MULTI: Multimodal Understanding Leaderboard with Text and Images

Zichen Zhu and **Yang Xu** and **Lu Chen**† and **Jingkai Yang** and **Yichuan Ma**
Yiming Sun and **Hailin Wen** and **Jiaqi Liu** and **Jinyu Cai**
Yingzi Ma and **Situo Zhang** and **Zihan Zhao** and **Liangtai Sun** and **Kai Yu**†
X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
{JamesZhutheThird, xuyang0112, chenlusz, kai.yu}@sjtu.edu.cn

Abstract

Rapid progress in multimodal large language models (MLLMs) highlights the need to introduce challenging yet realistic benchmarks to the academic community. Existing benchmarks primarily focus on simple natural image understanding, but MULTI emerges as a cutting-edge benchmark for MLLMs, offering a comprehensive dataset for evaluating MLLMs against understanding complex figures and tables, and scientific questions. This benchmark, reflecting current realistic examination styles, provides multimodal inputs and requires responses that are either precise or open-ended, similar to real-life school tests. It challenges MLLMs with a variety of tasks, ranging from formula derivation to image detail analysis, and cross-modality reasoning. MULTI includes over 18,000 questions, with a focus on science-based QA in diverse formats. We also introduce MULTI-ELITE, a 500-question subset for testing the extremities of MLLMs, and MULTI-EXTEND, which enhances In-Context Learning research with more than 4,500 knowledge pieces. Our evaluation indicates significant potential for MLLM advancement, with GPT-4V achieving a 63.7% accuracy rate on MULTI, in contrast to other MLLMs scoring between 31.3% and 53.7%. MULTI serves not only as a robust evaluation platform but also paves the way for the development of expert-level AI. Details and access are available at: https://OpenDFM.github.io/MULTI-Benchmark/.

Figure 1: An example of MULTI. English translations of Chinese text are shown for better readability. The markdown format remains as it is.

1 Introduction

The rapid advancement in large-scale language models (LLMs) has led to significant achievements in natural language processing and related disciplines. Yet, human communication and understanding extend beyond language, encompassing images, tables, mathematical and chemical formulas, graphs, diagrams, cartoons, posters, and other visual mediums. These play a crucial role in conveying information, particularly in scientific areas. Consequently, there's a growing interest in developing Multimodal LLMs (MLLMs) capable of processing and generating across various modalities, including visual ones, and performing tasks that require cross-modal reasoning.

Evaluating MLLMs presents unique challenges. Current benchmarks (Lu et al., 2022; Li et al., 2023c; Yue et al., 2023) either focus narrowly on natural scene images or are simplistic, failing to thoroughly assess the models' abilities. Many scientific benchmarks (Sun et al., 2023a; Huang et al., 2023) rely on multiple-choice questions with a single answer, which might not accurately gauge a model's comprehension and can lead to

†The corresponding authors are Lu Chen and Kai Yu.

superficial learning, i.e., the model will not look into other choices if the correct choice is straightforward. A more robust, detailed, and multi-faceted dataset is necessary to effectively evaluate MLLMs under diverse conditions and scenarios. Current benchmarks mentioned above are evaluated on English context, while the rapid progression of Chinese MLLMs highlights the need for a Chinese multimodal benchmark with both Chinese contents in text and image and brings new challenges to the community.

In this paper, we introduce MULTI, a novel benchmark named **M**ultimodal **U**nderstanding **L**eaderboard with **T**ext and **I**mages, specifically designed to evaluate multimodal LLMs on cross-modal questions. MULTI comprises 18,430 questions sourced from various educational and online materials, with most questions undergoing multiple rounds of human annotation for quality assurance. These questions span a range of scientific subjects like mathematics, physics, chemistry, biology, and computer science. They are crafted to test understanding and generation in various formats and complexity levels and are categorized into multiple-choice (select single or multiple answers), fill-in-the-blank, and open-ended questions.[1]

To further challenge multimodal LLMs, we developed two subsets within MULTI: MULTI-ELITE, consisting of 500 carefully selected tough questions, aiming to probe the limits of these models. MULTI-EXTEND, featuring 4,595 knowledge pieces, tests the models' learning capabilities and knowledge transfer. These subsets offer deeper insights into the strengths and weaknesses of multimodal LLMs, fostering new research avenues. An example of MULTI is shown in Figure 1. For more examples please refer to Appendix F.

