MV-Physics:Benchmark for Evaluating Multimodal Large Language Models in Physical Visual Scenes

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) show strong performance in visual reasoning, yet existing benchmarks for physics-related scenarios have significant limitations. To fill this gap, this paper proposes MV-Physics, a novel multi-dimensional benchmark tailored to physics visual scenarios, consisting of 8,011 junior high school physics questions from China's basic education system, covering 5 disciplinary subfields and 3 question types. A hierarchical, multi-dimensional evaluation frameworkspanning single- to multi-image reasoningis constructed to systematically assess MLLMs'physics visual reasoning abilities.Experimental results show Gemini 2.5 Pro achieves 88.34% accuracy in single-image tasks, while doubao-seed-1-6-vision-250815 reaches 89.96% in multi-image tasks. However, most models perform poorly on multiple-select questions, with accuracy below the 60% threshold. Moreover, analysis of reasoning efficiency in multi-image tasks indicates models still need substantial improvement in balancing accuracy and inference latency.This study provides a critical evaluation tool for advancing MLLMs in physics visual domains and serves as a "competence calibration metric" for intelligent teaching tools in basic education.

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) have become a research hotspot in the field of artificial intelligence. Their ability to process both visual and linguistic information has led to remarkable performance in tasks such as image understanding, visual question answering, and cross-modal reasoning(Bai et al., 2023; Meng et al., 2025; Liu et al.; Caffagni et al., 2024). However, in practical applications, many scenarios involve the application of physical knowledge, such as judging the physical properties of objects during robotic manipulation or predicting the physical laws governing vehicle motion in autonomous driving (Mason et al., 1989; Ren et al., 2024; Chu et al., 2024). These physical visual scenes impose higher demands on MLLMs, requiring not only robust visual perception and language understanding capabilities but also a solid foundation in physical knowledge and the reasoning ability to effectively integrate visual information with physical principles.Therefore, evaluating an MLLMs performance in physical-visual tasks has become crucial, and standardized testing similar to school examinations has proven to be an effective method for measuring model capabilities.

Although numerous multimodal reasoning evaluation benchmarks targeting science, mathematics, and general disciplines have emerged (e.g., ScienceQA (Lu et al., 2022), MMMU(Yue et al., 2024)), which have set benchmarks for MLLM evaluation, most of these related works merely scratch the surface in physical domain assessment. Specifically, they suffer from limited data scale (most contain only hundreds of test questions), lack fine-grained classification of physical sub-disciplines (e.g., mechanics, electromagnetism) and question types (e.g., multiple-select questions, fill-in-the-blank questions), and fail to conduct systematic and targeted evaluation and analysis on the unique reasoning requirements and core competency dimensions of the physical domain.

Thus, constructing a multi-dimensional evaluation benchmark specifically for physical-visual scenarios holds significant practical significance for accurately measuring the performance of existing models and promoting the development of MLLMs in the physical-visual field. Based on the above background, this paper proposes the MV-Physics evaluation benchmark. Our contributions are as follows:

- We construct a high-quality physics-visual reasoning QA dataset, which contains diverse physics scenario questions and provides fine-grained dual annotations (corresponding physics sub-discipline and question type) for each question.

- We design a multi-dimensional evaluation framework spanning hierarchical scenarios from single- to multi-visual inputs to comprehensively assess MLLMs overall performance in physics-visual tasks, and innovatively propose two specialized studies: topic-image relevance analysis for single-image reasoning and reasoning efficiency evaluation for multi-image reasoning.

- We evaluate mainstream MLLMs on this benchmark, analyze their performance across dimensions, and summarize existing models common flaws and potential improvement directions.

2 Related Work

2.1 Development of MLLMs

As the understanding capabilities of large language models in pure text have deepened, exemplified by models like ChatGPT(Ouyang et al., 2022) and DeepSeek (DeepSeek-AI et al., 2025), multimodal large language models have also experienced explosive development. OpenAI's GPT-4 series was introduced in 2023(OpenAI et al., 2023) , and subsequent releases like GPT-4V(OpenAI, 2023)and GPT-4o (OpenAI, 2024)took significant steps in multimodal fusion; Alibaba Cloud's Qwen VL series has continuously innovated, achieving globally leading results with its innovative architectural design (Bai et al., 2023; Wang et al.; Bai et al.); Google's Gemini series has iterated and evolved at an astonishing pace(Team and Google) ; and xAI launched Grok-4, claimed to be the world's most powerful model(xAI, 2025) . These models have all demonstrated unprecedented breakthroughs in visual language understanding tasks.

2.2 Development of MLLM Evaluation Benchmarks

Building on the research paradigm of using academic exams as evaluation benchmarks for Vision-Language Models (VLMs), a series of benchmarks focusing on multimodal exam scenarios have been proposed successively. As a representative dataset in this field, ScienceQA(Lu et al., 2022)covers 26 K-12 science topics and requires models to integrate multimodal content understanding with external knowledge reasoning to derive answers. MMMU(Yue et al., 2024) further increases task difficulty by focusing on university-level interdisciplinary complex reasoning tasks. MathVista(Lu et al., 2024) specifically evaluates mathematical reasoning in visual contexts, while MV-MATH (Wang et al., 2025) proposes analyzing models' mathematical reasoning capabilities in multi-visual scenarios. M3Exam (Zhang et al., 2023) introduces the first multilingual multimodal exam benchmark; building on this, Exams-V (Das et al., 2024)expands beyond M3Exam's language coverage and adopts a more realistic multimodal integration format by directly embedding question text into images. EMMA(Hao et al., 2025) includes four types of multimodal reasoning problems (mathematics, physics, chemistry, and coding), achieving interdisciplinary coverage. However, even though some of these benchmarks include physics-related content, they do not conduct systematic and targeted evaluation and analysis for the unique reasoning requirements and capability dimensions of the physics domain.

PHYBench(Shi et al., 2025) , TPBench (Chung et al., 2025), and UGPhysics(Xu et al., 2025) contain abundant physics questions but are limited to text-only question answering without visual inputs. For evaluation benchmarks targeting physics visual question answering scenarios: Seephys(Xiang et al., 2025) proposes four evaluation modes, Physreason (Zhang et al., 2025) introduces an automatic scoring framework, PhyX(Shen et al., 2025)focuses on reasoning type analysis, OlympiadBench centers on competition-level questions (He et al., 2024), and PHYSICS targets doctoral qualifying exams (Feng et al., 2025). Nevertheless, their evaluations are restricted to single-image reasoning scenarios and do not involve multi-image collaborative reasoning tasks, making it difficult to match real-world physics application scenarios where multiple visual information sources interact. Although MM-PhyQA (Kapuriya et al., 2025)includes multi-image reasoning scenarios, it only focuses on physics knowledge reasoning at the Indian high school level.

To address existing gaps, this paper proposes the MV-Physics evaluation benchmark, tailored to Chinese junior high school physics exam scenar-

ios with comprehensive knowledge coverage and fine-grained evaluation dimension decomposition. The benchmark spans five core physics branches (mechanics, electromagnetism, optics, thermodynamics, acoustics) and three typical question types (single-choice, multiple-select, fill-in-the-blank), enabling precise performance analysis across knowledge modules and question types. Integrating single-image and multi-image reasoning tasks, it holistically assesses models' capability to understand and reason about complex physics visual information. Unlike existing high-difficulty-focused benchmarks, MV-Physics adopts a more discriminative difficulty range, aiming to identify optimal intelligent auxiliary tools for junior high school physics education, thereby bearing both evaluative and practical significance.

3 MV-Physics

3.1 Overview

This paper constructs MV-Physics, a finely annotated multimodal dataset for systematically evaluating Multimodal Large Language Models (MLLMs) in physics visual reasoning scenarios. Each sample integrates question-related images and textual stems to simulate real-world problem-solving, imposing stringent demands on models modality fusion capabilities. MV-Physics contains 8,011 Chinese junior high school physics exam questions, spanning five disciplinary branches (mechanics, electromagnetism, optics, thermodynamics, acoustics) and three question types (single-choice, multiple-select, fill-in-the-blank). For hierarchical evaluation of visual reasoning capabilities, the dataset is split into two subsets based on visual context complexity: single-image reasoning (5,582 questions, with fill-in-the-blank items targeting single-step reasoning) and multi-image reasoning (2,429 questions, mostly involving multi-step reasoning, with 28 images per task).

Figures 1 and 2 present detailed statistics and coverage of the two subsets, while Figure 3 displays partial question examples from MV-Physics.

The paper(Deng et al., 2025) notes that vision models tend to blindly trust textual data rather than visually extracted information, while Seephys (Xiang et al., 2025) points out that models perform poorly on questions with strong visual relevance. To assess the hierarchy of large models' image understanding capabilities, and given that image information is closely tied to question solving in multi-image reasoning scenarios, we categorize single-image reasoning questions into two levels based on the impact of images on problem solving:

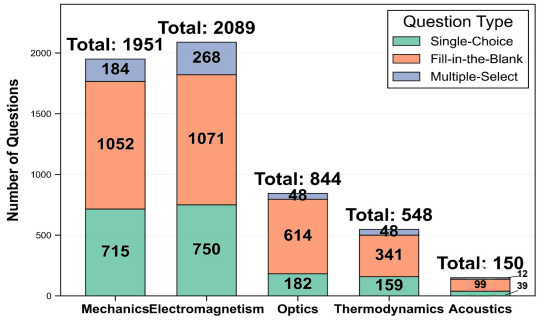


Figure 1: Data Distribution of Single-Image Reasoning: Disciplinary Branches and Question Types.

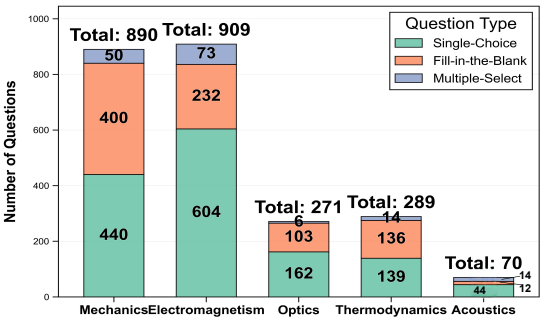


Figure 2: Data Distribution of Multi-Image Reasoning: Disciplinary Branches and Question Types.

- strongly relevant to images (strong): images contain core problem-solving information, making it impossible to solve the questions without visual data;
- weakly relevant to images (weak): images mainly serve as aids for model comprehension, with textual stems providing sufficiently detailed descriptions to enable or highly likely enable problem solving.

We conduct in-depth research on this classification and explore the underlying mechanisms, among which questions strongly relevant to images account for 78% of all single-image reasoning questions.

3.2 Data Construction

Data Collection: Our data was sourced from official educational resource websites of various provinces in China to ensure data quality. Initially, we collected physics test papers from differ-

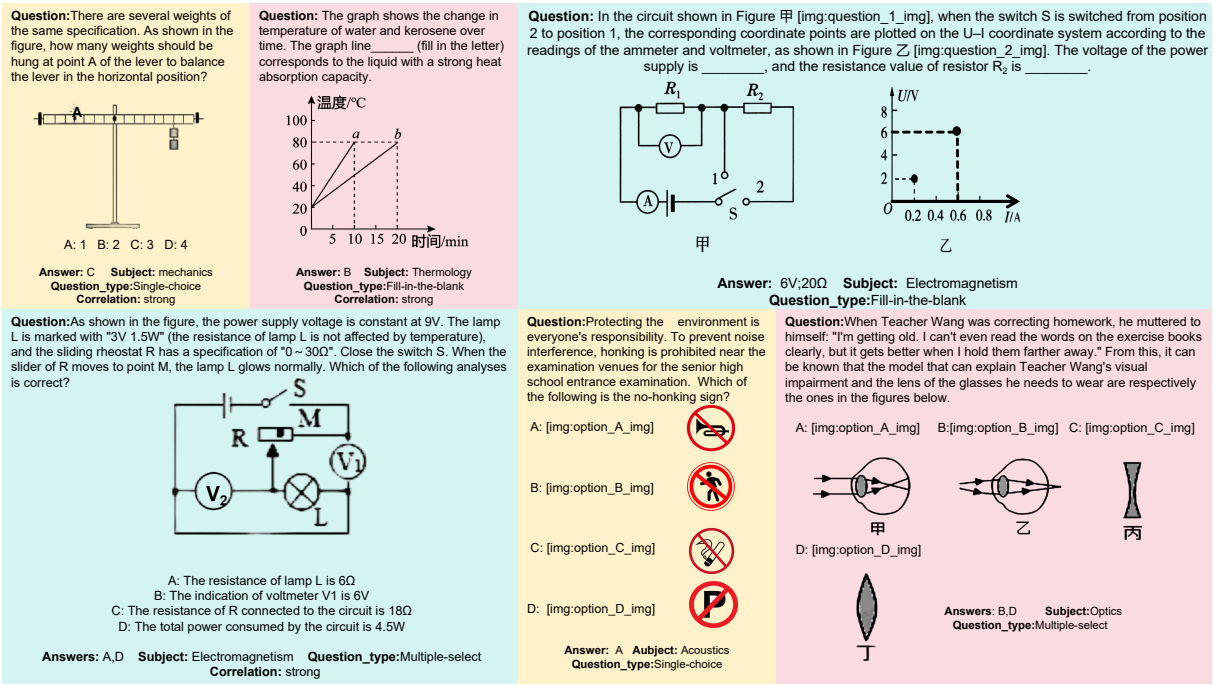


Figure 3: Question Examples.

ent regions and years across China, and manually screened out questions containing images.

Data Cleaning: To guarantee data quality, manual verification was conducted to filter out low-quality image-based questions, and image enhancement via scanning was performed on some materials. Automated methods were adopted to remove duplicate data. Additionally, the same questions were searched on other educational platforms to verify the correctness of answers.

Data Construction and Annotation: Images were stored in a unified format, and metadata was created for each question. This metadata includes a unique ID, file path to the question snapshot, subject, question type, and the correct answer for the question. A correlation label was added to questions in the single-image reasoning subset.

