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

Anonymous ACL submission

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) highlight strong image understanding, yet benchmarks for assessing practical problem-solving remain limited. To address this, we introduce MV-Physics, a benchmark for physical visual question answering, comprising 5582 authentic Chinese junior high physics problems with diverse scenarios and a five-dimensional evaluation framework. Experiments reveal significant differences among MLLMs in solving visual physics problems: Gemini 2.5 Pro (88.34%) and doubao Seed 1.6 (83.8%) have comprehensive advanced abilities, while most others fail. Electrometry shows high discriminative power; acoustics yields the best results. Multiple-choice questions are most challenging; models vary significantly in image relevance and text length adaptability. Top models excel in complex reasoning (S3) but underperform in simpler tasks (S1), indicating image understanding as a key bottleneck. This research provides a crucial evaluation tool and reference for MLLMs development in physical vision, facilitating relevant models optimization. It also offers a "calibration ruler" 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.

Currently, although some evaluation benchmarks for MLLMs exist(Yue et al., 2024; Tie et al., 2025) , significant shortcomings are evident in physical visual scenarios. On one hand, existing datasets have notable limitations; evaluation subjects are mostly concentrated in mathematics(Wang et al., 2025) , and when targeting physics, the problems are often too difficult to effectively differentiate model capability levels(He et al., 2024; Feng et al., 2025). On the other hand, existing evaluation schemes often rely on single dimensions, making it difficult to comprehensively assess the overall capabilities of MLLMs in physical visual scenes and precisely identify the weaknesses of various models (Xiang et al., 2025; Zhang et al., 2025). Therefore, constructing a multi-dimensional evaluation benchmark specifically for middle school physical visual scenes is of significant practical importance for accurately measuring the performance of existing models and promoting the development of MLLMs in the field of physical vision. Based on the above context, this paper proposes the MV-Physics evaluation benchmark. Our contributions are as follows:

- We constructed a novel physical visual reasoning question-answering dataset containing a large number of authentic and diverse physical scene problems.
- We designed a comprehensive evaluation scheme comprising five dimensions to assess the performance of MLLMs in physical visual scenes from various angles.

- We utilized this benchmark to evaluate multiple current mainstream MLLMs, analyzed their performance across different dimensions, and summarized common issues and directions for improvement among existing models.

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(**?**)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. With the continuous evolution of MLLMs, their application scenarios across various industries continue to expand, such as vertical practices in embodied intelligence and GUI understanding and navigation(Hong et al., 2024).

2.2 Development of MLLM Evaluation Benchmarks

PHYBench(Shi et al., 2025) , TPBench (Chung et al., 2025), and UGPhysics(Xu et al., 2025) contain relatively rich physics problems, but they are all pure text question-answering without visual input. EMMA (Hao et al., 2025)includes multimodal reasoning problems in mathematics, physics, chemistry, and code, but lacks specificity. Although there are some benchmarks for physical visual question answering scenarios, they fail to meet evaluation requirements. For instance, Seephys(Xiang et al., 2025) proposes four evaluation modes, Physreason (Zhang et al., 2025)proposes an automated scoring framework, and PhyX(Shen et al., 2025) focuses on reasoning type analysis, but their evaluation dimensions cannot precisely locate model weaknesses; OlympiadBench focuses on competition-level problems (He et al., 2024), PHYSICS focuses on doctoral qualification exams(Feng et al., 2025) , and MM-PhyQA focuses on multi-image physical knowledge reasoning at the Indian high school level (Kapuriya et al., 2025). The problem difficulty in these benchmarks is excessively high, resulting in extremely low overall accuracy rates for models, and such uniformly poor performance leads to no discriminative power,making it ineffective to compare model performance gradients.

In summary, existing evaluation benchmarks, whether in terms of problem coverage categories, quantity of problems, richness of annotations, or evaluation dimensions, struggle to reflect the true level of physical reasoning of multimodal large models in real-world scenarios.

3 MV-Physics

3.1 Overview

The MV-Physics dataset constructed in this paper is a meticulously annotated multimodal dataset designed to systematically evaluate the reasoning capabilities of Multimodal Large Models (MLLMs) in physical visual scenes. Each sample includes an image relevant to the problem and a textual question stem, totaling 5582 physics exam problems from Chinese junior high schools. It covers five disciplinary branches: mechanics, electromagnetism, optics, thermodynamics, and acoustics. Problem types include single-choice, multiple-choice, and fill-in-the-blank questions.

To comprehensively investigate the models' integrated problem-solving capabilities across different dimensions, the dataset features multi-dimensional annotations:

The paper(Deng et al., 2025) mentions that visual models might blindly trust textual data over visually extracted information. How would MLLMs perform in solving practical physics problems? Seephys (Xiang et al., 2025) noted that models perform poorly on strongly vision-related problems but did not conduct in-depth research or explore the underlying mechanisms. To detect the level of image understanding capability in large models, we categorized problems based on the degree of the image's impact on solving the problem into two levels: "Strong Image Relevance (strong)" and "Weak Image Relevance (weak)".

Many studies have shown that textual load significantly impacts the comprehension capabilities

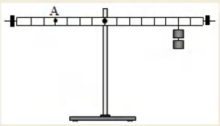
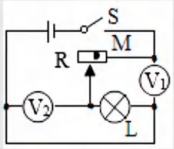
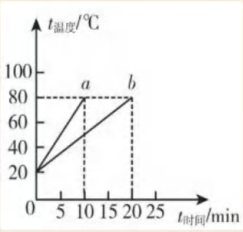
Single-choice	Multiple-choice	Fill-in-the-blank
Question: There are several weights of the same specification. As shown in the figure, how many weights should be hung at point A of the lever to balance the lever in the horizontal position?  A:1 B:2 C:3 D:4	**Question:**As shown in the figure, the power supply voltage is constant at 9V. The lamp L is marked with "3V 1.5W" (the resistance of lamp L is not affected by temperature), and the sliding rheostat R has a specification of "0 ~ 30Ω". Close the switch S. When the slider of R moves to point M, the lamp L glows normally. Which of the following analyses is correct?  A: The resistance of lamp L is 6Ω B: The indication of voltmeter V1 is 6V C: The resistance of R connected to the circuit is 18Ω D: The total power consumed by the circuit is 4.5W	**Question:** The graph shows the change in temperature of water and kerosene over time. The graph line_____ (fill in the letter) corresponds to the liquid with a strong heat absorption capacity. 
answer: C subject: mechanics question type: single choice correlation: strong task_level: S2 LC_Level: LC-2	correct_answers: A,D subject: Electromagnetism question_type: multiple_choice correlation: strong task_level: S3 LC_Level: LC-3	answer: B subject: Thermology question type: fill in the blank correlation: strong task_level: S2 LC_Level: LC-2

Figure 1: Examples of various problem types in MV-Physics.

of large models (Hu et al., 2024; Wang et al., 2024). To study the performance of MLLMs in middle school physics scenes under different textual loads, we divided them into three difficulty levels (LC-1, LC-2, LC-3) based on the length of the question stem. To precisely locate the model's reasoning capability level, we categorized problem difficulty into three levels (S1, S2, S3) based on the depth and complexity of the problems.

The proportion of the number of questions in each dimension (e.g., easy, medium, difficult) of the dataset is shown in Appendix I.

Additionally, to assist in verifying the model's understanding of physical symbols and circuit structures, a mini subset consisting of 50 circuit diagrams was extracted from the electromagnetism branch for targeted testing.

3.2 Data Construction

Data Collection:Our data was sourced from Xuekewang[1], a platform for sharing high-quality educational resources for primary and secondary schools, ensuring data correctness. Initially, physics test papers from various regions and years across China were collected. Manual screening was employed to organize qualified problems into JSON format. **Data Cleaning**:To ensure data quality, manual verification was used to filter out low-quality images. Automated methods were employed to remove duplicate data. Identical problems were searched on other educational platforms to verify answer correctness. The final dataset consists of 1845 single-choice questions, 3177 fill-in-the-blank questions, and 560 multiple-choice questions. **Data Annotation**:Based on the above data, we conducted a series of annotations. For controversial annotations, such as progressive structures divided according to the characteristics of different branches, two large models, KIMI and DeepSeek, were used for verification along with manual review to ensure annotation correctness.

4 Experimental Setup

We conducted comprehensive and extensive experiments on MV-Physics, testing 22 domestic and international large models via API (specific API matches are detailed in Appendix A): [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

[1] https://www.zxxk.com/

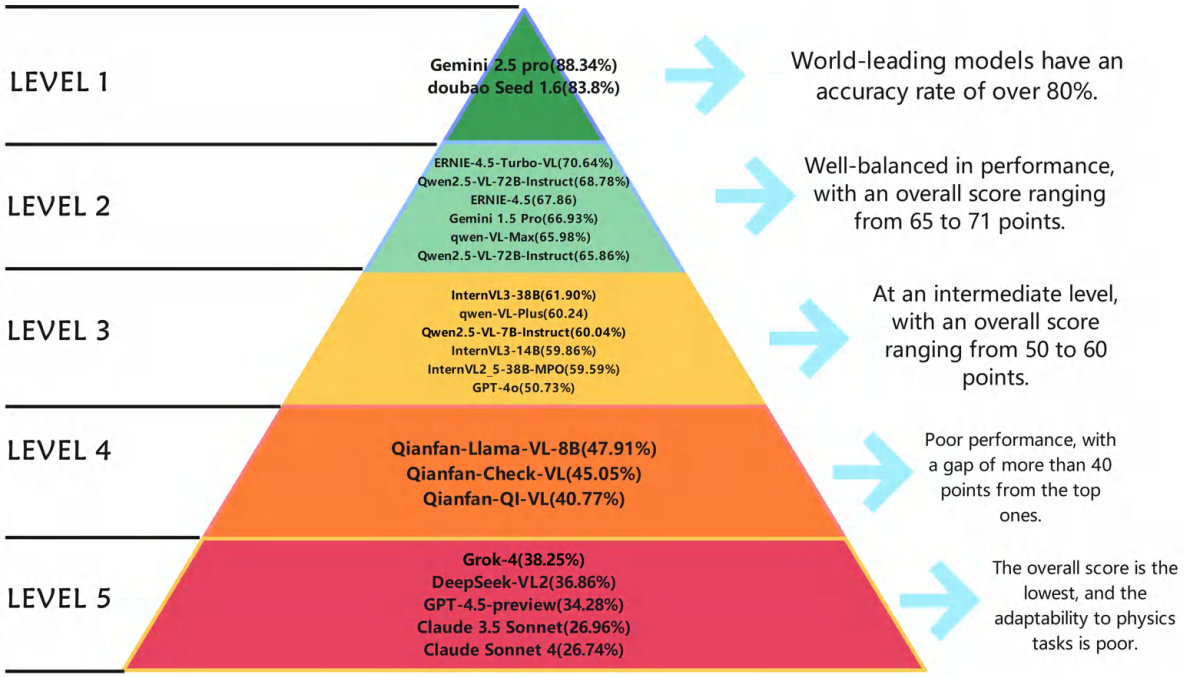


Figure 2: Performance gradient chart for all models.