We conducted comprehensive experiments on MULTI using leading-edge multimodal and single-modality LLMs (Bai et al., 2023a; Hu et al., 2023; Du et al., 2022; Sun et al., 2023b; OpenAI, 2022, 2023a,b). We analyzed their performance across various question types, subjects, and modalities, examining both their successes and shortcomings. Our findings reveal that multimodal LLMs still lag behind human performance in many aspects of MULTI, highlighting challenges like cross-modal alignment, logical reasoning, mathematical computations, and image comprehension. Results show that none of these models get half scores, not to mention the MULTI-ELITE set where GPT-4V only gets 14.0% of scores, while most of the other models get a score for near random. This indicates a large road for improvement.

In conclusion, We make the following contributions in this paper:

- We propose MULTI, a substantial and challenging multimodal benchmark focusing on Chinese scientific questions, designed to evaluate multimodal LLMs.
- We introduce MULTI-ELITE and MULTI-EXTEND sets to test models' bottleneck and in-context learning abilities, aiming for a more nuanced evaluation of multimodal LLMs.
- We present detailed experiments with various state-of-the-art multimodal and single-modality LLMs on MULTI, providing both qualitative and quantitative insights into their performance.
- We make the MULTI leaderboard, dataset, evaluation code, and the two subsets available to the research community, encouraging further participation and advancement in the field of multimodal LLMs.

[1]For the sake of simplifying writing, in the following paragraphs we may use the abbreviates multiple choices questions with a single answer as **SA** and those with multiple answers as **MA**. We also use **FB** for fill-in-the-blank questions, and **OP** for open-ended writing questions.

2 Related Works

Multimodal Large Language Models. With rapid progress has been made in instruction-tuned Large language models (LLMs) like FLAN-T5 (Chung et al., 2022), OPT-IML (Iyer et al., 2022), and Vicuna (Zheng et al., 2023), some studies work on introducing multimodal instruction tuning to enhance the open-source Multimodal Large Language Models (MLLMs), such as LLaVA (Liu et al., 2023c) and MiniGPT-4 (Zhu et al., 2023). Like LLMs, the primary trajectory to enhance the instruction-following capabilities of MLLMs is to improve the diversity, quality, and quantity of visual instruction data. Some works, such as SVIT (Zhao et al., 2023) and Instruction-BLIP (Dai et al., 2023), employ GPT-creation or collect traditional vision-language tasks to increase instruction diversity and quantity. Additionally, LRV (Liu et al., 2023b) includes both positive and negative instructions to mitigate multimodal hallucination. Another aspect of enriching visual instruction data is to introduce multilingual data. Visual instruction datasets like M^3IT (Li et al., 2023e) and MIMIC-IT (Li et al., 2023a) consist of multilingual instructions and tasks, which assist MLLMs in comprehending unseen instructions in different languages. VisualGLM (Du et al., 2022), VisCPM (Hu et al., 2023), QWen-VL (Bai et al., 2023a) and InternLM-XComposer-VL (Zhang et al., 2023a) train on both English and Chinese instructions, exemplifying these explorations in depth. GPT-4V(ision) (OpenAI, 2023b) and Gemini (Team, 2023), trained on a broad range of multilingual texts and images, demonstrate remarkable performance.

Benchmarks for Multimodal Large Language Models. In assessing MLLMs, traditional methods primarily rely on established vision-language (VL) benchmark datasets. Renowned benchmarks such as VQA (Goyal et al., 2017), OK-VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019), and MSCOCO (Lin et al., 2014) are tailored to specific VL tasks like image captioning, open-domain visual question answering, and visual reasoning. While the evaluation based on standard benchmark datasets yields significant insights into MLLMs' capabilities, these approaches may not entirely capture their comprehensive intelligence in real-world scenarios. Therefore, a diverse array of benchmarks has been developed to examine the varied real-world aspects of MLLMs. Benchmarks like LLaVA-Bench (Liu et al., 2023c), MMBench (Liu et al., 2023d), MM-VET (Yu et al., 2023), TouchStone (Bai et al., 2023b), MLLM-bench (Ge et al., 2023), and SEED-Bench(Li et al., 2023c,b), for instance, leverage GPT to evaluate the relevance and helpfulness of human-like long responses in the reality. POPE (Li et al., 2023f) and HallusionBench (Liu et al., 2023a) introduce various analytical criteria for the holistic evaluation of MLLMs' hallucinations. Furthermore, M3Exam (Zhang et al., 2023b), SciGraphQA (Li and Tajbakhsh, 2023), MathVista (Lu et al., 2023), AGIEval (Zhong et al., 2023), and MMMU (Yue et al., 2023) consider MLLMs as experts to extend the evaluation scope by incorporating advanced perception and reasoning within domain-specific knowledge, for instance, scientific questions and driving tests. The works most related to us are M3Exam, ScienceQA, SciEval (Sun et al., 2023a) and C-Eval (Huang et al., 2023). Our approach distinguishes itself by offering a broader spectrum of question types compared to the first two and supports a multimodal evaluation in contrast to the last two.