4 Experimental Setup

We conducted comprehensive and extensive experiments on MV-Physics. For the single-image reasoning scenario, we tested 22 domestic and international large models via their APIs (see Appendix A for detailed API mappings): [Gemini 2.5 pro (Team, 2025b), Gemini 1.5 Pro(Team and Others, 2024), doubao Seed 1.6(ByteDance, 2024), ERNIE-4.5-Turbo-VL, ERNIE-4.5(Team, 2024d), Qwen2.5-VL-72B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen2.5-VL-7B-Instruct(Yang et al., 2024) , Qwen-VL-Max, Qwen-VL-Plus(Bai et al., 2023), InternVL3-38B, InternVL3-14B, InternVL2.5-38B-MPO (Chen et al., 2023), Qianfan-Llama-VL-8B (Team, 2024b), Qianfan-Check-VL (Team, 2024a), Qianfan-QI-VL (Team, 2024c), GPT-4.5-preview (Team, 2025c), GPT-4o(OpenAI, 2024) , Grok-4 (xAI, 2025), DeepSeek-VL2 (Wu et al., 2024), Claude 3.5 Sonnet(Anthropic, 2024) , Claude Sonnet 4(Team, 2025a)].

For the multi-image collaborative reasoning scenario, we evaluated 28 domestic and international large models via their respective APIs (see Appendix A for detailed API mappings). Building on the model pool from the single-image reasoning task, we excluded five models: Gemini 1.5 Pro(Team and Others, 2024), Qianfan-Check-VL (Team, 2024a), Qianfan-QI-VL (Team, 2024c), GPT-4.5-preview (Team, 2025c), and Claude 3.5 Sonnet(Anthropic, 2024). The newly added models for this scenario include Claude Haiku 4.5(Anthropic, 2025) , GLM-4.5V(Zai Org, 2025), Llama-4-Maverick-17B-128E-Instruct (Meta, 2025), doubao-Seed-1.6-vision (ByteDance, 2024), Qwen3-VL-Plus, Qwen3-VL-32B-Thinking, Qwen3-VL-30B-A3B-Thinking, Qwen3-VL-8B-Thinking(Team, 2025d) , Qianfan-VL-8B(Baidu Qianfan Team, 2025b), Qianfan-VL-70B(Baidu Qianfan Team, 2025a) , and GPT5 (OpenAI, 2025).

Our evaluation was conducted across multiple dimensions to explore MLLM performance in solving problems spanning different disciplinary branches, question types, and levels of image-information relevance. In single-image reasoning experiments, we performed full-scale tests and supplementary assessments on the correlation: weak subsetwhere images act solely as auxiliary contextrequiring models to answer using only textual stems with images excluded.

Accuracy served as the primary evaluation metric: automated programming-based evaluation was applied to single-choice and multiple-select questions, while manual verification was used for fill-in-the-blank items. For complex multi-visual scenarios, an additional metricinference efficiencywas introduced. Since real-world model performance depends on balancing accuracy and response speed, average time per question was used to quantify efficiency, which directly reflects user waiting time from query submission to answer retrieval and aligns with practical application needs. A combined analysis of accuracy and average latency enables a more holistic assessment of a models viability in real deployment environments.

5 Results and Analysis

5.1 Single-Image Reasoning

Top-performing models (doubao-seed-1-6-250615, Gemini 2.5 Pro) achieved an overall score of over 80 points, outperforming the next tier by more than 10 points, which reflects their absolute advantage in the comprehensive task of physics discipline image understanding. The overall accuracy is presented in Table 1.

Disciplinary Difficulty and Model Adaptability: **Acoustics > Optics > Thermodynamics > Mechanics > Electromagnetism.** Electromagnetism, due to its complex physical laws, emerged as the domain with the highest discrimination, while acoustics, characterized by its fundamental and intuitive nature, became the field where models demonstrated the best performance.

The overall difficulty gradient across question types follows the order: **Multiple-select questions > fill-in-the-blank questions > single-choice questions**, with a notable bipolar distribution in model performance. Single-choice questions represented the easiest task for all models: nearly 50% of the evaluated models achieved an accuracy between 60% and 80%, reaching a moderate performance level but still leaving room for improvement, while the remaining models exhibited weak reasoning capabilities for this question type. Multiple-select questions were the most challenging category, marked by an extreme disparity in performance; the majority of models scored below the 60% accuracy threshold. For fill-in-the-blank questions, model performance was generally stable and clustered around a moderate level, with no significant polarization in results.Table 2 lists partial sample data.

Models	SC	MS	F
Qwen VL-Max	75.69	52.42	62.75
Qianfan-QI-VL	45.09	**0.54**	45.43
DeepSeek-VL2	**34.57**	15.36	42.06
InternVL3-38B	68.74	48.93	64.08
ERNIE-4.5	72.98	40.71	66.62
doubao-seed-1-6-250615	89.37	76.13	81.93
Claude 3.5 Sonnet	51.27	12.14	**24.76**
Claude Sonnet 4	50.08	11.61	24.88
Gemini 2.5 pro	**90.77**	**83.61**	**87.77**

Table 2: Examples of Model Performance Across Different Question Types in Single-Image Reasoning.

Based on the difference between the accuracy of weak questions with images removed and the accuracy of weak questions with images retained (values in parentheses), models can be divided into three categories.Table 3 lists partial sample data.

Image-dependent (negative difference, > -1) (12 models total): Auxiliary images enhance performance, indicating MLLMs lack adequate spatial imagination with text-only input. Extremely dependent models (e.g., InternVL2.5-38B-MPO: -38.02; Qianfan-Check-VL: -30.25) experience drastic performance drops without images, showing weak text-independent problem-solving ability. Slightly dependent models (e.g., ERNIE-4.5: -2.40) are minimally affected by image presence, with balanced text and image-text reasoning capabilities.

Image-interfered (positive difference, > 1) (5 models total): Auxiliary images reduce performance, as large models misinterpret images and prioritize perceived visual information when text-image contradictions arise. Extremely interfered models (e.g., Grok-4: +38.96; Claude 3.5 Sonnet: +20.94; Claude Sonnet 4: +20.42) perform far better without images, where visual information acts

Models	Single-Image Reasoning						Multi-Image Reasoning					
	overall	Mech	EM	Opt	Thermo	Acous	Overall	Mech	EM	Opt	Thermo	Acous
doubao-vision[1]	—	—	—	—	—	—	**89.96**	**90.90**	**85.51**	90.36	95.22	90.41
Qwen3-VL-32B-Thinking	—	—	—	—	—	—	87.40	87.53	78.49	**95.53**	**97.05**	**91.78**
doubao-seed[2]	83.80	82.70	81.30	88.54	86.82	94.63	85.48	84.33	81.31	90.87	91.62	91.78
Gemini 2.5 pro	**88.34**	**87.31**	**86.56**	**91.97**	**90.48**	**98.00**	83.82	84.32	77.50	88.47	90.04	91.78
Qwen3-VL-30B-A3B-Thinking	—	—	—	—	—	—	82.89	85.00	77.29	85.25	87.55	78.08
Qwen3-VL-Plus	—	—	—	—	—	—	80.36	80.53	72.06	87.57	89.44	86.30
GPT5	—	—	—	—	—	—	78.88	78.06	74.85	79.49	87.25	94.52
Qwen3-VL-8B-Thinking	—	—	—	—	—	—	79.98	71.22	78.84	83.57	88.28	73.97
Qwen2.5-VL-72B-Instruct	68.78	69.86	66.69	79.39	75.86	90.13	78.56	80.58	67.37	86.21	88.10	83.56
ERNIE-4.5-Turbo-VL	70.64	69.11	67.68	78.70	73.69	80.67	73.47	75.83	70.16	79.49	82.20	71.23
Qwen VL-Max	65.98	66.45	59.54	75.30	71.16	78.67	71.33	71.28	61.87	79.49	82.09	83.56
ERNIE-4.5	67.86	62.54	62.04	77.59	72.86	80.27	70.14	70.16	66.32	80.15	81.05	82.19
Gemini 1.5 Pro	66.93	64.74	63.83	76.84	69.77	81.86	—	—	—	—	—	—
GLM-4.5V	—	—	—	—	—	—	69.74	74.19	54.80	74.43	81.85	78.87
Grok-4	38.25	35.30	32.04	42.06	44.55	52.70	67.66	67.08	60.99	71.73	76.95	85.51
Qianfan-VL-70B	—	—	—	—	—	—	67.29	69.37	52.08	76.36	82.66	78.08
Llama-4-Maverick[3]	—	—	—	—	—	—	66.88	71.19	54.29	65.15	80.11	78.08
InternVL3-38B	61.90	63.82	55.83	73.48	76.58	85.71	65.90	68.07	49.73	73.48	84.27	76.71
Qwen2.5-VL-32B-Instruct	65.86	64.23	58.13	77.67	77.51	87.07	65.46	64.25	55.97	74.21	78.88	79.45
InternVL2.5-38B-MPO	59.59	59.37	52.09	67.58	70.82	82.31	62.36	65.45	47.79	70.45	74.73	69.86
Qwen2.5-VL-7B-Instruct	60.04	58.05	52.32	74.82	70.45	70.07	55.15	51.04	49.87	60.12	71.66	75.3
Claude Haiku 4.5	—	—	—	—	—	—	55.11	53.38	46.19	61.93	69.16	63.89
Qianfan-VL-8B	—	—	—	—	—	—	49.77	51.69	36.93	54.39	64.65	61.64
InternVL3-14B	59.86	59.42	51.17	70.73	72.49	81.00	60.63	63.40	46.40	65.92	74.90	76.71
Qwen-VL-Plus	60.24	60.10	52.80	71.41	67.97	74.67	62.11	60.19	54.97	76.80	77.81	83.00
Claude Sonnet 4	**26.74**	**25.70**	24.12	**28.47**	**31.86**	**43.44**	61.80	63.62	57.29	66.77	72.28	75.00
GPT-4o	50.73	48.09	44.90	58.87	62.78	77.03	54.96	61.27	42.57	55.69	58.96	78.08
Qianfan-Llama-VL-8B	47.91	44.40	41.01	64.61	55.95	67.35	49.58	50.68	36.60	53.64	67.44	65.75
Qianfan-Check-VL	45.05	44.14	37.42	57.66	54.56	58.50	—	—	—	—	—	—
Qianfan-QI-VL	40.77	38.07	35.93	49.64	51.02	57.14	—	—	—	—	—	—
DeepSeek-VL2	36.86	34.06	31.40	48.46	42.94	61.64	**38.00**	**36.58**	**34.06**	**35.00**	**57.79**	**54.79**
GPT-4.5-preview	34.28	33.31	31.17	36.50	40.88	52.27	—	—	—	—	—	—
Claude 3.5 Sonnet	26.96	27.49	**22.92**	28.57	32.09	44.09	—	—	—	—	—	—

[1] doubao-seed-1-6-vision-250815
[2] doubao-seed-1-6-251015
[3] Llama-4-Maverick-17B-128E-Instruct
Note: '—' indicates no available data for the corresponding reasoning task.

Table 1: Model Accuracy in Single-Image and Multi-Image Reasoning

Models	Overall	strong	weak	weak_np
Qwen VL-Plus	60.24	54.87	80.13	79.54 (-0.59)
Qwen2.5-VL-7B-Instruct	60.04	54.43	80.70	80.10 (-0.60)
Qianfan-Llama-VL-8B	47.91	41.37	72.17	70.12 (-2.05)
DeepSeek-VL2	36.86	32.24	53.96	42.94 (-11.02)
InternVL3-38B	61.90	58.23	85.79	79.63 (-6.16)
ERNIE-4.5-Turbo-VL	70.64	66.12	88.00	89.10 (+1.10)
doubao-seed-1-6-250615	83.80	80.57	95.33	95.62 (+0.29)
Claude Sonnet 4	26.74	22.46	42.31	62.73 (+20.42)
Gemini 2.5 pro	88.34	86.28	95.85	95.68 (-0.17)

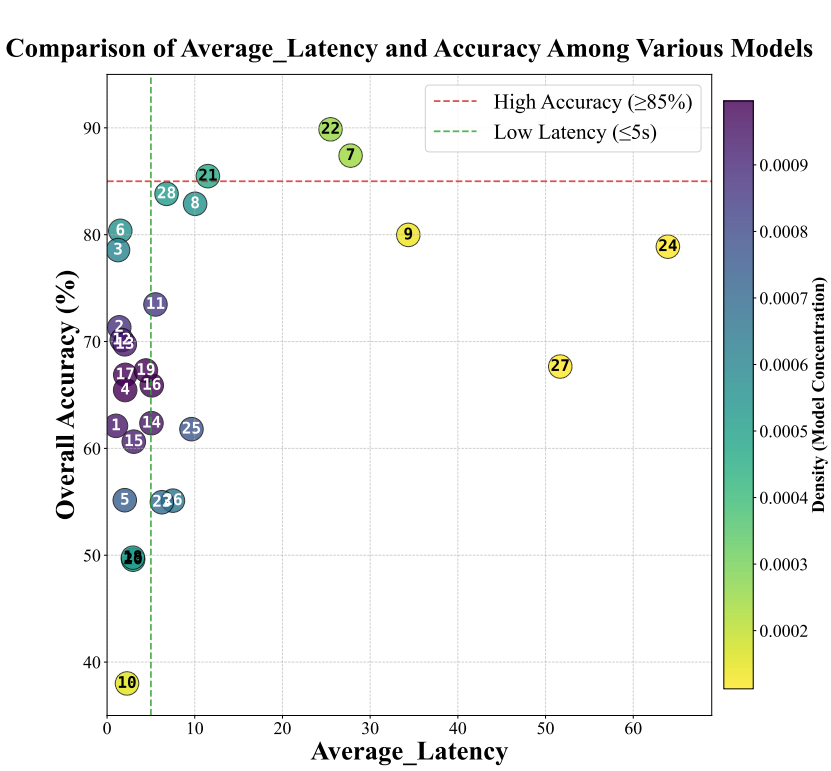
Table 3: Model accuracy rates across different image relevance conditions.

as a disruptive factor. Slightly interfered models (e.g., Qwen VL-Max: +2.68; ERNIE-4.5-Turbo-VL: +1.10) show marginally better performance with text alone, with text capabilities slightly exceeding image-text fusion abilities. Image-independent (difference: -1 to +1) (5 models total): MLLMs exhibit strong spatial imagination with text-only inputa trait particularly prominent in top models (e.g., Gemini 2.5 pro: -0.17; doubao: +0.29), whose differences are near zero, indicating balanced performance.

