{A}_c$. Examples include a feasible position offset or thermodynamic state inconsistent with the observed process, versus a state violating a required momentum or energy relation. The taxonomy describes construction intent rather than a post-hoc proof of realizability. Current metadata does not expose a verified class, predicate, perturbation magnitude, and tolerance for every candidate, so we do not report per-class proportions.

```python
def generate_distractors(final_state, perturbation_axes, n=3):
    """Produce controlled, visually similar candidate perturbations."""
    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: Controlled distractor generation (abridged).

Stage 4: Batch Orchestration and Filtering

Scenes are generated in batches by per-subfield orchestrators (`batch_generate_scenes_*.py`). The intended validation workflow separates ground-truth physical-consistency checks, visual-readability checks, and independent human review of complete items. However, the current artifact does not retain validator outputs, reviewer assignments, or adjudication records sufficient to verify complete execution of that workflow, so these steps are not used as empirical validation evidence here. We do not apply the ground-truth transition predicate to require every distractor to be a feasible continuation, because inadmissible states are an intentional distractor class.

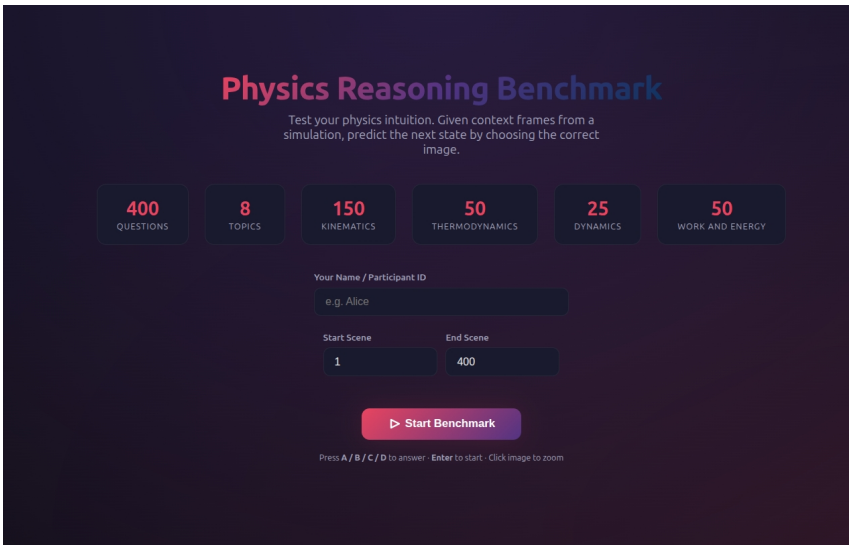


Figure 6: Interface illustration associated with the human-reference evaluation.

Artifact-level candidate audit. A machine-readable audit of all 400 released items confirms four choice IDs and four unique image paths per item, exactly one `is_correct` marker matching `correct_answer`, and complete archive coverage for all 1,600 candidate images. Within every item, the four candidate files also have distinct SHA-256 hashes. However, each choice record stores only `choice_id`, `image_path`, and `is_correct`; it does not store a verified distractor class, candidate state, violated predicate, perturbation magnitude and units, tolerance, transition residual, or validator outcome. The auxiliary candidate-text map contains four distinct descriptions for 350 items, but all four descriptions are identical in 50 Kinematics items. Thus the current artifact supports structural and image-uniqueness checks, not a candidate-level audit of physical admissibility or transition consistency. A future release should store the missing fields and validator version for every candidate before reporting class-specific proportions.

A.2 Human Reference Results

The full-protocol result table includes two aggregate human-reference envelopes, reported as descriptive context for the model results. Because the analysis uses per-subfield aggregate extrema, these values should be interpreted as reference ranges rather than participant-level means or population estimates. No human values are reported for PhysCot^S or PhysCot^T.