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)].

Our evaluation encompasses multiple dimensions, exploring the differential effectiveness of MLLMs in solving problems across different disciplinary branches and problem types, as well as the impact of image relevance and textual load on problem-solving ability. Furthermore, we evaluate model performance across gradient tasks to precisely locate their problem-solving capabilities.

During the experiments, overall testing was conducted, and a subset annotated as having weak image relevance (correlation: weak) was selected. For this subset, images were removed, and models were required to answer based solely on the textual stem information. We designed appropriate prompts for single-choice, multiple-choice, and fill-in-the-blank requirements (detailed prompts in Appendix B), collected output results, and compiled results across dimensions.

We used the accuracy rate of model responses as the evaluation metric. For single-choice and multiple-choice questions, automated evaluation was implemented through programmed scripts. For fill-in-the-blank questions, results were determined through manual verification.

Finally, separate testing was conducted on the mini-set (50 circuit symbol understanding and circuit analysis questions) to provide auxiliary validation of the large models' image recognition capabilities.

5 Results and Analysis

5.1 Main Results

Leading models (doubao Seed 1.6, Gemini 2.5 pro) achieved an overall accuracy exceeding 80%, outperforming the second tier by over 10 percentage points, reflecting an absolute advantage in comprehensive tasks involving image understanding in physics. Model performance is strongly correlated with parameter count/version (e.g., Qwen2.5-VL series: 72B > 32B > 7B). Large-parameter models perform more stably in complex tasks.

Subject Difficulty and Model Adaptability: Acoustics > Optics > Thermodynamics > Mechanics > Electromagnetism (ranked by top model scores). Electromagnetism, due to its complex laws, becomes the domain with the highest differentiation, while Acoustics, being fundamental and intuitive, is the domain where models perform best.

Models	overall	Mech	EM	Opt	Thermo	Acous
Gemini 2.5 pro	**88.34**	**87.31**	**86.56**	**91.97**	**90.48**	**98.00**
doubao Seed 1.6	**83.80**	**82.70**	**81.30**	**88.54**	**86.82**	**94.63**
ERNIE-4.5-Turbo-VL	70.64	69.11	67.68	78.70	73.69	80.67
Qwen2.5-VL-72B-Instruct	68.78	69.86	66.69	79.39	75.86	90.13
ERNIE-4.5	67.86	62.54	62.04	77.59	72.86	80.27
Qwen-VL-Max	65.98	66.45	59.54	75.30	71.16	78.67
Qwen2.5-VL-32B-Instruct	65.86	64.23	58.13	77.67	77.51	87.07
Gemini 1.5 Pro	66.93	64.74	63.83	76.84	69.77	81.86
InternVL3-38B	61.90	63.82	55.83	73.48	76.58	85.71
Qwen-VL-Plus	60.24	60.10	52.80	71.41	67.97	74.67
Qwen2.5-VL-7B-Instruct	60.04	58.05	52.32	74.82	70.45	70.07
InternVL3-14B	59.86	59.42	51.17	70.73	72.49	81.00
InternVL2_5-38B-MPO	59.59	59.37	52.09	67.58	70.82	82.31
GPT-4o	50.73	48.09	44.90	58.87	62.78	77.03
Qianfan-Llama-VL-8B	47.91	44.40	41.01	64.61	55.95	67.35
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
GPT-4.5-preview	34.28	33.31	31.17	36.50	40.88	52.27
Grok-4	38.25	35.30	32.04	42.06	44.55	52.70
Claude 3.5 Sonnet	26.96	27.49	**22.92**	28.57	32.09	44.09
Claude Sonnet 4	**26.74**	**25.70**	24.12	**28.47**	**31.86**	**43.44**

Table 1: Overall accuracy of various models.

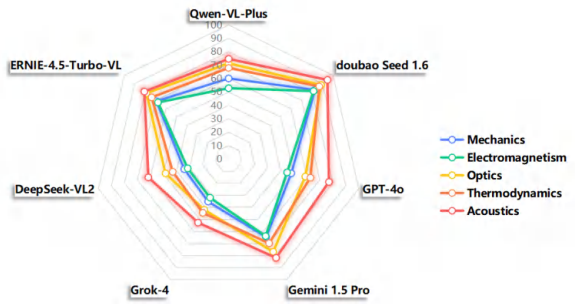


Figure 3: Example of model performance across different disciplinary branches.

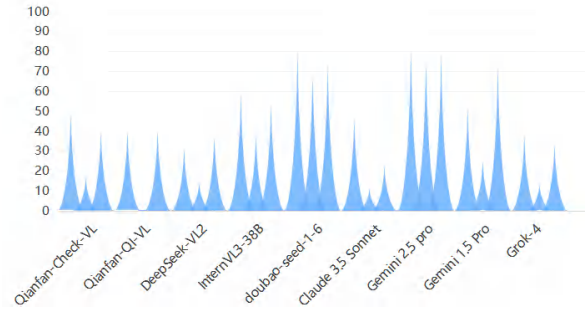


Figure 4: Example chart of model performance across problem types.Complete data in Appendix D.

Optimization Direction: Low-tier models need to strengthen basic science knowledge (e.g., force analysis in mechanics, laws of electromagnetism). Mid-tier models need to improve adaptability to high-difficulty scenarios.A more detailed analysis is provided in Appendix C.

5.2 Performance by Question Type

Overall Performance: Significant Differentiation in Problem Type Difficulty and Model Capability

Across problem types, difficulty ranks as:**Multiple-choice > Fill-in-the-blank > Single-choice**, with severe polarization. SSingle-choice questions are the easiest for most models, with nearly half achieving 60%80% accuracy. Multiple-choice is the most challenging, with most models below 600% accuracy, exposing significant reasoning gaps. Fill-in-the-blank performance lies in between, with relatively stable results.Detailed data can be found in Appendix C.

Significant Multimodal Advantage of Leading Models: Gemini 2.5 Pro leads all types (90.77% single, 83.61% multiple, 87.77% fill-in),

followed closely by Doubao Seed 1.6 (89.37%, 76.13%, 81.93%). In stark contrast, models like DeepSeek-VL2 (34.57% single), Claude 3.5 Sonnet (24.76% fill-in), and Qianfan-QI-VL (only 0.54% multiple) perform poorly.

Multiple-choice is a Common Challenge: Multiple-choice remains a key challenge even for top models, more authentically reflecting reasoning ability than single-choicewhere guessing plays a rolemaking it crucial for future optimization.

5.3 Performance by Image Relevance

We divided the problems into two categories: strong image relevance and weak image relevance. The detailed explanations are as follows:

Strong: The image contains essential information for solving the problem; solving the problem is impossible without the image.

Weak: The image primarily aids model understanding of the problem; the textual stem describes the image sufficiently, making it highly probable to solve the problem based on text alone.

In the "weak scenario," the change in accuracy before and after removing the image (with image w vs. no image w_np) reflects the model's dependence on "image assistance" and the strength of its "pure textual stem comprehension ability."

Models	Overall	strong	weak	weak_np
Qwen-VL-Plus	60.24	54.87	80.13	79.54 (-0.59)
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)
Claude 3.5 Sonnet	26.96	22.75	42.22	63.16(+20.94)
Gemini 2.5 pro	88.34	86.28	95.85	95.68 (-0.17)
Grok-4	38.25	32.09	35.77	74.73(+38.96)

Table 2: Model accuracy rates across different image relevance conditions. Complete data is available in Appendix E.

Note: **Overall:** Accuracy on the full dataset. **strong:** Accuracy on problems where the image is *essential* for solving the problem. **weak:** Accuracy on problems where the image provides *auxiliary* information. **weak_np:** Accuracy on the *weak* subset when answering based *solely on the text* (image removed). **Difference** (in parentheses, blue for negative, red for positive): *weak_np - weak* accuracy. A positive value indicates better performance without the image.

Based on the difference weak_np accuracy - weak accuracy (value in parentheses), models can be categorized into three types: Image-Dependent (Difference negative and < -1): 12 models. Auxiliary images improve performance, and MLLMs lack sufficient spatial imagination capability from text input alone. Extreme dependence: InternVL2_5-38B-MPO (-38.02), Qianfan-Check-VL (-30.25) performance plummets without images, indicating extremely weak independent text-based problem-solving ability. Slight dependence: ERNIE-4.5 (-2.40) the presence of the image has minimal impact, indicating balanced text and multimodal capabilities.

Image-Interfered (Difference positive and > 1): 5 models. Auxiliary images reduce performance. The large model misunderstands the image and, when textual and image information conflict, chooses the image information as it understands it. Extreme interference: Grok-4 (+38.96), Claude 3.5 Sonnet (+20.94), Claude Sonnet 4 (+20.42) performance is far better without images, indicating image information becomes an interfering factor. Slight interference: Qwen-VL-Max (+2.68), ERNIE-4.5-Turbo-VL (+1.10) performance is slightly better without images, indicating text capability is slightly stronger than multimodal fusion capability.

Unaffected by Images (Difference between -1 and +1): 5 models. MLLMs possess strong spatial imagination capability from text input. This characteristic is particularly evident in top models: Gemini 2.5 pro (-0.17), doubao (+0.29, nearly even) differences close to 0.

5.4 Performance by Textual Load

This study aims to investigate the impact of question stem text length (counting only the problem statement; option content for choice questions is excluded) on the problem-solving ability of MLLMs, focusing on the association mechanism between linguistic load and model task performance.

We constructed a three-level linguistic load classification system (LC level) based on the number of characters in the question stem, with specific classification criteria as follows:

LC-1 (Low Linguistic Load): Question stem text length 30 characters.

LC-2 (Medium Linguistic Load): 30 characters < Question stem text length 65 characters.

LC-3 (High Linguistic Load): Question stem text length > 65 characters.

