MME-SCI: A Comprehensive and Challenging Science Benchmark for Multimodal Large Language Models

Jiacheng Ruan[1*], **Dan Jiang**[1*], **Xian Gao**[1], **Ting Liu**[1], **Yuzhuo Fu**[1†], **Yangyang Kang**[2,3†]

[1] Shanghai Jiao Tong University
[2] Zhejiang University
[3] Bytedance
jackchenruan@sjtu.edu.cn

Abstract

Recently, multimodal large language models (MLLMs) have achieved significant advancements across various domains, and corresponding evaluation benchmarks have been continuously refined and improved. In this process, benchmarks in the scientific domain have played an important role in assessing the reasoning capabilities of MLLMs. However, existing benchmarks still face three key challenges: **1)** Insufficient evaluation of models' reasoning abilities in multilingual scenarios; **2)** Inadequate assessment of MLLMs' comprehensive modality coverage; **3)** Lack of fine-grained annotation of scientific knowledge points. To address these gaps, we propose MME-SCI, a comprehensive and challenging benchmark. We carefully collected 1,019 high-quality question-answer pairs, which involve 3 distinct evaluation modes. These pairs cover four subjects, namely mathematics, physics, chemistry, and biology, and support five languages: Chinese, English, French, Spanish, and Japanese. We conducted extensive experiments on 16 open-source models and 4 closed-source models, and the results demonstrate that MME-SCI is widely challenging for existing MLLMs. For instance, under the Image-only evaluation mode, o4-mini achieved accuracy of only 52.11%, 24.73%, 36.57%, and 29.80% in mathematics, physics, chemistry, and biology, respectively, indicating a significantly higher difficulty level compared to existing benchmarks. More importantly, using MME-SCI's multilingual and fine-grained knowledge attributes, we analyzed existing models' performance in depth and identified their weaknesses in specific domains. For example, in questions related to "Magnetic Field", o4-mini correctly answered only 5 out of 33 questions, thereby fine-grainedly exposing the model's vulnerabilities. These findings highlight the urgent need to enhance the scientific reasoning capabilities of MLLMs.

Code — https://github.com/JCruan519/MME-SCI

Introduction

In recent years, Multimodal Large Language Models (MLLMs) have achieved breakthrough progress in tasks such as visual question answering and multimodal reasoning (Wang et al. 2025c; Li et al. 2025b; Qu et al. 2025), with representative models including the GPT series (Achiam et al. 2023; OpenAI 2025), Qwen-VL series (Wang et al. 2024b; Bai et al. 2025), and InternVL series (Chen et al. 2024a; Zhu et al. 2025). However, as shown in Figure 1, compared with the rapid improvement of model capabilities, existing multimodal evaluation benchmarks—especially those for scientific-related scenarios—have shown a trend of "being unable to keep up". Specifically, on MMMU (Yue et al. 2024), a comprehensive benchmark based on university multidisciplinary problems, InternVL3-78B has achieved an accuracy of 72.2%, and Qwen2.5-VL-72B has also reached 68.2%. On AI2D (Kembhavi et al. 2016), a benchmark for scientific chart, even smaller models such as InternVL3-8B and Qwen2.5-VL-7B have achieved accuracies of 85.1% and 84.3%, respectively. These results indicate that the performance of advanced MLLMs on existing mainstream scientific benchmarks has approached a saturation level. Thus,

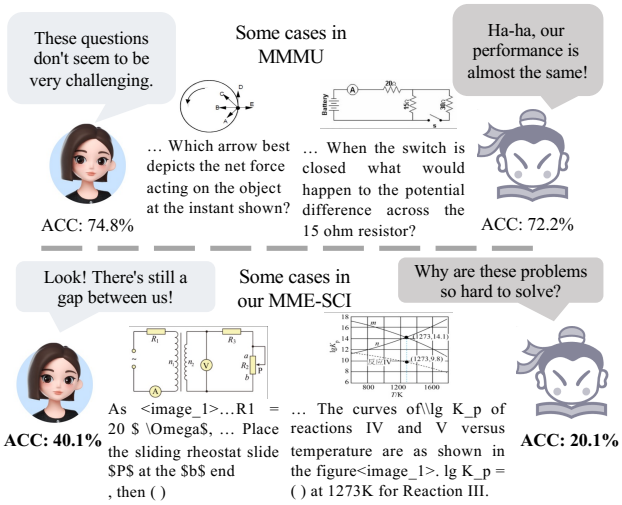

Figure 1: Comparison between prevalent benchmarks (e.g., MMMU) and MME-SCI. Existing benchmarks for MLLMs have become saturated and fail to distinguish performance differences among models. ACC stands for accuracy.

*These authors contributed equally.
†Yuzhuo Fu and Yangyang Kang are co-corresponding authors.

Benchmark	ML.	CMC.	MD.	FKP.
GAOKAO-Bench (Zhang et al. 2023)	✓	✗	✓	✗
MathVerse (Zhang et al. 2024)	✗	✓	✗	✓
MATH-Vision (Wang et al. 2024a)	✗	✗	✗	✓
MMMU (Yue et al. 2024)	✗	✗	✓	✓
EMMA (Hao et al. 2025)	✗	✗	✓	✓
GeoSense (Xu et al. 2025b)	✓	✗	✗	✓
PhyX (Shen et al. 2025)	✗	✗	✗	✓
VisioMath (Li et al. 2025a)	✗	✗	✗	✗
MME-SCI (Ours)	✓	✓	✓	✓

Table 1: Comparison between our MME-SCI and others. **ML.** denotes Multilingual, **CMC.** signifies Comprehensive Modality Coverage, **MD.** represents Multidisciplinary, and **FKP.** stands for Fine-grained Knowledge Points.

there is an urgent need to design a challenging benchmark to continuously drive breakthroughs in models' scientific reasoning and cross-modal understanding capabilities.