For more detailed analysis and complete data, see Appendix B.

5.2 Multi-image Reasoning

In the comprehensive multi-image understanding task, top-performing models exhibited a significant performance gap. doubao-seed-1-6-251015, doubao-seed-1-6-vision-250815, and Qwen3-VL-32B-Thinking all achieved an overall score of over 5 points, among which doubao-seed-1-6-vision-250815 reached a comprehensive accuracy rate of as high as 89.96%. This fully demonstrates its absolute competitiveness in physics visual problem reasoning tasks and highlights its core advan-



Number-Model Mapping Table

Number	Model
1	Qwen VL-Plus
2	Qwen VL-Max
3	Qwen2.5-VL-72B-Instruct
4	Qwen2.5-VL-32B-Instruct
5	Qwen2.5-VL-7B-Instruct
6	Qwen3-VL-Plus
7	Qwen3-VL-32B-Thinking
8	Qwen3-VL-30B-A3B-Thinking
9	Qwen3-VL-8B-Thinking
10	DeepSeek-VL2
11	ERNIE 4.5 Turbo VL
12	ERNIE 4.5
13	GLM-4.5V
14	InternVL2.5-38B-MPO
15	InternVL3-14B
16	InternVL3-38B
17	Llama-4-Maverick-17B-128E-Instruct
18	Qianfan-VL-8B
19	Qianfan-VL-70B
20	Qianfan-Llama-VL-8B
21	doubao-seed-1-6-251015
22	doubao-seed-1-6-vision-250815
23	GPT-4o
24	GPT5
25	Claude Sonnet 4
26	Claude Haiku 4.5
27	Grok-4
28	Gemini 2.5 pro

Figure 4: Scatter Plot of Average_Latency vs. Accuracy

tages in multimodal information integration and physical logic analysis.The overall accuracy is presented in Table 1.

Disciplinary Difficulty and Model Adaptability: **Acoustics > Optics > Thermodynamics > Mechanics > Electromagnetism**, which is consistent with the results of single-image reasoning.

Models	SC	MS	F
doubao-vision[1]	**88.00**	78.85	**91.04**
Qwen3-VL-32B-Thinking	86.33	**79.19**	88.16
ERNIE 4.5 Turbo VL	70.12	51.59	75.71
GLM-4.5V	60.54	38.71	74.59
InternVL2.5-38B-MPO	61.08	23.08	64.56
Qianfan-VL-8B	47.74	19.87	51.84
DeepSeek-VL2	**25.80**	**14.94**	**43.65**

[1] doubao-vision is the abbreviation of doubao-seed-1-6-vision-250815.

Table 4: Examples of Model Performance Across Different Question Types in Multi-Image Reasoning.

Contrary to the conventionally accepted question-type difficulty hierarchy, multiple-select questions exhibit the largest difficulty gap in multi-image reasoning tasks, with fill-in-the-blank questions slightly less challenging than single-choice ones. Multiple-select questions remain the definitive high-difficulty category: over 70% of models score below 60% accuracy, the top accuracy only reaches 79.19%, and the accuracy gap between top and bottom models exceeds 60 percentage points, indicating a notable difficulty barrier. Of the 28 tested models, only 5 perform better on single-choice than fill-in-the-blank questions, reflecting that in multi-image scenarios, models are more adept at information extraction for structured fill-in-the-blank tasks than option discrimination for single-choice ones. Table 4 lists partial sample data.

Analysis of Inference Efficiency: The evaluation results reveal a critical performance trade-off: while doubao-seed-1-6-vision-250815 (89.86% accuracy) and Qwen3-VL-32B-Thinking (87.40% accuracy) achieve high precision, their poor inference efficiencywith average per-question response times of 25.48 s and 27.77 s respectivelyrestricts real-world deployment potential. By contrast, Qwen3-VL-Plus delivers a favorable balance of 80.36% accuracy and 1.51 s average latency, demonstrating superior overall cost-effectiveness.The overall accuracy and of average latency are illustrated in Figure 4.

When selecting Vision-Language Models (VLMs), sole pursuit of maximum accuracy is

inadvisable; instead, inference efficiency and practical scenario requirements must be comprehensively weighed. Certain models boost processing speed substantially with minimal accuracy compromises via optimized architectures and efficient reasoning strategies, making them more practically valuable for real-time response scenarios. Balancing accuracy and response time will remain a core direction for future model optimization.

For more detailed analysis and complete data, refer to Appendix C.

5.3 Supplementary Analysis

From the disciplinary perspective, electromagnetism emerges as a universal bottleneck: even in multi-image tasks, the top model doubao-seed-1-6-vision-250815 only achieves 85.51% accuracy in this field, far below its performance in other branches.

In terms of question types, the high difficulty of multiple-select questions demonstrates cross-task consistency. In single-image tasks, over 70% of models score below 60% on these questions with an extreme performance gap; in multi-image tasks, this difficulty barrier is further amplified, with the accuracy difference between top and bottom models exceeding 60 percentage points.

Data-wise, this evaluation is based on a Chinese physics dataset, whose unique linguistic system imposes a notable threshold for multimodal models. Models optimized for Chinese semantics or integrated with large-scale Chinese disciplinary corpora (e.g., the doubao series) achieve precise alignment with domain terminology and physical laws, thus attaining higher accuracy.

Regarding model openness (open-source vs. closed-source), closed-source models show bipolar performance: Gemini 2.5 Pro, doubao-seed-1-6-250615 and doubao-seed-1-6-vision-250815 outperform most open-source models by a significant margin, while Claude Sonnet 4 and GPT-4o lag behind most open-source counterparts. Among open-source models, Qwen3-VL-32B-Thinking stands out with 87.40% accuracy in multi-image reasoning, yet most others remain at the mid-to-lower performance level.

6 Conclusion and Outlook

To evaluate multimodal large language models (MLLMs) on physics reasoning tasks, this paper proposes MV-Physics, a novel benchmark for physics visual question answering scenarios, which comprises a large annotated corpus of Chinese junior high school physics questions for comprehensive assessment of inter-model performance disparities.

Experiments on 22 mainstream MLLMs for single-image reasoning reveal a bipolar performance distribution: Gemini 2.5 Pro and doubao-seed-1-6-250615 demonstrate strong all-around problem-solving capabilities, while many others fail to meet passing scores. An in-depth analysis of performance variations is conducted to explore the mechanisms underlying question-image correlation. For multi-image reasoning tasks, tests on 28 mainstream MLLMs show doubao-seed-1-6-vision-250815 ranks first; however, inference efficiency analysis indicates substantial room for improvement in balancing model accuracy and response time.

Future work will expand MV-Physics in depth and breadth by incorporating senior high school physics questions and complex interdisciplinary scenarios (e.g., physics-mathematics-engineering integrated problems), to better assess MLLMs capabilities in advanced physics knowledge applications.

Limitations

The current MV-Physics benchmark covers only Chinese junior high school physics questions. While effective for basic scenarios, it lacks advanced/senior high school content, interdisciplinary problems (e.g., physics-chemistry-biology) and multilingual support, limiting its ability to evaluate models advanced and cross-lingual physics capabilities.

Experiments only tested mainstream multimodal models, excluding niche, domain-optimized and rapidly updated versions, so the generalization of conclusions needs further verification.

Ethical Consideration

Copyright and Licensing: All data in MV-Physics are collected from public sources.

Ethics and Data Privacy: All testing instances in MV-Physics are carefully scrutinized to exclude any examples with ethical concerns. Since all the data are collected from exampapers there is no privacy issue.

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-07-05.

Anthropic. 2025. Claude haiku 4.5. Hybrid reasoning large language model supporting text and image inputs, with a knowledge cutoff date of February 2025. Accessible via Claude.ai, Anthropic API, Amazon Bedrock and Google Vertex AI.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baidu Qianfan Team. 2025a. Qianfan-vl-70b visual-language model. Model ID: am-xe9vyup68g5b; 16K+ context window, Chinese-oriented, trained on Kunlun chips. Baidu Qianfan Model Center.

Baidu Qianfan Team. 2025b. Qianfan-vl-8b visual-language model. Model ID: am-xe9vyup68g5b; 16K+ context window, Chinese-oriented, trained on Kunlun chips. Baidu Qianfan Model Center.

ByteDance. 2024. doubao. https://www.bytedance.com/. Accessed: 2025-7-18.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara, and 1 others. 2024. The (r)evolution of multimodal large language models: A survey.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and Others. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Hongqing Chu, Dele Meng, Siwen Huang, Mengjian Tian, Jia Zhang, Bingzhao Gao, and Hong Chen. 2024. Autonomous high-speed overtaking of intelligent chassis using fast iterative model predictive control. *IEEE Transactions on Transportation Electrification*, pages 1244–1256.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. 2025. Theoretical physics benchmark (tpbench) a dataset and study of ai reasoning capabilities in theoretical physics.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7746–7762.

DeepSeek-AI DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint*, arXiv:2503.02199.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025. Physics: Benchmarking foundation models on university - level physics problem solving. *arXiv preprint arXiv:2503.21821*.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? : An enhanced multimodal reasoning benchmark. *Preprint*, arXiv:2501.05444. Project page: https://emma-benchmark.github.io/.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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 3789–3805, Bangkok, Thailand. Association for Computational Linguistics.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Janak Kapuriya, Chhavi Kirtani, Apoorv Singh, Jay Saraf, Naman Lal, Jatin Kumar, AdarshRaj Shivam, Astha Verma, Avinash Anand, and RajivRatn Shah. 2025. Mm-phyrlhf: Reinforcement learning framework for multimodal physics question-answering. *Preprint*, arXiv:2404.12926. PDF direct download: https://arxiv.org/pdf/2404.12926, Language: English.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, Yining Hua, and Alibaba Group. Qilin-med-vl: Towards chinese large vision-language model for general healthcare.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Matthew T. Mason, J. Kenneth Salisbury, and Joey K. Parker. 1989. Robot hands and the mechanics of manipulation. *Journal of Dynamic Systems, Measurement, and Control*, pages 119–119.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning.

Meta. 2025. Llama-4-maverick-17b-128e-instruct. 400B parameter multimodal model with mixture-of-experts architecture, supporting multilingual text and image inputs, and applicable for commercial and research use.

OpenAI. 2023. Gpt-4v(vision): System card. *OpenAI Technical Report*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o.

OpenAI. 2025. Gpt-5: Next-generation general-purpose large language model. Advanced multimodal model with enhanced reasoning, long-context understanding, and real-time knowledge integration, accessible via ChatGPT and OpenAI API.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and 1 others. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744.

Pengzhen Ren, Kaidong Zhang, Hetao Zheng, Zixuan Li, Yuhang Wen, Fengda Zhu, Mas Ma, and Xiaodan Liang. 2024. Damworld: Progressive reasoning with world models for robotic manipulation.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. 2025. Phyx: Does your model have the "wits" for physical reasoning? *arXiv preprint arXiv:2505.15929*. Project page: https://phyx-benchmark.github.io.

Qiu Shi, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, and 33 others. 2025. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *Preprint*, arXiv:2504.16074. PDF direct download: https://arxiv.org/pdf/2504.16074; Code repository: https://github.com/phybench-official/phybench, 500 original physics problems covering high school to olympiad difficulty.

Anthropic Team. 2025a. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic, Inc. Accessed: 2025-07-03.

Baidu Qianfan Team. 2024a. Qianfan-check-vl: Built-in image quality detection model (e-commerce scenario) on baidu qianfan platform. Accessed from Baidu Qianfan Model Center (Built-in Models List); Focus on e-commerce image quality inspection.

Baidu Qianfan Team. 2024b. Qianfan-llama-vl-8b: Built-in visual-language model on baidu qianfan platform. Accessed from Baidu Qianfan Model Center (Built-in Models List).

Baidu Qianfan Team. 2024c. Qianfan-qi-vl: Built-in visual inspection model (image quality) on baidu qianfan platform. Accessed from Baidu Qianfan Model Center (Built-in Models List); Supports AIGC defect detection, watermark recognition, etc.

Baidu Research Team. 2024d. Ernie-4.5 technical report. Technical report, Baidu, Inc. Accessed: 2025-07-03.

Gemini Team. 2025b. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google. Accessed: 2025-07-03.

Gemini Team and Google Google. Gemini: A family of highly capable multimodal models.

Gemini Team and Others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. Accessed: 2025-07-01.

OpenAI Team. 2025c. Introducing gpt-4.5: Advanced language and vision capabilities. Official release page for GPT-4.5, including capability descriptions and access details.

Qwen Team. 2025d. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Guiyao Tie, Xueyang Zhou, Tianhe Gu, Ruihang Zhang, Chaoran Hu, Sizhe Zhang, Mengqu Sun, Yan Zhang, Pan Zhou, and Lichao Sun. 2025. Mmlu-reason: Benchmarking multi-task multi-modal language understanding and reasoning. *arXiv preprint arXiv:2505.16459*.

Peijie Wang, Zhongzhi Li, Fei Yin, Dekang Ran, and Chenglin Liu. 2025. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Data and code available at: https://eternal8080.github.io/MV-MATH.github.io/.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

xAI. 2025. Grok 4. `https://x.ai/news/grok-4`. Accessed: 2025-09-08.

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. 2025. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*, San Diego, USA. Neural Information Processing Systems Foundation.

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. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. 5,520 problems (English/Chinese); MARJ evaluation pipeline; Codes and data available at https URL.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 24 others. 2024. Qwen2.5 technical report.

Xiangchen Yue, Zhang Zheng, Xiaoding Zhang, Yuxuan Huang, Yijiang Xu, Haotian Wang, Qing Liu, Weitao Chen, Wenhai Zhang, Xizhou Zhu, and 1 others. 2024. 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 (CVPR)*, pages 9556–9567.