3 The MULTI Benchmark

3.1 Overview

We propose MULTI, a `Multimodal Understanding Leaderboard with Text and Images`, which can serve as a challenging and diverse benchmark for the MLLM community. The detailed statistics are provided in Appendix B.

3.2 Data Construction Process

To build MULTI, we follow several principles: 1) We use markdown and LaTeX formula format to annotate each question with high quality and precision. 2) We revise and refine each question multiple times to prevent data leakage and increase difficulty. 3) We rate every question based on its difficulty and content richness.

Data Source We collected more than 2.7M raw data from the Internet, ranging from exams and quizzes from Chinese junior and senior schools and several society exams. We design an algorithm to pick out a proportion of the questions as the fundamental data of our benchmark. The selection is based on the questions' text length, number of images, corresponding subjects, and knowledge pieces, to reach a higher diversity of questions and coverage of knowledge. The details are presented in Appendix D. We also collected questions from internal exams and practises of several top universities. After the selection, we obtain over 18k questions as the raw data.

Data Pre-process The raw data range from HTML, photocopy, hand script, and plain text, and we preprocess some of them to reduce the load of further annotation. We remove most HTML tags indicating irrelevant content of the question such as alignment, color, etc. We reserve tags for underlining (<u> </u>), and we transfer several tagged styles including bold, italic, and tabular data into markdown format. For some complex tables that cannot be well converted, we save them as a screenshot picture after rendering with HTML. For photocopy and hand script, we adopt OCR tools to convert text content, crop images, and figures, and integrate them into markdown. We further transcript most of the math functions and chemistry structures into LaTeX format, with a small portion remaining as images.

Data Annotation An online platform has been developed for the annotation of data, employing skilled human annotators. These annotators, primarily undergraduate students from top universities in China familiar with exam quizzes and markdown rules, undertake this comprehensive task covering various aspects from formatting to semantic analysis:

- **Format Level**. Tasks at this level involve the removal of residual HTML tags and the conversion of content into markdown format (refer to examples (1) and (3) in Figure 5 of Appendix F). This includes transforming complex mathematical and chemical equations, usually in HTML, into LaTeX. For this purpose, Mathpix [2] is utilized for efficiency. The annotators also correct any character-level errors in text and formulas, often resulting from OCR inaccuracies.
- **Content Level**. Annotators split the raw content into distinct sub-questions, segregating parts like the question, answer, and analysis (if presented in raw data). We divide the question content into general and specific parts. The general part includes the problem introduction, background information, or instructions applicable across all sub-questions, while the specific part contains details unique to each sub-question. Annotators also standardize typesetting and image placement, ensuring a consistent format across questions of the same type (e.g., for multiple-choice questions with a single image, the format follows `problem content(general) + question content(specific) + [MASK] + [IMAGE_1] + choices`).
- **Label Level**. Annotators evaluate each question's difficulty and quality. A question is considered of higher quality if it includes comprehensive content, multiple images, or detailed explanations. Difficulty assessment is subjective. These evaluations aid in curating our MULTI-ELITE dataset. Annotators also verify information like question type, educational level, and related knowledge pieces.
- **Semantic Level**. At this stage, annotators are advised to identify and correct both superficial errors (e.g., empty/duplicate choices, incomplete mathematical functions such as between `$32$`, `$3^2$`, `$\sqrt[3]{2}$`, `$3\sqrt{2}$`, `$\frac{3}{2}$`) and more profound errors relating to factual accuracy and logical reasoning, such as content that is lacking or leads to inconclusive results. Those questions with profound errors will be dropped.