A high-quality scientific evaluation benchmark is the key premise for accurately assessing the reasoning capabilities of MLLMs in scientific domains. The design of such a benchmark should meet the following three core characteristics: **1) Multilingual adaptability.** Currently, the training corpus for MLLMs is predominantly in English, leading to performance differences when the same model is applied to non-English contexts. More importantly, multilingual scenarios are more effective in verifying whether MLLMs have truly mastered reasoning abilities, rather than relying on specific linguistic contexts. Furthermore, Cross-linguistic consistency is essential for supporting global scientific collaboration. **2) Comprehensive modality coverage.** Most existing benchmarks tend to focus on image-text hybrid scenarios and lack systematic testing of MLLMs on image-only and text-only modes. Evaluations with diverse modalities are essential to reflect the robustness and applicability of MLLMs in real-world applications. **3) Fine-grained knowledge points annotation.** Current benchmarks are still deficient in the systematic annotation of knowledge points, making it difficult for evaluation results to provide targeted feedback and thereby limiting in-depth analysis of models' potential flaws and disciplinary weaknesses.

To address the aforementioned gap, we introduce MME-SCI, a comprehensive and highly challenging multimodal evaluation benchmark. This benchmark consists of 1,019 high-quality question-answer pairs that have undergone manual selection, covering three question types: single-choice, multiple-choice, and fill-in-the-blank. The content spans four core subjects: mathematics, physics, chemistry, and biology. Additionally, it provides five language versions (e.g., Chinese, English, French, Spanish, and Japanese) to support multilingual evaluation. Furthermore, we have annotated each question with the corresponding knowledge point, covering a total of 63 fine-grained concepts, and designed three input modalities: text-only, image-only, and image-text hybrid. These designs collectively offer a comprehensive and detailed evaluation framework for systematically assessing the reasoning capabilities of MLLMs across diverse linguistic environments and knowledge systems.

We conducted comprehensive experiments of 16 open-source and 4 closed-source models on the MME-SCI benchmark. Taking the Image-only evaluation mode as an example: Qwen2.5VL-72B, the top-performing open-source model, achieved an accuracy of only 19.43%, while Doubao-Seed-1.6 (ByteDance 2025b), an advanced representative among closed-source models, reached an accuracy of merely 41.32%. These results demonstrate that MME-SCI poses significant challenges to existing MLLMs. Furthermore, leveraging the advantages of MME-SCI, such as multilingual support and fine-grained knowledge points, we are able to conduct in-depth analysis of the performance of existing models, thereby accurately revealing their shortcomings in language consistency and domain-specific applications.

The main contributions of this paper could be summarized as follows:

- We propose MME-SCI, a comprehensive multimodal scientific benchmark. This benchmark contains 1,019 manually curated samples, supports 5 languages and 3 input modalities, and covers 63 knowledge concepts in mathematics, physics, chemistry, and biology. It effectively addresses the limitations of existing benchmarks in multilingual adaptability, comprehensive modality coverage, and fine-grained knowledge annotation.
- Evaluation results based on 20 MLLMs show that MME-SCI poses significant challenges to current models and breaks the performance saturation of existing mainstream scientific evaluation benchmarks.
- Leveraging the strengths of MME-SCI in multilingual support and fine-grained annotation, we perform in-depth analyses and accurately identify the limitations of models in cross-lingual consistency, modality-specific reasoning, and discipline-specific knowledge, thereby offering targeted guidance for improving MLLMs.

Related Works

Multimodal Large Language Models

With the rapid advancement of LLMs and visual foundation models, MLLMs have demonstrated remarkable breakthroughs in multimodal tasks. The open-source community has taken the lead in promoting technological implementation through lightweight solutions: LLaVA (Liu et al. 2023) freezes the CLIP (Radford et al. 2021) for image encoding and injects visual prompts into the LLM decoder (Chiang et al. 2023) via a lightweight projection layer to achieve cross-modal alignment; the Qwen-VL series (Bai et al. 2023; Wang et al. 2024b; Bai et al. 2025) enhances spatiotemporal perception by progressively upgrading visual encoders, introducing dynamic resolution mechanisms, and multimodal rotational position encoding, while expanding training data scale to enable the evolution from single-image understanding to unified image-video processing; the InternVL series (Chen et al. 2024c,b,a; Zhu et al. 2025) has transitioned from the initially complex architecture based on BLIP-2 (Li et al. 2023) improvements to the simple 'ViT-MLP-LLM' framework, achieving performance close to closed-source models

