OmniPhys: A Unified Multimodal Benchmark for Physics Understanding and Generation

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated strong abilities in solving diverse visual and textual reasoning tasks. However, their development in the physics domain is significantly hindered by the lack of a comprehensive benchmark. To fill this gap, we introduce **OmniPhys**, a large-scale benchmark for multimodal physics understanding and reasoning, covering middle school through university-level problems. OmniPhys consists of **13,146** questions and **17,567** images, accompanied by detailed annotations that support fine-grained analysis of reasoning processes and knowledge usage. Beyond conventional evaluation, OmniPhys is a benchmark that systematically evaluates multimodal outputs in physics domain, including models' ability to generate structured physics diagrams, which constitute a fundamental component of authentic physics problem solving. Extensive evaluations reveal critical gaps in the capabilities of current MLLMs, especially in complex reasoning and visual generation. To address this, we release **OmniPhys** to serve as a foundational resource for advancing multimodal intelligence in physics and scientific domains. Codes and data are available at https://anonymous.4open.science/r/Omni-Phys-main-34A9/.

1 Introduction

The rapid evolution of Large Language Models (LLMs) (Jaech et al., 2024; OpenAI, 2023; Gheorghe Comanici et al., 2025; Grattafiori et al., 2024; Liu et al., 2024) and Multimodal Large Language Models (MLLMs) (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023a) has revolutionized language understanding and visual reasoning across general-purpose tasks. However, applying these models to specialized scientific domains, such as mathematical problem-solving(Lu et al., 2023; Yang et al., 2024b), physics problem-solving(Ding et al., 2023; Jaiswal et al., 2024) and diagram interpretation (Masry et al., 2022; Methani et al., 2020), remains challenging. **Physics**, in particular, stands out as a prototypical multimodal arena, demanding the rigorous integration of textual descriptions, visual diagrams, and symbolic logic to achieve accurate reasoning.

Physics underpins all branches of the natural sciences(Feynman, 1967; Smith, 2007). While specific datasets have emerged to benchmark physical reasoning (Ding et al., 2023; Anand et al., 2024; Xu et al., 2025; Luo et al., 2025), existing works rarely satisfy three critical criteria simultaneously: (1) **Cross-stage knowledge fusion**, spanning the full spectrum from middle school to university levels; (2) **Multimodal input comprehension**, requiring the interpretation of complex textual and visual cues; and (3) **Multimodal output generation**, assessing the model's ability to actively synthesize diagrams rather than merely selecting options. The absence of such a unified benchmark hinders a systematic evaluation of current models' capabilities.

To address these challenges, we present **OmniPhys**, a unified Chinese benchmark designed to assess physics mastery from secondary education to university levels. As shown in Figure 1, the dataset features a rigorous combination of textual, visual, and symbolic inputs, covering five major physical disciplines including mechanics, electromagnetism, and optics. To guarantee difficulty and pedagogical validity, all questions are meticulously curated from contemporary examination papers and authoritative textbooks, undergoing a strict multi-stage filtering process. Distinctively, OmniPhys introduces a pioneering subset for multimodal output tasks, specifically designed to assess the capabilities of MLLMs in physics diagram understanding and editing. To establish a benchmark baseline, we conducted comprehensive evaluations on OmniPhys. Our empirical results reveal that despite recent advancements, significant challenges persist for both proprietary and open-source MLLMs. Models across both categories demonstrate substan-

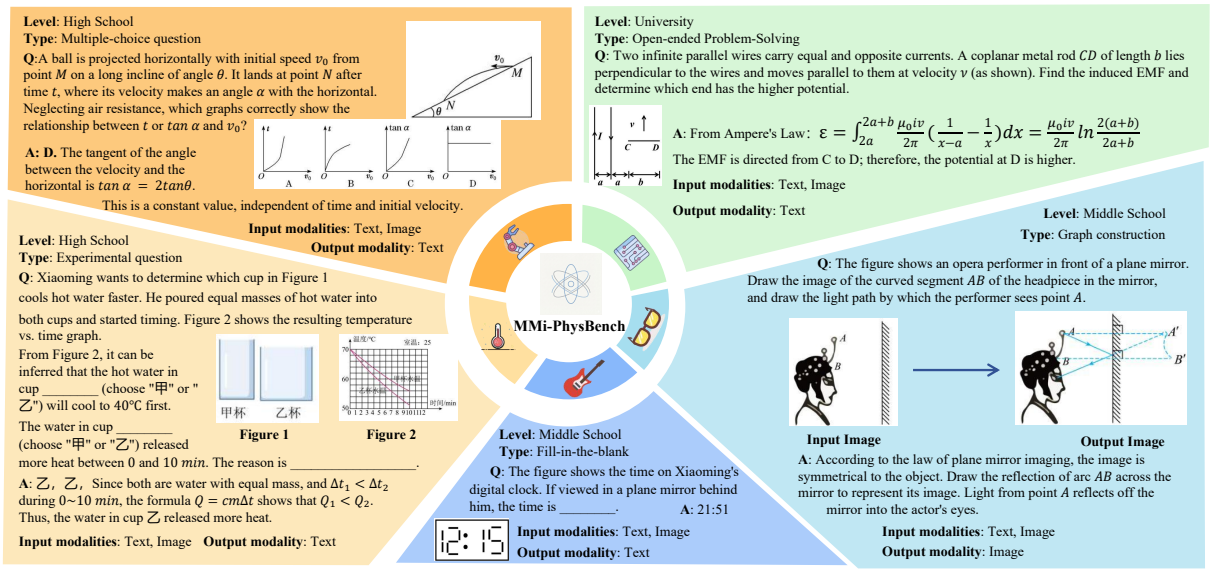

Figure 1: **Overview of the OmniPhys Dataset.** The benchmark encompasses five major physics disciplines, illustrated by representative samples: **Mechanics**, **Electromagnetism**, **Optics**, **Thermodynamics**, and **Acoustics**.

tial room for improvement in handling multimodal physics reasoning across diverse educational stages. Our main contributions are summarized as follows:

Holistic Benchmark. We present OmniPhys, an authentic dataset covering middle school to university physics to evaluate cross-stage reasoning transfer.

Generative Subset. We design a novel subset for multimodal output tasks, assessing the models' capability to synthesize and edit physics diagrams.

Comprehensive Evaluation. Extensive experiments with SOTA MLLMs reveal persistent challenges in conceptual mastery, positioning OmniPhys as a rigorous baseline for future research.

2 Related Works

2.1 Multimodal Large Language Models

Building on the reasoning prowess of LLMs (Zhao et al., 2025; OpenAI, 2023) and the alignment strategies of foundational vision-language models (Radford et al., 2021; Wang et al., 2022), recent MLLMs have achieved significant breakthroughs by synergizing visual encoders with LLM backbones. Representative works, such as BLIP-2 (Li et al., 2023b), Flamingo (Alayrac et al., 2022), and LLaVA (Liu et al., 2023), demonstrate that instruction tuning effectively transfers LLM reasoning capabilities to multimodal scenarios. The field has further evolved towards supporting arbitrary interleaved inputs and outputs, exemplified by unified models like GPT-4V (Wu et al., 2024) and Gemini (Gheorghe Comanici et al., 2025). Despite these advancements, MLLMs remain susceptible to hallucinations (Bai et al., 2024) and often falter in tasks requiring precise numerical calculation or multi-step logical deduction (Yan et al., 2025). These persistent limitations underscore the necessity for challenging benchmarks to probe deep multimodal reasoning.