Zhipu AI Zai Org. 2025. Glm-4.5v. Open-source visual reasoning model with MIT license, which refreshes 41 SOTA records in multimodal reasoning and supports commercial use.

Ge Zhang, Haonan Li, Swaroop Mishra, Yiming Yu, Pengfei Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Proceedings of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

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. *Preprint*, arXiv:2502.12054. 1,200 problems (25% knowledge-based, 75% reasoning-based); Physics Solution Auto Scoring Framework; Version v2; Code/data: https://dxzxy12138.github.io/PhysReason.

A API Summary Table for Each Model

Tables 5 and 6 present the APIs invoked by all evaluated large models.

API Platform	Models & API Endpoint
Alibaba DashScope	Models: Qwen-VL-Plus; Qwen-VL-Max; Qwen2.5-VL-72B-Instruct Endpoint: `https://dashscope.aliyuncs.com/compatible-mode/v1`
Baidu Qianfan	Models: Qwen2.5-VL-32B-Instruct; Qwen2.5-VL-7B-Instruct; Qianfan-Llama-VL-8B; Qianfan-Check-VL; Qianfan-QI-VL DeepSeek-VL2; InternVL3-38B; InternVL3-14B; InternVL2.5-38B-MPO; ERNIE-4.5-Turbo-VL; ERNIE-4.5 Endpoint: `https://qianfan.baidubce.com/v2/chat/completions`
Yunwu AI	Models: doubao Seed 1.6; GPT-4o; GPT-4.5-preview; Claude 3.5 Sonnet Claude Sonnet 4; Grok-4 Endpoint: `https://yunwu.ai/v1`
Shubiaobiao	Models: Gemini 2.5 pro; Gemini 1.5 Pro Endpoint: `https://api.shubiaobiao.cn/v1`

Table 5: Summary of API Platforms, Corresponding Models, and Endpoints in Single-image Reasoning.

API Platform	Models & API Endpoint
Alibaba DashScope	Models: Qwen VL-Plus; Qwen VL-Max; Qwen2.5-VL; Qwen3-VL-Plus; Qwen3-VL-32B-Thinking; Qwen3-VL-30B-A3B-Thinking; Qwen3-VL-8B-Thinking Endpoint: https://dashscope.aliyuncs.com/compatible-mode/v1
Baidu Qianfan	Models: DeepSeek-VL2; ERNIE 4.5 Turbo VL; ERNIE 4.5; GLM-4.5V; InternVL2.5-38B-MPO; InternVL3-14B; InternVL3-38B llama-4-maverick-17b-128e-instruct; Qianfan-VL-8B; Qianfan-VL-70B; qianfan-llama-vl-8b Endpoint: https://qianfan.baidubce.com/v2/chat/completions
Volces Ark	Models: doubao-seed-1-6-251015; doubao-seed-1-6-vision-250815 Endpoint: https://ark.cn-beijing.volces.com/api/v3
Yunwu AI	Models: GPT-4o; GPT5; Claude Sonnet 4; Claude Haiku 4.5; Grok-4 Endpoint: https://yunwu.ai/v1
Shubiaobiao	Models: Gemini 2.5 pro Endpoint: https://api.shubiaobiao.cn/v1

Table 6: Summary of API Platforms, Corresponding Models, and Endpoints in Multi-image Reasoning.

B More Information on Evaluation Results in the Single-Image Reasoning Scenario

B.1 Performance by Discipline

(1) Mechanics:

Top-tier models: Gemini 2.5 pro (87.31%), Doubao (82.70%), with scores exceeding 80;

Mid-tier models: ERNIE-4.5-Turbo-VL (69.11%), Qwen2.5-VL-72B (69.86%), with scores ranging from 65 to 70;

Lower-performing models: Claude 3.5 Sonnet (27.49%), GPT-4.5-preview (33.31%), with scores below 35, showing poor foundational knowledge in mechanics.

Disciplinary characteristics: Mechanics is a fundamental science discipline. Top-tier models hold obvious advantages, while lower-scoring models lack mastery of core knowledge such as force analysis and motion laws.

(2) Electromagnetism:

Top-tier models: Gemini 2.5 pro (86.56%), Doubao (81.30%), with scores exceeding 80, indicating a profound understanding of complex electromagnetism laws (e.g., electromagnetic fields, circuit analysis);

Second-tier models: ERNIE-4.5-Turbo-VL (67.68%), Qwen2.5-VL-72B (66.69%), with scores ranging from 65 to 70, performing slightly weaker in high-difficulty tasks (e.g., comprehensive problems on electromagnetic induction); Lower-performing models: Claude Sonnet 4 (24.12%), DeepSeek-VL2 (31.40%), with scores below 35, showing vague understanding of core electromagnetism concepts.

Disciplinary characteristics: Electromagnetism is a relatively difficult field in science, and advantages are concentrated among top-tier models.

(3) Optics:

Top-tier models: Gemini 2.5 pro (91.97%), Doubao (88.54%), with scores exceeding 85, leading the second tier by more than 10 points and demonstrating accurate mastery of optical laws such as refraction, reflection, and wave phenomena;

Mid-tier models: Qwen2.5-VL-72B (79.39%), ERNIE-4.5-Turbo-VL (78.70%), with scores ranging from 75 to 80;

Lower-performing models: GPT-4.5-preview (36.50%), Claude 3.5 Sonnet (28.57%), with scores below 30, having obvious knowledge gaps in optics.

Disciplinary characteristics: Optics is the field where top-tier models show the most prominent advantages (Gemini scores over 90), possibly because optical laws are highly intuitive, making it easier for models to learn through images/data.

(4) Thermodynamics:

Top-tier models: Gemini 2.5 pro (90.48%), Doubao (86.82%), with scores exceeding 85, showing solid mastery of thermodynamic laws and temperature changes;

Second-tier models: InternVL3-38B (76.58%), Qwen2.5-VL-32B (77.51%), with scores ranging from 75 to 80, demonstrating good stability;

Lower-performing models: DeepSeek-VL2 (42.94%), Qianfan-QI-VL (51.02%), with scores below 55, showing weak foundational knowledge in thermodynamics.

Disciplinary characteristics: Thermodynamics is of moderate difficulty. Top-tier models have obvious advantages, mid-tier models perform stably, and lower-scoring models lack understanding of abstract concepts (e.g., entropy change, heat transfer).

(5) Acoustics:

Top-tier models: Gemini 2.5 pro (98.00%), Doubao (94.63%), with scores exceeding 90,

achieving nearly perfect mastery in this fieldtheir performance here is the best among all disciplines;

Second-tier models: Qwen2.5-VL-72B (90.13%), InternVL3-38B (85.71%), with scores ranging from 85 to 90;

Lower-performing models: Claude Sonnet 4 (43.44%), GPT-4.5-preview (52.27%), with scores below 55, showing insufficient knowledge reserves in acoustics.

Disciplinary characteristics: Acoustics is the least difficult discipline overall. Top-tier models score close to full marks, reflecting their excellent learning outcomes regarding basic physical phenomena in acoustics (e.g., the generation and propagation of sound).

B.2 Performance Analysis by Question Type

The accuracy of all models across different question types is presented in the Table 7.

Models	SC	MS	F
Qwen-VL-Plus	68.71	23.79	61.74
Qwen-VL-Max	75.69	52.42	62.75
Qwen2.5-VL-72B-Instruct	75.67	50.27	68.05
Qwen2.5-VL-32B-Instruct	70.92	46.96	66.28
Qwen2.5-VL-7B-Instruct	65.74	29.11	62.24
Qianfan-Llama-VL-8B	51.42	24.46	50.02
Qianfan-Check-VL	54.35	18.21	44.47
Qianfan-QI-VL	45.09	0.54	45.43
DeepSeek-VL2	34.57	15.36	42.06
InternVL3-38B	68.74	48.93	64.08
InternVL3-14B	65.41	41.61	59.87
InternVL2.5-38B-MPO	65.25	40.00	59.77
ERNIE-4.5-Turbo-VL	77.34	44.55	71.51
ERNIE-4.5	72.98	40.71	66.62
doubao Seed 1.6	89.37	76.13	81.93
GPT-4o	51.50	27.91	54.34
GPT-4.5-preview	44.92	13.93	33.67
Claude 3.5 Sonnet	51.27	12.14	24.76
Claude Sonnet 4	50.08	11.61	24.88
Gemini 2.5 pro	90.77	83.61	87.77
Gemini 1.5 Pro	57.62	27.50	80.70
Grok-4	41.63	14.46	37.33

Table 7: Total Accuracy of Models Across Different Question Types in Single-Image Reasoning..

Single-Choice Questions: Option Prompts Reduce Difficulty, Top-Tier Models Demonstrate Concentrated Advantage

Core Characteristic:The "choose one out of four" nature of single-choice questions reduces the demand for "unaided reasoning" capability from the models, making them the easiest question type for models to tackle successfully.

Top-Tier Models: Gemini 2.5 pro leads with an accuracy of 90.77%, followed closely by Doubao Seed 1.6 at 89.37%. Their scores far exceed those of other models, with a gap of over 12 percentage points compared to the second runner-up (ERNIE-4.5-Turbo-VL, 77.34%), demonstrating an absolute advantage in the ability for "single-option judgment + interference option exclusion".

Bottom-Tier Models: DeepSeek-VL2 (34.57%) and Grok-4 (41.63%) scored the lowest, even falling below 1.5-1.7 times the random probability rate (25%), indicating an extremely weak understanding of the basic logic and option associations in single-choice questions.

Pattern: Model performance shows a positive correlation with parameter count (e.g., Qwen2.5-VL series: 72B (75.67%) > 32B (70.92%) > 7B (65.74%)). Larger models exercise more precise control over details in single-choice questions.

Multiple-Select Questions: Associative Judgment of Multiple Options Becomes a Bottleneck, Revealing Extreme Model Disparities

Core Characteristic:Multiple-Select questions require "identifying all correct options," necessitating simultaneous handling of rules such as "logical relationships between options" and "loss of points for both incorrect selections and omissions." This places the highest demand on model rigor and multi-dimensional reasoning ability, making it the "gold standard" question type for differentiating model performance.

Top-Tier Models: Gemini 2.5 pro holds a significant lead with an accuracy of 83.61%, while Doubao Seed 1.6 follows in second place with 76.13%. They are the only two models exceeding 70% accuracy in multiple-select questions, reflecting their extremely strong capability for logical association judgment regarding "multiple correct options."

Bottom-Tier Models: Qianfan-QI-VL ranks last with an accuracy of merely 0.54% (almost entirely incorrect). The Claude series (11.61%-12.14%) and GPT-4.5-preview (13.93%) all scored below 15%, indicating a complete failure to understand the logic of "multiple correct options coexisting" in multiple-select questions, frequently resulting in "omissions and misselections."

Notable Phenomenon: Mid-tier models (e.g., Qwen2.5-VL-72B, InternVL3-38B) have multiple-select scores concentrated between 40%-50%, residing in a stage where they "can identify some correct options but cannot get all correct," reflecting insufficient mastery over "option completeness."

Fill-in-the-Blank Questions: Option-Free Design Tests Direct Output Capability and Uncovers Model Biases

Core Characteristic: Without option prompts, fill-in-the-blank questions require models to generate direct answers (e.g., terms, numerical values, short phrases), demanding strong capabilities in accurate knowledge recall and question-stem logic conversion. Their accuracy typically falls between that of single-choice and multiple-select questions.

Top-Tier Models: Gemini 2.5 Pro (87.77%) and Doubao Seed 1.6 (81.93%) retain the top two positions, with a narrow gap relative to the third-place model (Gemini 1.5 Pro, 80.70%), demonstrating stable performance in generating correct answers without aids.

Special Case Model: Gemini 1.5 Pro shows performance bias: its fill-in-the-blank score (80.70%, second only to top models) far outpaces its single-choice (57.62%) and multiple-select (27.50%) accuracy, indicating strengths in direct answer generation but weaknesses in option judgmentpotentially linked to the models adaptation to varying output formats (options vs. direct text).

Bottom-Tier Models: Claude 3.5 Sonnet (24.76%) and Claude Sonnet 4 (24.88%) score the lowest, reflecting weak ability to convert question-stem information into accurate direct answers and failure to extract key input details effectively.

B.3 Question-Image Correlation

Models	Overall	strong	weak	weak-np
Qwen-VL-Plus	60.24	54.87	80.13	79.54 (-0.59)
Qwen-VL-Max	65.98	61.16	83.87	86.55 (+2.68)
Qwen2.5-VL-72B-Instruct	68.78	63.71	87.59	85.38 (-2.21)
Qwen2.5-VL-32B-Instruct	65.86	60.38	86.21	82.00 (-4.21)
Qwen2.5-VL-7B-Instruct	60.04	54.43	80.70	80.10 (-0.60)
Qianfan-Llama-VL-8B	47.91	41.37	72.17	70.12 (-2.05)
Qianfan-Check-VL	45.05	38.69	68.65	38.40 (-30.25)
Qianfan-QI-VL	40.77	36.10	58.22	40.05 (-18.17)
DeepSeek-VL2	36.86	32.24	53.96	42.94 (-11.02)
InternVL3-38B	61.90	58.23	85.79	79.63 (-6.16)
InternVL3-14B	59.86	53.78	82.40	79.75 (-2.62)
InternVL2.5-38B-MPO	59.59	53.34	82.72	44.70 (-38.02)
ERNIE-4.5-Turbo-VL	70.64	66.12	88.00	89.10 (+1.10)
ERNIE-4.5	67.86	61.23	84.14	81.74 (-2.40)
doubao Seed 1.6	83.80	80.57	95.33	95.62 (+0.29)
GPT-4o	50.73	45.17	71.15	71.11 (-0.04)
GPT-4.5-preview	34.28	30.16	49.34	44.24 (-5.10)
Claude 3.5 Sonnet	26.96	22.75	42.22	63.16 (+20.94)
Claude Sonnet 4	26.74	22.46	42.31	62.73 (+20.42)
Gemini 2.5 pro	88.34	86.28	95.85	95.68 (-0.17)
Gemini 1.5 Pro	66.93	62.14	84.41	77.02 (-7.39)
Grok-4	38.25	32.09	35.77	74.73 (+38.96)

Table 8: Total Accuracy Table of Correlation Between Questions and Images.