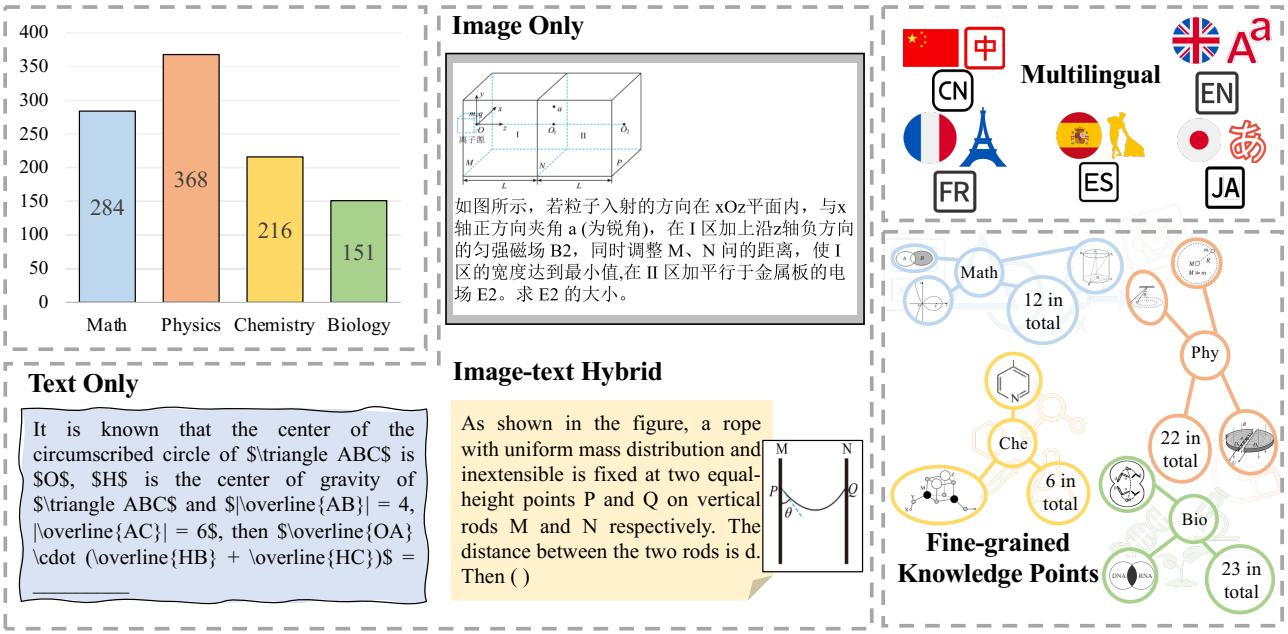

Figure 2: Overview of MME-SCI. This benchmark consists of 1,019 manually and carefully selected questions, covering four subjects, with the characteristics of multilingual support, full-modal coverage, and fine-grained knowledge points.

like GPT-4V (Yang et al. 2023) through gradual upgrades in model scaling, data optimization, and inference strategies during testing.

Science-related Benchmarks

With the continuous evolution of MLLMs capabilities, developing benchmarks that match their performance has become a critical research (Li et al. 2024b,c). Benchmarks in the multimodal field can be mainly categorized into natural domain and scientific domain: the former focuses on evaluating MLLMs' ability to understand natural images (Wang et al. 2025a; Ye et al. 2024; Guo et al. 2024), while the latter centers on reasoning tasks in scientific disciplines. Specifically, MMMU (Yue et al. 2024) is a benchmark covering 6 disciplines and 30 courses, containing 11.5K university-level questions; EMMA (Hao et al. 2025) is an enhanced multimodal reasoning benchmark that includes questions in mathematics, physics, chemistry, and coding, aiming to assess models' visual reasoning abilities; GeoSense (Xu et al. 2025b) is a bilingual benchmark that evaluates MLLMs' geometric reasoning capabilities through a five-level hierarchical framework based on geometric principles and 1,789 questions. In addition, other multimodal benchmarks (Jiang et al. 2025; Yao et al. 2025; Guo et al. 2025) have also promoted the development of MLLMs. However, with the rapid iteration of MLLMs, these benchmarks have gradually become saturated. As shown in Table 1, their core limitation lies in that existing benchmarks cannot fully cover the multi-dimensional capabilities of MLLMs. In contrast, our MME-SCI contains 1,019 manually curated question-answer pairs, simultaneously featuring characteristics such as Multilingual, Comprehensive modality coverage, Multi-disciplinary, and Fine-grained knowledge points. This not only poses more rigorous challenges to existing MLLMs but also enables more dimensional performance analysis.

MME-SCI

Overview of MME-SCI

As illustrated in Figure 2, we introduce MME-SCI, a comprehensive and challenging benchmark for scientific domain. It comprises 1,019 manually curated questions, systematically covering four subjects (mathematics, physics, chemistry, and biology) at the Chinese high school level. These four subjects contain 12, 22, 6, and 23 knowledge points, respectively, based on high school textbook systems and suggestions from subject experts. This enables MLLMs to accurately identify their weaknesses in specific knowledge points. In terms of question timeliness, 83.3% of the questions are sourced from 2025, 16.2% from 2024, and the remaining 0.5% from years prior to 2024, ensuring the novelty of the data sources.

Given that the training corpora of current MLLMs are mostly centered on English scenarios, MME-SCI is specifically equipped with 5 language versions, including Chinese, English, French, Spanish, and Japanese, to comprehensively evaluate the cross-lingual reasoning capabilities of models. In terms of modal design, MME-SCI encompasses three evaluation modes: text-only, image-only and image-text hybrid. Specifically, text-only mode are designed to assess language comprehension abilities, image-only mode emphasize the evaluation of visual semantic parsing capabilities, and image-text mode require the integration of visual information and textual logic. These tasks collectively form a multi-

