MDK12-Bench: A Multi-Discipline Benchmark for Evaluating Reasoning in Multimodal Large Language Models

Pengfei Zhou[1,2*], Xiaopeng Peng[4*], Fanrui Zhang[5,2*], Zhaopan Xu[6,3], Jiaxin Ai[7,2], Yansheng Qiu[7,3], Wangbo Zhao[1], Jianjun Song[8], Chuanhao Li[3], Weidong Tang[9], Zhen Li[3], Haoquan Zhang[3], Zizhen Li[2], Xiaofeng Mao[3], Yukang Feng[2], Jianwen Sun[2], Kai Wang[1], Xiaojun Chang[5], Wenqi Shao[3], Yang You[1†], Kaipeng Zhang[2,3†]

[1]National University of Singapore
[2]Shanghai Innovation Institute
[3]Shanghai Artificial Intelligence Laboratory
[4]Rochester Institute of Technology
[5]University of Science and Technology of China
[6]Harbin Institute of Technology
[7]Wuhan University
[8]Renmin University of China
[9]Xi'an University of Electronic Science and Technology

Abstract

Multimodal large language models (MLLMs), which integrate language and visual cues for problem-solving, are crucial for advancing artificial general intelligence (AGI). However, current benchmarks for measuring the intelligence of MLLMs suffer from limited scale, narrow coverage, and unstructured knowledge, offering only static and undifferentiated evaluations. To bridge this gap, we introduce MDK12-Bench, a large-scale multidisciplinary benchmark built from real-world K–12 exams spanning six disciplines with 141K instances and 6,225 knowledge points organized in a six-layer taxonomy. Covering five question formats with difficulty and year annotations, it enables comprehensive evaluation to capture the extent to which MLLMs perform over four dimensions: 1) difficulty levels, 2) temporal (cross-year) shifts, 3) contextual shifts, and 4) knowledge-driven reasoning. We propose a novel dynamic evaluation framework that introduces unfamiliar visual, textual, and question form shifts to challenge model generalization while improving benchmark objectivity and longevity by mitigating data contamination. We further evaluate knowledge-point reference-augmented generation (KP-RAG) to examine the role of knowledge in reasoning. Key findings reveal strengths and limitations in current MLLMs in multiple aspects and provide guidance for enhancing model reasoning, robustness, and AI-assisted education.

Dataset and Code — https://github.com/LanceZPF/MDK12

Introduction

Problem-solving is a core aspect of intelligence (Sternberg 1982; Lohman and Lakin 2011), requiring reasoning, abstraction, knowledge integration, and adaptability to novel and increasingly difficult challenges. Achieving Artificial General Intelligence (AGI) requires models to go beyond excelling at isolated tasks, where they must demonstrate the ability to generalize knowledge, integrate information across modalities, and reason effectively under diverse contexts. Recent advances in Multimodal Large Language Models (MLLMs) (Radford et al. 2021; Li et al. 2022; Liu et al. 2024a; Alayrac et al. 2022) have advanced these capabilities, driving interest in rigorous benchmarks. Measuring such abilities demands a multidimensional and fine-grained evaluation to understand the extent to which MLLMs' intelligence has developed and guide improvements in model adaptability and training.

However, most existing benchmarks generally remain confined to single text modality (Hendrycks et al. 2021b,a; Arora, Singh et al. 2023) or narrow domains, such as mathematics and physics (Zhang et al. 2024; He et al. 2024) and medicine (Sun et al. 2024b). While efforts have been made in developing multimodal multidisciplinary benchmarks (Yue et al. 2024; Li et al. 2025; Hao et al. 2025), they are limited in data size, diversity, and granularity. As shown in Table 1, the most recent multidisciplinary benchmarks consist of at most 21.1K instances and limited question types. Additionally, the lack of fine-grained annotations, such as difficulty level and knowledge, hinders systematic evaluation of model robustness, generalization to distributional shifts, and knowledge utilization in problem solving. Moreover, static evaluations adopted by current benchmarks are susceptible to data contamination, where benchmark data may enter the training corpora for new-generation MLLMs and become obsolete.

To address these limitations, we introduce MDK12-Bench, a large-scale multidisciplinary benchmark curated from real-world K–12 exams to evaluate MLLM problem-solving across diverse domains. It spans six subjects: Mathematics, Physics, Chemistry, Biology, Geography, and Information Science. Following multiple rigorous filtering and processing

*These authors contributed equally.
†Corresponding authors

Benchmarks	#Instances	#Images	Question Form	Answer Explanation	Knowledge Taxonomy	Diff Eval	Dyn Eval	RAG Eval	Temp Eval
ScienceQA (2022)	21.2K	10.3K	MC	✓	✗	✗	✗	✗	✗
OlympiadBench (2024)	8.4K	4.8K	Open	✗	✗	✗	✗	✗	✗
MMMU (2024)	11.5K	12.3K	MC, Open	✗	✗	✗	✗	✗	✗
EMMA (2025)	2.7K	3K	MC, Open	✗	✗	✗	✗	✗	✗
K12Vista (2025)	33K	NA	MC, Fill, Open	✗	✗	✗	✗	✗	✗
MDK12-Bench	**141.3K**	**105.2K**	**SC, MC, Fill, T/F, Open**	✓	✓	✓	✓	✓	✓

Table 1: Compared with existing multimodal education-related benchmarks, MDK12-Bench includes more comprehensive question coverage, multi-layer knowledge taxonomy, and detailed explanations. It also features cross-difficulty evaluation (Diff Eval), dynamic evaluation (Dyn), knowledge-point reference-augmented generation evaluation (RAG), and cross-year temporal evaluation (Temp). SC: single-choice, MC: multiple-choice; Open: Open-ended; Fill: fill-in-the-blank; T/F: true or false.

stages, the benchmark comprises 141K instances mapped to 6,225 human-annotated knowledge points organized in a six-layer taxonomy, making it the largest benchmark of its kind. MDK12-Bench includes five question types, instance-level difficulty annotations, and detailed explanations, enabling fine-grained analysis of model reasoning, generalization across task difficulty and time, and the effects of knowledge-augmented generation.