2.2 Benchmark for Physics Reasoning

As research on physical reasoning continues to advance, the field calls for comprehensive benchmarks to facilitate systematic evaluation.

Table 1 summarizes representative physical reasoning datasets. Early research primarily focused on text-only modalities (Ding et al., 2023; Jaiswal et al., 2024; Xu et al., 2025; Zheng et al., 2025) or general K-12 benchmarks with limited physics subsets (Hendrycks et al., 2020; Li et al., 2024; Zhong et al., 2024; Huang et al., 2023; Zhang et al., 2023). While recent works have introduced image inputs (He et al., 2024; Huang et al., 2024; Li et al., 2025a; Zhou et al., 2025), they still suffer from incomplete educational coverage and a lack of multimodal outputs (Guo et al., 2025). To bridge these gaps, we introduce **OmniPhys**, a unified multimodal benchmark designed to rigorously evaluate capabilities in both physics understanding and generation.

3 The OmniPhys Benchmark

We introduce **OmniPhys**, a comprehensive multimodal physics benchmark spanning junior high

Dataset	Size(Physics)	Lang.	Education Stage				Image Modality	
			PS	MS	HS	Uni	Input	Output
Text-Only Benchmarks								
PhysQA (Ding et al., 2023)	1,008	EN	✗	✓	✗	✗	✗	✗
C-Eval (Huang et al., 2023)	601	ZH	✗	✓	✓	✓	✗	✗
MMLU (Hendrycks et al., 2020)	548	EN	✗	✗	✓	✓	✗	✗
GAOKAO (Zhang et al., 2023)	111	ZH	✗	✗	✓	✗	✗	✗
AGIEval (Zhong et al., 2024)	200	EN	✗	✗	✓	✗	✗	✗
UGPhysics (Xu et al., 2025)	11,040	ZH/EN	✗	✗	✗	✓	✗	✗
PHYSICS (Zheng et al., 2025)	16,568	ZH/EN	✗	✗	✓	✓	✗	✗
Multimodal Input Benchmarks								
TheoremQA (Chen et al., 2023)	131	EN	✗	✗	✗	✓	✓	✗
Multi-Physics (Luo et al., 2025)	1,412	ZH	✗	✗	✓	✗	✓	✗
OlympiadBench (He et al., 2024)	2,334	ZH/EN	✗	✗	✓	✓	✓	✗
MM-PhyQA (Anand et al., 2024)	4,500	EN	✗	✗	✓	✗	✓	✗
PhysicsArena (Dai et al., 2025)	5,103	EN	✗	✗	✓	✗	✓	✗
MM-Eureka (Meng et al., 2025)	500	EN	✗	✗	✓	✗	✓	✗
PhysReason (Zhang et al., 2025)	1,200	EN	✗	✗	✗	✓	✓	✗
See-Phys (Xiang et al., 2025)	2,000	EN	✓	✓	✓	✓	✓	✗
K12Vista (Li et al., 2025a)	6,600	EN	✓	✓	✓	✗	✓	✗
MDK12-Bench (Zhou et al., 2025)	8,542	EN	✓	✓	✓	✗	✓	✗
OmniPhys (Ours)	**13,146**	**ZH**	✓	✓	✓	✓	✓	✓

Table 1: **Comparison with Existing Physics Datasets.** Our **OmniPhys** distinguishes itself by supporting *multimodal outputs* and covering the full spectrum of educational stages. (Lang.: Language, PS: Primary School, MS: Middle School, HS: High School, Uni: University)

Category	Count
Physics Domains	
Mechanics	6,410
Electromagnetism	4,865
Optics	852
Thermodynamics	586
Acoustics	425
Educational Stages	
Junior High School	4,324
Senior High School	8,317
University	505
Task Types	
Objective (e.g., MCQ)	9,327
Open-Ended (e.g., Calculation)	3,558
Multimodal Generation	261
Total Questions	**13,146**
Total Images	**17,567**

Table 2: **Statistics of OmniPhys.** The benchmark features a hierarchical distribution of difficulty spanning three educational stages.

to university curricula. It challenges MLLMs to synergize diagrams, formulas, and text for rigorous, visually grounded reasoning.

3.1 Data Sources and Distribution

OmniPhys is meticulously constructed from a diverse array of authoritative educational resources, ranging from standard textbooks to teacher-curated exercise collections, spanning the full spectrum from middle school to university levels. We employ a multi-stage processing pipeline that leverages state-of-the-art OCR tools, specifically MinerU (Niu et al., 2025) and DeepSeek-OCR (Wei et al., 2025), to analyze original PDFs. The processed data is organized into a structured `JSONL` format, where each instance is comprehensively annotated with problem texts, diagrams, answers, and step-by-step reasoning chains.

Table 2 details the data distribution. The data composition is strategically aligned with real-world physics curricula. Reflecting educational standards, Mechanics and Electromagnetism dominate the distribution. Structurally, OmniPhys adopts a difficulty pyramid, which spans from a Junior and Senior foundation to a University-level challenging set. This hierarchical design allows us to probe the upper bounds of model reasoning in complex scenarios, going beyond average-case performance.

3.2 Data Preprocessing and Quality Filtering

To ensure the correctness, clarity, and reliability of **OmniPhys**, we implement a multi-stage data preprocessing and filtering pipeline.

Completeness Screening. Each sample is strictly required to adhere to a valid `JSON` schema and contain all mandatory fields, including the problem statement, reference answer, and detailed

solution. Furthermore, we prune samples exhibiting insufficient textual length, effectively filtering out incomplete or low-quality content.

Semantic Deduplication. We employ an embedding-based deduplication strategy using the `bge-small-zh-v1.5` encoder (Xiao et al., 2023). We compute dense vector representations for all problem texts and identify duplicates based on a cosine similarity threshold of 0.95. This rigorous process eliminated 12.1% of redundant instances to yield the final valid dataset.

Visual Dependency Filtering. To ensure OmniPhys targets genuine multimodal reasoning, we categorize instances into three levels. **Level 1 (Text-Solvable)** problems contain only decorative images where all necessary information is fully specified in the text. **Level 2 (Text-Descriptive)** problems include images that convey information, but the textual description completely duplicates the visual content. **Level 3 (Image-Essential)** problems require direct interpretation of the image, as critical quantities or relationships are only available in the diagram. To ensure robust classification, we employ a diverse judge panel consisting of DeepSeek-V3(Liu et al., 2024), Qwen2.5(Team, 2025c), and GPT-3.5(Ye et al., 2023). Only samples classified as **Level 3** by all three judges are admitted into the final dataset. The detailed prompting strategy is provided in Appendix A.1.