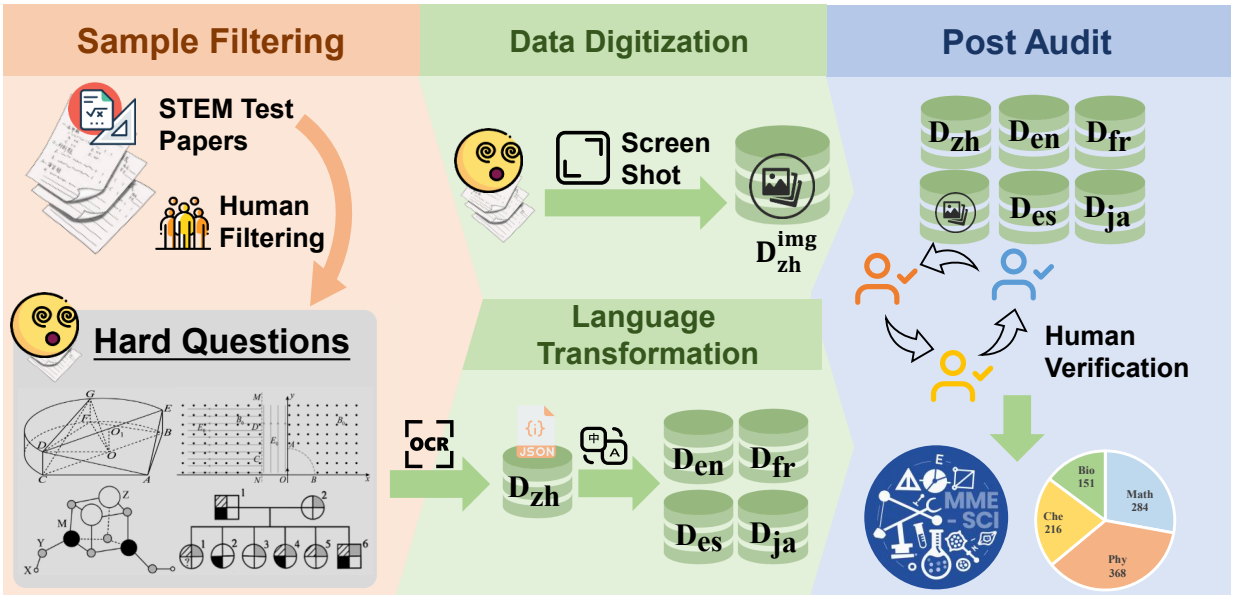

Figure 3: Construction pipeline of MME-SCI, which consists of three stages. Overall, a total of approximately 300 person-days were spent on problem selection, digitization, language conversion, and post-verification.

dimensional evaluation system that covers full-modal reasoning abilities.

To reduce the complexity of evaluation, all questions are designed as single/multiple-choice or fill-in-the-blank items, paired with concise and verifiable answers. This design not only supports automated scoring but also facilitates manual verification. By incorporating multilingual scenario design, full-modal evaluation, and fine-grained knowledge points coverage, MME-SCI aims to serve as a core benchmark for systematically enhancing MLLMs' complex reasoning capabilities in scientific domains.

Data Curation Process

As illustrated in Figure 3, the construction pipeline of MME-SCI consists of four core steps, namely sample filtering, data digitization, language transformation, and post-auditing, which are specified as follows:

Sample Filtering We recruited 3 evaluation volunteers (2 senior undergraduates and 1 graduate student), all of whom ranked within the top 0.1% in terms of average scores in China's College Entrance Examination (Gaokao). The volunteers were asked to solve questions from mock exam papers of high school science subjects[1], and to filter out the questions they answered incorrectly or found confusing.

Data Digitization and Language Transformation After manually selecting high-difficulty questions, we recruited five annotators to process the data. Specifically, we used GPT-4o and OCR tools to extract the questions and answers, converting them into JSON format to form the $\mathbf{D_{zh}}$. $\mathbf{D_{zh}}$ contains 805 questions that require image-based understanding (such as problems in analytical geometry and circuit analysis) and 214 text-only questions. Subsequently, we took screenshots of all the questions in $\mathbf{D_{zh}}$, creating the $\mathbf{D_{zh}^{img}}$ for use in the image-only evaluation mode. Examples of this can be found in Figure 2, with more samples provided in the Appendix E. We then translated the Chinese content in $\mathbf{D_{zh}}$ into English, French, Spanish, and Japanese, resulting in $\mathbf{D_{en}}$, $\mathbf{D_{fr}}$, $\mathbf{D_{es}}$, and $\mathbf{D_{ja}}$, respectively.

Post-Audit In the final stage, we recruited three reviewers to perform cross-validation on the OCR results in $\mathbf{D_{zh}}$, the integrity of the screenshots in $\mathbf{D_{zh}^{img}}$, and the language conversion results. If one reviewer identifies an error, the other two will conduct a secondary verification. In the secondary verification process, if any reviewer identifies an error, a revision will be made to ensure the quality of the data.

Differences from the Existing Benchmark

Compared with existing benchmarks, MME-SCI exhibits significant advantages in multilingual adaptability, full-modal coverage, and fine-grained annotation of knowledge points. Existing benchmarks are mostly limited to a single language environment (e.g., MMMU (Yue et al. 2024) only supports English, while GAOKAO-Bench (Zhang et al. 2023) is primarily in Chinese), making it difficult to evaluate models' reasoning capabilities in cross-lingual scientific scenarios. In contrast, our MME-SCI comprehensively covers five languages, enabling systematic assessment of models' understanding and reasoning of scientific concepts across different linguistic contexts.

[1]Mock high school exam papers were chosen instead of Gaokao papers because Gaokao papers have a wide dissemination range and low accessibility, making them more likely to be included in the training data of MLLMs.