Furthermore, we propose a dynamic evaluation framework that introduces MLLMs to unseen visual and textual shifts, providing a rigorous test of model generalization to contextual changes. This approach also promotes more objective evaluation and enhances the long-term validity of the benchmark. Additionally, we introduce knowledge-point reference-augmented generation (KP-RAG), which examines the impact of knowledge on model reasoning and problem solving.

We conducted extensive experiments on state-of-the-art MLLMs, including proprietary and open-source models, evaluating both chat-oriented and reasoning-focused variants. Key findings include: **1) Model size**: Larger models outperform smaller ones across disciplines and difficulty levels, improving perception via richer multimodal representations but offering little gain in reasoning accuracy or answer completeness due to the lack of explicit reasoning-oriented training. **2) Reasoning advantage**: Reasoning-optimized models achieve higher overall and reasoning accuracy than chat models but show limited benefits in visual perception. **3) Discipline difficulty**: Models struggle more in Math and Physics (7.6% below average) than in other disciplines (3.7% above average). **4) Harder and newer exams**: Accuracy drops by 8.3% on harder and 12.6% on newer exams. **5) Dynamic perturbations**: Dynamic evaluation reduces average performance by 13.7%, which may stem from overfitting to static pretraining data. **6) Limited KP-RAG gains**: KP-RAG improves accuracy by 7% on easy exams but only 2% on hard ones, indicating reasoning rather than factual knowledge limits performance on harder tasks. Our main contributions include:

- **Large-Scale, Diverse Benchmark.** We present MDK12-Bench, a multidisciplinary benchmark for evaluating MLLM problem-solving on real-world K-12 exams. It comprises 141K instances linked to 6,225 knowledge points, structured in a six-layer taxonomy, with diverse question formats, difficulty labels, and explanations.
- **Dynamic and Knowledge-Referenced Methods.** We propose a dynamic evaluation method to challenge generalization under contextual shifts. We also introduce a knowledge-point reference-augmented generation (KP-RAG) pipeline to enhance answer generation and examine the role of knowledge in problem solving and reasoning.
- **Comprehensive Multi-Dimensional Evaluation.** We evaluate state-of-the-art MLLMs across difficulty, temporal, contextual, and knowledge dimensions, showing notable accuracy drops on harder and newer exams, as well as dynamic contextual changes.
- **Extensive Leaderboard.** We present detailed rankings and analyses of both proprietary and open-source MLLMs. Findings on results highlight the value of the MDK-benchmark in advancing our understanding of both the capabilities and critical limitations of current MLLMs.

Related Works

Benchmarking MLLMs. Evaluating the intelligence of MLLM has been challenging (Zhou et al. 2025a). The early benchmarks used text-only exams, for example, GSM-8K (Cobbe et al. 2021) and MATH (Hendrycks et al. 2021b) for math reasoning, as well as MMLU (Hendrycks et al. 2021a), GPQA (Rein et al. 2025) and AGIEval (Zhong et al. 2024) for multitask accuracy. Multimodal benchmarks later emerged, such as MathVerse (Zhang et al. 2024) for visual math, SciEval (Sun et al. 2024a) for physics, MM-Bench (Liu et al. 2024b) for basic multimodal understanding, and EXAMS-V (Das et al. 2024) for multilingual testing. More recent benchmarks (Wang et al. 2024b; Lu et al. 2022; Yue et al. 2024; Hao et al. 2025; He et al. 2024; Li et al. 2025) starts to span multiple subjects. However, these efforts remain limited in scale, diversity of questions, fine-grained annotations, and multidimensional evaluation. Few addresses differentiated difficulty (Ding et al. 2024) or provide limited knowledge taxonomy (Wang et al. 2024b; Huang et al. 2024). Our MDK12-Bench addresses these gaps with a comprehensive large-scale dataset organized in a six-level knowledge taxonomy. It provides instance-level grade and difficulty labels, five question formats, temporal annotations, detailed knowledge points, and answer explanations.

Dynamic Evaluation. Building a strong benchmark requires not only rich and well-curated data, but also carefully de-