Difficulty Screening. We implement a model-based difficulty screening. We employ two representative lightweight MLLMs: Qwen2.5-VL-3B(Team, 2025b) and MiniCPM-V-2.6(Yao et al., 2024). We adopt an adversarial filtering criterion: any problem correctly solved by both models is deemed insufficiently challenging and is subsequently discarded. The process is parallelized using the `vLLM` framework on 2 NVIDIA RTX 4090 GPUs. This step refines the dataset distribution, removing easy problems to better reflect meaningful cross-stage physics mastery.

Data Leakage Prevention. We specifically selecting materials from frontline educators and physically digitized examinations that are effectively insulated from public indexing. To empirically verify this isolation, we adopt a search-based filtering protocol. We eliminate samples where `GPT-4` (OpenAI, 2025a) retrieves the exact solution via web browsing or exhibits inconsistent responses when the search function is toggled. Finally, manual verification is conducted on the remaining subset to guarantee a valid zero-shot evaluation.

Meanwhile, we manually checked our collected materials to ensure that they do not contain personally identifying information, references to real individuals, or offensive or sensitive content.

4 Experiments

4.1 Experimental Setup

We evaluate a diverse suite of state-of-the-art MLLMs on **OmniPhys**, spanning both proprietary and open-source categories. The models included in the evaluation are detailed below.

Proprietary Models. We prioritize recently released models to benchmark the upper bounds of current capabilities. Our evaluation covers a broad spectrum of high-performance systems, including GPT-5.2(OpenAI, 2025c), GPT-5.1(OpenAI, 2025b), GPT-4o(OpenAI, 2025a), and o4-mini(OpenAI, 2025d), alongside competitive counterparts such as Gemini-3-pro(Deepmind, 2025), Gemini-2.5-Flash(Gheorghe Comanici et al., 2025), Claude-4.5-sonnet(Anthropic, 2025), Grok-4(Xai, 2025), Doubao-Seed-1.6(Seed, 2025), GLM-4.6v(AI, 2025), and Qwen3-VL-Plus(Bai et al., 2025). All proprietary models are accessed via their official APIs.

Open-source Models. We assess the full spectrum of the Qwen2.5-VL(Team, 2025b) series (3B, 7B, 32B, 72B) and its successor Qwen3-VL(Bai et al., 2025) (2B, 8B, 32B, 235B), alongside InternVL-3.5(Wang et al., 2025) at 2B, 14B, and 38B scales. To further ensure architectural diversity, we incorporate representative models such as Kimi-VL-A3B(Team et al., 2025), DeepSeek-VL-7B(Lu et al., 2024), LLaVA-OneVision-1.5-8B(An et al., 2025), and Phi-4-Multimodal(Abdin et al., 2024). The Qwen3-VL-235B-a22b is accessed via its official API, while all other open-source models are evaluated locally using 8 NVIDIA RTX 4090 and 2 NVIDIA RTX PRO 6000 GPUs, under identical inference settings.

4.2 Dual-Track Reasoning Evaluation

To achieve a granular assessment of multimodal reasoning capabilities, we propose the **Dual-Track Reasoning Evaluation (DTRE)** framework. We categorize problems into two distinct streams: Objective Tasks with deterministic outputs (e.g., multiple-choice, fill-in-the-blank) and Open-Ended Tasks requiring step-by-step derivation (e.g., calculation, experimental design).

To quantify model performance, we formalize

Model	Acoustics			Optics			Mechanics			Thermodynamics			Electricity			Overall				
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	P_{obj}	P_{open}
Proprietary / Closed-source MLLMs																				
GPT-5.2	67.48	81.65	67.78	65.15	79.73	64.42	66.33	79.67	72.30	69.89	82.13	73.84	60.15	77.05	66.34	64.15	78.85	69.42	39.36	29.45
GPT-5.1	58.68	72.24	55.40	52.77	66.09	49.32	52.30	67.91	54.47	59.87	73.50	52.04	45.48	64.07	50.95	50.25	66.68	52.59	23.53	12.87
GPT-4o	46.21	64.07	47.36	41.15	55.99	35.03	39.41	54.01	40.33	49.02	60.46	46.76	36.63	52.24	38.28	39.04	53.93	39.50	12.28	4.61
o4-mini	62.39	68.86	58.46	59.19	67.79	49.08	64.85	72.01	60.01	70.76	77.39	63.86	59.55	67.78	52.63	62.75	70.36	56.48	39.34	21.50
Claude-4.5-Sonnet	59.56	72.98	66.46	56.50	72.72	57.27	54.24	71.73	62.99	56.07	69.90	64.02	45.44	65.11	59.14	51.31	69.30	61.11	28.42	24.42
Gemini-3-Pro	77.89	80.00	78.92	**79.83**	**82.01**	**69.81**	**83.63**	85.30	**80.63**	**86.51**	**88.43**	**81.47**	**78.86**	80.84	**73.71**	**81.66**	83.49	**77.15**	**69.82**	**49.30**
Gemini-2.5-Flash	67.57	81.86	77.00	68.24	81.81	64.54	69.50	84.33	74.49	78.07	87.42	75.19	64.56	80.98	70.33	67.93	83.01	72.20	46.31	35.89
Grok-4	48.30	65.13	59.33	50.45	67.70	50.55	47.91	66.14	51.76	56.13	70.63	54.89	42.64	62.82	48.60	46.53	65.21	50.60	18.96	13.80
Doubao-Seed-1.6	**79.80**	**84.60**	**81.66**	74.75	81.04	59.34	82.64	**86.51**	72.03	82.35	87.21	75.05	75.21	**81.75**	68.95	79.35	**84.41**	70.16	61.42	32.80
GLM-4.6v	70.85	79.05	73.42	69.24	78.67	62.24	75.05	82.59	70.98	76.30	83.15	71.81	68.16	77.48	66.10	72.13	80.44	68.49	52.12	33.14
Qwen3-VL-Plus	68.71	83.98	74.66	67.71	80.10	52.86	70.98	83.01	68.32	75.71	84.05	67.86	65.55	79.02	65.90	68.96	81.42	66.34	45.58	29.60
Open-source MLLMs																				
Qwen2.5-VL-3B	29.97	28.59	16.83	23.65	21.62	5.84	24.06	18.63	10.04	29.71	26.56	12.71	21.57	16.16	9.92	23.45	18.42	9.88	0.96	0.11
Qwen2.5-VL-7B	46.50	40.30	26.42	40.44	41.56	14.96	39.72	39.00	20.94	44.44	46.55	22.24	36.92	37.29	19.45	39.05	38.89	20.03	6.74	1.21
Qwen2.5-VL-32B	56.32	68.27	60.37	53.13	65.41	37.29	56.54	68.18	56.13	63.63	72.71	53.05	48.79	62.34	49.31	53.79	66.05	51.99	25.42	14.08
Qwen2.5-VL-72B	60.98	69.80	60.42	58.11	66.42	35.20	61.16	67.91	48.41	70.04	74.20	51.03	55.30	63.98	45.18	59.19	66.67	46.41	26.40	9.61
Qwen3-VL-2B	39.36	52.22	29.08	37.55	50.22	20.84	31.70	48.45	26.78	41.00	55.26	25.26	29.97	46.60	25.59	32.00	48.25	25.84	6.82	2.61
Qwen3-VL-8B	54.88	69.63	46.58	52.79	67.33	31.26	51.83	67.92	43.40	60.34	71.91	46.41	46.77	65.25	42.39	50.43	67.08	42.35	22.90	7.93
Qwen3-VL-32B	66.03	77.92	**70.00**	63.97	75.81	48.42	64.68	78.03	64.20	72.70	80.76	64.25	57.17	73.60	60.21	62.25	76.36	61.57	36.08	22.57
Qwen3-VL-235B-a22b	**74.57**	**84.87**	69.96	**68.45**	**79.64**	**54.37**	**70.72**	**83.19**	**69.13**	**75.02**	**85.03**	**69.99**	**66.12**	**79.53**	**66.30**	**69.15**	**81.72**	**67.05**	**47.16**	**29.01**
Kimi-VL-A3B	41.39	45.54	26.75	37.20	43.55	16.03	35.28	39.47	21.88	43.19	48.39	23.41	32.13	38.75	21.09	34.70	39.97	21.28	5.61	0.79
InternVL3.5-2B	30.82	42.24	28.61	31.79	41.00	12.92	25.98	35.47	18.73	34.12	43.84	18.29	23.75	34.55	17.88	25.99	35.98	18.04	3.06	0.53
InternVL3.5-14B	50.40	55.46	46.34	47.59	54.47	22.79	40.95	52.77	32.88	52.51	59.89	36.11	36.23	49.27	30.20	40.31	51.96	31.36	12.75	3.68
InternVL3.5-38B	51.02	63.30	39.46	47.61	60.75	28.65	41.91	57.22	38.09	57.15	65.45	39.23	37.29	53.33	35.67	41.41	56.49	36.54	13.32	4.10
Deepseek-VL-7B	24.09	13.66	11.92	23.47	15.23	7.56	22.25	12.35	6.25	24.54	17.35	9.72	22.89	13.48	7.92	22.69	13.19	7.21	1.06	0.14
LLaVA-OneVision-1.5-8B	38.26	48.45	30.38	38.50	44.77	21.81	36.23	39.14	27.42	41.92	43.40	29.94	33.43	40.77	26.01	35.65	40.45	26.61	7.26	2.53
Phi-4-Multimodal	17.24	18.50	12.91	19.76	16.83	5.31	18.71	14.24	5.23	21.43	20.67	5.72	20.23	15.18	6.20	19.44	15.10	5.70	0.53	0.14