The total accuracy of all models regarding question-image correlation is presented in Table 8. Strong Scenario (Image as Essential Information):

Top Performance: Gemini 2.5 pro (86.28) and doubao Seed 1.6 (80.57) significantly outperformed other models, indicating their exceptional ability to understand and integrate images serving as core information.

Poor Performance: The Claude series (22.46-22.75) and Grok-4 (32.09) scored below 30, revealing significant shortcomings in handling tasks where the image is essential.

Weak Scenario (Image as Auxiliary Information):

Top Performance: Gemini 2.5 pro (95.85) and doubao (95.33) achieved near-perfect scores, demonstrating highly efficient integration of textual and visual information for problem-solving with image assistance.

Poor Performance: Grok-4 (35.77) and the Claude series (42.22-42.31) scored below 50, indicating weak performance even with the aid of images.

Weak Scenario with Image Removed (Text-Only):

Top Performance: Doubao (95.62), Gemini 2.5 pro (95.68), and ERNIE-4.5-Turbo-VL (89.10) maintained text-only performance close to their weak scenario scores, indicating extremely strong text comprehension capabilities.

Anomalous Performance: InternVL2.5-38B-MPO (44.70) experienced a drastic drop of 38.02 points compared to its weak scenario performance (82.72), making it the only model scoring below 50 in the text-only condition and revealing an extreme dependence on auxiliary images. Conversely, Grok-4 (74.73) saw a sharp increase of 38.96 points compared to its weak scenario performance (35.77), performing far better with text alone, indicating that images act as a severe interference for this model.

Summary of Representative Model Characteristics:

Gemini 2.5 pro:Top-tier across all scenarios (Strong 86.28, Weak 95.85, No Image 95.68). Balances image-text integration and pure text capabilities with no obvious weaknesses, making it the model with optimal comprehensive performance.

Doubao Seed 1.6:Performance in the Weak scenario (95.33) is nearly identical to the No Image scenario (95.62). Pure text capability is slightly stronger than image-text integration, indi-

cating extremely high text comprehension precision. ERNIE-4.5-Turbo-VL: Performance slightly improved (+1.10) after image removal, highlighting its strong pure text capability and suitability for scenarios without images.

InternVL2.5-38B-MPO: Exhibits extreme dependence on auxiliary images. Performance collapses without images, indicating a critical deficiency in text-only capability.

Grok-4:Experiences the most severe image interference. Its pure text capability (74.73) far exceeds its performance in image-included scenarios (35.77), suggesting a failure in its image-text fusion mechanism.

Core Conclusions:

Top models (Gemini 2.5 pro, doubao) achieve dual strength in both "image-text integration" and "pure text" processing, demonstrating balanced and stable capabilities.

Most models depend on auxiliary images (negative difference value), but the degree of dependency varies significantly. Models with extreme dependency require focused optimization of their text-based capabilities.

A few models (Grok-4, Claude series) are interfered with by images. Their image-text information filtering mechanisms need improvement to prevent invalid images from negatively impacting judgment.

Scores in the Strong scenario are generally lower than in the Weak scenario, indicating that models find processing tasks where the image is essential more challenging. This represents a common direction for future optimization.

Cross-Scenario Core Conclusion: The Strong Scenario Acts as a Magnifying Glass for Models' "True Capabilities"

The "Stability Barrier" of Top-Tier Models: Across all question types, the performance gap between top-tier models (Gemini 2.5 pro, Doubao) in the general scenario and the Strong scenario remains stable, with minimal decay (most differences < 5 points), even achieving "zero decay" (e.g., Gemini 2.5 pro in fill-in-the-blank). This stability allows them to widen the gap with other models in both scenarios, consolidating their leading positions.

The "Scenario Dependency Dilemma" of Mid-Tier Models: Mid-tier models generally experience decay in the Strong scenario (especially in fill-in-the-blank, difference of 5-7 points), leading to an expanded gap with the top tier. This reflects that their "ability to utilize image information" is a weakness, and they can only maintain competitiveness in the general scenario (where images are auxiliary).

The "Solidification of Dual Disadvantages" for Bottom-Tier Models:Bottom-tier models perform poorly in both scenarios, and their decay in the Strong scenario is slightly greater than that of the top tier. This leads to a further increase in the absolute performance gap with the top tier (particularly evident in single-choice and fill-in-the-blank questions), indicating that they not only have weak foundational capabilities but also lack "anti-interference" or "image information utilization" abilities. This gap appears "irreversible."

Tables 9 and 10 respectively present the comparison between the overall accuracy and the accuracy under strong correlation of models across different disciplinary branches and question types.

Single-Choice Questions: Significant Gap in General Scenario, Slightly Narrowed Gap in Strong Scenario, Top-Tier Model Stability Dominates

1. General Scenario (Single-Choice)

Highest Score (Gemini 2.5 pro): 90.77%; Lowest Score (DeepSeek-VL2): 34.57% Maximum Gap: 56.2 points. The average score of the top 5 models (80.4%) was 35.7 points higher than that of the bottom 5 models (44.7%).

Characteristic: The gap between top-tier models (Gemini 2.5 pro, doubao) and mid-tier models (e.g., ERNIE-4.5-Turbo-VL) exceeds 12 points, and the gap between mid-tier and bottom-tier models exceeds 25 points, indicating a clear hierarchy.

2. Strong Scenario (Single-Choice Strong) Highest Score (Gemini 2.5 pro): 89.32%; Lowest Score (DeepSeek-VL2): 30.98% Maximum Gap: 58.34 points (slightly larger than the general scenario). The average score of the top 5 models (78.1%) was 36.9 points higher than that of the bottom 5 models (41.2%).

Characteristic: Top-tier models show minimal decay (Gemini 2.5 pro difference 1.45 points, Doubao 2.32 points) and maintain a gap of over 10 points with mid-tier models. Bottom-tier models decay slightly more than the top tier (e.g., DeepSeek-VL2 difference 3.59 points, Claude 3.5 Sonnet 6.21 points), leading to a slight widening of the absolute gap between the top and bottom tiers in the Strong scenario.

3. Scenario Comparison: Strong Gap Stability, Top-Tier "Anti-Decay" Ability Determines Hierar-

Models	Mech		Electrom		Opt		Thermo		Acou	
	all	str	all	str	all	str	all	str	all	str
Qwen-VL-Plus	60.1	53.3	52.8	51.1	71.4	64.5	68.0	64.4	74.7	63.4
Qwen-VL-Max	66.5	60.2	59.5	58.1	75.3	69.0	71.2	67.1	78.7	71.8
Qwen2.5-VL-72B	69.9	62.9	66.7	60.3	79.4	72.6	75.9	69.0	90.1	80.3
Qwen2.5-VL-32B	64.2	58.1	58.1	56.3	77.7	69.5	77.5	73.7	87.1	74.7
Qwen2.5-VL-7B	58.1	51.6	52.3	50.6	74.8	66.1	70.5	67.0	70.1	54.9
QF-Llama-VL-8B	44.4	36.4	41.0	38.9	64.6	55.9	56.0	50.8	67.4	47.9
QF-Check-VL	44.1	36.0	37.4	35.6	57.7	47.8	54.6	49.9	58.5	43.7
QF-QI-VL	38.1	32.4	35.9	34.5	49.6	41.6	51.0	47.0	57.1	49.3
DeepSeek-VL2	34.1	29.2	31.4	29.5	48.5	42.3	42.9	39.3	61.6	50.0
InternVL3-38B	63.8	57.2	55.8	53.9	73.5	63.2	76.6	73.0	85.7	74.7
InternVL3-14B	59.4	53.0	51.2	49.1	70.7	59.8	72.5	68.4	81.0	67.6
InternVL2.5-38B	59.4	51.9	52.1	50.1	67.6	56.7	70.8	66.7	82.3	66.2
ERNIE-4.5-Turbo	69.1	62.9	67.7	66.1	78.7	71.6	73.7	69.5	80.7	71.8
ERNIE-4.5	62.5	57.1	62.0	60.2	77.6	69.7	72.9	68.4	80.3	66.2
doubao Seed 1.6	82.7	78.0	81.3	80.4	88.5	83.4	86.8	84.5	94.6	91.4
GPT-4o	48.1	42.0	44.9	42.8	58.9	49.0	62.8	58.2	77.0	66.7
GPT-4.5-preview	33.3	27.7	31.2	29.6	36.5	32.2	40.9	36.4	52.3	39.7
Claude 3.5	27.5	21.9	22.9	21.2	28.6	22.7	32.1	82.8	44.1	39.7
Claude Sonnet 4	25.7	20.4	24.1	22.5	28.5	21.2	31.9	28.3	43.4	30.9
Gemini 2.5 pro	87.3	84.3	86.6	86.0	92.0	89.4	90.5	88.7	98.0	95.7
Gemini 1.5 Pro	64.7	58.7	63.8	62.5	76.8	67.2	69.8	66.5	81.9	75.4
Grok-4	35.3	29.2	32.0	30.5	42.1	37.8	44.6	40.2	52.7	42.0

Table 9: Accuracy of Various Models Across Different Disciplinary Branches (all=all scenarios, str=strong scenarios).

Models	SC		MS		F	
	all	strong	all	strong	all	strong
Qwen-VL-Plus	68.71	64.83	23.79	21.73	61.74	55.76
Qwen-VL-Max	75.69	71.57	52.42	50.50	62.75	57.24
Qwen2.5-VL-72B-Instruct	75.67	71.66	50.27	47.89	68.05	62.26
Qwen2.5-VL-32B-Instruct	70.92	66.78	46.96	44.27	66.28	59.91
Qwen2.5-VL-7B-Instruct	65.74	61.27	29.11	27.77	62.24	55.87
Qianfan-Llama-VL-8B	51.42	46.48	24.46	23.94	50.02	41.92
Qianfan-Check-VL	54.35	48.41	18.21	17.10	44.47	37.44
Qianfan-QI-VL	45.09	41.80	0.54	0.60	45.43	40.00
DeepSeek-VL2	34.57	30.98	15.36	15.49	42.06	36.43
InternVL3-38B	68.74	63.71	48.93	47.48	64.08	57.21
InternVL3-14B	65.41	60.43	41.61	39.24	59.87	52.84
InternVL2.5-38B-MPO	65.25	60.22	40.00	38.23	59.77	52.39
ERNIE-4.5-Turbo-VL	77.34	74.23	44.55	43.24	71.51	65.94
ERNIE-4.5	72.98	69.25	40.71	39.44	66.62	61.01
doubao Seed 1.6	89.37	87.05	76.13	75.67	81.93	77.81
GPT-4o	51.50	46.47	27.91	27.02	54.34	48.11
GPT-4.5-preview	44.92	41.12	13.93	13.48	33.67	29.22
Claude 3.5 Sonnet	51.27	45.06	12.14	11.87	24.76	20.72
Claude Sonnet 4	50.08	44.35	11.61	11.47	24.88	20.58
Gemini 2.5 pro	90.77	89.32	83.61	83.16	87.77	87.77
Gemini 1.5 Pro	57.62	52.64	27.50	26.36	80.70	76.49
Grok-4	41.63	38.25	14.46	14.08	37.33	32.13

Table 10: Comparison Between Overall Scores and Strong Scores Across Various Question Types.

Models	SC		MS		F	
	all	strong	all	strong	all	strong
Qwen2.5-VL-7B-Instruct	65.74	61.27 (-4.47)	29.11	27.77 (-1.34)	62.24	55.87 (-6.37)
InternVL3-14B	65.41	60.43 (-4.98)	41.61	39.24 (-2.37)	59.87	52.84 (-7.03)
ERNIE-4.5	72.98	69.25 (-3.73)	40.71	39.44 (-1.27)	66.62	61.01 (-5.61)
doubao Seed 1.6	89.37	87.05 (-2.32)	76.13	75.67 (-0.46)	81.93	77.81 (-4.12)
GPT-4o	51.50	46.47 (-5.03)	27.91	27.02 (-0.89)	54.34	48.11 (-6.23)
Claude 3.5 Sonnet	51.27	45.06 (-6.21)	12.14	11.87 (-0.27)	24.76	20.72 (-4.04)
Claude Sonnet 4	50.08	44.35 (-5.73)	11.61	11.47 (-0.14)	24.88	20.58 (-4.30)
Gemini 2.5 pro	90.77	89.32 (-1.45)	83.61	83.16 (-0.45)	87.77	87.7 (0)

Table 11: Example of Accuracy Differences Between Overall Accuracy and Strong Accuracy Across Various Question Types.

chy

The hierarchical ranking of models remains almost consistent (Top Mid Bottom) across both scenarios, indicating that the performance gap in single-choice questions is not affected by the scenario and is primarily determined by foundational capability. The Strong scenario only amplifies the "disadvantage of bottom-tier models" (greater decay at the bottom), while the strong "anti-decay" capability of top-tier models further consolidates their leading position.

Multiple-Select Questions: Extreme Gap in General Scenario, Slightly Narrowed Gap in Strong Scenario, Top-Tier "Zero Decay" Consolidates Advantage 1.General Scenario (Multiple-Select)

Highest Score (Gemini 2.5 pro): 83.61%; Lowest Score (Qianfan-QI-VL): 0.54Maximum Gap: 83.07 points. The average score of the top 5 models (60.5%) was 46.1 points higher than that of the bottom 5 models (14.4%), representing the largest gap among all question types.

Characteristic: The top two models (Gemini 2.5 pro, Doubao) lead the third place by over 25 points. Bottom-tier models are nearly ineffective (average score < 15%).

2.Strong Scenario (Multiple-Select Strong)

Highest Score (Gemini 2.5 pro): 83.16%; Lowest Score (Qianfan-QI-VL): 0.60Maximum Gap: 82.56 points (slightly smaller than the general scenario). The average score of the top 5 models (59.8%) was 45.5 points higher than that of the bottom 5 models (14.3%).

Characteristic: Decay of top-tier models approaches zero (Gemini 2.5 pro difference 0.45 points, Doubao 0.46 points), and the gap with mid-tier models remains stable. Isolated bottom-tier models show slight improvement (e.g., DeepSeek-VL2 difference -0.13 points), but due to their extremely low baseline scores (15%), the impact on the overall gap is negligible.