Available interface material. Figures 6 and 7 illustrate the web-interface design associated with the human-reference evaluation. The interface displays four context frames in temporal order and four image candidates labeled `A`–`D`. The screenshots document the presentation format used for this comparison.

Aggregation of human scores. *Human (Upper envelope)* records the highest displayed per-subfield accuracy and *Human (Lower envelope)* the lowest. Because the extrema may come from different individuals, the two rows describe a per-subfield range rather than the performance of a single best or weakest participant. Direct subtraction gives widths of 2.66, 0.00, 0.00, 4.00, 0.00, 2.00, 4.00, and 2.00 percentage points: the range is 0.00–4.00 points, and six of eight subfields are at most 2.67 points wide. Participant-level dispersion and agreement are outside the scope of this aggregate comparison.

Scope of interpretation. The supplementary artifact reports the aggregate reference scores rather than participant-level response records. Accordingly, we use the human results only as descriptive envelopes and do not draw conclusions about individual behavior, population-level performance, or inter-participant agreement.

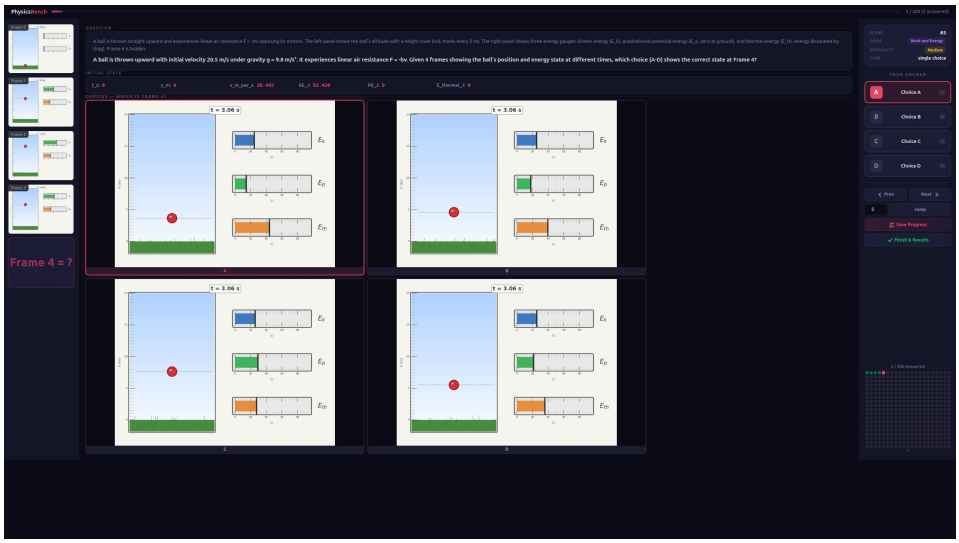


Figure 7: Question-answering interface illustration associated with the human-reference evaluation.

A.3 BENCHMARK EVALUATION PROTOCOLS

PHYSCOT is administered to all evaluated models through three controlled diagnostic 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"]`: stored metadata 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). This field is exposed to the model only in PHYSCOT^T; PHYSCOT and PHYSCOT^S present image-only candidates.

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.

Repeated evaluation and scoring. For every model and protocol, the main tables aggregate three evaluations of the same benchmark questions. For a subfield with n questions, we sum its integer correct counts across the three passes and divide by $3n$; this is equivalent to averaging the three pass-level subfield accuracies. The reported Avg. is then the unweighted arithmetic mean over the eight subfields, so it is a macro-average that assigns equal weight to each subfield despite their different sizes. It must therefore be distinguished from single-pass micro accuracy, which divides the total number of correct answers by all 400 questions. A missing, failed, or unparseable response is counted as incorrect. The retained supplementary JSON files described below each record only one evaluation pass, not the three-pass average reported in the main tables.

Verified benchmark denominators. Direct counting of the released 400-item data file gives $n = (150, 25, 50, 25, 25, 50, 25, 50)$ for Kinematics, Dynamics, Work & Energy, Electrostatic Field, Magnetic Field, Combined Field, Electromagnetic Induction, and Thermodynamics, respectively. The corresponding three-pass denominators are $3n = (450, 75, 150, 75, 75, 150, 75, 150)$. Difficulty counts are 130 Easy, 135 Medium, and 135 Hard, and correct-answer positions are balanced