LC-2 > LC-3 > LC-1: 18 models. Most models perform better with medium-length or longer text. Short texts might increase comprehension difficulty due to low information density and lim-

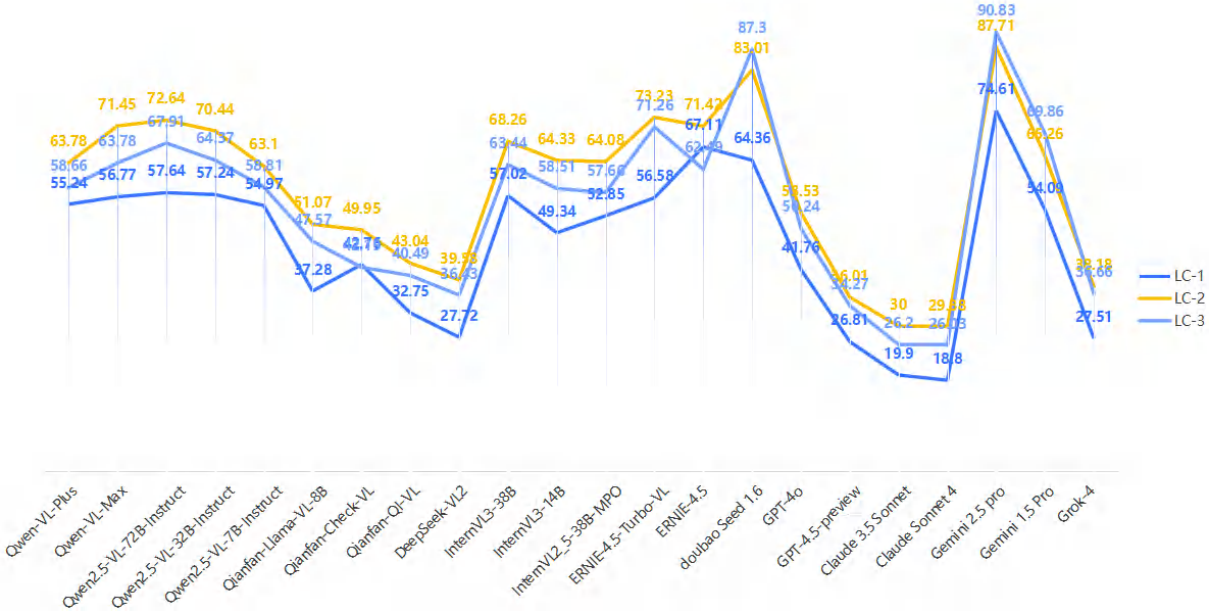


Figure 5: Accuracy rates of models across LC levels.More detailed data can be found in Appendix F.

ited context. This indicates that most models have reached a highly mature level in processing complex text, achieving a breakthrough from being "limited by text length" to being "empowered by text length." Notably, the performance difference between LC-2 and LC-3 for these models is small, and the slight score gap arises from the inherent difficulty gradient of the evaluation tasks (generally, increased text length often accompanies increased task complexity, such as more constraints and reasoning steps).

LC-3 > LC-2 > LC-1: 3 models. Notably, the leading models Gemini 2.5 pro (LC-1=74.61 LC-2=87.71 LC-3=90.83) and doubao Seed 1.6 (LC-1=64.36 LC-2=83.01 LC–3=87.30) are among them. This not only demonstrates the exceptional performance of top models in handling complex reasoning tasks but also highlights their advantage with long text. Firstly, they exhibit outstanding long-range dependency modeling, avoiding memory decay of early information. Secondly, they possess efficient redundant information filtering mechanisms, accurately locating core constraints. Thirdly, they have stronger adaptability to task complexity; when increased text length accompanies more reasoning steps and overlapping constraints, their multi-step logical processing capability is fully activated, whereas most models are prone to "information overload" and experience performance degradation at similar complexity levels.

5.5 Performance by Problem Difficulty Gradient

We categorized problem difficulty into three levels based on problem depth and complexity:

S1: Basic operations and direct recognition. S2: Simple reasoning and single dynamic change. S3: Complex reasoning and multi-step synthesis.

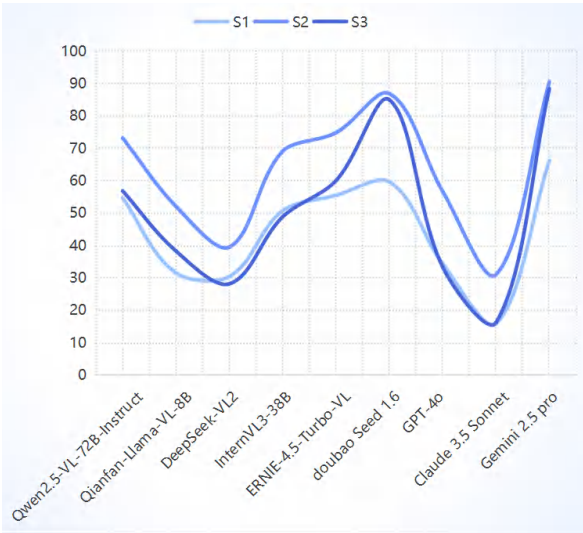


Figure 6: Model accuracy across difficulty gradient levels.Complete data in Appendix G.

All models exhibit consistent advantage on S2 tasks.This phenomenon is closely related to task characteristics. S1 tasks (e.g., reading values, identifying quantities) demand precise visual-text alignment and detailed recognition, such as

estimating instrument readings or judging anglesa challenge for most models. S3 tasks (e.g., multi-step calculations, dynamic analysis) require long-chain reasoning and multi-source integration, exposing most models' weakness in complex logic. In contrast, S2 tasks, supported by rich textual cues without deep multimodal integration, form a "comfort zone" for models. This suggests that while current MLLMs possess basic visual reasoning, their pure visual understanding remains limited. Incorporating textual information significantly enhances performance, confirming its auxiliary role in visual-physical semantic parsing.

A noteworthy performance differentiation exists in the data:**Most models show high convergence in accuracy on S1 and S3 tasks (small difference), while leading models (Gemini 2.5 pro, Doubao Seed 1.6) show significant differences their S3 task performance approaches the optimal level of S2, yet forms a clear gap with S1 tasks.**This characteristic reveals that leading models have significantly advanced capabilities in deep reasoning (especially complex logic chain construction and multi-step problem solving), enabling them to effectively handle the high complexity demands of S3 tasks. However, insufficient image understanding capability remains a common bottleneck for all models; even leading models have not yet broken through this technological limitation, reflecting that current multimodal models still have considerable room for improvement in fine-grained visual semantic parsing.

5.6 Auxiliary Evaluation - Symbol Understanding

We selected 50 electromagnetism circuit diagrams to evaluate large models' visual understanding through symbol recognition and circuit analysis tasks. Circuit diagrams are ideal for this purpose due to three key features: the standardized yet complex symbol system tests accurate recognition and logical reasoning; adjustable task difficulty from simple loops to multi-branch circuits enables layered assessment of parsing and reasoning ability; and their cross-modal natureintegrating visual symbols with physical principleseffectively examines visual-semantic mapping.

Evaluation revealed polarized outcomes: half of the models handled both simple and complex tasks robustly. Doubao Seed 1.6 and Gemini 2.5 Pro excelled in comprehension and explanation, while Qianfan-Check-VL and Qianfan-QI-VL failed even basic tasks. The remaining models managed simple symbol understanding but struggled with complex logical analysis, showing incomplete or interrupted reasoning.(comparisons between answers from Doubao Seed 1.6/Gemini 2.5 pro and general models are in Appendix H)

6 Conclusion and Outlook

This paper introduces MV-Physics, a specialized benchmark for physical visual question answering designed to evaluate the practical problem-solving capabilities of multimodal large language models (MLLMs). The benchmark consists of Chinese junior high physics questions annotated across multiple dimensionsincluding subject branch, question type, image relevance, text length, and difficultyto enable comprehensive ability assessment.

Evaluation of 22 mainstream MLLMs reveals a clear performance divide: Gemini 2.5 Pro and Doubao Seed 1.6 achieve strong results, whereas most models fall below a passing threshold.

Future work will extend MV-Physics to high school and interdisciplinary content, incorporate finer annotations such as conceptual abstraction and dynamic complexity, and provide ongoing evaluation to track MLLM progress in physical reasoning.

Limitations

The MV-Physics benchmark is limited to Chinese junior high physics content, lacking coverage of advanced or interdisciplinary topics, which restricts evaluation of complex reasoning.

The study evaluated only 22 mainstream multimodal models, omitting specialized or newer architectures, limiting generalizability.

Annotation relied on manual labeling, introducing potential subjectivity that may affect evaluation precision.

Data Availability

The original questions in this study are sourced from Xuekewang. We confirm their use is strictly limited to non-commercial academic research, consistent with the purposes for which we accessed the data. The MV-Physics benchmark dataset built from these questions will be publicly available upon the papers acceptance. Though we selected the Creative Commons Attribution 4.0 International (CC BY 4.0) license to facilitate academic sharing, this dataset is authorized solely for

academic researchany derived research outcomes shall not be used commercially, to comply with the original datas access conditions and respect its copyright.

References

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

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*.

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.

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, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad - level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

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*.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, NicholasJing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations.

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.

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.

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.

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

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

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.

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.

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*.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin 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)*. 45 pages; Accepted by CVPR 2025.

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.

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.

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

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 3: Summary of API Platforms, Corresponding Models, and Endpoints.

B Evaluation Prompt

#Chinese Prompt

单选题: "你是一个专业的问题解答助手。请根据问题，从给定的选项中选择唯一正确的答案，并只输出选项前的数字（0,1,2,3）。不要添加任何解释或额外文本。如果无法确定，输出-1。"

多选题:"你是一个专业的物理问题解答助手。请根据问题，从给定的选项中选择**所有正确的答案**,并只输出选项前的数字，用逗号分隔（如0,2,3）。不要添加任何解释或额外文本。

填空题:你是一个专业的问题解答助手。请根据问题，直接输出空白处应填写的答案内容。只输出答案本身，不要添加任何解释、单位或额外文本。如果无法确定答案，请输出'UNKNOWN'。"

Figure 7: Overall Evaluation Prompt in Chinese.

#English Translation

Single-choice question: "You are a professional assistant for solving problems. Please select the only correct answer from the given options based on the question, and output only the number before the option (0, 1, 2, 3). Do not add any explanations or additional text. If it cannot be determined, output -1."

Multiple-choice question: "You are a professional assistant for solving physics problems. Please select all correct answers from the given options based on the question, and output only the numbers before the options, separated by commas (e.g., 0,2,3). Do not add any explanations or additional text."

Fill-in-the-blank question: You are a professional assistant for solving problems. Please directly output the content that should be filled in the blank according to the question. Output only the answer itself, without any explanations, units, or extra text. If the answer cannot be determined, output 'UNKNOWN'."

Figure 8: Overall Evaluation Prompt's English Translation.