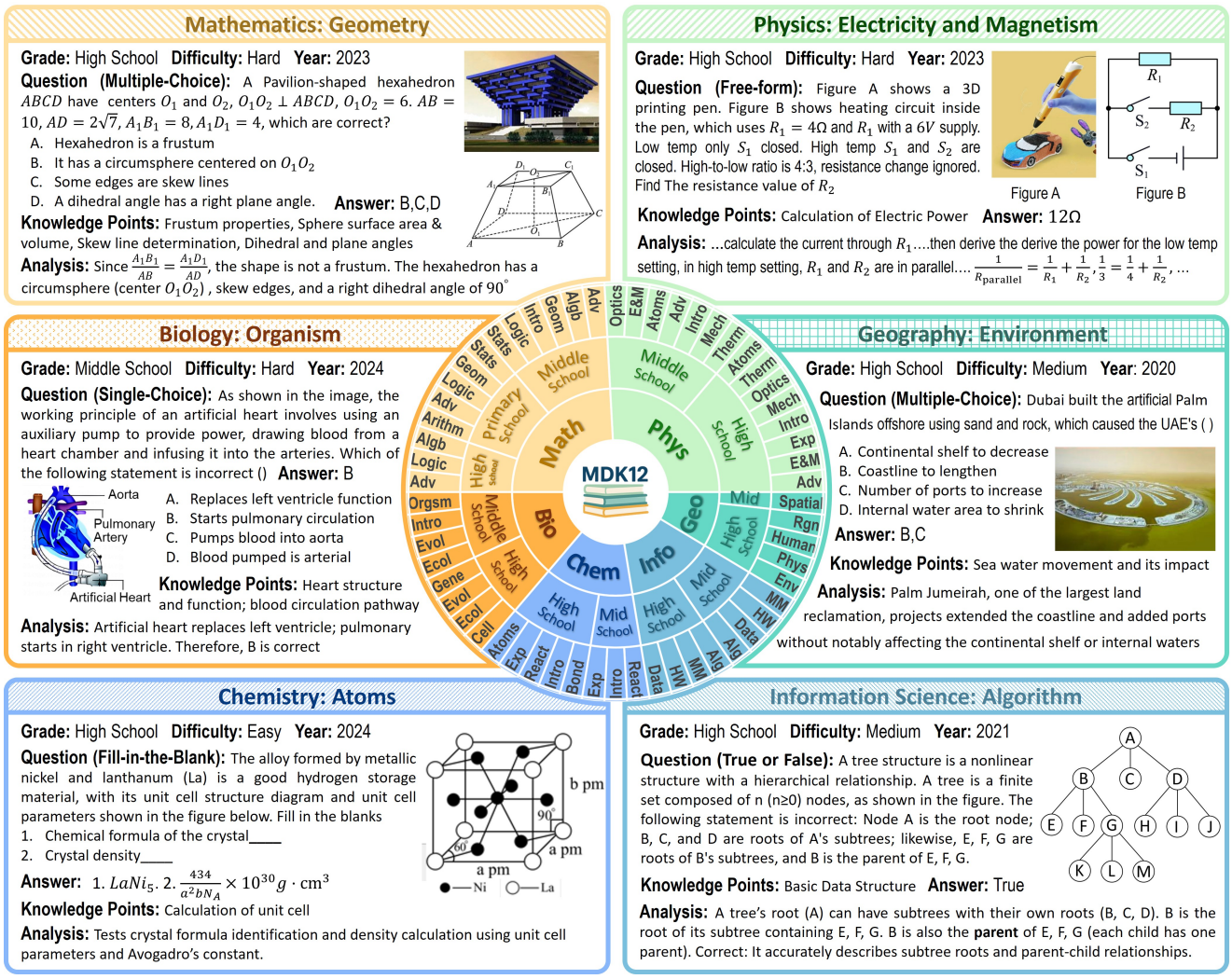

Figure 1: MDK12-Bench comprises 141K instances, spans 6 disciplines in K-12 education, and has a six-layer taxonomy hierarchy: subject, grade, subfield, curriculum, topics, and knowledge points, where the three rings showcase the first three layers. Examples illustrate the representative exam questions and fine-grained annotations.

signed rigorous evaluation methods. As pretraining corpora expand, benchmarks face a growing risk of data contamination (Xu et al. 2024; Chen et al. 2024a), where test content overlaps with training data, causing leaderboards to overstate true model capabilities. Most existing benchmarks rely on static evaluations, making them particularly vulnerable to such contamination and raising doubts about their objectivity and long-term validity. To mitigate contamination, adaptive evaluation methods have been explored. Meta-probing agents (Zhu et al. 2024) dynamically adjust the test content and difficulty during fine-tuning or domain adaptation, while other approaches alter visual and textual contexts to assess contamination effects (Yang et al. 2024). Beyond contamination, out-of-distribution visual changes may also affect the generalization and reasoning capabilities of MLLMs (Hu et al. 2025). Building on these insights, we propose a dynamic evaluation framework for MDK12-Bench that addresses both challenges. Unlike prior methods focused solely on contamination, our framework deliberately introduces novel visual, textual, and question-type variations during the test time, evaluating the generalization of models to unfamiliar conditions while ensuring sustained benchmark integrity over time.

MKD12-Benchmark

MKD12-Bench was curated with the participation of more than 20 researchers and several K-12 educators.

Data Collection. We follow 4 key principles: (1) covering multidisciplinary, real-world exams; (2) incorporating diverse visual contexts and question formats for comprehensive evaluation; (3) ensuring varied difficulty levels and broad regional coverage to reveal model limitations and reduce bias; and (4) supporting robust evaluations. Guided by these, we gathered 5.8M multimodal exam instances from open-access K–12 repositories spanning a wide range of grades, curriculum,