Model	$\mathbf{D_{zh}}$					$\mathbf{D_{zh}^{img}}$					$\mathbf{D_{en}}$	$\mathbf{D_{fr}}$	$\mathbf{D_{es}}$	$\mathbf{D_{ja}}$
	math	**phy**	**chem**	**bio**	**AVG.**	**math**	**phy**	**chem**	**bio**	**AVG.**	**AVG.**	**AVG.**	**AVG.**	**AVG.**
***Open-source LVLMs* < 10B (Small group)**														
InternVL3-2B	9.86	6.52	18.06	22.52	12.27	9.15	5.71	10.65	14.57	9.03	13.05	10.89	11.78	15.70
Qwen2.5VL-3B	10.56	7.88	18.98	25.17	13.54	8.10	7.61	14.35	21.85	11.29	16.68	14.03	14.62	11.58
InternVL3-8B	10.56	9.78	21.76	25.83	14.92	4.93	7.07	15.74	17.88	9.91	16.78	15.41	16.39	16.68
Qwen2.5VL-7B	8.80	11.41	22.69	25.83	15.21	10.21	13.32	23.61	27.15	16.68	15.80	15.80	15.21	15.70
Qwen2.5-Omni-7B	11.62	6.52	20.83	27.81	14.13	4.23	8.42	20.37	23.18	11.97	16.00	16.19	16.88	14.43
Ovis2-8B	15.14	8.15	18.06	27.81	15.11	10.92	6.52	12.96	18.54	10.89	16.00	14.72	15.11	10.89
Llava-OneVision-7B	8.10	7.88	15.74	20.53	11.48	2.46	12.23	6.94	17.22	9.13	13.74	13.64	13.54	10.30
Kimi-VL-A3B-Instruct	13.73	7.61	16.67	24.50	13.74	8.80	4.89	13.43	22.52	10.40	12.86	9.32	11.19	11.09
Kimi-VL-A3B-Thinking	23.94	10.33	22.22	35.76	20.41	22.97	8.97	18.98	29.14	17.98	21.10	22.18	21.59	13.25
Phi-4-multimodal-instruct	7.75	7.61	8.33	15.23	8.93	3.17	4.35	3.24	9.27	4.51	11.68	10.11	11.87	11.97
Avg. Performance	12.01	8.37	18.33	25.10	13.97	8.49	7.91	14.03	20.13	11.18	15.37	14.23	14.82	13.16
10B < *Open-source LVLMs* < 40B (Middle group)														
Ovis2-16B	15.85	7.88	20.83	28.48	15.90	9.86	6.52	18.06	19.87	11.87	17.08	15.01	16.09	15.11
Ovis2-34B	11.97	10.33	21.76	30.46	16.19	10.21	7.07	16.20	20.53	11.87	17.57	15.70	17.08	14.33
Skywork-R1V-38B	22.18	10.87	25.93	26.49	19.53	15.14	8.15	18.52	25.17	14.82	22.18	23.55	20.90	16.29
Avg. Performance	16.67	9.69	22.84	28.48	17.21	11.74	7.25	17.59	21.86	12.85	18.94	18.09	18.02	15.24
***Open-source LVLMs* > 40B (Large group)**														
Llava-OneVision-72B	13.93	10.35	20.83	32.45	16.86	4.58	7.69	9.30	11.41	7.71	18.84	14.92	16.98	16.88
Qwen2.5VL-72B	19.01	14.40	30.56	34.44	22.08	14.79	11.96	29.17	32.45	19.43	19.17	18.65	19.53	17.47
InternVL3-78B	16.25	11.72	26.39	27.15	18.39	15.30	12.98	21.30	25.83	17.33	20.12	20.12	19.45	19.43
Avg. Performance	16.40	12.16	25.93	31.35	19.11	11.56	10.88	19.92	23.23	14.82	19.38	17.90	18.65	17.93
Closed-source LVLMs														
Claude-4-Sonnet	21.13	17.12	31.48	33.77	23.75	19.43	13.59	23.61	21.19	18.47	22.96	22.28	24.14	22.37
Doubao-1.5-tv-pro	28.87	28.80	40.74	41.72	33.27	38.73	31.52	44.44	46.36	38.47	34.05	29.93	32.09	34.54
Doubao-Seed-1.6	44.37	31.52	42.13	41.72	38.86	49.30	33.15	42.13	45.03	41.32	40.14	37.88	38.76	36.41
o4-mini-20250416	50.35	26.36	38.89	35.10	37.00	52.11	24.73	36.57	29.80	35.62	33.17	29.54	31.11	29.64
Avg. Performance	36.18	25.95	38.31	38.08	33.22	39.89	25.75	36.69	35.59	33.47	32.58	29.91	31.53	30.74

Table 2: Results on MME-SCI. We present detailed comparative results for the $\mathbf{D_{zh}}$ and $\mathbf{D_{zh}^{img}}$ scenarios, including accuracy for four subjects. For the remaining four languages, we report the average accuracy (AVG.).

In terms of modal evaluation dimensions, existing benchmarks have limited coverage. For instance, VisioMath (Li et al. 2025a) focuses on scenarios with image options, while MATH-Vision (Wang et al. 2024a) centers on image-text hybrid scenarios. However, in practical applications, MLLMs not only process image-text hybrid inputs but also frequently need to handle pure image or pure text inputs. In contrast, our MME-SCI can simultaneously support independent testing for three types of modalities: text-only, image-only, and image-text hybrid. Furthermore, existing benchmarks generally adopt coarse-grained knowledge point classification, whereas MME-SCI annotates each question with fine-grained knowledge points (e.g., "Trigonometric Functions and Solving Triangle" in mathematics, "Regulation of Plant Life Activities" in biology). This design enables it to more accurately identify models' weaknesses in specific knowledge points compared to existing benchmarks, providing fine-grained diagnostic references for MLLMs.