#Chinese Prompt

单选题: "你是一个专业的物理问题解答助手。请根据提供的图像和问题，从给定的选项中选择唯一正确的答案，并只输出选项前的数字（0,1,2,3）。不要添加任何解释或额外文本。如果无法确定，输出-1。"

多选题:"你是一个专业的物理问题解答助手。请根据提供的图像和问题，从给定的选项中选择**所有正确的答案**，并只输出选项前的数字，用逗号分隔（如0,2,3）。不要添加任何解释或额外文本。如果无法确定，输出空列表。"

填空题:"你是一个专业的物理问题解答助手。请根据提供的图像和问题，直接输出空白处应填写的答案内容。只输出答案本身，不要添加任何解释、单位或额外文本。如果无法确定答案，请输出'UNKNOWN'。"

Figure 9: Prompt in Chinese for Image Removal Evaluation in Weak Scenarios.

#English Translation

Single-choice question: "You are a professional physics problem-solving assistant. Please select the only correct answer from the given options based on the provided image and question, and output only the number before the option (0, 1, 2, 3). Do not add any explanations or additional text. If it cannot be determined, output -1."

Multiple-choice question: "You are a professional assistant for solving physics problems. Please select all correct answers from the given options according to the provided image and question, and output only the numbers before the options, separated by commas (e.g., 0,2,3). Do not add any explanations or extra text. If it cannot be determined, output an empty list."

Fill-in-the-blank question: "You are a professional assistant for solving physics problems. Please directly output the content that should be filled in the blank according to the provided image and question. Output only the answer itself, without any explanations, units, or extra text. If the answer cannot be determined, output 'UNKNOWN'."

Figure 10: Prompt's English Translation for Image Removal Evaluation in Weak Scenarios.

C Detailed Analysis of Model Performance Across Disciplinary Branches

Performance by Discipline: Significant Differences in Disciplinary Difficulty and Differentiation of Models' Advantageous Fields

(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).

D Performance Analysis by Question Type

Models	SC	MC	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

SC=Single Choice,MC=Multiple Choice,F=Fill-in-the-blank

Table 4: Total Accuracy of Models Across Different Question Types.

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

Core Characteristic:Multiple-choice 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-choice 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-choice questions, frequently resulting in "omissions and misselections."

Notable Phenomenon: Mid-tier models (e.g., Qwen2.5-VL-72B, InternVL3-38B) have multiple-choice 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: Lack of Options Tests Direct Output Capability, Revealing Model "Bias"

Core Characteristic: Fill-in-the-blank questions offer no option prompts, requiring models to generate answers directly (e.g., terms, numerical values, short phrases). This demands high "accurate knowledge recall + question stem logic conversion" ability. Accuracy scores typically fall between those of single-choice and multiple-choice questions.

Top-Tier Models: Gemini 2.5 pro (87.77%) and Doubao Seed 1.6 (81.93%) remain in the top two positions, and the gap with the third-place model (Gemini 1.5 Pro, 80.70%) is relatively small, demonstrating stability in "outputting correct answers without aids."

Special Case Model:Gemini 1.5 Pro exhibits a "bias" in performance: its fill-in-the-blank score is 80.70% (second only to the top models), but its single-choice (57.62%) and multiple-choice (27.50%) scores are considerably lower. This suggests an advantage in "direct answer generation" tasks but a weakness in "option judgment" tasks, potentially related to the model's adaptation to different "output formats (options vs. direct text)."

Bottom-Tier Models: Claude 3.5 Sonnet (24.76%) and Claude Sonnet 4 (24.88%) scored the lowest, indicating a very weak ability for direct conversion from "question stem answer," and an inability to extract key information from the stem and output it accurately.

E Further Analysis on Image Relevance

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

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, indicating 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 utiliza-

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

Models	SC		MC		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 7: Comparison Between Overall Scores and Strong Scores Across Various Question Types.

Note: Overall Score refers to the accuracy rate in regular scenarios where images may serve as supplementary information (non-mandatory); Strong Score refers to the accuracy rate in scenarios where images are "mandatory problem-solving information" (problems cannot be solved without images).

Models	SC		MC		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 8: Example of Accuracy Differences Between Overall Accuracy and Strong Accuracy Across Various Question Types.

Note: Difference (Accuracy Score - Overall Accuracy): A negative value indicates performance degradation in Strong scenarios (the closer to 0, the stronger the stability); a positive value indicates performance improvement in Strong scenarios (rarely occurs, mostly in special cases).

tion" abilities. This gap appears "irreversible."

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 Hierarchy

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

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-Choice 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-choice (-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

Models	Overall	sc_w	sc_w_np	mc_w	mc_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 9: Total Table of Models' Accuracy Rates Across Various Question Types Before and After Image Removal in Weak Scenarios.

Models	sc_w	sc_w_np
Qwen VL-Plus	82.47	83.54 (+1.07)
Qwen VL-Max	90.35	90.42 (+0.07)
Qwen2.5-VL-32B-Instruct	85.61	87.76 (+2.15)
Qianfan-QI-VL	56.93	71.13 (+14.20)
InternVL3-38B	86.60	88.43 (+1.83)
InternVL2_5-38B-MPO	83.13	84.91 (+1.78)
ERNIE 4.5 Turbo VL	88.37	90.91 (+2.54)
ERNIE 4.5	86.25	87.00 (+0.75)
doubao Seed 1.6	97.25	97.51 (+0.26)
GPT-4o	69.08	71.99 (+2.91)
GPT-4.5-preview	57.95	61.96 (+4.01)
Claude 3.5 Sonnet	68.90	76.95 (+8.05)
Claude Sonnet 4	67.10	78.85 (+11.75)
Gemini-2.5-pro	95.76	96.27 (+0.51)
Grok-4	53.56	59.19 (+5.63)

Models	sc_w	sc_w_np
Qwen2.5-VL-72B-Instruct	89.88	89.47 (-0.41)
Qwen2.5-VL-7B-Instruct	81.64	81.14 (-0.50)
Qianfan-Llama-VL-8B	68.98	68.30 (-0.68)
Qianfan-Check-VL	75.44	74.18 (-1.26)
DeepSeek-VL2	47.38	43.98 (-3.40)
InternVL3-14B	83.13	82.04 (-1.09)
Gemini 1.5 Pro	75.18	73.71 (-1.47)

Models	mc_w	mc_w_np
Qwen VL-Plus	40.32	41.27 (+0.95)
Qwen VL-Max	67.74	68.25 (+0.51)
Qwen2.5-VL-7B-Instruct	39.68	45.90 (+6.22)
Qianfan-Llama-VL-8B	28.57	33.33 (+4.76)
Qianfan-QI-VL	0	1.59 (+1.59)
DeepSeek-VL2	14.29	15.87 (+1.58)
InternVL3-38B	60.32	61.40 (+1.08)
ERNIE 4.5 Turbo VL	54.84	66.67 (+11.83)
ERNIE 4.5	50.79	50.88 (+0.09)
Claude 3.5 Sonnet	14.29	30.16 (+15.84)
Claude Sonnet 4	12.70	26.98 (+14.28)
Grok-4	17.46	36.54 (+19.08)

Models	mc_w	mc_w_np
Qwen2.5-VL-72B-Instruct	69.35	67.86 (-1.49)
Qwen2.5-VL-32B-Instruct	68.25	50.79 (-17.46)
Qianfan-Check-VL	26.98	22.22 (-4.76)
InternVL3-14B	60.32	50.00 (-10.32)
InternVL2_5-38B-MPO	53.97	24.14 (-29.83)
doubao Seed 1.6	79.37	75.41 (-3.96)
GPT-4o	34.92	33.33 (-1.59)
GPT-4.5-preview	17.46	15.87 (-1.59)
Gemini-2.5-pro	87.10	85.25 (-1.85)
Gemini 1.5 Pro	36.51	30.16 (-6.35)

Models	fb_w	fb_w_np
Qwen VL-Max	81.62	85.97 (+4.35)
doubao Seed 1.6	95.80	96.42 (+0.62)
Claude 3.5 Sonnet	38.58	61.14 (+22.56)
Claude Sonnet 4	39.55	60.03 (+20.48)
Grok-4	55.15	86.07 (+30.92)

Models	fb_w	fb_w_np
Qwen VL-Plus	82.26	80.61 (-1.65)
Qwen2.5-VL-72B-Instruct	87.88	84.59 (-3.29)
Qwen2.5-VL-32B-Instruct	88.13	81.57 (-6.56)
Qwen2.5-VL-7B-Instruct	83.80	82.39 (-1.41)
Qianfan-Llama-VL-8B	77.79	73.87 (-3.92)
Qianfan-Check-VL	68.53	19.94 (-48.59)
Qianfan-QI-VL	64.12	27.10 (-37.02)
DeepSeek-VL2	61.21	44.70 (-16.51)
InternVL3-38B	87.57	76.34 (-11.23)
InternVL3-14B	83.94	81.02 (-2.92)
InternVL2_5-38B-MPO	85.03	77.58 (-7.45)
ERNIE 4.5 Turbo VL	90.60	89.82 (-0.78)
ERNIE 4.5	85.89	81.43 (-4.46)
GPT-4o	75.49	73.90 (-1.59)
GPT-4.5-preview	48.75	40.47 (-8.28)
Gemini-2.5-pro	96.66	96.24 (-0.42)
Gemini 1.5 Pro	94.69	83.01 (-11.68)

Figure 11: Comparison of Accuracy Increment of Models Before and After Image Removal in Weak Scenarios.

assistance. Positive Extreme (Significant improvement after removal):

Grok-4: Overall score increases significantly due to improvements in multiple-choice (+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-choice +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-Choice / 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 distract attention.

Multiple-Choice Questions (mc_w vs mc_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-choice 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-choice 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-choice 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-choice 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-choice (most complex) has the highest image dependency. Fill-in-the-blank, due to the "precise output" requirement, sees a polarized image role (assistance vs. interference).

F

Models	LC-1	LC-2	LC-3
Qwen-VL-Plus	55.24	63.78	58.66
Qwen-VL-Max	56.77	71.45	63.78
Qwen2.5-VL-72B-Instruct	57.64	72.64	67.91
Qwen2.5-VL-32B-Instruct	57.24	70.44	64.37
Qwen2.5-VL-7B-Instruct	54.97	63.10	58.81
Qianfan-Llama-VL-8B	37.28	51.07	47.57
Qianfan-Check-VL	42.76	49.95	42.19
Qianfan-QI-VL	32.75	43.04	40.49
DeepSeek-VL2	27.72	39.58	36.43
InternVL3-38B	57.02	68.26	63.44
InternVL3-14B	49.34	64.33	58.51
InternVL2_5-38B-MPO	52.85	64.08	57.66
ERNIE-4.5-Turbo-VL	56.58	73.23	71.26
ERNIE-4.5	67.11	71.42	62.49
doubao Seed 1.6	64.36	83.01	87.30
GPT-4o	41.76	53.53	50.24
GPT-4.5-preview	26.81	36.01	34.27
Claude 3.5 Sonnet	19.90	30.00	26.20
Claude Sonnet 4	18.80	29.88	26.03
Gemini 2.5 pro	74.61	87.71	90.83
Gemini 1.5 Pro	54.09	65.26	69.86
Grok-4	27.51	38.18	36.66

Table 10: Total Accuracy of Each Model Across Different Text Lengths.