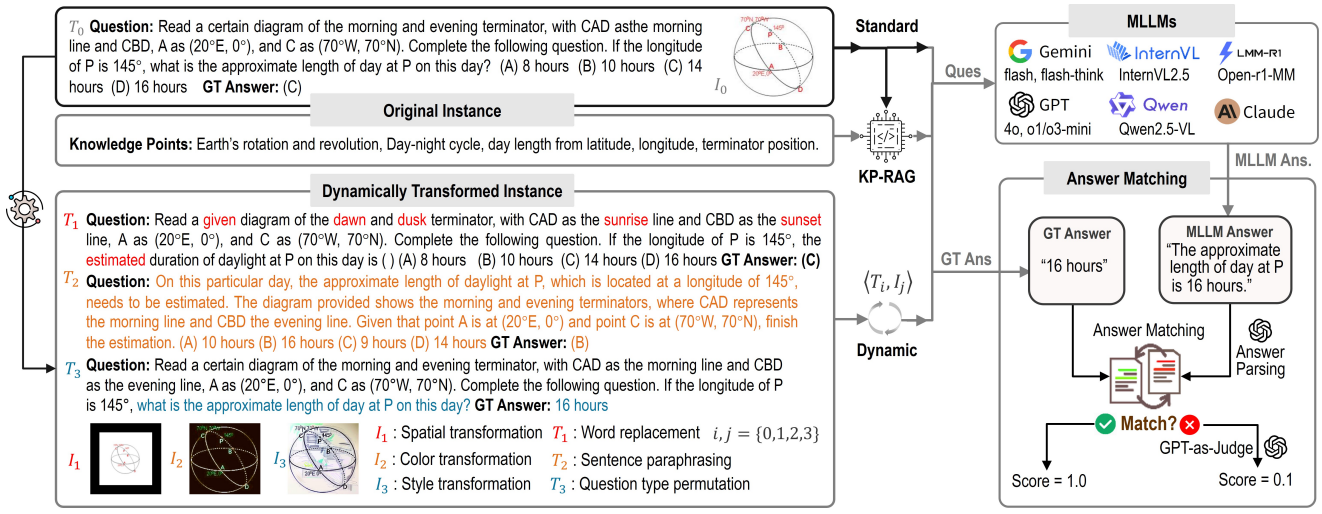

Figure 2: Our multi-dimensional evaluation pipeline comprises standard, KP-RAG, and dynamic evaluation methods.

MDK-Full Statistics	
Total instances	141,320
Text-only instances	77,857 (55.1%)
Multimodal instances	63,463 (44.9%)
Total images	105,218
Exam years coverage	12 years
MDK-Mini Statistics	
Total instances	14,856
Easy/Medium/Hard	4.952 each
Knowledge Taxonomy Statistics	
Level 1 Disciplines	6
Level 2 Grade	3 (math) or 2 (other)
Level 3 Subfield	36
Level 4 Curriculum	90
Level 5 Topics	499
Level 6 Knowledge points	6,225

Table 2: Key Statistics of MDK12-Bench

knowledge, question formats, and exam years.

Data Screening. The data screening of the initial collection is conducted in three stages to ensure high-quality, curriculum-aligned exam data. **Rule-based Filtering**: A comprehensive set of rules, designed by data experts, is applied to automatically remove low-quality or irrelevant instances. These rules cover text–image correspondence, image resolution and clarity, content completeness, metadata accuracy, structural and formatting consistency, semantic coherence, duplication and redundancy checks, logical soundness, and appropriate year coverage. After this step, the dataset is reduced to 4.2M instances. **GPT-based Filtering**: To further refine quality, GPT-4o is employed to automatically assess semantic consistency between questions and answers, reasoning soundness, factual correctness, language clarity, and overall content completeness. This automated review filters the dataset down to 0.6M instances. **Educator Filtering**: Finally, professional K–12 educators manually review the remaining data to ensure strict curriculum alignment, question–answer correctness, reasoning validity, and adherence to formatting standards. This validation results in a curated dataset of 0.2M instances.

Data Processing. Following screening, we applied rule-based parsing to convert each exam instance into a uniform structured data format and converted into JSON and TSV files. To ensure linguistic and formatting consistency the GPT-4o is used to translate Chinese text into English and verified by domain experts for technical accuracy. Data experts then inspected the processed dataset for translation fidelity, content completeness, unit and encoding consistency, equation formatting, question categorization, and overall compliance with formatting standards. Rules were established to automatically remove instances that failed to meet these criteria, yielding a final dataset of 141.3K exam instances. Each instance includes the following data fields: difficulty level, exam year, question form, question, answer, text, image, grade level, curriculum, topic, knowledge points, and answer explanation.

Knowledge Taxtonomy. We constructed a six-layer knowledge taxonomy from the processed data: Level 1 – disciplines, Level 2 – grades, Level 3 – subfields, Level 4 – curriculum, Level 5 – topics, and Level 6 – knowledge points. The benchmark covers six subjects: mathematics, physics, chemistry, biology, geography, and information science. Each subject includes middle school (K7–K9) and high school (K10–K12) grades, with mathematics additionally covering primary school (K1–K6). Data experts defined six subfields per subject, and GPT-4o was used to map 90 curricula to 36 subfields, followed by manual inspection. The complete list of subfields is shown in Fig. 1. Each processed question is linked to this taxonomy, enabling structured, fine-grained knowledge representation. The statistics are in the Table 2.

Data Statistics. The statistics of MDK12-Bench is summarized in Table 2, the full set of our benchmark **MDK12-Full** comprises 141,320 unique exam instances, including 77,857 (55.1%) text-only and 63,463 (44.9%) text–image pairs, totaling 105,218 images. Covering a 12-year span (2016–2025),

Models	Overall	Mathematics			Physics			Chemistry			Biology			Geography			Info Sci		
		Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
Gemini2-thinking	**67.8**	**68.7**	**67.9**	50.6	**69.3**	61.8	54.2	74.4	70.2	51.7	**74.1**	67.4	47.2	72.2	**77.5**	63.3	81.4	**77.3**	**73.1**
GPT-o1	65.5	56.2	50.1	**58.2**	60.9	58.4	52.5	71.8	76.5	60.5	59.7	**81.1**	55.7	**77.3**	70.7	67.1	**86.4**	66.0	69.8
GPT-o1-mini	62.4	53.3	47.8	55.6	58.1	54.3	49.2	68.9	71.7	57.4	57.4	78.2	53.4	73.5	67.3	64.1	82.3	63.4	66.5
Gemini2-flash	61.5	65.8	65.3	42.1	65.6	58.3	49.7	69.4	65.1	43.8	69.0	63.1	38.9	67.9	73.3	58.1	75.9	72.6	69.3
GPT-4o	59.0	55.1	53.4	45.3	59.3	54.7	49.6	64.3	64.0	60.3	67.2	67.8	**69.2**	66.2	57.4	53.3	68.4	62.6	70.1
Claude-3.7	58.2	54.4	51.2	43.3	56.3	53.1	47.6	62.4	61.3	57.7	64.5	65.2	66.8	63.6	55.4	51.5	66.2	60.7	68.4
Qwen2.5-VL-72B	67.5	67.0	63.1	55.0	68.9	63.8	57.9	76.7	**77.1**	**66.8**	72.7	71.5	68.0	71.2	67.3	68.7	69.3	65.1	65.4
InternVL2.5-MPO	65.2	55.6	50.9	43.2	64.6	59.4	53.7	**78.4**	73.1	60.5	73.9	70.5	63.6	72.9	73.6	67.5	76.2	68.3	67.3
InternVL2.5-78B	64.6	51.8	48.5	41.1	60.9	55.6	49.7	78.2	75.9	59.8	74.0	70.2	65.5	74.9	72.8	**69.0**	79.1	68.6	66.3
QVQ-72B	64.4	60.9	62.3	54.3	60.6	**68.8**	**68.7**	70.6	68.7	62.1	65.3	63.9	58.2	64.8	66.8	65.9	66.7	62.9	67.4
Qwen2.5-VL-7B	60.3	59.7	56.2	46.7	57.2	56.2	46.7	66.2	67.6	56.9	64.2	62.3	59.8	61.7	63.0	67.9	63.6	64.6	62.0
InternVL2.5-8B	54.6	46.1	40.5	35.9	51.3	45.0	38.8	63.9	65.3	51.1	59.4	56.7	54.3	64.3	58.1	58.5	73.5	62.1	57.4
Qwen2-VL-7B	45.4	42.3	37.9	31.6	45.7	39.1	33.8	53.0	50.4	41.3	52.1	46.8	49.7	43.0	47.8	41.5	55.4	51.9	53.3