Experiments

Evaluation details

We conducted extensive experiments on 16 open-source models and 4 closed-source models. Specifically, the open-source models, which range from 2B to 78B, including Llava-OneVision (7B/72B) (Li et al. 2024a), Qwen2.5VL (3B/7B/72B) (Bai et al. 2025), Qwen2.5-Omni (7B) (Xu et al. 2025a), Ovis2 (8B/16B/34B) (Lu et al. 2024), InternVL3 (2B/8B/78B) (Zhu et al. 2025), Kimi-VL-A3B-Instruct/Thinking (Team et al. 2025), Phi-4-multimodal-instruct (4.2B) (Abouelenin et al. 2025), Skywork-R1V-38B (Peng et al. 2025). For the closed-source models, we employed Claude-4-Sonnet (Anthropic 2025), Doubao-1.5-thinking-vision-pro (ByteDance 2025a), Doubao-Seed-1.6 (ByteDance 2025b), and o4-mini-20250416 (OpenAI 2025)

We adopt the 'LLM-as-a-Judge' paradigm (Zheng et al. 2023; Gu et al. 2024) and introduce different evaluation templates for various languages to assess the models. Unless specified otherwise, we configured the maximum number of new tokens to 8,192, and the temperature was set to 0. All experiments are conducted on 8×H20 GPUs. More details can be found in the Appendix.

Main Results

The experimental results on MME-SCI are shown in Table 2, from which the following conclusions can be drawn:

1) A significant gap between open-source and closed-source models. Compared to advanced closed-source mod-

els, the most powerful open-source models (Large group) demonstrates an average accuracy reduction of 13.94% across six scenarios. Specifically, in the $\mathbf{D}_{\mathbf{zh}}^{\mathbf{img}}$ scenario, which demands higher visual capabilities of the MLLM, the closed-source models achieved a gain of 125.84% over the Large group. These results highlight the considerable gap between existing open-source and closed-source models in scientific reasoning tasks, further confirming the challenges posed by MME-SCI to most existing MLLMs.

2) Models with reasoning capabilities have demonstrated significant competitive advantages. In the context of scientific disciplines, models typically require extensive and complex reasoning, which places higher demands on the model's reasoning capabilities. Compared to models that have not undergone specialized reasoning training, models with reasoning capabilities clearly perform better in such scenarios. For instance, the Thinking version of Kimi-VL-A3B achieves a 48.54% performance gain over its Instruct version on the $\mathbf{D}_{\mathbf{zh}}$, and even achieves an astonishing 72.88% gain in Image-only scenarios. Additionally, for MLLMs of similar parameter sizes, such as Ovis2-34B and Skywork-R1V-38B, the latter outperforms the former by 4.09% in average accuracy across six evaluation scenarios.

3) There are significant differences across subjects. Overall, all models perform noticeably better in the chemistry and biology domains than in mathematics and physics, as the latter require a higher level of reasoning capability. Furthermore, most models perform better in mathematics than in physics, which can be attributed to the fact that physics not only requires complex reasoning but also demands that the model understands the physical laws of the real world (Shen et al. 2025). Additionally, there are evident cases of specialization in individual models. For instance, Qwen2.5VL-72B performs only 0.66% worse than o4-mini in biology, but is 31.34% worse in mathematics, leading to an overall poorer performance. These findings provide valuable insights for improving future model training, particularly by adjusting the training data composition to enhance performance across multiple aspects.

4) Most models exhibit performance degradation when processing Image-only mode, with open-source models being particularly affected. Compared to the $\mathbf{D}_{\mathbf{zh}}$, the screenshot-based inputs in the $\mathbf{D}_{\mathbf{zh}}^{\mathbf{img}}$ pose greater challenges to open-source models—specifically, their accuracy in the Small group, Middle group, and Large group drop by 2.79%, 4.36%, and 4.29%, respectively. In contrast, closed-source models, particularly Doubao-1.5-think-vision-pro, achieve a 5.20% accuracy improvement on the $\mathbf{D}_{\mathbf{zh}}^{\mathbf{img}}$. This phenomenon indicates that closed-source models, perhaps benefiting from their OCR capabilities, are better able to adapt to image-only tasks, whereas open-source models still face significant challenges under such an evaluation mode. Thus, future work should focus on enhancing the ability of MLLMs to understand screenshot-based inputs, thereby improving their robustness in real-world application scenarios.

5) The visual reasoning ability of the Any2any model degrades. We specifically evaluated the Any2any model corresponding to Qwen2.5VL-7B, namely Qwen2.5-Omni-7B.

Model \ CoT	Doubao -Seed-1.6	Qwen2.5 VL-7B	Qwen2.5 VL-72B
Qwen2.5VL-7B	15.51(+0.30)	13.84(-1.37)	14.72(-0.49)
Qwen2.5VL-72B	22.22(+0.14)	20.12(-1.96)	19.92(-2.16)
Qwen2.5VL-7B	15.90(+0.69)	14.23(-0.98)	14.52(-0.69)
Qwen2.5VL-72B	22.50(+0.42)	19.53(-2.55)	19.53(-2.55)

Table 3: The performance on $\mathbf{D}_{\mathbf{zh}}$ using *concise* (rows 1 and 2) and *detailed* (rows 3 and 4) pre-knowledge descriptions as context. The baseline of Qwen2.5VL-7B/72B are 15.21% and 22.08%, respectively.

The results indicate that, compared to the original vision-language model, its extended Any2any model with additional modalities shows a degradation in visual reasoning ability, with a 4.71% drop on the $\mathbf{D}_{\mathbf{zh}}^{\mathbf{img}}$. Therefore, in future model expansions to include more modalities, it is crucial to minimize the degradation of the original capabilities in order to truly achieve an "omnidirectional" model.

Further Analysis

In this section, leveraging the characteristics of our MME-SCI, we have conducted in-depth analytical experiments on contextual information, language consistency, knowledge point differences, and error causes. Additional analytical experiments can be found in the Appendix C.