Based on comprehensive scores across the three text length categories, models can be classified into three types: "All-Rounders," "Specialized," and "Underperformers."

1.All-Rounders (Leading scores across all length categories) Representative Models: Gemini 2.5 pro, doubao Seed 1.6

Core Characteristics:Scores show an upward trend as text length increases (Gemini 2.5 pro: LC-1=74.61 LC-2=87.71 LC-3=90.83; Doubao: LC-1=64.36 LC-2=83.01 LC-3=87.30). They possess the strongest long-text (LC-3) processing capability. Notably, Gemini 2.5 pro's LC-3 score of 90.83 is the only one exceeding 90 among all models, reflecting its top-tier ability to understand long-text context and integrate information. They also maintain high scores on short text (LC-1) (Gemini 2.5 pro 74.61, doubao 64.36), showing no obvious weaknesses.

2.Specialized (Excelling in one length category, medium in others)

Representative Models: ERNIE-4.5, ERNIE-4.5-Turbo-VL, Qwen2.5-VL-72B-Instruct

Core Characteristics: ERNIE-4.5: Ranks second in short-text (LC-1=67.11) score among all models (second only to Gemini 2.5 pro), but performance slightly decreases on long text (LC-3=62.49), indicating a strength in short-text information extraction. ERNIE-4.5-Turbo-VL: Leads in medium-length (LC-2=73.23) and long-text (LC-3=71.26) scores, with medium performance on short text (LC-1=56.58), making it suitable for processing medium and longer text. Qwen2.5-VL-72B-Instruct: Maintains medium-to-upper level scores across all three lengths (LC-1=57.64 LC-2=72.64 LC-3=67.91), performs best on medium length (LC-2), and is overall balanced but not top-tier.

3. Underperformers (Low scores across all length categories)

Representative Models: Claude series (3.5 Sonnet, Sonnet 4), DeepSeek-VL2, Grok-4

Core Characteristics:Scores are below 40 across all lengths (Claude Sonnet 4: LC-1=18.80 LC-2=29.88 LC-3=26.03). This is the worst-performing echelon . The small fluctuation in scores across lengths indicates overall weak text comprehension capability, unaffected by length (due to insufficient foundational ability). Short-text (LC-1) scores are the lowest (Claude series all below 20), indicating particularly weak short-text information capture capability.

Correlation between Model Scale and Text Length Processing Capability

Large Models (Parameter scale 32B): e.g., Gemini 2.5 pro, Qwen2.5-VL-72B/32B-Instruct. Generally exhibit the pattern "Long-text score > Medium-length score > Short-text score." Performance improves with increasing length, reflecting the advantage of large models in modeling complex contexts.

Small/Medium Models (Parameter scale 14B):e.g., Qwen2.5-VL-7B-Instruct, InternVL3-14B. Overall scores are lower than large models, and long-text (LC-3) scores do not significantly exceed short-text scores, indicating limited long-text processing capability.

Core Conclusions

Long-text processing capability is a hallmark of top-tier models: Gemini 2.5 pro and doubao score highest on long text (LC-3), and their performance improves with length, reflecting their advantage in integrating complex semantics and multiple information points.

Short-text processing is not necessarily easier:Some models (e.g., Claude series) score lowest on short text, indicating that limited information and reliance on precise semantic matching in short text place higher demands on a model's detailed understanding.

Models	Overall	LC-1	LC-1_W	LC-1_S	LC-2	LC-2_W	LC-2_S	LC-3	LC-3_W	LC-3_S
Qwen-VL-Plus	60.24	55.24	89.74	52.03	63.78	84.70	57.99	58.66	76.77	53.31
Qwen-VL-Max	65.98	56.77	92.31	53.46	71.45	89.95	66.33	63.78	79.63	59.09
Qwen2.5-VL-72B-Instruct	68.78	57.64	94.87	54.18	72.64	92.92	67.02	67.91	83.90	63.18
Qwen2.5-VL-32B-Instruct	65.86	57.24	94.87	53.72	70.44	91.47	64.66	64.37	82.51	58.72
Qwen2.5-VL-7B-Instruct	60.04	54.97	92.31	51.45	63.10	85.38	56.92	58.81	77.20	53.30
Qianfan-Llama-VL-8B	47.91	37.28	87.18	32.61	51.07	78.57	43.51	47.57	67.42	41.49
Qianfan-Check-VL	45.05	42.76	76.92	39.57	49.95	75.11	43.02	42.19	64.20	35.66
Qianfan-QI-VL	40.77	32.75	32.20	32.83	43.04	61.77	40.23	40.49	56.57	37.15
DeepSeek-VL2	36.86	27.72	51.28	25.49	39.58	58.84	34.30	36.43	51.14	32.05
InternVL3-38B	61.90	57.02	97.44	53.24	68.26	91.24	61.93	6.42	81.81	56.66
InternVL3-14B	59.86	49.34	94.87	45.08	64.33	88.94	57.55	58.51	77.72	52.81
InternVL2_5-38B-MPO	59.59	52.85	97.44	48.68	64.08	88.02	57.49	57.66	78.67	51.42
ERNIE-4.5-Turbo-VL	70.64	56.58	94.87	53.00	73.23	90.62	68.40	71.26	85.92	66.92
ERNIE-4.5	67.86	67.11	92.31	64.75	71.42	90.55	66.14	62.49	79.75	57.42
doubao Seed 1.6	83.80	64.36	97.44	61.22	83.01	95.67	79.42	87.30	95.14	84.87
GPT-4o	50.73	41.76	71.79	38.94	53.53	77.70	46.80	50.24	67.05	45.23
GPT-4.5-preview	34.28	26.81	55.88	24.21	36.01	53.87	31.08	34.27	46.42	30.61
Claude 3.5 Sonnet	26.96	19.90	50.00	17.37	30.00	48.82	34.62	26.20	37.96	22.60
Claude Sonnet 4	26.74	18.80	58.62	15.54	29.88	48.07	24.68	26.03	38.07	22.37
Gemini 2.5 pro	88.34	74.61	97.44	72.46	87.71	96.12	85.36	90.83	95.60	89.38
Gemini 1.5 Pro	66.93	54.09	87.18	50.66	65.26	85.65	59.48	69.86	83.46	65.48
Grok-4	38.25	27.51	48.72	25.54	38.18	58.18	32.62	36.66	49.37	32.89

Table 11: Models' Accuracy Across Different Text Loads (Including Weak and Strong Scenarios).

Note: 1. LC-1/LC-2/LC-3 = different text load levels; 2. Subscripts W = Weak scenario, S = Strong scenario.
Data explanation: All accuracy values are percentages, and higher values indicate better model performance.

Models	Overall	LC-1	LC-1_W	LC-1_S	LC-2	LC-2_W	LC-2_S	LC-3	LC-3_W	LC-3_S
Qwen-VL-Plus	60.24	78.48	95.65	75.56	71.58	87.00	66.67	64.14	75.82	60.98
Qwen-VL-Max	65.98	79.75	100.00	76.30	79.69	95.00	74.80	71.08	71.08	67.61
Qwen2.5-VL-72B-Instruct	68.78	83.54	100.00	80.74	78.84	94.00	74.00	71.14	84.07	67.66
Qwen2.5-VL-32B-Instruct	65.86	77.07	95.65	73.88	74.42	90.86	69.23	66.43	78.69	63.11
Qwen2.5-VL-7B-Instruct	60.04	80.25	95.65	77.61	67.48	86.80	61.38	61.42	74.32	57.93
Qianfan-Llama-VL-8B	47.91	66.88	82.61	64.18	53.96	74.62	47.44	46.15	61.20	42.07
Qianfan-Check-VL	45.05	63.06	82.61	59.70	57.61	81.22	50.16	49.52	68.16	44.46
Qianfan-QI-VL	40.77	47.77	56.52	45.86	48.28	62.56	43.78	41.62	50.84	39.17
DeepSeek-VL2	36.86	41.40	47.83	40.30	35.65	54.87	29.65	32.28	39.34	30.37
InternVL3-38B	61.90	78.98	100.00	75.37	72.11	93.91	65.22	63.64	77.05	60.00
InternVL3-14B	59.86	73.89	95.65	70.15	69.67	89.00	63.30	59.67	74.32	55.85
InternVL_5-38B-MPO	59.59	75.80	100.00	71.64	68.57	89.85	61.86	60.14	73.77	56.44
ERNIE 4.5 Turbo VL	70.64	81.53	100.00	78.36	80.99	93.50	77.00	73.04	81.22	70.83
ERNIE 4.5	67.86	83.44	95.65	81.34	76.96	93.40	71.73	67.22	77.22	64.52
doubao Seed 1.6	83.80	90.20	100.00	88.46	89.07	98.50	85.95	89.51	95.48	87.82
GPT-4o	50.73	66.03	82.61	63.16	54.49	79.80	46.34	45.87	55.56	43.21
GPT-4.5-preview	34.28	53.51	66.67	51.04	47.08	63.76	41.56	40.71	49.14	38.39
Claude 3.5 Sonnet	26.96	52.87	57.14	52.05	54.43	73.74	45.61	46.30	62.75	41.21
Claude Sonnet 4	26.74	51.81	69.23	48.57	50.92	67.35	43.86	48.06	65.91	43.21
gemini-2.5-pro	88.34	91.03	100.00	89.47	89.91	96.50	87.77	91.56	94.38	90.78
Gemini 2.5 Flash	77.35	82.83	100.00	78.21	74.50	93.18	69.23	66.05	77.04	63.18
Gemini 1.5 Pro	66.93	68.99	91.30	65.19	60.43	82.09	53.50	52.80	65.57	49.33

Table 12: Accuracy of Each Model Across Different Text Lengths in Single Choice Questions(Including Weak and Strong Scenarios).