Table 3: **Main Results on OmniPhys.** Metrics: S_1 (Answer Accuracy), S_2/S_3 (Process Quality for Objective/Open-ended tasks), and P_{obj}, P_{open} (Strict Mastery Rates). **Formatting**: In each section (Closed-source MLLMs/Open-source MLLMs), the **best** score is bolded and the second best is underlined. The **global best** across all models is highlighted in **blue**.

two complementary metrics: **Result Score** (S_{res}) and **Process Score** (S_{proc}).

Result Score (S_{res}) Designed for **Objective Tasks**, this metric quantifies the accuracy of the final answer. We denote the ground truth set as $\mathcal{A}_{gt}$ and the predicted answer set as $\mathcal{A}_{pred}$. To ensure robustness against formatting variations, we apply a standard normalization function $\phi(\cdot)$.

For single-choice questions and single-slot fill-in-the-blank tasks, the scoring is binary:

$$S_{res} = \mathbb{I}(\phi(\mathcal{A}_{pred}) = \phi(\mathcal{A}_{gt})) \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

For multi-select questions and multi-slot tasks, we adopt a **strict partial credit mechanism** to penalize random guessing. A score is awarded if and only if the predicted set is a subset of the ground truth (i.e., no incorrect options are selected). The score is calculated as:

$$S_{res} = \begin{cases} \frac{|\phi(\mathcal{A}_{pred})|}{|\phi(\mathcal{A}_{gt})|} & \text{if } \phi(\mathcal{A}_{pred}) \subseteq \phi(\mathcal{A}_{gt}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Process Score (S_{proc}) To quantitatively assess the quality of the Chain-of-Thought (CoT) in Open-Ended Tasks, we define a metric based on the completeness of the logical derivation. Let the ground truth reasoning path be decomposed into a set of M **key reasoning steps**, denoted as $\mathcal{K} = \{k_1, k_2, ..., k_M\}$. We verify whether each key step k_i is explicitly or implicitly present and correctly applied in the model's reasoning path $\mathcal{R}_{pred}$. The process score is calculated as the recall rate of these key steps:

$$S_{proc} = \frac{1}{M} \sum_{i=1}^{M} \delta(k_i, \mathcal{R}_{pred}) \tag{3}$$

where $\delta(k_i, \mathcal{R}_{pred}) \in \{0, 1\}$ indicates whether the i-th key step is successfully recovered in the model's generation.

We implement the LLM-as-a-Judge framework(Li et al., 2025b). The final value is derived from the arithmetic mean of scores independently assigned by `DeepSeek-V3.2`(DeepSeek-AI, 2025b) and `GPT-4`(OpenAI, 2023). Detailed prompts for both inference and evaluation are provided in Appendix A.2 and Appendix A.3.

To quantify strict mastery, we define P_{obj} as the rate of instances achieving perfect alignment in both result and reasoning (i.e., $S_1 = S_2 = 1.0$), and P_{open} as the rate of flawless derivation in open-ended tasks (i.e., $S_3 = 1.0$).