Impacts of in-context learning setting

We use knowledge point descriptions as context to investigate the impact of prior knowledge on model performance. Specifically, we employ Doubao-Seed-1.6 and Qwen2.5-VL-7/72B as prior knowledge generators to generate both concise and detailed descriptions of knowledge points for each sample in the $\mathbf{D}_{\mathbf{zh}}$, with manual verification to ensure that these descriptions do not contain answer information. During the testing phase, the generated prior knowledge points are input into MLLMs as context along with the questions. Results in Tables 3 demonstrate that knowledge generated by stronger models can improve the performance of weaker models to a certain extent, while knowledge generated by weaker models exert a negative effect on stronger models, leading to performance degradation.

Impacts of various languages

As shown in Figure 4, we present the consistency results of responses from different models to the same question across five language scenarios. It can be observed that as model capabilities improve, the number of samples with consistent correct responses shows an increasing trend. This indicates that the enhancement of a model's fundamental capabilities enables it to perform more consistent reasoning in cross-lingual scenarios. However, from an overall data perspective, the MME-SCI dataset contains a total of 1,019 questions. Even the top-performing Doubao-Seed-1.6 only achieves a 13.84% of linguistically consistent correct responses, while the worst-performing Phi-4 achieves

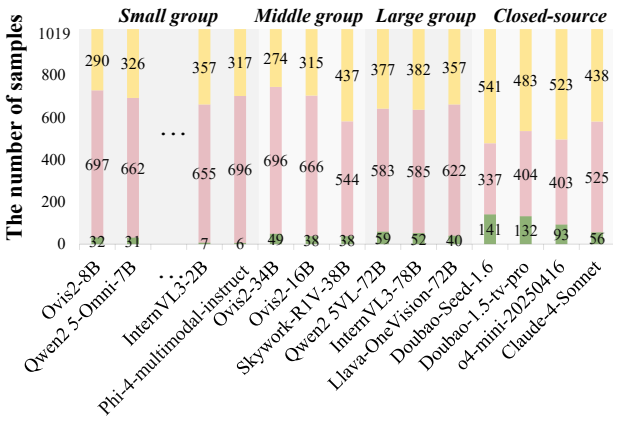

Figure 4: Consistency of responses across five languages. The green section at the bottom shows samples answered correctly consistently across all five languages; the middle red section shows those answered incorrectly consistently across all five. The yellow section at the top is the rest.

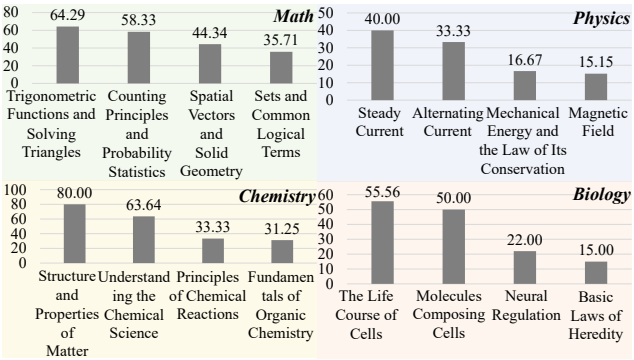

Figure 5: Performance of o4-mini on $\mathbf{D_{zh}}$ with fine-grained knowledge points. We present the top-2 and bottom-2 knowledge points in terms of accuracy across 4 subjects.

merely a 0.59% in this regard. This suggests that future MLLMs should focus more on learning truly reasoning abilities rather than merely adapting to specific languages.

Impacts of fine-grained knowledge points

As illustrated in Figure 5, we report the accuracy of o4-mini across individual Knowledge Points (KPs). The results reveal a significant KP-specific bias of o4-mini in the chemistry subject: while the model attains an accuracy of 80.00% on its dominant KP, its performance drops to only 31.25% on the "Fundamentals of Organic Chemistry". Such within-subject performance discrepancies directly contribute to the model's underperformance in the overall chemistry subject. Furthermore, the individual KP results in physics indicate that the model performs at a consistently low level in this subject. These observations can provide fine-grained and targeted guidance for subsequent model optimization, such as the introduction of additional physical reasoning training.

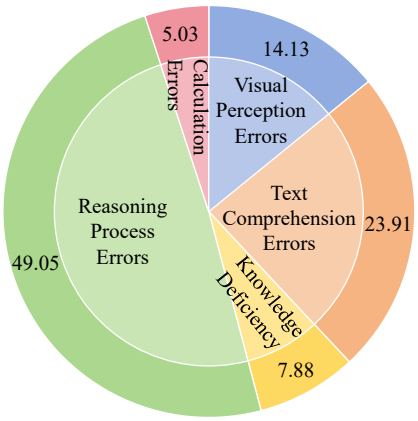

Figure 6: Error distribution on $\mathbf{D_{zh}}$ of Doubao-Seed-1.6. **Visual perception errors**: response biases caused by incorrect recognition/extraction of image information in questions (e.g., text, symbols, and shapes in the image). **Text comprehension errors**: errors due to misinterpretation of information, logic, implicit conditions, or explicit instructions in question text. **Knowledge deficiency**: errors resulting from lack of core domain knowledge (e.g., concepts, theorems) required for the question. **Calculation errors**: errors in operations, formula substitution, etc., despite correct understanding of question logic and knowledge. **Reasoning process errors**: errors from flawed logical deduction or causal judgment, despite correct information acquisition and knowledge possession. Furthermore, the erroneous cases can be found in the Appendix D.