3.Scenario Comparison: Gap Core Determined by "Absolute Advantage of Top Tier," Strong Scenario Fails to Alter Hierarchy

The gap between the top and bottom tiers exceeds 80 points in both scenarios, stemming primarily from the top-tier models' dual capability in processing "multiple options + image information," which bottom-tier models completely lack. The slight narrowing of the gap in the Strong scenario is merely due to random fluctuations in a few bottom-tier models and does not change the fundamental nature of "top-tier monopoly and bottom-tier failure."

Fill-in-the-Blank Questions: Significant Gap in General Scenario, Widened Gap in Strong Scenario, Image Dependency Amplifies Model Divergence

1.General Scenario (Fill-in-the-Blank)

Highest Score (Gemini 2.5 pro): 87.77%; Lowest Score (Claude 3.5 Sonnet): 24.76%

Maximum Gap: 63.01 points. The average score of the top 5 models (76.8%) was 41.8 points higher than that of the bottom 5 models (35.0%).

Characteristic: The top "three-high camp" (Gemini 2.5 pro, Doubao, Gemini 1.5 Pro) maintains a gap of over 9 points with mid-tier models, and the gap between mid-tier and bottom-tier models exceeds 25 points.

2.Strong Scenario (Fill-in-the-Blank Strong)

Highest Score (Gemini 2.5 pro): 87.77% (same as general scenario); Lowest Score (Claude 3.5 Sonnet): 20.72%

Maximum Gap: 67.05 points (4.04 points wider than the general scenario). The average score of the top 5 models (72.7%) was 41.6 points higher than that of the bottom 5 models (31.1%).

Characteristic: Among top-tier models, only

Gemini 2.5 pro achieved "zero decay" (difference 0 points). Doubao decayed by 4.12 points but still maintained a gap of over 15 points with mid-tier models. Mid-tier models decayed significantly (e.g., Qwen2.5-VL-32B difference 6.37 points, InternVL3-14B 7.03 points). Bottom-tier models decayed by around 4 points, leading to a widening gap between the top and mid-tiers, and between the mid and bottom tiers.

3.Scenario Comparison: Image Dependency Amplifies Gaps, "Zero Decay" Models Become Absolute Leaders

The gap widens in the Strong scenario for fill-in-the-blank questions primarily because mid-tier models lack sufficient capability for "precise extraction of image information" (decay 5-7 points), while top-tier models (especially Gemini 2.5 pro) can fully utilize image information, further widening the gap. Bottom-tier models, due to a "lack of dual text-image capabilities," perform extremely poorly in both scenarios, with a gap exceeding 60 points compared to the top tier, which is further exacerbated in the Strong scenario.

Overall Score Changes: Top Models Stable, Mid and Low-Tier Models Exhibit Significant Volatility

Top-Tier Models (Top 3 Overall Scores):Exhibit almost no fluctuation after image removal, indicating extremely strong pure text comprehension capabilities.

Gemini 2.5 pro (88.34): Overall score change is minimal (within ś0.5%), indicating that in weak scenarios, the image is merely "icing on the cake." Even without the image, it can solve problems stably based solely on the question stem, showing very low dependency on images.

Doubao Seed 1.6 (83.80):** Overall score change is +0.26%, almost unchanged, demonstrating extremely strong pure text question stem comprehension ability. Image assistance provides limited improvement.

ERNIE-4.5-Turbo-VL (70.64): Overall score increases +2.54% after image removal. Performance slightly improves, possibly because image assistance offers no additional benefit, and pure text problem-solving is more efficient for it.

Mid-Tier Models (Overall Scores 60-68): Exhibit differentiated volatility; some rely on image assistance. Qwen2.5-VL-72B (68.78): Overall score decreases -0.41%, a slight drop, indicating images provide mild assistance, and removal has little impact.

InternVL3-38B (61.90): Overall score increases +1.83% after image removal. Performance improves, possibly because images cause slight interference, and pure text solving is more stable.

Qwen2.5-VL-32B (65.86): Overall score is pulled down by a significant drop in multiple-select (-17.46%), but single-choice improves +2.15%. Performance is split, indicating high dependency on images for specific question types.

Low-Tier Models (Overall Score < 60):Exhibit extreme volatility; poor ability regarding "dependency on / resistance to interference from" image assistance. Positive Extreme (Significant improvement after removal):

Grok-4: Overall score increases significantly due to improvements in multiple-select (+19.08%) and fill-in-the-blank (+30.92%), total change +5.63%. This indicates images are "interference items" for it; without images, it can focus on the stem.

Claude Series: Claude 3.5 Sonnet overall +8.05%, Claude Sonnet 4 overall +11.75%, especially multiple-select +15.84% and +14.28% respectively. Image assistance is ineffective or even interfering.

Negative Extreme (Significant decline after removal): Qianfan-Check-VL: Fill-in-the-blank score drops -48.59% after removal. Even in weak scenarios (where solving without image should be possible), it overly depends on images, indicating very weak pure text comprehension.

Qianfan-QI-VL: Fill-in-the-blank score drops -37.02%, but single-choice improves +14.20%. Performance is contradictory, possibly due to vastly different image dependency across question types.

Analysis by Question Type: Single-Choice Most Stable, Multiple-Select / Fill-in-the-Blank More Volatile

Single-Choice Questions (sc_w vs sc_w_np): Overall the most stable; top models show almost no change.

Score changes for all models are within ś5%. Top models (Gemini 2.5 pro +0.51%, Doubao +0.26%) show almost no fluctuation, indicating single-choice stems provide sufficiently clear information, and image assistance plays the weakest role, aligning with the "solvable without image" characteristic of weak scenarios. Low-tier models like Qianfan-QI-VL (+14.20%) and Grok-4 (+5.63%) show significant improvement, possibly because simple stems mean images instead dis-

Models	Overall	sc_w	sc_w_np	ms_w	ms_w_np	fb_w	fb_w_np
Qwen-VL-Plus	60.24	82.47	83.54 (+1.07)	40.32	41.27 (+0.95)	82.26	80.61 (-1.65)
Qwen-VL-Max	65.98	90.35	90.42 (+0.07)	67.74	68.25 (+0.51)	81.62	85.97 (+4.35)
Qwen2.5-VL-72B-Instruct	68.78	89.88	89.47 (-0.41)	69.35	67.86 (-1.49)	87.88	84.59 (-3.29)
Qwen2.5-VL-32B-Instruct	65.86	85.61	87.76 (+2.15)	68.25	50.79 (-17.46)	88.13	81.57 (-6.56)
Qwen2.5-VL-7B-Instruct	60.04	81.64	81.14 (-0.50)	39.68	45.90 (+6.22)	83.80	82.39 (-1.41)
Qianfan-Llama-VL-8B	47.91	68.98	68.30 (-0.68)	28.57	33.33 (+4.76)	77.79	73.87 (-3.92)
Qianfan-Check-VL	45.05	75.44	74.18 (-1.26)	26.98	22.22 (-4.76)	68.53	19.94 (-48.59)
Qianfan-QI-VL	40.77	56.93	71.13 (+14.20)	0	1.59 (+1.59)	64.12	27.10 (-37.02)
DeepSeek-VL2	36.86	47.38	43.98 (-3.40)	14.29	15.87 (+1.58)	61.21	44.70 (-16.51)
InternVL3-38B	61.90	86.60	88.43 (+1.83)	60.32	61.40 (+1.08)	87.57	76.34 (-11.23)
InternVL3-14B	59.86	83.13	82.04 (-1.09)	60.32	50.00 (-10.32)	83.94	81.02 (-2.92)
InternVL2.5-38B-MPO	59.59	83.13	84.91 (+1.78)	53.97	24.14 (-29.83)	85.03	77.58 (-7.45)
ERNIE-4.5-Turbo-VL	70.643	88.37	90.91 (+2.54)	54.84	66.67 (+11.83)	90.60	89.82 (-0.78)
ERNIE-4.5	67.86	86.25	87.00 (+0.75)	50.79	50.88 (+0.09)	85.89	81.43 (-4.46)
doubao Seed 1.6	83.802	97.25	97.51 (+0.26)	79.37	75.41 (-3.96)	95.80	96.42 (+0.62)
GPT-4o	50.73	69.08	71.99 (+2.91)	34.92	33.33 (-1.59)	75.49	73.90 (-1.59)
GPT-4.5-preview	34.28	57.95	61.96 (+4.01)	17.46	15.87 (-1.59)	48.75	40.47 (-8.28)
Claude 3.5 Sonnet	26.96	68.90	76.95 (+8.05)	14.29	30.16 (+15.84)	38.58	61.14 (+22.56)
Claude Sonnet 4	26.74	67.10	78.85 (+11.75)	12.70	26.98 (+14.28)	39.55	60.03 (+20.48)
Gemini 2.5 pro	88.341	95.76	96.27 (+0.51)	87.10	85.25 (-1.85)	96.66	96.24 (-0.42)
Gemini 1.5 Pro	66.93	75.18	73.71 (-1.47)	36.51	30.16 (-6.35)	94.69	83.01 (-11.68)
Grok-4	38.25	53.56	59.19 (+5.63)	17.46	36.54 (+19.08)	55.15	86.07 (+30.92)

Table 12: Total Table of Models' Accuracy Rates Across Various Question Types Before and After Image Removal in Weak Scenarios.

tract attention.

Multiple-Select Questions (ms_w vs ms_w_np): Most volatile, best reflects image dependency.

Score changes are generally larger than for single-choice. Among top models: Gemini 2.5 pro (-1.85%), Doubao (-3.96%) show slight declines, indicating the higher complexity of multiple-select stems means image assistance has a certain role (albeit limited impact). Mid-/Low-tier models show significant differentiation: ERNIE-4.5-Turbo-VL (+11.83%), Claude 3.5 Sonnet (+15.84%) improve after removal, indicating images interfere with their multiple-select solving. Qwen2.5-VL-32B (-17.46%), InternVL2.5-38B-MPO (-29.83%) drop significantly, showing over-reliance on image assistance, unable to solve multiple-select based on the stem alone even in weak scenarios. Fill-in-the-Blank Questions (fb_w vs fb_w_np): Extreme changes concentrated; image role is polarized.

Top models change very little (Gemini 2.5 pro -0.42%, Doubao +0.62%), indicating they can accurately output answers based solely on the stem; image assistance is minimal. Low-tier models show extreme changes: Grok-4 (+30.92%), Claude 3.5 Sonnet (+22.56%) improve significantly after removal, indicating images are strong interference for their fill-in-the-blank solving (possibly due to complex image information they cannot extract effectively). Qianfan-Check-VL (-48.59%), Qianfan-QI-VL (-37.02%) drop significantly, indicating even in weak scenarios, they completely rely on details in the image (e.g., numbers, symbols) to answer correctly, lacking pure text comprehension ability.

Core Conclusions

Top Models' "Pure Text Comprehension Ability" is Overwhelming: Gemini 2.5 pro and Doubao show almost no fluctuation before and after image removal, proving their deep understanding of weak scenario stems. Images are merely auxiliary, meeting the core requirement of "solvable without image."

Mid-Tier Models' "Image Dependency" is Differentiated: Some models (e.g., ERNIE-4.5-Turbo-VL) remain stable without images, while others (e.g., Qwen2.5-VL-32B) plummet in multiple-select after removal due to over-reliance on images, requiring optimization of pure text logical reasoning ability.

Low-Tier Models are "Weak in Both Anti-Interference and Comprehension":They either improve after removal due to image interference (e.g., Grok-4, Claude series) or crash after removal due to over-dependency on images (e.g., Qianfan series), reflecting immature processing of both stems and images. Question Type Difficulty Positively Correlates with Image Role: Single-choice (easiest) has the lowest image dependency, multiple-select (most complex) has the highest image de-

pendency. Fill-in-the-blank, due to the "precise output" requirement, sees a polarized image role (assistance vs. interference). Table 12 presents the accuracy of each model across all question types before and after image removal in weak correlation scenarios.

C Additional Information on Evaluation Results in the Multi-Image Reasoning Scenario

C.1 Performance Across Disciplinary Branches

(1) Mechanics

Top-performing models: doubao-seed-1-6-vision-250815 (90.90%), Qwen3-VL-32B-Thinking (87.53%). Scoring above 85, these models exhibited exceptional reasoning capabilities in mechanics, with solid mastery of core knowledge including force analysis and motion laws.

Mid-tier models: Qwen2.5-VL-72B-Instruct (80.58%), Qwen3-VL-Plus (80.53%). With scores around 80, they could competently complete routine mechanics tasks.

Underperforming models: DeepSeek-VL2 (36.58%), Qianfan-Llama-VL-8B (50.68%). Scoring below 55, these models showed a critical lack of logical deduction ability for basic mechanical concepts and complex scenarios.

Disciplinary characteristics: As a foundational physics discipline, mechanics highlighted the advantages of top models. High-scoring models accurately disassembled the core logic of mechanical problems, while low-scoring models had distinct knowledge gaps in multi-object force analysis and complex motion process deduction.

(2) Electromagnetism

Top-performing models: doubao-seed-1-6-vision-250815 (85.51%), doubao-seed-1-6-251015 (81.31%). Surpassing 80 points, they demonstrated a thorough understanding of complex laws (e.g., electromagnetic field coupling, electromagnetic induction) and could tackle high-difficulty comprehensive electromagnetism tasks.

Second-tier models: Qwen3-VL-8B-Thinking (78.84%), Qwen3-VL-32B-Thinking (78.49%). Scoring 7580, they maintained stable performance on basic electromagnetism problems but underperformed slightly in comprehensive questions with cross-parameter interactions.

Underperforming models: DeepSeek-VL2 (34.06%), Qianfan-VL-8B (36.93%). Scoring below 35, these models had major blind spots in core electromagnetism concepts and laws.

Disciplinary characteristics: As a high-difficulty physics field with highly abstract knowledge and intricate interrelated laws, electromagnetism saw advantages concentrated among top models, while low-scoring models were barely capable of completing relevant reasoning tasks.