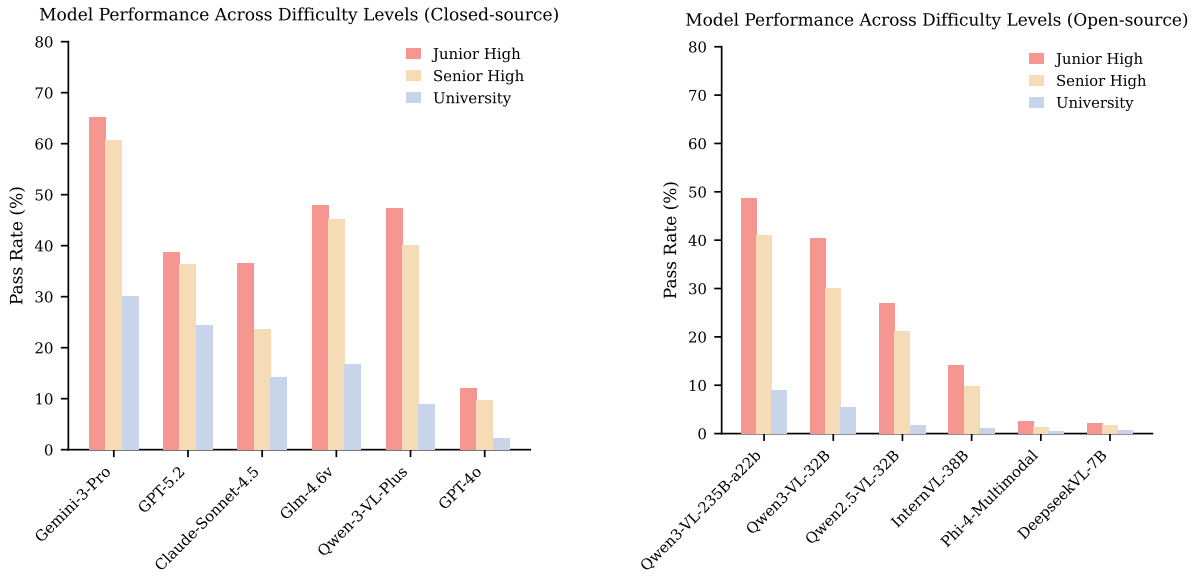

Figure 2: **Performance Degradation Across Educational Stages.** We report the pass rates (%) of 12 representative MLLMs across Junior High, Senior High, and University levels. Both closed-source (Left) and open-source (Right) models exhibit a consistent downward trend, validating the hierarchical difficulty design of OmniPhys.

4.3 Main Results Analysis

The main evaluation results on **OmniPhys** are presented in Table 3 and Figure 2. We summarize the key observations as follows:

Dominance of Proprietary Models. Proprietary systems maintain a clear performance advantage over open-source counterparts. Gemini-3-Pro establishes a new state-of-the-art, closely trailed by Doubao-Seed-1.6. Notably, these two models significantly outperform the flagship GPT-5.2 by over 15%, marking a divergence in the top-tier landscape. However, even the leading models struggle to achieve saturation, with strict mastery rates(P_{obj}) remaining below 70%. This performance gap underscores the substantial difficulty of OmniPhys, indicating that the benchmark effectively evaluates genuine reasoning capabilities rather than mere rote memorization.

Scaling Laws and Generational Gains in Open-Source Models. The open-source landscape demonstrates strict adherence to scaling laws while highlighting significant generational leaps. The Qwen3-VL family scales monotonically, culminating in the 235B model which validates the scaling hypothesis and notably surpasses the proprietary flagship GPT-5.2. Furthermore, architectural superiority proves pivotal: Qwen3-VL-32B significantly outperforms its predecessor and even eclipses the substantially larger Qwen2.5-VL-72B. These findings confirm that algorithmic efficiency is as decisive as raw parameter scale, positioning OmniPhys as a critical testbed for scrutinizing the efficacy of scaling strategies and architectural innovations within the open-source community.

Validating Difficulty Hierarchy. Figure 2 reveals a universal inverse correlation between model performance and educational stages. Across all evaluated models, accuracy peaks at the Junior High level and degrades monotonically through Senior High, reaching its nadir at the University level. This consistent performance drop empirically validates the quality and hierarchical design of OmniPhys, confirming that the benchmark effectively captures the escalating cognitive complexity and reasoning depth required by advanced physics curricula.

4.4 Research on Multimodal Outputs

Current benchmarks predominantly focus on text-based tasks, overlooking the critical capacity to *visualize* and *construct* physical scenarios. Since drawing diagrams is a fundamental demonstration of understanding, we extend our evaluation to multimodal generation, a domain rarely explored in existing physics benchmarks.

To operationalize this, we define the **Physics Diagram Editing** task. Unlike standard retrieval, this requires the model to transform a multimodal input (I_{in}, T) into a target state (I_{out}) governed by strict physical laws. Our pilot experiments reveal a criti-

Model	Evaluation Scores	
	Human Eval.	Model Eval.
Nano Banana	0.23	0.67
Doubao-Seedream-4.0	0.18	0.59
Gemini-3-pro-Image	0.29	0.73
GPT-Image-1	0.27	0.75
Qwen-Image-Edit-Plus	0.10	0.51

Table 4: Performance comparison in multimodal output settings.

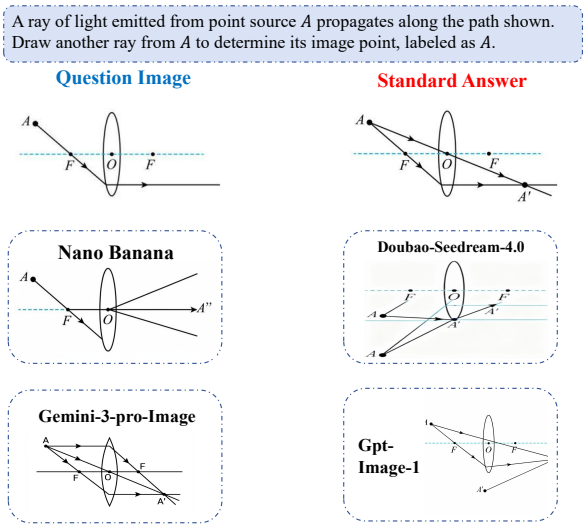

Figure 3: A case study of multimodal outputs in **Omni-Phys** dataset.

cal divergence: while current models exhibit high visual fidelity in general editing, they frequently violate physical constraints—failing to maintain vector directionality during coordinate transformations or preserve the topological integrity of rigid bodies. This gap underscores the necessity of our generative evaluation, exposing blind spots in physical reasoning that text-only metrics fail to capture.

To rigorously assess generation quality, we employed a dual-track strategy: (1) **Human Evaluation**, where three physics graduate students graded physical correctness, and (2) **Automated Evaluation**, utilizing `GPT-5.1` to score instruction adherence on a discrete $\{0, 0.5, 1\}$ scale. Detailed annotation protocols and the specific judging prompts are provided in Appendix A.4 and Appendix B.

As shown in Table 4, the model evaluator exhibits **excessive optimism**, consistently inflating scores (0.51-0.75) compared to human judgment (0.10-0.29). This significant divergence indicates that the current MLLM-as-a-judge paradigm lacks the robustness required for rigorous physical verification. It suggests that relying solely on automated

Model	Full Set	Test-Mini	Drop
Doubao-seed-1.6	76.81	22.45	-70.8%
Gemini-3-pro	80.41	34.66	-56.9%
GLM-4.6v	71.12	19.90	-72.0%
GPT-5.2	65.60	21.82	-66.7%
Qwen3-vl-plus	68.24	18.76	-72.5%
Average	**72.44**	**23.52**	**-67.8%**

Table 5: Comparison of **Average Scores** (scaled to 0-100) on the full dataset versus the Test-Mini set. The significant score decline (avg. -67.8%) confirms the **elevated difficulty** of the subset, validating its **distinct research value** as a rigorous probe for robust physical reasoning.

judges is currently insufficient; specifically, the accurate discrimination of visual quality in multimodal outputs remains heavily dependent on human evaluators. Conversely, the low human ratings attest to the high difficulty and quality of our dataset, offering substantial headroom for future research and model development.