with 100 each for A–D. All 49 displayed result rows across the three main tables—47 model/random rows and two legacy human rows—lie on their applicable integer-count grids, and every displayed Avg. matches the unweighted mean of its eight subfield entries within rounding tolerance. This verifies arithmetic consistency. The supplementary release includes all item-level execution records that remain available. Original records for the remaining passes and model–protocol pairs were not retained and cannot be reconstructed from the current materials, so pass-level variance and confidence intervals are unavailable.

Audit of retained single-pass logs. The supplementary directory retains 16 JSON files across six model folders and three protocol scripts. Each JSON file is one 400-question evaluation pass and must not be read as an average over repeated runs. The scripts now mirror the formal definitions in this appendix: `reasoning` uses all four context frames in temporal order with image-only candidates (PhysCot), `perception` uses only the first context frame with image-only candidates (PhysCot^S), and `modeling` uses all four context frames in temporal order with candidate-state text (PhysCot^T). We recomputed the displayed single-pass macro-averages from the integer per-topic counts stored in each retained file's `statistics` object.

Table 7: Audit of retained results. Each column reports the eight-subfield macro-average from one evaluation pass; bracketed values are saved records with a non-empty execution-error field, all scored incorrect. Main-table values are separately reported three-pass averages.

Model folder	**PhysCot** *one pass*	**PhysCot^S** *one pass*	**PhysCot^T** *one pass*
Opus-4.6-thinking	45.08 [0]	38.83 [12]	70.50 [0]
GPT-5.4	–[†]	50.33 [1]	65.33 [1]
Kimi-K2.5	38.00 [56]	21.83 [86]	58.25 [50]
MiniMax-VL-01	20.25 [0]	19.67 [0]	50.67 [0]
Gemma-3-27B-it	–[‡]	22.58 [0]	43.75 [0]
Gemma-4-31B-it	38.83 [1]	39.50 [0]	59.58 [0]

These retained values constitute the complete available first-pass provenance. Most covered model–protocol pairs are close to the corresponding three-pass table entries, although Opus-4.6-thinking under PhysCot^T differs by 6.39 percentage points (70.50% versus 76.89%). This bounded consistency check cannot recover unretained passes or provide pass-level uncertainty. We therefore do not pool single-pass records with the main-table averages. Comparisons among the three formal PhysCot protocols use three-pass macro-averages; the distractor analysis in Table 4(b) and the shuffled/last-frame entries in Table 4(c) are separately labeled legacy panel-specific diagnostics and are not used to recompute or validate the main-table results.

[†]The supplied GPT-5.4 PhysCot file had only 342 item records despite aggregate metadata claiming 400 questions, so it was removed rather than presented as a raw pass. [‡]The supplied Gemma-3 PhysCot file identified its internal model as `glm-4.6v`, so it was also removed. These deletions leave the corresponding main-table cells without retained pass-level provenance; the three-pass aggregate values are unchanged.

Protocol summary. Table 8 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):
```