Models	Overall	LC-1	LC-1_S	LC-1_W	LC-2	LC-2_S	LC-2_W	LC-3	LC-3_S	LC-3_W
Qwen-VL-Plus	60.24	30.56	29.41	50.00	21.95	20.36	36.00	24.55	21.90	42.86
Qwen-VL-Max	65.98	47.22	47.06	50.00	52.03	49.77	72.00	53.43	51.65	65.71
Qwen2.5-VL-72B-Instruct	68.78	39.89	38.24	50.00	48.78	46.15	72.00	53.07	50.83	68.57
Qwen2.5-VL-32B-Instruct	65.86	50.00	50.00	50.00	51.01	48.42	73.08	42.96	39.67	65.71
Qwen2.5-VL-7B-Instruct	60.04	27.78	26.47	50.00	30.77	29.41	42.31	27.80	26.45	37.14
Qianfan-Llama-VL-8B	47.91	13.89	11.76	50.00	24.70	24.43	26.92	25.63	25.21	28.57
Qianfan-Check-VL	45.05	11.11	11.76	0	22.27	20.36	38.46	15.52	14.88	20.00
Qianfan-QI-VL	40.77	0	0	-	0.81	0.78	0	0.36	0	0
DeepSeek-VL2	36.86	8.33	8.82	0	18.62	19.91	7.69	13.36	12.40	20.00
InternVL3-38B	61.90	52.78	52.94	50.00	55.47	54.30	65.38	42.60	40.50	57.14
InternVL3-14B	59.86	44.44	44.12	50.00	46.56	44.34	65.38	36.82	33.88	57.14
InternVL2_5-38B-MPO	59.59	55.56	55.88	50.00	44.94	43.44	57.69	33.57	30.99	51.43
ERNIE-4.5-Turbo-VL	70.64	44.44	44.12	50.00	44.67	13.12	57.69	44.44	43.22	52.94
ERNIE-4.5	67.86	58.33	58.82	50.00	43.72	42.53	53.85	35.74	33.88	48.57
doubao Seed 1.6	83.80	60.61	61.29	50.00	66.95	67.14	65.38	87.45	86.76	91.43
GPT-4o	50.73	25.00	23.53	50.00	30.89	30.00	38.46	25.63	24.79	31.43
GPT-4.5-preview	34.28	16.67	14.71	50.00	12.96	12.67	15.38	14.44	14.05	17.14
Claude 3.5 Sonnet	26.96	11.11	8.82	50.00	10.12	9.95	11.54	14.08	14.05	14.29
Claude Sonnet 4	26.74	11.11	8.82	50.00	10.93	10.41	15.38	12.27	12.81	8.57
Gemini 2.5 pro	88.34	69.70	70.97	50.00	78.69	78.08	84.00	90.00	89.78	91.43
Gemini 1.5 Pro	66.93	25.00	26.47	0	27.94	26.70	38.46	27.44	26.03	37.14
Grok-4	38.25	16.67	14.71	50.00	11.74	10.41	23.08	16.61	17.36	11.43

Table 13: Accuracy of Each Model Across Different Text Lengths in Multiple Choice Questions(Including Weak and Strong Scenarios).

Note: There are only 2 questions for LC-1_W, which is not included in the overall pattern).

Models	Overall	LC-1	LC-1_S	LC-1_W	LC-2	LC-2_S	LC-2_W	LC-3	LC-3_S	LC-3_W
Qwen-VL-Plus	60.24	44.70	42.40	85.71	67.83	61.90	88.26	61.09	54.96	79.55
Qwen-VL-Max	65.98	44.32	42.00	85.71	69.30	64.08	87.32	62.07	56.42	79.02
Qwen2.5-VL-72B-Instruct	68.78	44.70	42.00	92.86	73.42	67.35	94.37	68.60	63.16	84.93
Qwen2.5-VL-32B-Instruct	65.86	46.39	43.37	100	72.06	65.67	94.31	66.16	59.84	85.13
Qwen2.5-VL-7B-Instruct	60.04	43.46	40.65	92.86	67.82	61.54	89.42	62.09	55.65	81.15
Qianfan-Llama-VL-8B	47.91	22.81	18.47	100	55.45	45.91	88.63	51.04	43.89	72.51
Qianfan-Check-VL	45.05	34.98	32.53	78.57	50.53	43.78	73.93	42.83	35.13	65.92
Qianfan-QI-VL	40.77	28.52	27.71	42.86	49.63	44.15	68.75	45.69	40.03	62.76
DeepSeek-VL2	36.86	22.09	19.67	64.29	48.60	42.72	68.90	41.57	36.12	57.82
InternVL3-38B	61.90	44.49	41.37	100	68.26	61.42	91.94	64.69	57.79	85.34
InternVL3-14B	59.86	35.36	31.73	100	64.33	56.63	91.00	61.02	54.53	80.45
InternVL2_5-38B-MPO	59.59	38.78	35.34	100	65.18	58.00	90.05	59.98	52.48	82.45
ERNIE-4.5-Turbo-VL	70.64	43.35	40.56	92.86	73.83	68.56	91.94	74.19	68.94	89.96
ERNIE-4.5	67.86	58.56	56.63	92.86	73.86	68.53	92.42	64.22	58.01	82.89
doubao Seed 1.6	83.80	49.81	46.99	100	81.91	77.58	96.71	86.36	83.31	95.28
GPT-4o	50.73	29.66	28.11	57.14	58.64	52.27	80.57	55.61	49.52	73.82
GPT-4.5-preview	34.28	16.67	15.20	42.86	35.03	30.19	51.64	35.31	31.11	47.86
Claude 3.5 Sonnet	26.96	10.23	8.40	42.86	26.85	22.51	41.78	25.70	21.91	37.07
Claude Sonnet 4	26.74	9.47	7.20	50.00	27.57	23.01	43.19	25.66	21.65	37.68
Gemini 2.5 pro	88.34	65.53	63.6	100	88.15	85.52	97.18	90.63	88.69	96.33
Gemini 1.5 Pro	66.93	48.20	45.19	92.86	80.87	76.43	95.45	84.67	81.41	94.41
Grok-4	38.25	16.67	15.20	42.86	40.59	34.52	61.50	38.54	33.81	52.75

Table 14: Accuracy of Each Model Across Different Text Lengths in Fill-in-the-Blank Questions(Including Weak and Strong Scenarios).

Model scale positively correlates with long-text ability: Large models generally excel at long text, while small/medium models have limited performance on long text, suggesting that increasing parameter scale may be an effective path to enhancing long-text processing capability.

G Models' Accuracy Across Difficulty Gradients

Models	S1	S2	S3
Qwen-VL-Plus	50.71	66.50	50.33
Qwen-VL-Max	55.58	69.99	53.34
Qwen2.5-VL-72B-Instruct	54.68	73.17	56.85
Qwen2.5-VL-32B-Instruct	54.05	70.52	51.21
Qwen2.5-VL-7B-Instruct	52.55	63.89	46.47
Qianfan-Llama-VL-8B	31.53	52.11	38.32
Qianfan-Check-VL	52.55	63.89	46.47
Qianfan-QI-VL	37.07	43.24	31.44
DeepSeek-VL2	30.40	39.53	28.19
InternVL3-38B	50.72	69.06	48.84
InternVL3-14B	46.57	64.98	43.87
InternVL2_5-38B-MPO	44.77	64.73	44.44
ERNIE-4.5-Turbo-VL	55.50	74.80	60.02
ERNIE-4.5	65.59	69.22	51.56
doubao Seed 1.6	59.64	86.86	84.95
GPT-4o	34.42	56.69	33.37
GPT-4.5-preview	23.25	37.83	24.97
Claude 3.5 Sonnet	15.98	31.06	16.24
Claude Sonnet 4	14.34	30.82	17.20
Gemini 2.5 pro	66.19	90.54	88.37
Gemini 1.5 Pro	56.51	70.45	56.63
Grok-4	25.36	40.08	26.24

Table 15: Models' Accuracy Across Difficulty Gradients.

H Example of Circuit Comprehension Answer and Dataset composition ratios

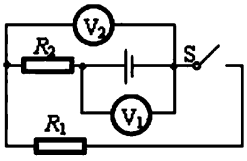


Figure 12: Simple Circuit Diagram.

符号解释：
1.R_1:电阻符号，表示电路中的一个电阻元件。电阻是阻碍电流流动的物理量，其值通常用欧姆（Ω）来表示。
2.R_2:电阻符号，表示电路中的另一个电阻元件，与R_1并联连接。
3.V_1:电压表符号，用于测量电路中某一点的电压。在图中，V_1测量的是电阻R_1两端的电压。
4.V_2:电压表符号，用于测量电路中另一点的电压。在图中，V_2测量的是电阻R_2两端的电压。
5.S:开关符号，表示电路中的一个开关元件。开关可以控制电路的通断，当开关S闭合时，电路导通；当开关S断开时，电路断开。
6.电源:图中未直接显示电源的具体符号，但根据电路图的上下文，电源的存在可以通过电压表V_1和V_2的读数来推断。电源提供电能，使电流能够在电路中流动。
7.箭头:在电压表V_1和V_2的连接线上有箭头，表示电压的方向。电压表的正极通常标记为“+”，负极标记为“-”。箭头指向电压表的正极端，表示电压的方向是从正极到负极。
综上所述，图中的符号分别代表了电阻、电压表、开关以及电源等基本电路元件，并通过这些元件展示了电路的基本结构和功能。

Figure 13: General Model Response in Chinese for Simple Circuit Diagram.

R_1:
Physical Meaning: R is the symbol for Resistance, and the subscript "1" is used to distinguish it from other resistors in the circuit. The rectangular box represents the standard circuit symbol for a fixed resistor.
Contextual Meaning: This is the first resistor in the circuit, connected in series with resistor R_2. When the switch is closed, current flows through it, producing a voltage drop.
R_2:
Physical Meaning: R is the symbol for Resistance, and the subscript "2" is used to distinguish it from other resistors in the circuit. The rectangular box represents the standard circuit symbol for a fixed resistor.
Contextual Meaning: This is the second resistor in the circuit, connected in series with R_1. Voltmeter V_2 is connected in parallel with it to measure the voltage across it.
Power Supply Symbol (Two Parallel Lines of Unequal Length):
Physical Meaning: This is the circuit symbol for a direct current (DC) power supply, typically representing a battery or battery pack. The longer line represents the positive terminal (+), and the shorter line represents the negative terminal (-).
Contextual Meaning: This is the energy source for the entire circuit, providing electromotive force (EMF) to drive the current.
V_1 (Circled V_1):
Physical Meaning: V is the symbol for a Voltmeter, and the subscript "1" is used to distinguish it from other voltmeters in the circuit. A voltmeter is used to measure the potential difference (voltage) between two points in a circuit.
Contextual Meaning: Voltmeter V_1 is connected in parallel with the power supply. Therefore, it measures the terminal voltage of the power supply. When switch S is open, it measures the electromotive force (EMF) of the power supply; when switch S is closed, it measures the total voltage across the external circuit.
V_2 (Circled V_2):
Physical Meaning: V is the symbol for a Voltmeter, and the subscript "2" is used to distinguish it from other voltmeters in the circuit.
Contextual Meaning: Voltmeter V_2 is connected in parallel with resistor R_2. Its function is to measure the voltage drop across resistor R_2 when current flows through the circuit.
S (Breakpoint with a Slash):
Physical Meaning: S is the common identifier for a Switch. The symbol in the diagram represents a single-pole single-throw (SPST) switch in the open state.
Contextual Meaning: Switch S is the controller of the circuit. In the open state shown in the diagram, the circuit is incomplete, and no current flows. Closing switch S completes the circuit, allowing current to flow.
Straight Lines:
Physical Meaning: Straight lines represent wires, which are pathways for electric current.
Contextual Meaning: In an ideal circuit diagram, wires are considered ideal conductors with zero resistance. These wires connect the power supply, resistors, switch, and voltmeters to form a complete circuit.
Black Dots (Line Intersections):
Physical Meaning: Black dots represent nodes or junctions in the circuit.
Contextual Meaning: This indicates that at these points, multiple wires are electrically connected, allowing current to split or converge.