(3) Optics

Top-performing models: Qwen3-VL-32B-Thinking (95.35%), doubao-seed-1-6-251015 (90.87%). Scoring above 90over 10 points higher than the second tierthey precisely mastered light refraction, reflection, and imaging laws, and efficiently completed vision-related tasks such as optical path analysis.

Mid-tier models: Gemini 2.5 Pro (88.47%), Qwen3-VL-Plus (87.57%). Scoring 8590, they possessed strong capabilities in solving optical problems.

Underperforming models: DeepSeek-VL2 (35.00%), GPT-4o (55.69%). Scoring below 60, these models had obvious deficiencies in optical knowledge reserves and image-related reasoning capabilities.

Disciplinary characteristics: Optical laws are highly intuitive and closely linked to image features, which aligns with the technical strengths of multimodal models. Thus, top models delivered particularly outstanding performance in this field, with a performance ceiling significantly higher than that of other disciplines.

(4) Thermodynamics

Top-performing models: Qwen3-VL-32B-Thinking (97.05%), doubao-seed-1-6-vision-250815 (95.22%). With scores exceeding 95, they demonstrated solid mastery of thermodynamic laws, heat conduction, and entropy change, and could accurately solve diverse thermodynamics problems.

Second-tier models: doubao-seed-1-6-251015 (91.62%), Gemini 2.5 Pro (90.04%). Scoring around 90, they maintained stable thermodynamics reasoning performance.Underperforming models: DeepSeek-VL2 (57.79%), GPT-4o (58.96%). With scores below 60, these models lacked the ability to comprehend abstract thermodynamic concepts and apply relevant laws.

Disciplinary characteristics: Moderate difficulty with a highly logical knowledge system

and moderate scenario relevance. Top models fully leveraged their reasoning advantages, mid-tier models delivered consistent performance, and only low-scoring models showed distinct gaps in knowledge and capabilities.

(5) Acoustics

Top-performing models: GPT5 (94.52%), Qwen3-VL-32B-Thinking (91.78%), doubao-seed-1-6-251015 (91.78%), Gemini 2.5 Pro (91.78%). With multiple models scoring over 90, acoustics emerged as the best-performing discipline overall, with near-perfect mastery of basic physical phenomena (sound generation, propagation, resonance).

Second-tier models: Qwen3-VL-Plus (86.30%), Grok-4 (85.51%). Scoring 8590, they had a solid acoustics foundation.

Underperforming models: DeepSeek-VL2 (54.79%), Claude Haiku 4.5 (63.89%). With scores below 70, these models lacked sufficient acoustics knowledge reserves and basic law application abilities.

Disciplinary characteristics: The least difficult physics discipline overall, featuring intuitive knowledge scenarios and simple laws that match multimodal models basic learning capabilities. Thus, performance gaps across model tiers were relatively narrow, and top models achieved near-perfect scores.

C.2 Performance Across Question Types

The accuracy of all models across different question types is presented in the Table 13.Dominance of Top Models Across All Question Types: Further Amplified Advantages in Fill-in-the-Blank Questions

Top models not only achieved high-score breakthroughs across all question types but also further expanded their advantages in fill-in-the-blank questions. Leading models such as doubao-seed-1-6-vision-250815 and Qwen3-VL-32B-Thinking all attained an accuracy rate of over 88% in fill-in-the-blank questions, which was 2-3 percentage points higher than their performance in single-choice questions. They even broke through the 75% threshold in the most difficult Multiple-select questions, forming an all-excellent pattern of "leading in fill-in-the-blanks, stable in single-choice, and up to standard in Multiple-select".

Mid-tier models, although also able to maintain the advantage of better performance in fill-in-the-blank questions than in single-choice questions, generally scored below 60% in Multiple-select questions, dragging down their overall performance.

Models	SC	MS	F
doubao-vision[1]	88.00	78.85	91.04
Qwen3-VL-32B-Thinking	86.33	79.19	88.16
doubao-seed[2]	84.20	76.97	86.34
Qwen3-VL-30B-A3B-Thinking	82.02	73.83	83.86
Qwen3-VL-Plus	82.03	59.87	80.50
Qwen3-VL-8B-Thinking	79.81	69.54	80.49
Qwen2.5-VL-72B-Instruct	77.78	52.87	79.99
ERNIE 4.5 Turbo VL	70.12	51.59	75.71
GLM-4.5V	60.54	38.71	74.59
ERNIE 4.5	72.89	46.79	70.11
Qianfan-VL-70B	64.65	41.03	69.44
Llama-4-Maverick[3]	66.84	34.62	68.29
InternVL3-38B	63.27	41.03	67.99
Qwen VL-Max	73.40	49.04	71.51
Qwen2.5-VL-32B-Instruct	69.99	33.12	65.13
InternVL2.5-38B-MPO	61.08	23.08	64.56
InternVL3-14B	61.22	38.46	61.37
Qwen VL-Plus	62.73	42.04	62.75
Qwen2.5-VL-7B-Instruct	67.60	32.48	51.31
Qianfan-VL-8B	47.74	19.87	51.84
Qianfan-Llama-VL-8B	48.32	17.95	51.44
DeepSeek-VL2	25.80	14.94	43.65

[1] doubao-seed-1-6-vision-250815
[2] doubao-seed-1-6-251015
[3] Llama-4-Maverick-17B-128E-Instruct

Table 13: Examples of Model Performance Across Different Question Types in Multi-Image Reasoning.

Bottom-tier models ranked last across all question types; even though their accuracy in fill-in-the-blank questions was slightly higher than that in single-choice questions, both scores were below 55%, rendering them of no practical value for problem- solving.

(1) Single-Choice Questions: Option Hints Lower Reasoning Thresholds, Advantages of Top Models Concentrated

Core characteristics: Thanks to the "four-option single-answer" hint mechanism, single-choice questions significantly reduce the requirement for models' "unassisted reasoning" ability, only needing to complete single-option judgment and interference item elimination, making it the easiest question type for models to break through.

Top models: doubao-seed-1-6-vision-250815 ranked first with an accuracy rate of 88.00%, followed by Qwen3-VL-32B-Thinking with 86.33%. Both scores far exceeded those of other models, and opened a gap of 1.7-3.8 percentage points with the third-place model (doubao-seed-1-6-251015, 84.20%), reflecting their absolute advantage in the

ability of "single-option judgment + interference item elimination".

Bottom models: DeepSeek-VL2 (25.80%) and Qianfan-Llama-VL-8B (48.32%) scored the lowest. Among them, DeepSeek-VL2's score was close to the probability of random selection, indicating that it had an extremely weak understanding of the basic physical logic of single-choice questions and the corresponding association of options.

Rule: There is a positive correlation between parameter size and performance (e.g., Qwen2.5-VL series: 72B (77.78%) > 32B (69.99%) > 7B (67.60%)). Large-parameter models have more precise control over the details of single-choice questions and better identification of interference items.

(2) Multiple-select Questions: Correlation Judgment of Multiple Options Becomes a Bottleneck, Model Gaps Extreme

Core characteristics: Multiple-select questions require "identifying all correct options", needing to simultaneously handle rules such as "logical correlation between options" and "score deduction for multiple or missing selections". They impose the highest requirements on models' rigor and multi-dimensional reasoning ability, making them the "gold standard question type" for distinguishing model performance.

Top models: Qwen3-VL-32B-Thinking took a leading position with an accuracy rate of 79.19% by a significant margin, followed by doubao-seed-1-6-vision-250815 with 78.85%. Together with doubao-seed-1-6-251015 (76.97%), they were the only three models with scores exceeding 75% in Multiple-select questions, reflecting their strong ability to judge the logical correlation of "multiple correct options".

Bottom models: DeepSeek-VL2 ranked last with an accuracy rate of 14.94% (almost all wrong answers), while Qianfan-Llama-VL-8B (17.95%) and Qianfan-VL-8B (19.87%) both scored below 20%. This indicates that they completely failed to understand the "coexistence of multiple options" logic of Multiple-select questions, often experiencing problems such as "missing selections and wrong selections".

Special phenomenon: Mid-tier models (e.g., Qwen2.5-VL-72B-Instruct, ERNIE 4.5 Turbo VL) had Multiple-select scores concentrated in the 45%-55% range, staying in the stage of "being able to judge some correct options but not all", reflecting their insufficient control over "option completeness".

(3) Fill-in-the-Blank Questions: No Option Hints Test Direct Output Ability, Models Show Obvious "Partiality"

Core characteristics: Without option hints, fill-in-the-blank questions require models to directly generate answers (such as physical terminology, precise numerical values), imposing high requirements on the ability of "accurate knowledge memory + question stem logic conversion", and the overall scores are better than those of single-choice and Multiple-select questions.

Top models: doubao-seed-1-6-vision-250815 (91.04%) and Qwen3-VL-32B-Thinking (88.16%) still ranked in the top two, and the gap with the third-place model (doubao-seed-1-6-251015, 86.34%) was small, reflecting their stability in "unassisted output of correct answers".

Special models: GLM-4.5V showed obvious "partiality": it scored 74.59% in fill-in-the-blank questions (at the upper-middle level), but had low scores in single-choice (60.54%) and Multiple-select (38.71%) questions. This indicates that it has advantages in "direct answer generation" tasks but performs poorly in "option judgment" tasks, which may be related to the model's adaptability to "output forms (free text vs. option selection)".

Bottom models: DeepSeek-VL2 (43.65%) and Qwen2.5-VL-7B-Instruct (51.31%) scored the lowest, reflecting their extremely weak ability to directly convert "question stem to answer", and their failure to extract key physical information from the question stem and output it accurately.

C.3 Analysis of Inference Efficiency

1.**Performance Across Question Types**

The accuracy and average inference latency of each model across different question types are presented in Tables 14. **(1) Single-Choice Questions Latency Characteristics**

Overall range: The average latency was 10.2 seconds, making it the least time-consuming type among the three question categories. The latency of 90% of models was concentrated in the range of 0.7-15 seconds.

Ultra-fast models: Qwen VL-Plus (0.77s), Qwen VL-Max (0.86s), with latency below 1 second;

Ultra-slow models: Grok-4 (47.38s), GPT5 (45.68s), Qwen3-VL-8B-Thinking (32.66s), with latency exceeding 30 seconds, mainly due to the overhead of deep reasoning.

Models	SC	SC Average Latency (s)	MS	MS Average Latency (s)	F	F Average Latency (s)
doubao-seed-1-6-vision-250815	88.00	21.0625	78.85	43.0032	91.04	29.3202
Qwen3-VL-32B-Thinking	86.33	24.8638	79.19	44.5538	88.16	29.5289
Gemini 2.5 pro	77.32	6.6680	55.41	6.8930	87.58	6.9464
doubao-seed-1-6-251015	84.20	10.5745	76.97	14.5535	86.34	12.4338
Qwen3-VL-30B-A3B-Thinking	82.02	10.2305	73.83	12.6185	83.86	9.2783
Qwen3-VL-Plus	82.03	1.1423	59.87	1.2317	80.58	2.1385
Qwen3-VL-8B-Thinking	79.81	32.6604	69.54	58.8859	80.49	32.7605
Qwen2.5-VL-72B-Instruct	77.78	0.9196	52.87	1.1079	79.99	1.7898
GPT5	83.15	45.6797	70.70	89.4148	77.59	88.1585
ERNIE 4.5 Turbo VL	70.12	5.0719	51.59	5.5866	75.71	6.2060
GLM-4.5V	60.54	1.8703	38.71	2.1015	74.59	2.2403
Qwen VL-Max	73.40	0.8558	49.04	1.1044	71.51	2.2907
ERNIE 4.5	72.89	1.3867	46.79	1.8207	70.11	2.1040
Qianfan-VL-70B	64.65	4.4940	41.03	3.8679	69.44	4.4322
Llama-4-Maverick-17B-128E-Instruct	66.84	1.5407	34.62	1.9425	68.29	2.9345
InternVL3-38B	63.27	5.0179	41.03	4.6886	67.99	5.2924
Grok-4	71.34	47.3759	50.99	64.1286	67.06	56.3086
Qwen2.5-VL-32B-Instruct	69.99	1.3216	33.12	1.4643	65.13	3.2382
InternVL2.5-38B-MPO	61.08	4.0335	23.08	3.8466	64.56	6.8790
Qwen VL-Plus	62.73	0.7668	42.04	1.2103	62.75	1.3718
InternVL3-14B	61.22	2.8656	38.46	2.8829	61.37	3.3608
Claude Sonnet 4	60.77	7.3812	51.32	20.4842	62.64	11.2882
GPT-4o	58.82	4.7445	58.82	7.4911	54.59	8.4012
Claude Haiku 4.5	54.77	5.8317	33.79	13.9116	56.15	9.1020
Qianfan-VL-8B	47.74	2.9509	19.87	2.7836	51.84	2.9617
Qianfan-Llama-VL-8B	48.32	2.9548	17.95	2.8433	51.44	3.0150
Qwen2.5-VL-7B-Instruct	67.60	1.6333	32.48	1.6502	51.31	2.6758
DeepSeek-VL2	25.80	1.7760	14.94	2.0168	43.65	3.1006

Table 14: Scatter Plot of Average Latency vs. Accuracy Across Different Question Types

Performance-Efficiency Trade-off

Optimal balance models: Qwen2.5-VL-72B-Instruct (accuracy 77.78%, latency 0.92s), achieving near-real-time response in the high-accuracy range;

Performance-priority models: doubao-seed-1-6-vision-250815 (accuracy 88.00%, latency 21.06s), topping the accuracy rankings with latency only half of that of GPT5;

Efficiency-priority models: Qwen VL-Plus (accuracy 62.73%, latency 0.77s), serving as the lowest-cost solution for real-time scenarios. **(2) Multiple-select Questions Latency Characteristics**

Overall range: The average latency was 17.6 seconds, the most time-consuming type among the three question categories, with over 50% of models having latency > 10 seconds.

Regular-latency models: Qwen VL-Max (1.10s), Qwen2.5-VL-72B-Instruct (1.11s), maintaining the advantage of low latency; Ultra-slow models: GPT5 (89.41s), Qwen3-VL-8B-Thinking (58.89s), Grok-4 (64.13s), with latency close to 1 minute.