Table 8: The three PHYSCOT evaluation protocols.

Protocol	Context	Candidate	Primary Diagnostic Focus
PHYSCOT	4 frames	image only	Multi-image reasoning
PHYSCOT^S	1 frame	image only	Perception-focused
PHYSCOT^T	4 frames	image + text	Text-augmented physical modeling

```
    """PhysCot (full): multi-image context + image-only candidates."""
    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({"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-focused). This protocol supplies only the first context frame, substantially reducing temporal information and requiring the model to use fine-grained visual features from a single image. It is the same first-frame-only run referenced in Section 4.3, not a separate evaluation. Candidates remain image-only, so neither symbolic candidate-state descriptions nor multi-frame reasoning are available. The resulting condition emphasizes visual state perception under limited temporal evidence, but because the target is still a future state, its performance also reflects physical prediction from incomplete observations. Accordingly, differences between PHYSCOT^S and the full setting are interpreted as diagnostic effects of additional temporal context rather than as the isolated causal contribution of visual perception.

```
def build_query_physcot_s(self, question):
    """PhysCot^S: single context frame + image-only candidates.
    Perception-focused condition with reduced 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 (Text-augmented Physical Modeling). This protocol supplies all four context frames and candidate-state text. For 350 items, the four descriptions are distinct and can reduce candidate-side visual extraction demands; for 50 Kinematics items, all four candidate descriptions are identical and provide no candidate-side textual disambiguation. PhysCot^T should therefore be interpreted as a mixed text-augmented condition rather than a uniform state-text oracle. It probes whether the model can benefit from the supplied candidate text where it is discriminative while applying the shared physical scaffold to select the consistent outcome.

```python
def build_query_physcot_t(self, question):
    """PhysCot^T: multi-image context + text-augmented candidates.
    Reduces candidate-side visual ambiguity by exposing per-candidate
    physical state descriptions; probes 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 $\{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$ 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. The theoretical random-guess rate is 25%; the displayed empirical random baseline is 24.83%.

Baseline scope. Random guessing calibrates chance performance but does not test whether the benchmark admits shortcuts. The following diagnostic baselines are therefore important but are not included in the current results: *candidate-only*, which removes all context frames to test option-side artifacts; *last-frame visual similarity*, which selects the candidate nearest to the final observed frame under a fixed image metric; *coordinate extrapolation*, which predicts the next simulator state using a simple constant-velocity or constant-acceleration rule; *text-only*, which removes images while retaining the question and scaffold; and a *numeric-state oracle*, which supplies structured context and candidate states to reduce visual extraction demands. Until these controls are measured, PhysCot accuracy should be interpreted as end-to-end performance under the stated protocol, not as proof that low-level matching, textual leakage, or simple trajectory heuristics cannot contribute.

External-validity scope and stress-test plan. The current 400 items are sampled from fixed simulator and renderer families and do not constitute an out-of-distribution evaluation. A future external-validity suite should report each shift separately: (i) held-out physical-parameter ranges and combinations; (ii) camera, layout, palette, resolution, rendering-style, and visual-noise shifts; (iii) independently implemented simulators or renderers; (iv) compositional systems that combine processes from multiple subfields; and (v) expert-verified real scientific image sequences. The first four preserve known simulator states and permit controlled scoring, whereas real imagery introduces annotation and observability differences and should therefore be reported as a separate transfer set rather than pooled with synthetic accuracy. No such result is included here, so present scores support comparisons only within the stated synthetic distribution.

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 9: 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

Shared evaluation instruction. In addition to the subfield-specific blocks below, every model receives the same instruction structure:

```
You are a physics expert analyzing simulation frames.
Physics topic: {topic}.
{scenario-specific instructions}
{question}
{initial-state quantities and key-physics checklist}

Inspect the supplied context frame(s), reason step by step
using the relevant physical principles, compare choices A-D,
and select the candidate consistent with the observed sequence.

Answer format:
ANSWER: [A/B/C/D]
```

```
REASONING: [step-by-step explanation]
```

Listing 9: Shared physics-scaffolded evaluation prompt.

The reported benchmark therefore measures reasoning under explicit physics scaffolding. It does not test whether a model can independently identify or discover the relevant law from an unconstrained prompt. Because a minimal-instruction variant has not yet been evaluated, differences attributable specifically to the checklist, equations, or quantitative initial-state block remain unknown.

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.
```

Table 10: Prospective operational predicates for overlapping diagnostic error factors.

Factor	Operational criterion
Perception (P_i)	r_i explicitly assigns a rendered observable (e.g., position, direction, gauge value, or field orientation) a value inconsistent with M_i or the verified visual record. Candidate choice alone is insufficient.
Reasoning (R_i)	r_i correctly states the observables needed for the decision but applies a law, conservation relation, force direction, state constraint, or transition rule that rejects y_i^* or supports $\hat{y}_i$. If correct visual extraction is not evidenced, R_i remains unresolved.
Temporal integration (T_i)	r_i explicitly reverses or ignores a multi-frame trend/order, or a matched order/context control changes the error while single-frame observables remain correctly reported. Without sequence-specific evidence, T_i is not assigned.

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 accompanies the textual candidates so that physical reasoning is evaluated with explicit candidate-state descriptions; in PhysCot and PhysCot^S it supplies consistent physics scaffolding without revealing the target frame's state variables.

A.5 DIAGNOSTIC ERROR FACTORS AND REANNOTATION PROTOCOL

Section 4.3 reports only legacy aggregate counts for overlapping diagnostic labels. The current artifact does not contain the item-level assignments, label intersections, annotator identities or expertise, decision logs, or agreement records needed to reproduce those counts. We therefore treat them as provisional marginals rather than a validated error decomposition. The protocol below is a prospective specification for reannotation; it is not claimed to have generated the current aggregates.

Evidence and operational predicates. For an incorrect response i, let $E_i = (X_i, M_i, r_i, \hat{y}_i, y_i^*)$ contain the presented frames, verified simulator/render metadata, available model rationale, selected candidate, and correct candidate. Reannotation should evaluate the following binary predicates from recorded evidence; a wrong answer alone is insufficient to trigger any factor.