Figure 3 presents a case study revealing that despite high visual fidelity, model-generated outputs frequently deviate from strict physical correctness. Specifically, Nano banana and Doubao-seedream-4.0 fail to correctly grasp the concept of a convex lens focus, while Gemini-3-pro-Image and gpt-Image-1 introduce erroneous, redundant line segments. These discrepancies highlight a significant capability gap in both multimodal understanding and precise generation, indicating substantial room for improvement toward human-level rigor.

5 Ablation Study

To enable efficient yet rigorous experimentation, we constructed a *Test-Mini* set (10% of samples) using an adversarial hardness-based strategy. Instead of random sampling, we prioritized instances with high empirical failure rates (75%) and reasoning complexity (25%) based on the consensus of five SOTA models. This adversarial selection dramatically amplified the challenge: the **average model score plummeted** from 72.44 (full set) to 23.52 (mini set) as shown in Table 5, while the all-model failure rate surged from 1.6% to 15.5%. This significant score decline ensures high discriminative power for our ablation studies. Implementation details are provided in Appendix C.

We evaluate a diverse suite of models, including deep reasoning architectures such as DeepSeek-R1-Distill-Qwen-7B(DeepSeek-AI, 2025a) and

Multimodal Large Language Models (MLLMs)							Large Language Models (LLMs)				
Model	**Text+Img**		**Text+Caption**		**Text-only**		**Model**	**Text+Caption**		**Text-only**	
	P_{obj}	P_{open}	P_{obj}	P_{open}	P_{obj}	P_{open}		P_{obj}	P_{open}	P_{obj}	P_{open}
GPT-5.2	2.78	3.02	4.95	1.08	5.82	0.86	Deepseek-V3.2	5.58	1.08	5.70	1.73
GPT-5.1	2.55	1.08	4.00	0.86	2.55	1.08	GPT-3.5-Turbo	1.94	0.63	1.45	0.22
o4-mini	7.23	1.96	6.55	0.86	6.79	1.73	Qwen2.5-32B	3.64	0.02	3.24	0.22
Gemini-3-Pro	25.82	9.50	16.12	4.10	15.88	6.26	Qwen2.5-7B	1.82	0.05	0.97	0.05
Gemini-2.5-Flash	11.88	2.16	8.12	1.73	5.70	1.30	Qwen2.5-Math-7B	1.21	0.22	1.04	0.02
GPT-4o	3.45	0.26	2.55	0.05	1.33	0.05	InternLM-Chat-20B	1.09	0.05	1.09	0.00
Qwen3-VL-Plus	3.03	0.43	5.58	0.22	7.03	1.30	InternLM-Math-20B	1.29	0.12	0.48	0.00
GLM-4.6v	4.00	0.43	6.42	0.43	8.12	1.08	DeepSeek-R1-Distill-Qwen-7B	2.91	0.12	4.00	0.05
Claude-4.5-Sonnet	3.03	0.65	2.18	1.30	3.88	0.86	P1-30B-A3B	10.67	1.73	9.82	2.59
Avg.	**7.09**	**2.17**	**6.27**	**1.18**	**6.34**	**1.61**	***Avg.***	**3.35**	**0.45**	**3.09**	**0.54**

Table 6: Ablation study results comparing MLLMs (left) and LLMs (right). The columns are arranged by decreasing modal information: from Text+Image to Text-only. We report Objective (P_{obj}) and Open-ended (P_{open}) Accuracy.

P1-30B-A3B(Team, 2025a), as well as variants fine-tuned on mathematics like InternLM-Math-20B(Cai et al., 2024) and Qwen2.5-Math-7B(Yang et al., 2024a) or physics like P1-30B-A3B(Team, 2025a). These models are assessed across three settings to quantify visual dependency: **Text+Img**, which utilizes the original multimodal input; **Text+Caption**, where diagrams are replaced by textual descriptions; and **Text-only**, which requires the model to perform blind inference based solely on the question text.

As shown in Table 6, the adversarial *Test-Mini* set imposes a significantly higher difficulty than the main benchmark, serving as a rigorous probe for deep reasoning capabilities. On average, the **Text+Img** setting yields the highest performance, with MLLMs consistently outperforming LLMs, validating the indispensable role of visual constraints in physics problems. However, we also observe a counter-intuitive phenomenon where certain models perform better in the **Text-only** setting. This suggests that for specific architectures, visual inputs or captions may be misinterpreted as distractor noise rather than helpful context, highlighting persistent challenges in cross-modal alignment that require further investigation.

6 Conclusion

In this paper, we present **OmniPhys**, a comprehensive and high-quality multimodal benchmark designed to rigorously evaluate physics reasoning across educational stages from junior high to university. By employing the Dual-Track Reasoning Evaluation (DTRE) protocol, we move beyond surface-level answer matching to assess the fidelity of the complete reasoning process. Our extensive experiments reveal that OmniPhys poses a significant challenge to current state-of-the-art models, highlighting substantial headroom for improvement. Furthermore, we pioneer the evaluation of multimodal generation through the Physics Diagram Editing task and validate the critical necessity of visual data through in-depth ablation studies. Collectively, these contributions establish OmniPhys as a robust foundation for advancing the next generation of physically grounded AI.

7 Limitations

Our current work presents a foundational step with identifying limitations that guide future research. A primary limitation is the geographic bias of our current data sources, which are predominantly sourced from Chinese educational materials. To enhance global representativeness, future iterations will incorporate diverse international curricula, such as A-Level, IPhO and so on. In addition, further investigation into robust methods for multimodal outputs, especially for diagram-oriented tasks, is a direction for out future work. Regarding evaluation, methodologies for evaluating multimodal physical outputs remain underexplored. A key challenge is the tendency of MLLM-based judges to overestimate the quality of generated content, often overlooking subtle physical inconsistencies. To address this, we aim to establish a robust and effective evaluation framework specifically tailored for multimodal settings. Currently, failure analysis is limited in depth; we therefore plan to conduct a rigorous error attribution study to distinguish whether failures stem from visual perception or logical reasoning, providing more granular insights into MLLM mechanisms in physics reasoning.

References

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

Zhipu AI. 2025. Glm-4.6v system card. https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4.6v. Released on December 11, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. 2024. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–64. Springer.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*.

Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and 1 others. 2025. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*.

Google Deepmind. 2025. Gemini3 – our most intelligent ai model that brings any idea to life. https://deepmind.google/models/gemini/.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. 2023. Using large language model to solve and explain physics word problems approaching human level. *arXiv preprint arXiv:2309.08182*.

Richard Feynman. 1967. The character of physical law (1965). *Cox and Wyman Ltd., London*.

Eric Bieber Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report / arXiv.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-lin Li, Xinjie Lin, Jinnian Zhang, Xin-Sheng Chen, Yi Zhang, and 1 others. 2025. Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. *arXiv preprint arXiv:2505.16770*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, and 1 others. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. 2024. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*.