Table 3: Performance of MLLMs across six disciplines and three difficulty levels, and average over all

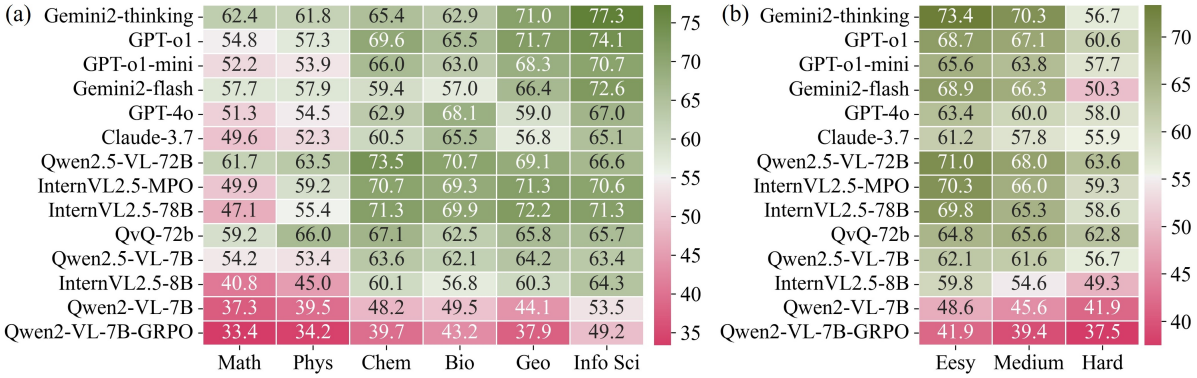

Figure 3: Comparison of average model accuracy with respect to (a) six disciplines and (b) exam difficulty levels

it includes five question formats: single-choice, multiple-choice, fill-in-the-blank, true/false, and open-ended. To support lightweight evaluation, we introduce **MDK12-Mini**, consisting of 10% of MDK12-Full uniformly sampled across easy, medium, and hard levels. Knowledge points are uniformly sampled to ensure each subset instance is linked to at least one unique knowledge point. In the Supplementary Material, we provide evidence demonstrating that MDK12-Mini yields evaluation results comparable to those of MDK12-Full.

MLLM Evaluation

Evaluation Methods

As illustrated in Fig. 2, the evaluation is based on answer matching which comprises three steps: (1) Input question into the MLLMs to generate an answer; (2) GPT-4o is prompt to parse and extract the MLLM response as its final answer; and (3) MLLM answer is compared with the ground truth answer. compared against the ground truth. Exact matches score 1.0, while partial matches are graded by GPT-as-Judge using predefined rules (e.g., 0.5 for one of two filled blanks, m/n for m correct out of n choices). In **Standard** evaluation, the MLLM is given the original question and evaluated against its ground truth answer. **Dynamic** evaluation transforms both the question and its ground truth, while **KP-RAG** evaluation enriches the question with relevant knowledge points, prompting the model to elaborate on these points and answer using both the question and expanded knowledge.

The dynamic evaluation framework introduces controlled perturbations to images and texts, including both questions and ground-truth answers, curating new test samples during MLLM evaluation. A GPT-based judge checks whether dynamic text or image alterations alter the meaning of the original question or answer and removes samples with significant changes. The dynamic transform includes visual and textual transforms as detailed below.

Image bootstrapping strategies apply spatial, color, and style variations to increase visual diversity and difficulty in visual recognition and reasoning. They preserve semantics while making the context unfamiliar to models. **Spatial.** We pad the original image with colors uniformly sampled from black, white, and grey. The padding width is proportional to the image dimension along each side, with the ratios uniformly sampled in the range between 10% and 20%. The image padding allows the evaluation of a model's ability to recognize layout changes and apply its structural knowledge to compare, contrast, and reason about image layout changes. **Color.** In this step, the colors of the original image were

Model	Overall			Easy			Medium			Hard		
	Standard	Dynamic	Δ	Standard	Dynamic	Δ	Standard	Dynamic	Δ	Standard	Dynamic	Δ
Gemini2-thinking	**58.1**	41.6	16.5	**66.7**	**43.8**	22.9	**57.0**	**44.8**	12.2	**51.5**	36.2	15.3
Gemini2-flash	56.4	**47.0**	9.4	66.6	50.1	16.4	54.7	46.1	8.6	48.9	**44.5**	4.4
GPT-4o	51.2	40.9	10.3	54.1	35.7	18.5	53.7	51.3	2.4	35.4	34.8	0.6
Claude-3.7	46.7	31.4	15.3	49.2	32.3	16.9	50.2	36.3	13.9	40.5	25.2	15.3
InternVL2.5-8B	41.7	26.1	15.6	48.5	23.5	25.0	44.1	27.5	16.6	38.4	30.8	7.7
Qwen2-VL-7B	27.3	26.1	1.2	31.8	34.6	-2.8	25.5	25.4	0.0	25.6	20.5	5.0

Table 4: Accuracy of MLLMs on standard vs. dynamic evaluation; Δ shows their difference. **Best** and second best highlighted.