Primary, secondary, and unresolved labels. The prospective multilabel annotation evaluates P_i, R_i, and T_i independently. For a mutually exclusive primary summary, the deterministic precedence is perception, then temporal integration, then reasoning, reflecting the earliest evidenced failure in the perception–integration–inference chain; every other positive predicate is retained as a secondary factor. Cases with no supported predicate are unresolved. This precedence is a reporting convention, not a causal claim. It has not been applied to the current aggregate counts, which cannot be converted into primary counts without item-level intersections.

Annotation decision process. Annotators first compare explicit visual claims in r_i with M_i, then test sequence-specific claims against frame order and any matched control, and finally test stated

physical inferences against the supplied law and verified transition. They record supporting evidence spans or control outcomes for every positive predicate. When r_i omits the information needed to distinguish perception from reasoning, the relevant predicate is unresolved rather than inferred from the selected candidate.

Representative cases. In a kinematics or dynamics item, selecting a candidate with the wrong position or motion direction is labeled perception-related when the discrepancy is directly visible. Selecting a visually plausible candidate after correctly identifying the trajectory but applying the wrong momentum or energy relation is labeled reasoning-related. In a field item, confusing the curvature direction or field orientation is perception-related when the visual cue is misread, whereas using the wrong Lorentz-force direction after reading the cue is reasoning-related. In a thermodynamics item, misreading a gauge value is perception-related, while selecting a state that violates the ideal-gas relation after identifying the displayed variables is reasoning-related. A failure to compare the direction or rate of change across ordered frames may additionally receive a temporal-integration label.

Required reliability and intersection reporting. No human-agreement statistic is reused for the diagnostic error labels. Reannotation should use at least two independent annotators with graduate-level physics training, preserve pre-adjudication labels, and report raw agreement plus per-predicate and primary-label Cohen's κ before third-annotator adjudication. The release should include item-level evidence, unresolved counts, and the full intersections $N_P, N_R, N_T, N_{PR}, N_{PT}, N_{RT}, N_{PRT}$. None of these records or statistics is presently available; consequently, Table 4(a) remains exploratory and cannot support causal or prevalence claims.