Chong Li, Chenglin Zhu, Tao Zhang, Mingan Lin, Zenan Zhou, and Jian Xie. 2025a. K12vista: Exploring the boundaries of mllms in k-12 education. *arXiv preprint arXiv:2506.01676*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

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

Zhongze Luo, Zhenshuai Yin, Yongxin Guo, Zhichao Wang, Jionghao Zhu, and Xiaoying Tang. 2025. Multi-physics: A comprehensive benchmark for multimodal llms reasoning on chinese multi-subject physics problems. *arXiv preprint arXiv:2509.15839*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

OpenAI. 2023. Gpt-4 technical report. Technical report.

OpenAI. 2025a. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf. Released on December 11, 2025.

OpenAI. 2025b. Gpt-5.1 system card. https://openai.com/index/introducing-gpt-5-1/. Released on December 11, 2025.

OpenAI. 2025c. Gpt-5.2 system card. https://openai.com/index/introducing-gpt-5-2/. Released on December 11, 2025.

OpenAI. 2025d. O4-mini system card. https://platform.openai.com/docs/models/o4-mini. Released on December 11, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

ByteDance Seed. 2025. Seed1.6 – tech introduction. https://seed.bytedance.com/en/seed1_6.

George Smith. 2007. Newton's philosophiae naturalis principia mathematica.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

P1 Team. 2025a. P1: Mastering physics olympiads with reinforcement learning.

Qwen Team. 2025b. Qwen2.5-vl.

Qwen Team. 2025c. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22227–22238.

Xai. 2025. Grok-4 system card. https://x.ai/news/grok-4. Released on December 11, 2025.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, and 1 others. 2025. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2025. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11798–11827.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024a. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Shenghe Zheng, Qianjia Cheng, Junchi Yao, Mengsong Wu, Haonan He, Ning Ding, Yu Cheng, Shuyue Hu, Lei Bai, Dongzhan Zhou, and 1 others. 2025. Scaling physical reasoning with the physics dataset. *arXiv preprint arXiv:2506.00022*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, and 1 others. 2025. Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. *arXiv preprint arXiv:2504.05782*.

A Prompt Usage

A.1 Visual Dependency Annotation

Prompt for Visual Dependency Annotation

You are an analyst of multimodal physics problem datasets. You are given access only to the problem text, without seeing the accompanying image. Based on the textual description, infer the importance of the image for solving the problem.

Image importance levels are defined as follows:

Level 1 (Text-Only Solvable): The image is purely decorative or illustrative. All information required to solve the problem is fully described in the text, and the image is not necessary.

Level 2 (Text-Descriptive): The image contains information, but the text fully restates or describes all visual content. The image serves only as an aid for understanding and is not strictly required for reasoning.

Level 3 (Image-Essential): The text refers to the image (e.g., "as shown in the figure"), and critical information such as numerical values, geometric relationships, or measurement readings is available only in the image. The image is essential for solving the problem.

Please output only a single integer (1, 2, or 3) corresponding to the image importance level.

Problem text: {problem_text}

A.2 Model Inference Prompt

Prompt for Model Inference

I will present you with a physics problem. Please read and solve it. Adhere strictly to the following output format:

”’ Reasoning Process: [Solve step-by-step logically. Limit each step to no more than 30 words, focusing only on core deductions without redundant explanations.]

Answer: [Output your final answer. Do not add extra content.] ”’

[Input Image]

{problem_text}

A.3 Automated Evaluation Prompts

We employ a dynamic prompting strategy based on the question type. The evaluator (LLM-as-a-Judge) receives specific instructions for objective and open-ended tasks respectively.

Prompt for Objective Tasks (Dual-Track Evaluation)

You are a rigorous grader. Please conduct a dual-track evaluation for this objective task: assess both the correctness of the final result and the validity of the reasoning process.

[Question Data] [Question]: {question} [Standard Answer]: {std_answer} [Explanation]: {std_explanation}

[Student Response] {model_answer}

[Evaluation Tasks] Please complete the following two parts and merge them into a JSON output:

1. **Process Analysis (process_eval)**: - Decompose the standard solution into m key steps. - Count how many steps (n) the student correctly completed. - Process Score = n/m. - If the student provides the correct answer without reasoning steps, treat it as a "reasoning shortcut" (assign score

based on context, typically penalized or full if trivial).
2. **Result Analysis (result_eval)**: - Ignore the process and strictly check if the final answer matches the standard. - For Single Choice / True-False: 1.0 for match, 0.0 for mismatch. - For Multi-Select / Fill-in-the-Blank: Score = (Count of Correct Slots) / (Total Required Slots). - Ignore format variations (e.g., "A" vs. "A.").
[Output Requirement]
Output ONLY in JSON format without Markdown tags: { "process_eval": { "reason": "Brief analysis (Total m steps, Correct n steps)", "total_steps": m, "correct_steps": n, "process_score": 0.0 to 1.0 }, "result_eval": { "reason": "Brief justification for the result", "score": 0.0 to 1.0 } }

Prompt for Open-Ended Tasks (Process-Only Evaluation)

You are an expert physics evaluator. Your task is to assess the student's reasoning logic step-by-step.
[Question Data]
[Question]: {question} [Standard Answer]: {std_answer} [Explanation]: {std_explanation}
[Student Response]
{model_answer}
[Scoring Criteria] 1. Decompose the complete resolution process into m **Key Reasoning Steps**. 2. Determine how many of these steps (n) are correctly included in the student's response. 3. The final Process Score is n/m.
[Output Requirement]
Output ONLY in JSON format without Markdown tags: { "reason": "Step-by-step analysis of the derivation", "total_steps": m (integer), "correct_steps": n (integer), "process_score": Calculated result of n/m (0.0 to 1.0) }

A.4 Automated Evaluation Prompts for Multimodal Generation

We utilize a Vision-Language Model (e.g., GPT-5.1 or GPT-4o) as the evaluator to assess the quality of generated physics diagrams. The judge receives the problem text, the context image (if available), and the generated diagram, then assigns a discrete score based on physical fidelity.

Prompt for Diagram Editing Evaluation

You are a professional and rigorous physics instructor. Your task is to grade physics diagrams drawn by students based on specific problem requirements.
[Input Data]
1. **Problem Description**: Describes the physical scenario or process to be drawn. 2. **Context Image** (Optional): Provides background information (if any). 3. **Student Answer Image**: The diagram generated by the student.
[Evaluation Task]
Based on physical principles, comprehensively evaluate whether the student's image accurately and completely fulfills the problem requirements. The diagrams may involve mechanics analysis, motion trajectories, optical paths, circuit connections, or electromagnetic field distributions.
[Scoring Criteria]

- **1.0 (Fully Correct)**: The image perfectly reflects the problem requirements. All key physical elements (e.g., vector directions, points of application, trajectory shapes, light paths, circuit connections, physical labels) are accurate and adhere to physical laws.
- **0.5 (Partially Correct)**: The image captures the core physical concept but contains defects in details. Examples: Main structure is correct but minor labels are missing; key vectors are roughly correct in direction but deviate significantly in angle or proportion; general trend is correct but local errors exist; or unnecessary misleading lines

are included.