Error Analysis

To thoroughly analyze the defects of MLLMs, we examined the incorrect samples of the Doubao-Seed-1.6 on the $\mathbf{D_{zh}}$ and categorized the error causes, with the results shown in Figure 6. Statistics indicate that reasoning process errors are the most frequent, accounting for 49.05%, which suggests that the reasoning process is the core link where MLLMs are most prone to mistakes (Ma et al. 2023; Wang et al. 2025b; Song et al. 2025). In contrast, calculation errors only account for 5.03%, indicating that Doubao-Seed-1.6 has already acquired reliable computing capabilities.

Conclusions

In this paper, we construct MME-SCI, a comprehensive and challenging benchmark characterized by multilingual support, comprehensive modality coverage, multidisciplinary integration, and multi-knowledge point inclusion. We evaluate 20 popular MLLMs on this benchmark, and the results demonstrate that MME-SCI is not only highly challenging but also capable of effectively distinguishing the performance differences among various models. Furthermore, leveraging its inherent characteristics, MME-SCI can help researchers precisely identify the shortcomings of models in language consistency, modal adaptability, reasoning ability, and command of knowledge points. We hope that our MME-SCI will provide directional guidance and research inspiration for the development of MLLMs in the new era.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61977045).

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Claude 4: Sonnet. Accessed: 2025-07-20.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

ByteDance. 2025a. doubao-1.5-thinking-vision-pro. Accessed: 2025-07-20.

ByteDance. 2025b. doubao-seed-1.6. Accessed: 2025-07-20.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Guo, M.-H.; Xu, J.; Zhang, Y.; Song, J.; Peng, H.; Deng, Y.-X.; Dong, X.; Nakayama, K.; Geng, Z.; Wang, C.; et al. 2025. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. *arXiv preprint arXiv:2505.02018*.

Guo, X.; Zhang, R.; Duan, Y.; He, Y.; Zhang, C.; Liu, S.; and Chen, L. 2024. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv e-prints*, arXiv–2411.

Hao, Y.; Gu, J.; Wang, H. W.; Li, L.; Yang, Z.; Wang, L.; and Cheng, Y. 2025. Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark. *arXiv preprint arXiv:2501.05444*.

Jiang, D.; Zhang, R.; Guo, Z.; Li, Y.; Qi, Y.; Chen, X.; Wang, L.; Jin, J.; Guo, C.; Yan, S.; et al. 2025. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, 235–251. Springer.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, C.; Zhang, T.; Wang, M.; and Huang, H. 2025a. VisioMath: Benchmarking Figure-based Mathematical Reasoning in LMMs. *arXiv preprint arXiv:2506.06727*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Lu, W.; Fei, H.; Luo, M.; Dai, M.; Xia, M.; Jin, Y.; Gan, Z.; Qi, D.; Fu, C.; et al. 2024b. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Li, L.; Chen, G.; Shi, H.; Xiao, J.; and Chen, L. 2024c. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*.

Li, Y.; Liu, Z.; Li, Z.; Zhang, X.; Xu, Z.; Chen, X.; Shi, H.; Jiang, S.; Wang, X.; Wang, J.; et al. 2025b. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ma, Q.; Zhou, H.; Liu, T.; Yuan, J.; Liu, P.; You, Y.; and Yang, H. 2023. Let's reward step by step: Step-Level reward model as the Navigators for Reasoning. *arXiv preprint arXiv:2310.10080*.

OpenAI. 2025. o3-o4-mini-system-card. Accessed: 2025-07-20.

Peng, Y.; Wang, P.; Wang, X.; Wei, Y.; Pei, J.; Qiu, W.; Jian, A.; Hao, Y.; Pan, J.; Xie, T.; et al. 2025. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*.

Qu, X.; Li, Y.; Su, Z.; Sun, W.; Yan, J.; Liu, D.; Cui, G.; Liu, D.; Liang, S.; He, J.; et al. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shen, H.; Wu, T.; Han, Q.; Hsieh, Y.; Wang, J.; Zhang, Y.; Cheng, Y.; Hao, Z.; Ni, Y.; Wang, X.; et al. 2025. PhyX: Does Your Model Have the" Wits" for Physical Reasoning? *arXiv preprint arXiv:2505.15929*.

Song, M.; Su, Z.; Qu, X.; Zhou, J.; and Cheng, Y. 2025. PRMBench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Wang, F.; Wang, H.; Guo, Z.; Wang, D.; Wang, Y.; Chen, M.; Ma, Q.; Lan, L.; Yang, W.; Zhang, J.; et al. 2025a. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14325–14336.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; Gao, Z.; Chen, L.; Chen, Z.; Zhu, J.; Zhao, X.; Liu, Y.; Cao, Y.; Ye, S.; Zhu, X.; et al. 2025b. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025c. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xu, L.; Zhao, Y.; Wang, J.; Wang, Y.; Pi, B.; Wang, C.; Zhang, M.; Gu, J.; Li, X.; Zhu, X.; et al. 2025b. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Yao, H.; Huang, J.; Qiu, Y.; Chen, M. K.; Liu, W.; Zhang, W.; Zeng, W.; Zhang, X.; Zhang, J.; Song, Y.; et al. 2025. MMReason: An Open-Ended Multi-Modal Multi-Step Reasoning Benchmark for MLLMs Toward AGI. *arXiv preprint arXiv:2506.23563*.

Ye, J.; Wang, G.; Li, Y.; Deng, Z.; Li, W.; Li, T.; Duan, H.; Huang, Z.; Su, Y.; Wang, B.; et al. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 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*, 9556–9567.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Qiao, Y.; et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 169–186. Springer.

Zhang, X.; Li, C.; Zong, Y.; Ying, Z.; He, L.; and Qiu, X. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.