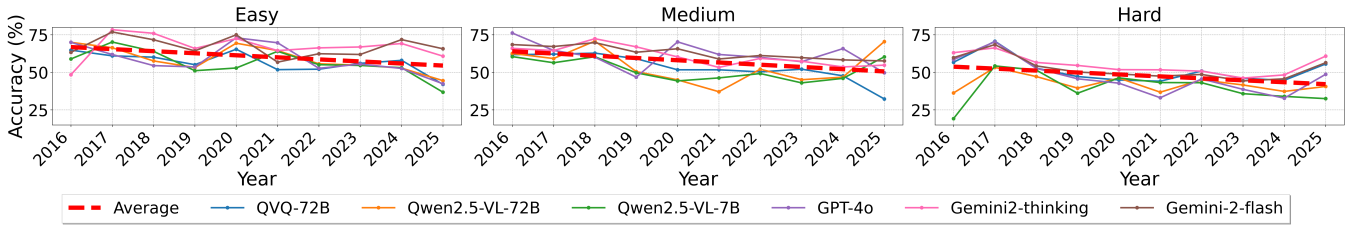

Figure 4: Model accuracy across exam years.

inverted. Salt-and-pepper noise of random noise density is also added. This transformation assesses the model's ability to utilize its familiar visual knowledge to identify and reason about visual entities when an image experiences significant color distortions and random visual artifacts. **Style.** Using the Flux-Dev (Black Forest Labs 2024) model, we introduce subtle style variations without significantly altering its key visual elements and question semantics. This tests whether models can rely on their physical appearance understanding and knowledge to recognize and reason over unseen style variations.

We also introduce three textual bootstrapping methods to modify questions while preserving the answer's correctness.**Word Substitution.** We replace certain keywords with synonyms or contextually related expressions. This tests how well a model can maintain an accurate understanding when familiar terms are changed, thus assessing vocabulary sensitivity and semantic generalization. **Sentence Paraphrasing.** We rephrase entire sentences through variations in sentence structure, word order, or style. This checks whether a model can consistently capture the underlying meaning even when the surface form of the text is altered. **Question Type Conversion.** We convert a question from one format to another, such as turning a multiple-choice problem into a fill-in-the-blank.

Experiments

Baselines and Experimental Setup

We conduct systematic and fine-grained evaluation of a set of ten existing MLLMs from multiple dimensions. The baseline MLLMs including **Proprietary** models: Gemini-2.0-flash-exp (Team et al. 2023), Gemini-2.0-flash-thinking-exp (Team et al. 2023), GPT-4o (OpenAI 2024a), GPT-o1-mini (OpenAI 2024b), Claude-3.7-Sonnet (Anthropic 2023). **Open-source** models: Qwen2.5-VL (Bai et al. 2025), InternVL2.5 (Chen et al. 2024b), QVQ-72B-preview (Team 2024), InternVL2.5-78B-MPO (Wang et al. 2024a). All experiments are performed in a zero-shot setup and evaluated by accuracy metric, demonstrating MLLMs' ability to generalize in multi-disciplinary problem solving without few-shot prompting or model fine-tuning. In the following discussions, we also compare **Reasoning** models with their **Chat** counterparts, such as Gemini2-thinking vs. Gemini2-flash, GPT-o1 vs. GPT-4o, and InternVL2.5-MPO vs. InternVL2.5-78B.

Cross-Discipline and Difficulty Results

Table 3 presents performance comparison across six disciplines on the easy, medium, and hard subsets of MDK12-Mini. **Cross-Disciplines** performance is shown in Fig. 3(a), illustrates the average accuracy of each model. Models consistently perform worse in Math and Physics, with scores 7.6% lower than the overall average of 59.8%. In contrast, Chemistry, Biology, Geography, and Information Science achieve an average score that 3.7% higher than the overall average. **Cross-Difficulty** performance is presented in Fig.3(b), which shows the average accuracy of each model across the three difficulty levels. All models show decreased accuracy on harder exam questions, with an average drop of 8.3% compared to easier ones. Larger models outperform smaller ones across disciplines and difficulty levels. Separately, reasoning-oriented models also achieve higher accuracy than their chat-focused counterparts, a trend more evident in the Gemini and GPT series than in the Intern series. The evaluation on MDK-Full is provided in the Supplementary Material and shows results consistent with those of MDK-Mini.