Task-specificity: For the same model, latency in the more complex Multiple-select questions was generally 30%-50% higher than that in single-choice questions (e.g., GPT4o increased from 4.74s to 7.49s).

Performance-Efficiency Trade-off

Optimal balance models: Qwen3-VL-30B-A3B-Thinking (accuracy 73.83%, latency 12.62s), controlling latency within 15 seconds in the high-accuracy range;

Performance-priority models: doubao-seed-1-6-vision-250815 (accuracy 78.85%, latency 43.00s), leading in accuracy with latency less than 50% of that of GPT5;

Efficiency-priority models: Qwen VL-Max (accuracy 49.04%, latency 1.10s), acting as the basic fallback solution for this scenario.

(3)Fill-in-the-Blank Questions Latency Characteristics

Overall range: The average latency was 15.8 seconds, between that of single-choice and Multiple-select questions, with 70% of models having latency concentrated in the range of 2-20 seconds.

Regular-latency models: Qwen VL-Plus (1.37s), ERNIE 4.5 (2.10s), retaining the low-latency attribute of basic models;

Ultra-slow models: GPT5 (88.16s), Qwen3-VL-8B-Thinking (32.76s), Grok-4 (56.31s), with deep reasoning being the core factor of time consump-

tion.

Performance-Efficiency Trade-off

Optimal balance models: Gemini 2.5 Pro (accuracy 87.58%, latency 6.95s), achieving regular latency in the top-tier accuracy range and serving as the benchmark for comprehensive tasks;

Performance-priority models: doubao-seed-1-6-vision-250815 (accuracy 91.04%, latency 29.32s), taking a leading position in accuracy by a significant margin with latency only 1/3 of that of GPT5;

Efficiency-priority models: Qwen VL-Plus (accuracy 62.75%, latency 1.37s), suitable for lightweight comprehensive reasoning scenarios.

(4)Question-Type-Specific Model Selection Recommendations

Single-choice questions: Select Qwen VL-Plus for real-time scenarios, and doubao-seed-1-6-vision-250815 for high-accuracy scenarios;

Multiple-select questions: Select Qwen VL-Max for real-time scenarios, and doubao-seed-1-6-vision-250815 for high-accuracy scenarios;

Fill-in-the-blank questions: Select Gemini 2.5 Pro for comprehensive reasoning scenarios, and doubao-seed-1-6-vision-250815 for extreme-performance scenarios.

2.Performance Across Disciplinary Branches

The accuracy and average inference latency of each model across different disciplinary branches are presented in Tables 15.

(1) Mechanics

Latency Characteristics

Overall range: The average latency of the mechanics discipline was 7.42 seconds, making it the least time-consuming branch among the five disciplinary categories. The latency of 75% of models was concentrated in the range of 0.7 to 6 seconds, with no models exhibiting extremely high latency.

Ultra-fast models: Qwen VL-Plus (0.74s), Qwen2.5-VL-72B-Instruct (0.84s), with latency below 1 second, which can meet the judgment requirements of real-time mechanics scenarios.

Slow models: Grok-4 (39.07s), GPT5 (23.52s), with latency exceeding 20 seconds, mainly due to the computing overhead of deep physical modeling.

Performance-Efficiency Trade-off

Optimal balance model: Qwen2.5-VL-72B-Instruct (accuracy 80.58%, latency 0.84s), achieving near-real-time response in the high-accuracy range and serving as the benchmark for routine mechanics tasks.

Performance-priority model: doubao-seed-1-6-vision-250815 (accuracy 90.90%, latency 9.44s), leading in accuracy by a significant margin with latency only half of that of GPT5.

Efficiency-priority model: Qwen VL-Plus (accuracy 60.19%, latency 0.74s), which is the lowest-cost solution for real-time lightweight mechanics scenarios.

(2)Electromagnetism

Latency Characteristics

Overall range: The average latency of electromagnetism was 16.85 seconds, the most time-consuming branch among the five disciplinary categories. Over 40% of models had latency greater than 10 seconds, and there were a large number of ultra-slow models.

Regular-latency models: Qwen VL-Plus (1.00s), Qwen2.5-VL-72B-Instruct (1.16s), where basic models still maintained the advantage of low latency.

Ultra-slow models: GPT5 (71.03s), Qwen3-VL-8B-Thinking (49.73s), Grok-4 (59.93s), with latency close to 1 minute, and the step overhead of multi-formula simultaneous reasoning was significant.

Task specificity: For the same model, the latency in electromagnetism was generally 2 to 5 times higher than that in mechanics. For example, Qwen3-VL-32B-Thinking increased from 15.35s to 32.62s, confirming the additional computing consumption caused by the association of abstract formulas in electromagnetism.

Performance-Efficiency Trade-off

Optimal balance model: Gemini 2.5 pro (accuracy 77.50%, latency 6.86s), controlling latency within 7 seconds in the high-accuracy range, which is the optimal solution among overseas models.

Performance-priority model: doubao-seed-1-6-251015 (accuracy 81.31%, latency 14.03s), leading in accuracy with latency only 1/5 of that of GPT5, and is the first choice for core electromagnetism tasks.

Efficiency-priority model: Qwen VL-Max (accuracy 61.87%, latency 1.25s), suitable for real-time lightweight electromagnetism scenarios.

(3)Optics

Latency Characteristics

Overall range: The average latency of optics was 9.27 seconds, between that of mechanics and electromagnetism. The latency of 60% of models was concentrated in the range of 1 to 10 seconds,

Models	Mech	A_Latency (s)	EM	A_Latency (s)	Opt	A_Latency (s)	Thermo	A_Latency (s)	Acou	A_Latency (s)
doubao-seed-1-6-vision	90.90	9.4405	85.51	34.1424	90.36	15.0538	95.22	23.4388	90.41	18.0429
Qwen3-VL-32B-Thinking	87.53	15.3478	78.49	32.6164	95.35	16.1658	97.05	28.3493	91.78	25.2591
Gemini 2.5 pro	84.32	6.3173	77.50	6.8569	88.47	7.0611	90.04	6.7825	91.78	6.3666
doubao-seed-1-6-251015	84.33	6.7669	81.31	14.0326	90.87	8.5984	91.62	11.1084	91.78	8.6114
Qwen3-VL-30B-A3B-Thinking	85.00	6.8142	77.29	12.8640	85.25	6.3780	87.55	9.4288	78.08	7.4445
Qwen3-VL-Plus	80.53	1.0989	72.06	1.4774	87.57	1.5951	89.44	1.5655	86.30	1.4302
Qwen2.5-VL-72B-Instruct	80.58	0.8431	67.37	1.1605	86.21	1.3115	88.10	1.2870	83.56	1.4227
GPT5	78.06	23.5228	74.85	71.0282	79.49	45.0606	87.25	68.8744	94.52	53.7182
ERNIE 4.5 Turbo VL	75.83	4.6221	70.16	5.4817	79.49	5.4898	82.20	5.6957	71.23	5.3082
Qwen VL-Max	71.28	0.8542	61.87	1.2523	79.49	1.6611	82.09	1.5480	83.56	1.1810
ERNIE 4.5	70.16	1.4746	66.32	1.6208	80.15	1.8213	81.05	1.7079	82.19	1.6598
Qianfan-VL-70B	69.37	4.3760	52.08	4.4557	76.36	4.2805	82.66	4.3233	78.08	4.8648
Grok-4	67.08	39.0687	60.99	59.9304	71.73	40.1439	76.95	51.0964	85.51	41.6265
InternVL3-38B	68.07	5.1950	49.73	5.1221	73.48	4.9844	84.27	4.9560	76.71	5.5581
Llama-4-Maverick-17B	71.19	1.5907	54.29	1.9709	65.15	2.3518	80.11	2.1280	78.08	2.0673
Qwen3-VL-8B-Thinking	71.22	16.7019	78.84	49.7331	83.57	12.7195	88.28	28.9687	73.97	28.9564
GLM-4.5V	74.19	1.8934	54.80	1.9742	74.43	2.1299	81.85	2.0625	78.87	1.9685
Qwen2.5-VL-32B-Instruct	64.25	1.4308	55.97	1.8710	74.21	2.0888	78.88	2.3388	79.45	1.9358
InternVL2.5-38B-MPO	65.45	4.6171	47.79	4.7664	70.45	5.2776	74.73	5.2208	69.86	5.3819
InternVL3-14B	63.40	2.7359	46.40	3.0332	65.92	3.0698	74.90	3.0210	76.71	3.2230
Qwen VL-Plus	60.19	0.7401	54.97	1.0037	76.80	1.0657	77.81	1.0489	83.00	0.9437
GPT-4o	61.27	5.3663	42.57	5.8503	55.69	7.1202	58.96	6.7367	78.08	5.3058
Claude Sonnet 4	63.62	5.8219	57.29	10.4975	66.77	8.4549	72.28	9.4574	75.00	9.3733
Claude Haiku 4.5	53.38	4.8268	46.19	7.6655	61.93	5.2518	69.16	8.4227	63.89	6.9839
Qwen2.5-VL-7B-Instruct	51.04	1.6346	49.87	1.7600	60.12	1.8906	71.66	2.3134	75.34	2.1474
Qianfan-VL-8B	51.69	2.9117	36.93	2.9655	54.39	2.9618	64.65	2.8577	61.64	3.1396
Qianfan-Llama-VL-8B	50.68	2.9463	36.60	3.0089	53.64	2.9224	67.44	2.8837	65.75	3.1700
DeepSeek-VL2	36.58	1.7472	34.06	2.0807	35.00	2.4446	57.79	2.4035	54.79	2.4250

Table 15: Scatter Plot of Average Latency vs. Accuracy Across Different Disciplinary Subfields.

and the overall time consumption distribution was relatively balanced.

Regular-latency models: Qwen VL-Plus (1.07s), Qwen VL-Max (1.66s), where basic models retained the low-latency attribute. Ultra-slow models: GPT5 (45.06s), Grok-4 (40.14s), with latency exceeding 40 seconds, and the overhead of combining deep visual feature extraction with physical laws was relatively large.

Performance-Efficiency Trade-off

Optimal balance model: Qwen3-VL-Plus (accuracy 87.57%, latency 1.60s), achieving regular latency in the top-tier accuracy range and serving as the benchmark for general optics tasks.

Performance-priority model: Qwen3-VL-32B-Thinking (accuracy 95.35%, latency 16.17s), topping the accuracy rankings with latency only 1/3 of that of GPT5.

Efficiency-priority model: Qwen VL-Plus (accuracy 76.80%, latency 1.07s), suitable for real-time optics judgment scenarios.

(4)Thermodynamics

Latency Characteristics

Overall range: The average latency of thermodynamics was 14.71 seconds, second only to electromagnetism. 50% of models had latency greater than 10 seconds, and deep state derivation was the core factor of time consumption.

Regular-latency models: Qwen VL-Plus (1.05s), Qwen2.5-VL-72B-Instruct (1.29s), where basic models could still achieve low-latency responses.

Ultra-slow models: GPT5 (68.87s), Grok-4 (51.10s), Qwen3-VL-8B-Thinking (28.97s), and multi-round state hypothesis verification led to a surge in time consumption.

Performance-Efficiency Trade-off

Optimal balance model: Gemini 2.5 pro (accuracy 90.04%, latency 6.78s), achieving regular latency in the high-accuracy range and being the optimal solution for comprehensive thermodynamics tasks.

Performance-priority model: Qwen3-VL-32B-Thinking (accuracy 97.05%, latency 28.35s), leading in accuracy by a significant margin and suitable for scientific research scenarios without real-time requirements.

Efficiency-priority model: Qwen VL-Plus (accuracy 77.81%, latency 1.05s), suitable for real-time thermodynamics state judgment scenarios.

(5)Acoustics

Latency Characteristics

Overall range: The average latency of acoustics was 12.15 seconds, between that of optics and thermodynamics. The latency of 70% of models was concentrated in the range of 1 to 10 seconds, and

the preprocessing overhead of waveform analysis was controllable.

Regular-latency models: Qwen VL-Plus (0.94s), Qwen VL-Max (1.18s), where basic models maintained the advantage of low latency.

Ultra-slow models: GPT5 (53.72s), Grok-4 (41.63s), with latency exceeding 40 seconds, and the association derivation between waveform frequency and physical laws was time-consuming.

Model type difference: The self-developed model doubao-seed-1-6-251015 (8.61s) had a latency only 1/6 of that of GPT5 with the same performance, showing a significant efficiency advantage.

Performance-Efficiency Trade-off

Optimal balance model: Gemini 2.5 pro (accuracy 91.78%, latency 6.37s), achieving regular latency in the top-tier accuracy range and serving as the benchmark for comprehensive acoustics tasks.

Performance-priority model: doubao-seed-1-6-251015 (accuracy 91.78%, latency 8.61s), with accuracy equal to that of GPT5 and latency only 1/6 of that of GPT5, highlighting a prominent efficiency advantage.

Efficiency-priority model: Qwen VL-Plus (accuracy 83.00%, latency 0.94s), suitable for real-time acoustics scenarios.

(6)Core Disciplinary Conclusions and Overall Model Selection Recommendations

Core Conclusions

The reasoning complexity of disciplines is irrelevant to time consumption. From the previous experimental conclusions, the difficulty ranking of disciplines is acoustics > optics > thermodynamics > mechanics > electromagnetism, while the average time consumption ranking is electromagnetism > thermodynamics > acoustics > optics > mechanics.

Overall Disciplinary Model Selection Recommendations

Mechanics: Select Qwen VL-Plus for real-time scenarios and doubao-seed-1-6-vision-250815 for high-accuracy scenarios. Electromagnetism: Prioritize doubao-seed-1-6-251015.

Optics: Select Qwen3-VL-Plus for general scenarios and Qwen3-VL-32B-Thinking for scientific research scenarios.

Thermodynamics: Select Gemini 2.5 pro for comprehensive tasks and Qwen3-VL-32B-Thinking for deep reasoning tasks.

Acoustics: Select Gemini 2.5 pro for balanced scenarios and doubao-seed-1-6-251015 for high-performance scenarios.