A.6 Benchmark Samples

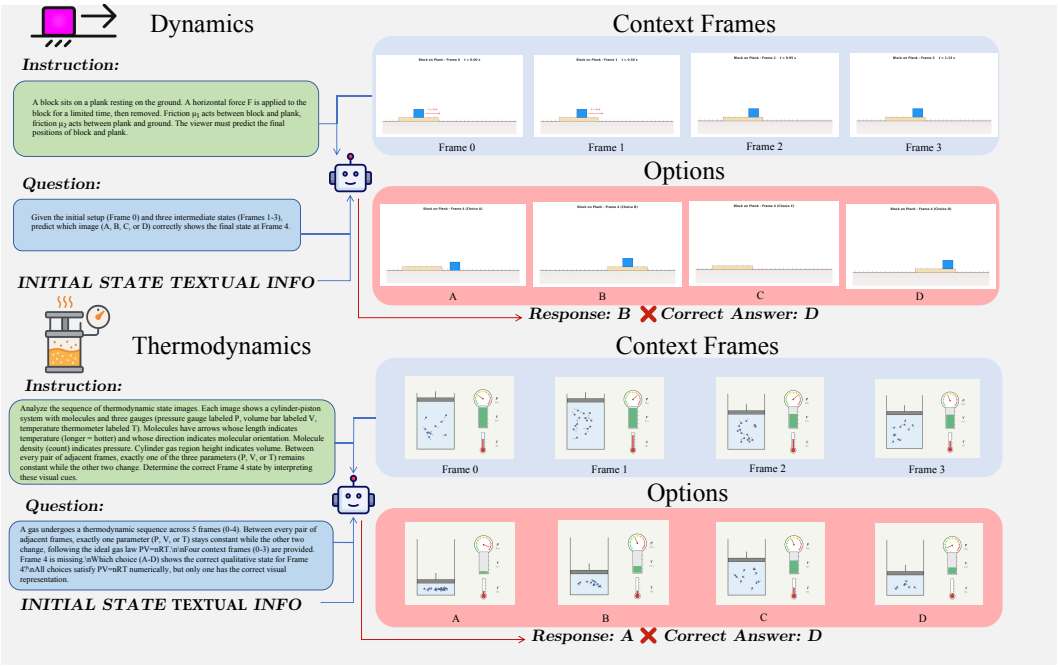


Figure 8: Sample Questions from Dynamics and Thermodynamics

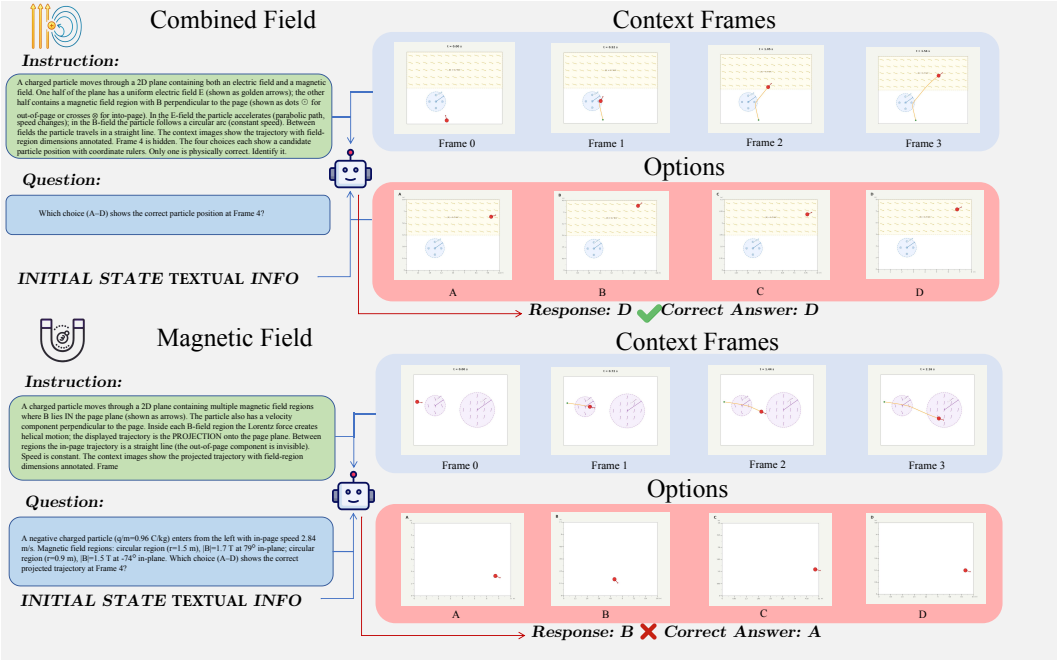


Figure 9: Sample Questions from Combined Field and Magnetic Field

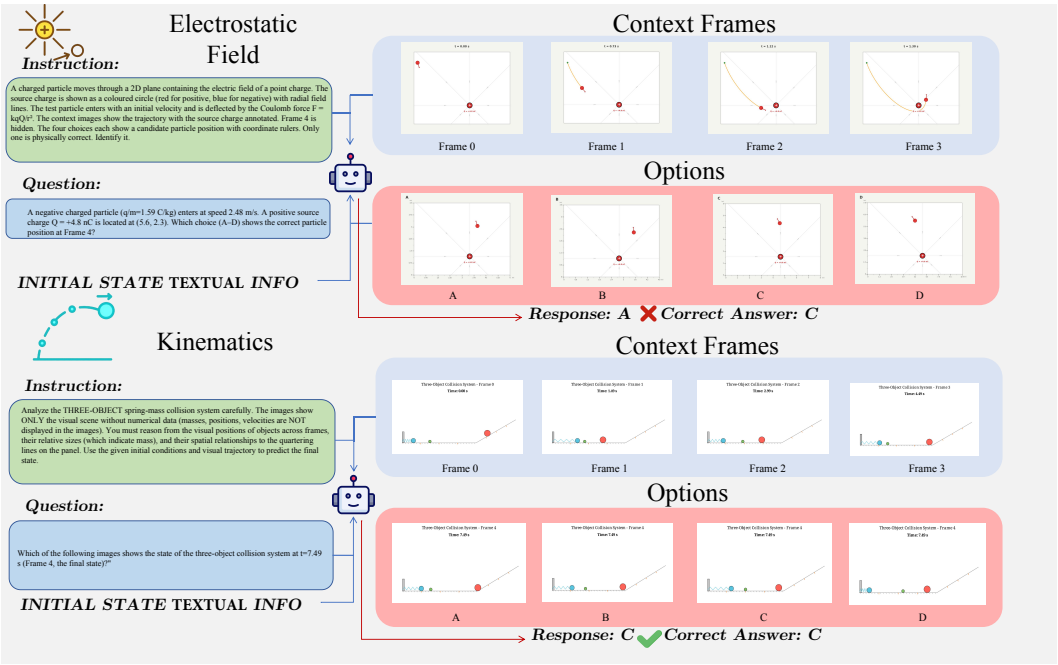


Figure 10: Sample Questions from Electrostatic Field and Kinematics

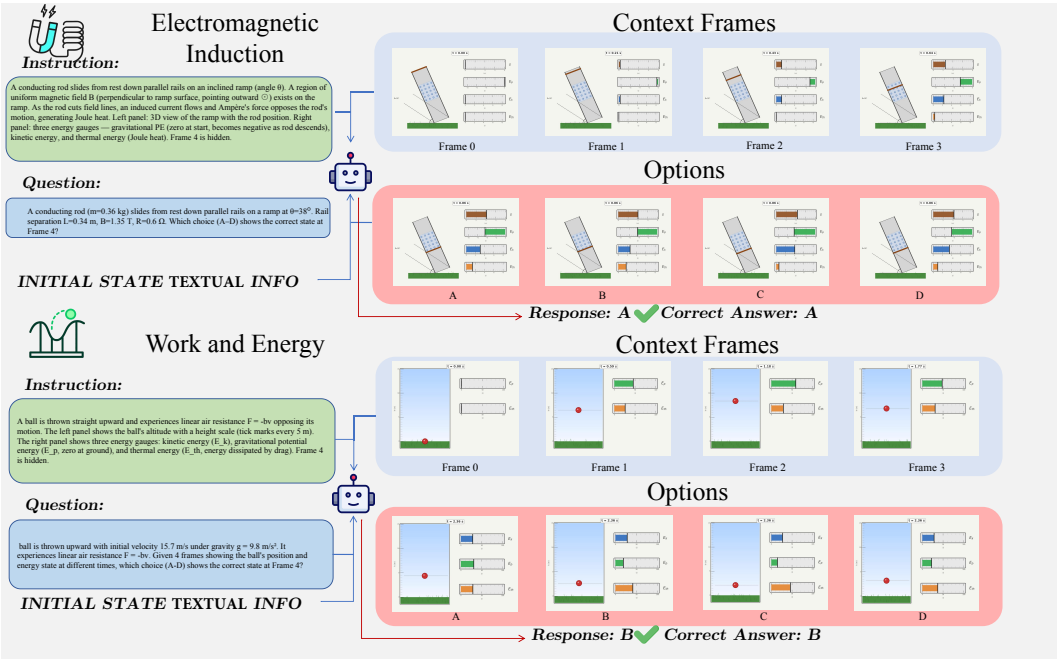


Figure 11: Sample Questions from Electromagnetic Induction and Work & Energy

A.7 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. Models are evaluated with the provider or

repository default decoding settings available at test time. Because hosted defaults and endpoints may change, this choice supports procedural comparability across the reported runs but does not by itself guarantee exact version-level reproduction; the retained outputs and their coverage are reported in Appendix A.3.

Table 11: Full benchmark results on the **PhysCot** protocol. Bold marks all tied best scores and underlining all tied second-highest distinct scores within each model group. Human rows reproduce legacy aggregate extrema for descriptive scale only.