Cross-Year Evaluation Results

A year-by-year accuracy breakdown across difficulty levels is shown in Fig. 4, which showcases the temporal shifts in model performance relative to exam year. While accuracy

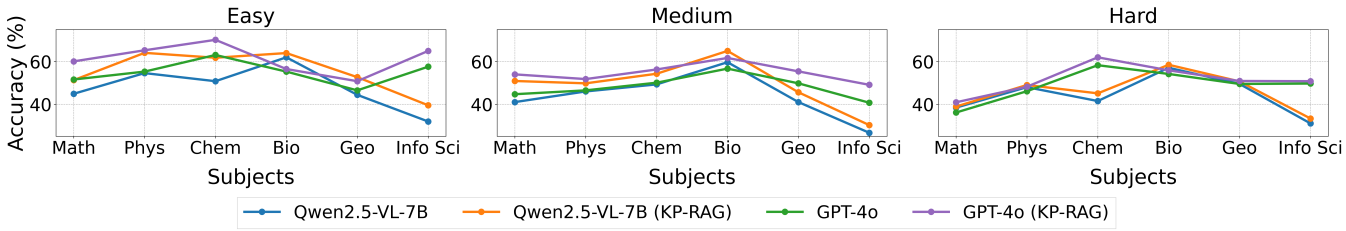

Figure 5: Comparison of model accuracy with vs. without KP-RAG.

naturally declines with increasing difficulty, we observe a further performance drop on newer exams across all difficulty levels, with accuracy gaps of 12.3%, 13.4%, and 11.6% between the oldest and newest exams for easy, medium, and hard levels, respectively. This temporal trend may stem from distributional shifts where newer exams introduce updated novel concepts or rephrased questions that differ from the training data of models. As a result, models face reduced familiarity and limited exposure, leading to lower accuracy even at similar difficulty levels.

Dynamic Evaluation Results

We sampled 50% of MDK12-Mini's instances (695 easy, 818 medium, 1124 hard) as the standard set and generated a dynamic query set using three textual and three visual bootstrapping methods. Table 4 shows that dynamic evaluation reduces model performance by an average of 13.7%, indicating its generalization limitations. Despite strong baseline performance, leading models especially reasoning model (e.g., Gemini-think) are more sensitive to contextual shifts possibly due to their stronger context-aware capability, more complex reasoning chain, or overfitting to massive static pretraining corpora, which dynamic perturbations easily disrupt.

KP-RAG Evaluation Results

We compare the accuracy of Qwen2.5-VL-7B and GPT-4o with and without knowledge-point referenced generation (KP-RAG) in Fig. 5. It is observed that incorporating KP-RAG improves model accuracy by an average of 6.9%, 6.0%, and 2.1% on easy, medium, and hard exams, respectively. The larger gains on easy and medium exams likely arise because these questions are less reasoning-intensive and more knowledge-retrieval driven, allowing explicit knowledge-point augmentation to benefit the model. In contrast, harder exams often require multi-step reasoning, abstract problem-solving, or cross-knowledge integration, where simply adding related knowledge points offers limited improvement.

Error Analysis

We analyze 100 sampled errors from five models: Gemini2-thinking, Gemini2-flash, InternVL2.5-MPO, InternVL2.5-78B, and InternVL2.5-8B, and categorized them into five types (Fig. 6): Question Misunderstanding, Reasoning Error, Visual Comprehension Error, Incomplete Answers, and Other Errors. Reasoning models (Gemini-think and InternVL2.5-MPO) reduce reasoning errors by 12% and 22% compared to their chat counterparts (Gemini2-flash and InternVL2.5-78B),

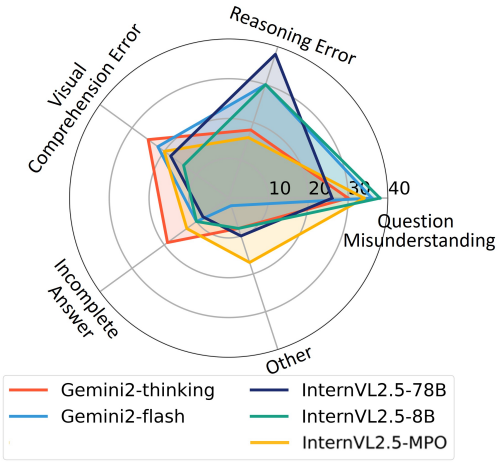

Figure 6: Error analysis of five representative models.

likely due to RL-based reasoning-oriented post-training, but exhibit 2% higher visual errors and more incomplete answers (12%, 5%) due to vision encoder limits and the 2048-token cutoff. Larger models (e.g., InternVL2.5-78B) reduce visual and other errors by 3% and 1% compared to InternVL2.5-8B but show no improvement in reasoning, question understanding, or incomplete answers. This suggests that scaling mainly enhances perception through richer multimodal representations but little to improve reasoning accuracy, prompt interpretation, instruction-following, or token-length limitations.

Conclusion

We present MDK12-Bench, a comprehensive multimodal benchmark for evaluating the problem-solving intelligence of MLLMs across diverse disciplines and dimensions based on real-world K–12 exams. We propose a dynamic framework that applies textual and visual bootstrapping strategies to assess model generalization and rigorously mitigate data contamination. Experimental results reveal limitations of state-of-the-art MLLMs, including high sensitivity to contextual changes, poor generalization to novel and complex tasks, and limited benefits from knowledge augmentation in reasoning-intensive problems. These findings affirm MDK12-Bench as an essential foundation for diagnosing model strengths and limitations and for guiding the development of robust and generalizable multimodal intelligence towards improved adaptability, reasoning, and knowledge integration.

Acknowledgments

This paper is supported by the National Key R&D Program of China No.2022ZD0160101, and also sponsored by the NUS Startup Grant (Presidential Young Professorship), Singapore MOE Tier-1 Grant, ByteDance Grant, NUS ARTIC Grant, Apple Grant, Alibaba Grant, and Google Grant.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Anthropic. 2023. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.anthropic.com/news/claude-3-family. Accessed: 2025-03-24.

Arora, D.; Singh, H.; et al. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7527–7543.

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*.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux. Accessed: 2024-11-05.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are We on the Right Way for Evaluating Large Vision-Language Models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Das, R. J.; Hristov, S. E.; Li, H.; Dimitrov, D. I.; Koychev, I.; and Nakov, P. 2024. EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models. *arXiv preprint arXiv:2403.10378*.