Figure 14: General Model Response 's English translation for Simple Circuit Diagram.

电源符号（长竖线+短竖线）：物理意义为直流电源，提供电路的电动势（电压）；上下文含义是电路的能源，长竖线代表正极，短竖线代表负极，为整个电路提供电压。 R_1：物理意义为电阻元件，表征对电流的阻碍作用（单位：Ω）；上下文含义是电路中的定值电阻之一，与R_2串联，其两端电压等于电源电压与R_2两端电压之差（由V_1和V_2读数计算）。 R_2：物理意义为电阻元件，表征对电流的阻碍作用（单位：Ω）；上下文含义是电路中的定值电阻之一，与R_1串联，其两端电压由电压表V_2直接测量。 V_1：物理意义为电压表，用于测量电路两点间电压（单位：V）；上下文含义是并联在电源两端，测量电源输出电压（电路总电压）。 V_2：物理意义为电压表，用于测量电路两点间电压（单位：V）；上下文含义是并联在电阻R_2两端，测量R_2两端的电压降。 S：物理意义为单刀单掷开关，用于控制电路通断；上下文含义是串联在电源正极与电路之间，闭合时电路接通（电流流过R_1和R_2），断开时电路断路（无电流）。	Power Supply Symbol (long line + short line): Physical meaning: Represents a DC power source, providing electromotive force (voltage) to the circuit.Contextual meaning: Serves as the energy source of the circuit. The long line denotes the positive terminal, and the short line denotes the negative terminal, supplying voltage to the entire circuit. R_1: Physical meaning: A resistive component, representing opposition to electric current (unit: Ω).Contextual meaning: One of the fixed resistors in the circuit, connected in series with R_2. The voltage across it equals the difference between the power supply voltage and the voltage across R_2 (calculated using readings from V_1 and V_2). R_2: Physical meaning: A resistive component, representing opposition to electric current (unit: Ω).Contextual meaning: One of the fixed resistors in the circuit, connected in series with R_1. The voltage across it is directly measured by voltmeter V_2. V_1: Physical meaning: A voltmeter, used to measure the voltage between two points in the circuit (unit: V).Contextual meaning: Connected in parallel across the power supply, measuring the output voltage (total circuit voltage). V_2: Physical meaning: A voltmeter, used to measure the voltage between two points in the circuit (unit: V).Contextual meaning: Connected in parallel across resistor R_2, measuring the voltage drop across R_2. S: Physical meaning: A single-pole single-throw (SPST) switch, used to control the circuit's on/off state.Contextual meaning: Connected in series between the positive terminal of the power supply and the circuit. When closed, the circuit is energized (current flows through R_1 and R_2); when open, the circuit is disconnected (no current flows).

Figure 15: Response from doubao seed 1.6 for Simple Circuit Diagram.

R1： 物理意义:R是电阻（Resistance）的符号，下标“1”用于区分电路中的其他电阻。矩形框是固定电阻器的标准电路符号。 上下文含义:这是电路中的第一个电阻器，它与电阻器R_2串联连接。当开关闭合时，电流会流过它并产生电压降。 R_2： 物理意义:R是电阻（Resistance）的符号，下标“2”用于区分电路中的其他电阻。矩形框是固定电阻器的标准电路符号。 上下文含义:这是电路中的第二个电阻器，与R_1串联。电压表V_2与其并联，用于测量其两端的电压。 电源符号 (两条长短不一的平行线): 物理意义:这是直流（DC）电源的电路符号，通常代表电池或电池组。长线代表正极（+），短线代表负极（-）。 上下文含义: 这是整个电路的能量来源，提供电动势（EMF）来驱动电流。 V_1 (带圆圈的V_1): 物理意义:V是电压表（Voltmeter）的符号，下标“1”用于区分电路中的其他电压表。电压表用于测量电路中两点之间的电势差（电压）。 上下文含义:电压表V_1与电源并联。因此，它测量的是电源的端电压。当开关S断开时，它测量的是电源的电动势（EMF）；当开关S闭合时，它测量的是外电路的总电压。 V_2 (带圆圈的V_2): 物理意义: V是电压表（Voltmeter）的符号，下标“2”用于区分电路中的其他电压表。 上下文含义: 电压表V_2与电阻器R_2并联。它的作用是测量当电路中有电流通过时，电阻器R_2两端的电压降。 S (带斜线的断点): 物理意义: S是开关（Switch）的常用标识符。图中的符号是一个单刀单掷开关（SPST），并且处于断开状态 上下文含义: 开关S是电路的控制器。在图中所示的断开状态下，电路不完整，没有电流流过。闭合开关S将完成电路，使电流开始流动。 直线: 物理意义: 直线代表导线（Wire），即电流的通路。 上下文含义: 在理想电路图中，导线被认为是理想导体，其电阻为零。这些导线将电源、电阻、开关和电压表连接成一个完整的电路 黑点 (线路交汇处): 物理意义: 黑点表示电路的节点（Node）或连接点（Junction）。 上下文含义: 这表明在这些点上，多条导线被电性连接在一起，电流可以在此分流或汇合。	R_1: Physical Meaning: R is the symbol for Resistance, and the subscript "1" is used to distinguish it from other resistors in the circuit. The rectangular box represents the standard circuit symbol for a fixed resistor. Contextual Meaning: This is the first resistor in the circuit, connected in series with resistor R_2. When the switch is closed, current flows through it, producing a voltage drop. R_2: Physical Meaning: R is the symbol for Resistance, and the subscript "2" is used to distinguish it from other resistors in the circuit. The rectangular box represents the standard circuit symbol for a fixed resistor. Contextual Meaning: This is the second resistor in the circuit, connected in series with R_1. Voltmeter V_2 is connected in parallel with it to measure the voltage across it. Power Supply Symbol (Two Parallel Lines of Unequal Length): Physical Meaning: This is the circuit symbol for a direct current (DC) power supply, typically representing a battery or battery pack. The longer line represents the positive terminal (+), and the shorter line represents the negative terminal (-). Contextual Meaning: This is the energy source for the entire circuit, providing electromotive force (EMF) to drive the current. V_1 (Circled V_1): Physical Meaning: V is the symbol for a Voltmeter, and the subscript "1" is used to distinguish it from other voltmeters in the circuit. A voltmeter is used to measure the potential difference (voltage) between two points in a circuit. Contextual Meaning: Voltmeter V_1 is connected in parallel with the power supply. Therefore, it measures the terminal voltage of the power supply. When switch S is open, it measures the electromotive force (EMF) of the power supply; when switch S is closed, it measures the total voltage across the external circuit. V_2 (Circled V_2): Physical Meaning: V is the symbol for a Voltmeter, and the subscript "2" is used to distinguish it from other voltmeters in the circuit. Contextual Meaning: Voltmeter V_2 is connected in parallel with resistor R_2. Its function is to measure the voltage drop across resistor R_2 when current flows through the circuit. S (Breakpoint with a Slash): Physical Meaning: S is the common identifier for a Switch. The symbol in the diagram represents a single-pole single-throw (SPST) switch in the open state. Contextual Meaning: Switch S is the controller of the circuit. In the open state shown in the diagram, the circuit is incomplete, and no current flows. Closing switch S completes the circuit, allowing current to flow. Straight Lines: Physical Meaning: Straight lines represent wires, which are pathways for electric current. Contextual Meaning: In an ideal circuit diagram, wires are considered ideal conductors with zero resistance. These wires connect the power supply, resistors, switch, and voltmeters to form a complete circuit. Black Dots (Line Intersections): Physical Meaning: Black dots represent nodes or junctions in the circuit. Contextual Meaning: This indicates that at these points, multiple wires are electrically connected, allowing current to split or converge.

Figure 16: Response from Gemini 2.5 Pro for Simple Circuit Diagram.

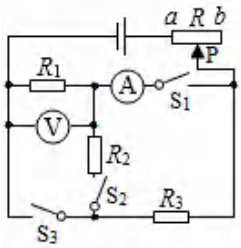


Figure 17: Complex Circuit Diagram.

符号解释: 1. R_1: 电阻符号，表示电路中的一个电阻元件。电阻用于限制电流的流动，其值通常用欧姆（Ω）来表示。 2. V: 电压表符号，用于测量电路中某两点之间的电势差（电压）。在图中，电压表连接在电阻R_1和R_2之间，用于测量这两个电阻之间的电压。 3. A: 电流表符号，用于测量通过电路中某点的电流强度。在图中，电流表连接在开关S_1闭合时的电路路径上，用于测量流过电阻R_1的电流。 4. S_1: 开关符号，表示可以控制电路通断的装置。在图中，开关S_1控制着整个电路的通断，当S_1闭合时，电流可以通过整个电路；当S_1断开时，电路被切断。 5. S_2: 开关符号，表示可以控制电路通断的装置。在图中，开关S_2控制着电阻R_2所在的支路，当S_2闭合时，电流可以通过电阻R_2；当S_2断开时，电阻R_2所在支路被切断。 6. S_3: 开关符号，表示可以控制电路通断的装置。在图中，开关S_3控制着电阻R_3所在的支路，当S_3闭合时，电流可以通过电阻R_3；当S_3断开时，电阻R_3所在支路被切断。 7. R: 变阻器符号，表示可以调节电阻大小的元件。在图中，变阻器R调节整个电路的总电阻，从而改变电路中的电流和电压。 8. a, b: 点标记，表示电路中的两个端点。在图中，a和b是电源的两端，电源提供电能，使电路中的电流能够流动。 9. 电源: 表示提供电能的装置，图中电源由a、b两点组成，为整个电路提供电压。 综上所述，图中的符号分别代表了电阻、电压表、电流表、开关、变阻器以及电源等基本电路元件，它们共同构成了一个简单的电路系统，用于测量和控制电路中的电流和电压。	Symbol Explanation: 1.R_1: Resistor symbol, representing a resistive component in a circuit. Resistors are used to limit the flow of electric current, and their values are typically measured in ohms (Ω). 2.V: Voltmeter symbol, used to measure the potential difference (voltage) between two points in a circuit. In the diagram, the voltmeter is connected between resistors R_1 and R_2 to measure the voltage across them. 3.A: Ammeter symbol, used to measure the current intensity at a specific point in the circuit. In the diagram, the ammeter is connected along the path when switch S_1 is closed, measuring the current flowing through resistor R_1. 4.S_1: Switch symbol, representing a device that controls the circuit's on/off state. In the diagram, switch S_1 controls the entire circuit—when closed, current flows through the circuit; when open, the circuit is interrupted. 5.S_2: Switch symbol, representing a device that controls the circuit's on/off state. In the diagram, switch S_2 controls the branch containing resistor R_2—when closed, current flows through R_2; when open, the branch is disconnected. 6.S_3: Switch symbol, representing a device that controls the circuit's on/off state. In the diagram, switch S_3 controls the branch containing resistor R_3—when closed, current flows through R_3; when open, the branch is disconnected. 7.R: Rheostat symbol, representing a component whose resistance can be adjusted. In the diagram, the rheostat R regulates the total resistance of the circuit, thereby altering the current and voltage. 8.a, b: Point markers, representing two endpoints in the circuit. In the diagram, a and b are the terminals of the power supply, which provides electrical energy to enable current flow in the circuit. 9. Power Supply: Represents the device providing electrical energy. In the diagram, the power supply consists of points a and b, supplying voltage to the entire circuit. In conclusion, the symbols in the diagram represent basic circuit components such as resistors, voltmeters, ammeters, switches, rheostats, and power supplies. Together, they form a simple circuit system for measuring and controlling current and voltage in the circuit.