Categories	**Classical Mechanics**			**Electromagnetism**				**Thermo.**	*Avg.*↑
LMM *(LLM)*	*Kine.*↑	*Dyn.*↑	*W&E*↑	*Elec.F*↑	*Mag.F*↑	*Comb.F*↑	*EM Ind.*↑	*Thermo.*↑	
HUMAN (LOWER ENVELOPE)	92.67%	100.00%	94.00%	96.00%	92.00%	94.00%	96.00%	94.00%	94.83%
HUMAN (UPPER ENVELOPE)	95.33%	100.00%	94.00%	100.00%	92.00%	96.00%	100.00%	96.00%	96.67%
Proprietary LMMs:									
OPUS-4.6-THINKING	35.78%	44.00%	30.67%	18.67%	**69.33%**	**65.33%**	50.67%	42.67%	44.64%
GPT-5.4	**37.56%**	44.00%	54.00%	53.33%	38.67%	33.33%	48.00%	50.00%	44.86%
GEMINI-3.1-PRO	27.78%	**77.33%**	40.00%	21.33%	49.33%	57.33%	60.00%	**62.67%**	**49.47%**
QWEN-3.6-PLUS	22.00%	69.33%	**68.67%**	21.33%	48.00%	44.67%	**70.67%**	42.67%	48.42%
QWEN-3.5-PLUS	16.44%	68.00%	60.67%	10.67%	36.00%	56.00%	48.00%	50.67%	43.31%
GLM-5	33.78%	16.00%	20.67%	21.33%	40.00%	26.00%	18.67%	14.67%	23.89%
DOUBAO-SEED-2-0-PRO-260215	34.89%	34.67%	45.33%	**76.00%**	46.67%	27.33%	37.33%	58.67%	45.11%
Open-source LMMs:									
KIMI-K2.5	22.89%	49.33%	58.67%	28.00%	20.00%	29.33%	**57.33%**	37.33%	37.86%
MINIMAX-VL-01	9.78%	28.00%	23.33%	22.67%	10.67%	22.67%	20.00%	26.67%	20.47%
GEMMA-3-27B-IT	**33.11%**	33.33%	42.00%	24.00%	16.00%	16.00%	10.67%	18.67%	24.22%
GEMMA-4-31B-IT	25.78%	16.00%	53.33%	**54.67%**	45.33%	32.00%	34.67%	48.67%	**38.81%**
GLM-4.6V-FLASH	24.00%	33.33%	21.33%	36.00%	26.67%	17.33%	20.00%	31.33%	26.25%
Qwen2.5-VL-7B-Instruct	32.44%	22.67%	20.00%	28.00%	16.00%	32.00%	26.67%	16.00%	24.22%
Qwen3-VL-8B-Instruct	9.33%	33.33%	20.67%	18.67%	13.33%	**44.00%**	10.67%	10.00%	20.00%
QWEN3.5-27B	20.44%	57.33%	41.33%	9.33%	22.67%	32.00%	32.00%	46.00%	32.64%
QWEN3.5-397B-A17B	24.89%	**64.00%**	**68.00%**	8.00%	28.00%	42.00%	16.00%	**52.67%**	37.94%
QWEN3.6-35B-A3B	23.11%	52.00%	44.00%	4.00%	29.33%	43.33%	34.67%	43.33%	34.22%
INTERNVL3_5-8B	19.33%	36.00%	26.00%	29.33%	17.33%	15.33%	20.00%	26.00%	23.67%
INTERNVL3_5-GPT-OSS-20B-A4B-PREVIEW	30.67%	9.33%	22.00%	33.33%	13.33%	25.33%	24.00%	22.67%	22.58%
INTERNVL3_5-30B-A3B	14.44%	32.00%	15.33%	44.00%	16.00%	35.33%	18.67%	25.33%	25.14%
INTERNVL3_5-38B	25.33%	14.67%	34.00%	36.00%	33.33%	26.00%	24.00%	18.00%	26.42%
INTERNVL3_5-241B-A28B	19.33%	20.00%	36.00%	29.33%	13.33%	30.67%	22.67%	20.00%	23.92%
LLAVA-ONEVISION-1.5-8B-BASE	25.56%	13.33%	16.00%	28.00%	**48.00%**	24.67%	1.33%	2.67%	19.94%
LLAVA-ONEVISION-1.5-8B-INSTRUCT	30.22%	28.00%	26.00%	10.67%	20.00%	15.33%	52.00%	29.33%	26.44%
random guess	24.67%	22.67%	24.00%	26.67%	22.67%	26.00%	26.67%	25.33%	24.83%

A.8 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.9 SOCIAL IMPACT

Advancing science and education. A controlled diagnostic framework for physical reasoning offers evidence about whether visual perception or physical world modeling may bottleneck a model, while complementary shortcut baselines are needed for stronger attribution. Such evidence may guide controlled model analysis, but the present synthetic evaluation does not establish readiness for physics tutoring, high-stakes scientific use, or reasoning over real experimental imagery.