Ding, M.; Deng, C.; Choo, J.; Wu, Z.; Agrawal, A.; Schwarzschild, A.; Zhou, T.; Goldstein, T.; Langford, J.; Anandkumar, A.; et al. 2024. Easy2Hard-Bench: Standardized Difficulty Labels for Profiling LLM Performance and Generalization. *Advances in Neural Information Processing Systems*, 37: 44323–44365.

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. In *International Conference on Machine Learning*.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 3828–3850.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hu, W.; Gu, J.-C.; Dou, Z.-Y.; Fayyaz, M.; Lu, P.; Chang, K.-W.; and Peng, N. 2025. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models. In *The Thirteenth International Conference on Learning Representations*.

Huang, Z.; Wang, Z.; Xia, S.; Li, X.; Zou, H.; Xu, R.; Fan, R.-Z.; Ye, L.; Chern, E.; Ye, Y.; et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37: 19209–19253.

Jin, K.; Wang, Y.; Santos, L.; Fang, T.; Yang, X.; Im, S. K.; and Oliveira, H. G. 2025. Reasoning or Not? A Comprehensive Evaluation of Reasoning LLMs for Dialogue Summarization. *Expert Systems with Applications*, 299: 129831.

Lan, G.; Inan, H. A.; Abdelnabi, S.; Kulkarni, J.; Wutschitz, L.; Shokri, R.; Brinton, C. G.; and Sim, R. 2025. Contextual Integrity in LLMs via Reasoning and Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Li, C.; Zhu, C.; Zhang, T.; Lin, M.; Zhou, Z.; and Xie, J. 2025. K12Vista: Exploring the Boundaries of MLLMs in K-12 Education. *arXiv preprint arXiv:2506.01676*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233.

Lohman, D. F.; and Lakin, J. M. 2011. Intelligence and reasoning. *The Cambridge handbook of intelligence*, 419–441.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

OpenAI. 2024a. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2025-03-08.

OpenAI. 2024b. GPT-o1-mini: A Cost-Effective Reasoning Model. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/. Accessed: 2025-03-08.

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.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2025. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Sternberg, R. J. 1982. Reasoning, problem solving, and intelligence. *Handbook of human intelligence*, 225–307.

Sun, L.; Han, Y.; Zhao, Z.; Ma, D.; Shen, Z.; Chen, B.; Chen, L.; and Yu, K. 2024a. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19053–19061.

Sun, Q.; Qiu, Z.; Ye, H.; and Wan, Z. 2019. Multinational Corporation Location Plan under Multiple Factors. In *Journal of Physics: Conference Series*, volume 1168, 032012.

Sun, Y.; Wu, H.; Zhu, C.; Zheng, S.; Chen, Q.; Zhang, K.; Zhang, Y.; Wan, D.; Lan, X.; Zheng, M.; et al. 2024b. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, 56–73.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, Q. 2024. QVQ: To See the World with Wisdom. https://qwenlm.github.io/blog/qvq-72b-preview/. Accessed: 2025-03-29.

Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; and Dai, J. 2024a. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442*.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024b. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *International Conference on Machine Learning*, 50622–50649.

Wen, J.; Cui, J.; Zhao, Z.; Yan, R.; Gao, Z.; Dou, L.; and Chen, B. M. 2023. SyreaNet: A Physically Guided Underwater Image Enhancement Framework Integrating Synthetic and Real Images. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5177–5183.

Xu, R.; Wang, Z.; Fan, R.-Z.; and Liu, P. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Yang, Y.; Zhang, S.; Shao, W.; Zhang, K.; Bin, Y.; Wang, Y.; and Luo, P. 2024. Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping. *arXiv preprint arXiv:2410.08695*.

Yu, J.; Zhou, S.; Yang, D.; Li, S.; Wang, S.; Hu, X.; Xu, C.; Xu, Z.; Shu, C.; and Yuan, Z. 2025. Mquant: Unleashing the inference potential of multimodal large language models via static quantization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1783–1792.

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.

Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; and Wei, X. 2025a. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*.

Zeng, S.; Qi, D.; Chang, X.; Xiong, F.; Xie, S.; Wu, X.; Liang, S.; Xu, M.; and Wei, X. 2025b. JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation. *arXiv preprint arXiv:2509.22548*.

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.

Zhang, X.; Liu, H.; Xue, L.; Li, X.; Guo, W.; Yu, S.; Ru, J.; and Xu, H. 2021. Multi-objective collaborative optimization algorithm for heterogeneous cooperative tasks based on conflict resolution. In *International Conference on Autonomous Unmanned Systems*, 2548–2557.

Zhao, Z.; and Chen, B. M. 2023. Benchmark for Evaluating Initialization of Visual-Inertial Odometry. In *2023 42nd Chinese Control Conference (CCC)*, 3935–3940.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *NAACL-HLT (Findings)*.

Zhou, P.; Peng, X.; Song, J.; Li, C.; Xu, Z.; Yang, Y.; Guo, Z.; Zhang, H.; Lin, Y.; He, Y.; et al. 2025a. OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 56–66.

Zhou, P.; Peng, X.; Zhang, F.; Xu, Z.; Ai, J.; Qiu, Y.; Li, C.; Li, Z.; Li, M.; Feng, Y.; et al. 2025b. Mdk12-bench: A comprehensive evaluation of multimodal large language models on multidisciplinary exams. *arXiv preprint arXiv:2508.06851*.

Zhou, S.; Wang, S.; Yuan, Z.; Shi, M.; Shang, Y.; and Yang, D. 2025c. GSQ-Tuning: Group-Shared Exponents Integer in Fully Quantized Training for LLMs On-Device Fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Zhu, K.; Wang, J.; Zhao, Q.; Xu, R.; and Xie, X. 2024. Dynamic Evaluation of Large Language Models by Meta Probing Agents. In *Forty-first International Conference on Machine Learning*.