- **0.0 (Incorrect)**: The image contains fundamental physical errors or omits the most critical information, failing to reflect the problem requirements. Examples: Depicting the wrong physical phenomenon; missing core elements required by the stem; or generating an image completely irrelevant to the problem.

[Output Requirement] Please output strictly in JSON format (no Markdown tags): { "reasoning": "Concise justification pointing out specific merits or demerits", "score": 1.0 or 0.5 or 0.0 }
[Input Sequence] [Problem Description]: {question_text}
[Context Image]: (Input Image Here)
[Student Answer Image]: (Generated Image Here)

B Annotation Interface Details

To ensure the quality and consistency of our evaluation, we utilized the Label Studio platform for manual annotation. Figure 4 illustrates the annotation interface used by human evaluators to assess the model's multimodal outputs.

The interface was carefully designed to present all relevant information required for reliable judgment within a single view. Specifically, it simultaneously displays (1) the original problem statement, (2) the original input image, (3) the ground-truth reference image provided by the dataset, and (4) the image generated by the evaluated model. This design allows annotators to directly compare the model output with the reference solution in the context of the original task, thereby making the scoring process both efficient and less prone to omission or misinterpretation.

For each sample, annotators were asked to assign a quality score based on the correctness, completeness, and visual faithfulness of the generated image with respect to the reference answer. Since all relevant inputs and outputs are visible in a unified interface, the evaluation process is straightforward to operate and reduces unnecessary cognitive load on the annotators.

The annotation was conducted by three graduate students specializing in physics, all of whom had prior experience with problem solving and diagram interpretation in the target domain. To improve reliability, each sample was independently annotated by all three evaluators, and the final human evaluation score reported in our experiments is computed as the average of the three scores. This aggregation strategy helps mitigate individual bias and increases the robustness of the evaluation results. The annotators were compensated with standard research assistant stipends in accordance with institutional guidelines.

C Test-Mini Construction and Ablation Details

In this section, we provide a detailed elaboration on the construction process of the *Test-Mini* set and the specific settings used for the ablation study.

C.1 Hardness-Based Selection Methodology

To ensure the *Test-Mini* set effectively probes the upper limits of multimodal reasoning, we implemented a multi-dimensional hardness scoring mechanism. For each candidate question x_i in the full dataset (Without multimodal outputs, $N = 12,885$), we computed a hardness score $H(x_i)$ based on two factors: empirical model failure and reasoning complexity.

Metric	Full Set	Test-Mini
Sample Size	12,885	1,288
Avg. Reasoning Length (chars)	563	979
Visual Necessity Score	8.98	8.99
All-Model Failure Rate	1.6%	15.5%
Avg. Model Accuracy	**72.44%**	**23.52%**

Table 7: Comparison of key statistics between the full dataset and the selected Test-Mini subset. The subset demonstrates significantly higher complexity and lower model solvability.

The score is defined as:

$$H(x_i) = w_1 \cdot \mathcal{F}(x_i) + w_2 \cdot \mathcal{C}(x_i) \qquad (4)$$

where:

- $\mathcal{F}(x_i)$ represents the **Empirical Failure Rate**. It is calculated as $\mathcal{F}(x_i) = 1 - \frac{1}{|M|}\sum_{m \in M} S(m, x_i)$, where M is the set of five baseline models and $S(m, x_i) \in [0, 1]$ is the normalized score of model m on question x_i.

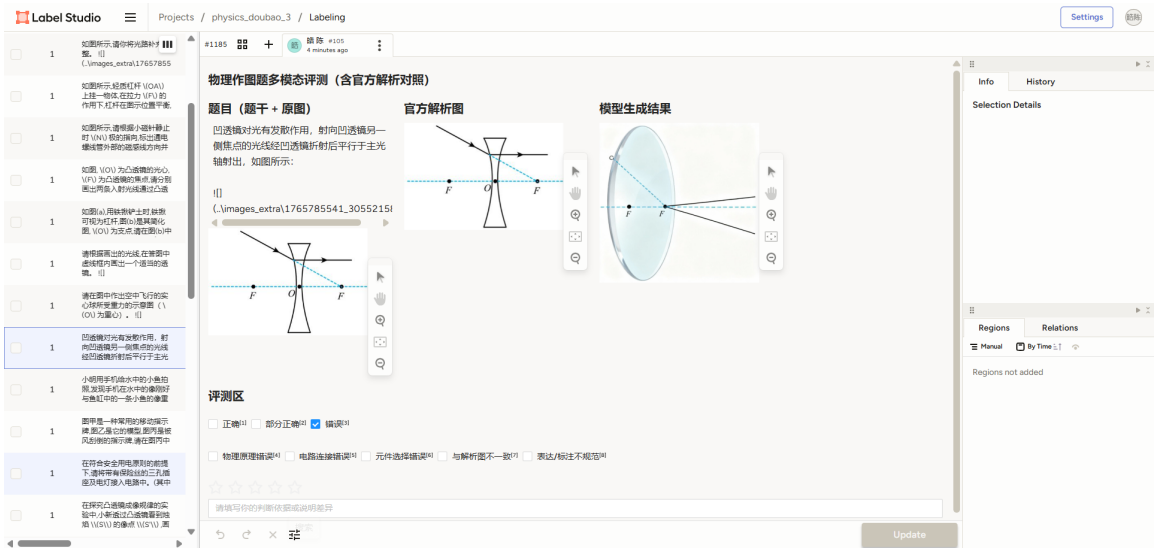

Figure 4: Screenshot of the Label Studio interface for human evaluation of multimodal outputs.

- $\mathcal{C}(x_i)$ represents the **Reasoning Complexity**, quantified by the percentile rank of the character length of the ground truth explanation (Chain-of-Thought).

- We set the weights to $w_1 = 0.75$ and $w_2 = 0.25$, prioritizing empirical difficulty while accounting for logical depth.

The baseline models set M includes five state-of-the-art closed-source models: *Doubao-seed-1.6*, *Gemini-3-pro-preview*, *GLM-4.6v*, *GPT-5.2*, and *Qwen3-vl-plus*. We selected the top 10% of samples ranked by $H(x_i)$ to form the *Test-Mini* set.

C.2 Statistics and Performance Gap

The selection process resulted in a subset with significantly higher difficulty. As shown in Table 7, the *Test-Mini* set exhibits a sharp increase in the "All-Model Failure Rate" (questions where no model answered correctly) from 1.6% to 15.5%, and a substantial increase in the average reasoning chain length.

Table 5 details the performance drop for each baseline model. The consistent degradation across all models (ranging from 56.9% to 72.5%) confirms that the difficulty of the *Test-Mini* set is not biased towards a specific architecture but stems from the inherent complexity of the physics problems.

D AI Assistant Usage Disclosure

We used AI assistants (Gemini/ChatGPT) to assist with writing script code for data processing and for polishing the language of the manuscript to improve readability. All scientific claims, experimental designs, and final text were manually verified and revised by the authors.