Figure 18: General Model Response for Complex Circuit Diagram.

电源符号（两条不等长平行线）：提供直流电压的装置，是电路中电能的来源，维持电路持续电流。 R_1：定值电阻，电路中的固定电阻元件，用于阻碍电流，具体阻值需结合题目条件确定。 R_2：定值电阻，另一固定电阻元件，与电路中其他元件形成串联或并联关系（取决于开关状态）。 R_3：定值电阻，第三个固定电阻元件，通常位于特定支路中，参与电路的分压或限流。 A（电流表）：测量电流的仪器，串联在电路中，用于测量通过其所在支路的电流大小，单位为安培（A）。 V（电压表）：测量电压的仪器，并联在电路两点间，用于测量该两点间的电压差，单位为伏特（V）。 S_1：开关，控制滑动变阻器所在支路的通断，闭合时该支路接通，断开时支路断开。 S_2：开关，控制R_2所在支路的通断，闭合时R_2接入电路，断开时R_2被断路。 S_3：开关，通常为总开关或控制第三条支路，闭合时对应支路接通，断开时电路断开。 滑动变阻器（a、b接线柱）：a和b为滑动变阻器的两个固定接线柱，接入电路时提供总电阻长度。 P（滑片箭头）：滑动变阻器的可移动滑片，通过沿a-b方向滑动改变接入电路的电阻长度，从而调节电路电阻大小。 导线（直线）：连接电路元件的导体，形成电流通路，理想情况下电阻为0。	Power Supply Symbol (Two Unequal Parallel Lines): A device providing DC voltage, serving as the source of electrical energy in the circuit and sustaining continuous current flow. R_1: Fixed resistor, a constant resistive component in the circuit used to impede current. Its specific resistance value depends on the given conditions. R_2: Fixed resistor, another constant resistive component forming series or parallel connections with other elements in the circuit (depending on the switch states). R_3: Fixed resistor, the third constant resistive component, typically located in a specific branch to participate in voltage division or current limiting. A (Ammeter): An instrument for measuring current, connected in series within the circuit to measure the current intensity in its branch, with units in amperes (A). V (Voltmeter): An instrument for measuring voltage, connected in parallel between two points in the circuit to measure the voltage difference, with units in volts (V). S_1: Switch, controlling the on/off state of the branch containing the sliding rheostat. When closed, the branch is connected; when open, the branch is disconnected. S_2: Switch, controlling the on/off state of the branch containing R_2. When closed, R_2 is connected to the circuit; when open, R_2 is disconnected. S_3: Switch, typically serving as the main switch or controlling a third branch. When closed, the corresponding branch is connected; when open, the circuit is disconnected. Sliding Rheostat (Terminals a and b): a and b are the two fixed terminals of the sliding rheostat. When connected to the circuit, they provide the total resistance length. P (Slider Arrow): The movable slider of the sliding rheostat. By sliding along the a-b direction, it changes the resistance length connected to the circuit, thereby adjusting the circuit's resistance. Wires (Straight Lines): Conductors connecting circuit components to form current paths. Ideally, they have zero resistance.

Figure 19: Response from doubao seed 1.6 for Complex Circuit Diagram.

电源符号 (左上角): 物理意义是直流电源，为整个电路提供持续的电压和能量。图中长线代表电源正极，短线代表负极。 R:物理意义是滑动变阻器（Sliding Rheostat），是一种阻值可以连续变化的电阻器。在电路中，它用于调节电路中的总电流或分配电压。 a, b: 物理意义是滑动变阻器R的两个固定接线端。 P: 物理意义是滑动变阻器R的滑片（Wiper/Slider）。通过移动滑片P在电阻丝上的位置，可以改变接入电路部分的电阻值 指向P的箭头: 物理意义是可变性符号，表明该元件（滑动变阻器）的参数是可调节的，箭头指示了滑片P的可移动性。 R1: 物理意义是定值电阻器（Fixed Resistor）。在此电路中，它与电压表并联，与R2串联（当S2闭合时）。 R2: 物理意义是定值电阻器。它与开关S2串联，控制其是否接入电路。 \n- R3: 物理意义是定值电阻器。它与开关S3串联，构成一个独立的支路。 V (带圆圈): 物理意义是电压表（Voltmeter）。它用于测量电路中某两点间的电压差。在此图中，它并联在电阻R1的两端，用于测量R1上的电压。 A (带圆圈): 物理意义是电流表（Ammeter）。它用于测量通过电路某处的电流强度。在此图中，它串联在包含滑动变阻器和开关S1的支路中，用于测量该支路的电流。 S1: 物理意义是开关（Switch）。它用于控制电流表和滑动变阻器所在支路的通断。 S2: 物理意义是开关。它用于控制电阻R2是否接入电路。 S3: 物理意义是开关。它用于控制包含电阻R3的支路的通断。 直线: 物理意义是导线（Wire）。理论上视为无电阻的导体，用于连接各个电路元件，构成闭合回路。 黑点: 物理意义是节点（Node）或连接点。它表示三条或更多导线在此处电气连接，是电流分流或汇合的位置。	Power Supply Symbol (Upper Left Corner): Physically represents a DC power supply, providing continuous voltage and energy to the entire circuit. The longer line denotes the positive terminal, and the shorter line denotes the negative terminal. R: Physically represents a sliding rheostat, a resistor whose value can be continuously adjusted. In the circuit, it is used to regulate the total current or distribute voltage. a, b: Physically represent the two fixed terminals of the sliding rheostat R. P: Physically represents the slider/wiper of the sliding rheostat R. By moving the position of slider P along the resistance wire, the resistance value of the connected part can be changed. Arrow Pointing to P: Physically represents a variability symbol, indicating that the parameter of this component (the sliding rheostat) is adjustable. The arrow indicates the movability of slider P. R1: Physically represents a fixed resistor. In this circuit, it is connected in parallel with the voltmeter and in series with R2 (when S2 is closed). R2: Physically represents a fixed resistor. It is connected in series with switch S2, controlling whether it is接入电路 (connected to the circuit). R3: Physically represents a fixed resistor. It is connected in series with switch S3, forming an independent branch. V (Circled): Physically represents a voltmeter. It is used to measure the voltage difference between two points in the circuit. In this diagram, it is connected in parallel across resistor R1 to measure the voltage across R1. A (Circled): Physically represents an ammeter. It is used to measure the current intensity at a specific point in the circuit. In this diagram, it is connected in series in the branch containing the sliding rheostat and switch S1, measuring the current in that branch. S1: Physically represents a switch. It controls the on/off state of the branch containing the ammeter and the sliding rheostat. S2: Physically represents a switch. It controls whether resistor R2 is connected to the circuit. S3: Physically represents a switch. It controls the on/off state of the branch containing resistor R3. Straight Lines: Physically represent wires. Theoretically considered as conductors with zero resistance, used to connect various circuit components and form a closed loop. Black Dots: Physically represent nodes or connection points. They indicate where three or more wires are electrically connected, serving as points for current splitting or converging.

Figure 20: Response from Gemini 2.5 Pro for Complex Circuit Diagram.

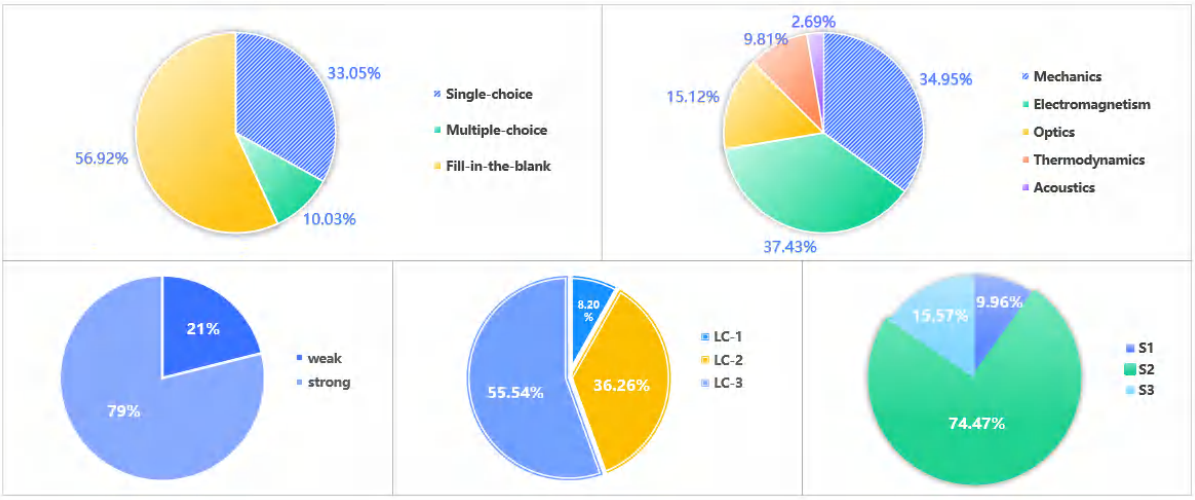


Figure 21: Dataset composition ratios. From top to bottom, left to right: proportion of problem types in MV-Physics; proportion of disciplinary branches in MV-Physics; proportion of image relevance levels in MV-Physics; proportion of textual load levels in MV-Physics; proportion of problem difficulty gradient levels in MV-Physics.