Data Post-process To collect more challenging data for our benchmark, we adopt several data post-process strategies:

- **Formation**. During the data preprocessing stage and annotation stage, we formatted the questions in a render-friendly manner, meanwhile, reducing the content similarity that the MLLMs are trained on. During this stage, we assess if there are any omissions or missing elements.

[2] https://mathpix.com/snipping-tool

- **Disambiguration**. For blank-filling questions containing multiple [MASK]s, we manually modify those with parallel relations into two sub-questions (refer to example (5) in Figure 6 of Appendix F), this will simply give out a unique fixed answer.
- **Distillation**. This is completed during our annotation process. We reduce assistance information so that the answer must depend on more detailed analysis (refer to example (4) in Figure 6 of Appendix F). In this way, we greatly enhance question difficulty.
- **Transformation**. We randomly modify the questions such as from single-choice to blank-filling (refer to example (2) in Figure 6 of Appendix F), or convert certain kinds of single-choice questions into multiple-choice ones (refer to example (1) and (5) in Figure 6 of Appendix F). Lots of single-choice questions have a list of options and the choices are presented as the combination of those options where only one is correct. We transform those questions into multiple-choice questions where the options become new choices and the correct answer corresponds to the combinations. In this way we successfully increase the scale of multiple-choice questions, improving the diversity of the questions.

The details of our data annotation platform are outlined in Appendix E, which fully meets our requirements for data annotation. Initially, we extracted a total of 2.7 million questions from the internet. Through an algorithmic selection in the preprocessing stage, we narrowed this down to 18,000 questions with wide coverage. During the construction, we conducted two rounds of data annotation and three rounds of automatic checking to ensure the granularity and credibility of every question in our set. In the first round of annotation, we filtered out and modified questions based on predefined criteria. The second round of data annotation focused more on semantic analysis and data enhancement. This post-processing stage significantly increased the number of MA questions by 3.22 times, and the total point proportion of non-SA questions rose from 26.0% to 40.1%. We also removed over 800 similar questions. We provide detailed examples of the data annotation and post-process in Appendix F.

Data Distribution Our benchmark showcases a remarkable diversity in the choice architecture of multiple-choice questions, encompassing options that range from 2 to as many as 13. Furthermore, it includes questions that vary in the number of correct answers, from questions with a singular correct option to those with multiple correct choices. We provide the distribution of choices in multiple-choice questions as shown in Table 10.

In addition to multiple-choice questions, our benchmark also includes a substantial number of fill-in-the-blank and open-ended questions, creating a diverse and comprehensive range of testing scenarios. Moreover, we have incorporated unique open-response questions that require creative answers, such as drawings. It is important to note that these open-response questions are not included in our formal evaluation and scoring procedures; they are primarily proposed for qualitative research and development in the field of MLLMs. Our benchmark is carefully designed to thoroughly assess and enhance the ability of MLLMs to process and respond to various question types, resembling real-world learning scenarios.

3.3 The MULTI-ELITE Set

We have selected an additional set of 500 questions to create the advanced dataset. This set is comprised of objective questions, i.e. multiple-choice and fill-in-the-blank questions. The questions are averagely distributed in all of the subjects and education levels, evaluated as with high difficulty and quality by annotators, and with rich text and image content. The evaluation results presented in § 4 have also been considered in this selection process.

3.4 The MULTI-EXTEND Background Knowledge Dataset

External knowledge is crucial to provide critical information that can assist in solving questions using the In-Context Learning (ICL) abilities. Some of the raw questions retrieved from the Internet have corresponding knowledge pieces attached. We also collect more knowledge pieces for uncovered questions with the assistance of LLMs and search engines (e.g. New Bing[3], Wikipedia[4]). We conduct annotations on these knowledge pieces to confirm the correctness of the content and present them in the MULTI-EXTEND dataset. This dataset consists of about 4.6K knowledge pieces, it is designed to test the in-context learning abilities and knowledge transfer skills of models. This dataset provides comprehensive insights into the capabilities and limitations of multimodal LLMs, opening new pathways for research exploration.

3.5 Comparison with Existing Benchmarks

MULTI demonstrates a comprehensive blend of features that surpasses existing benchmarks in several dimensions. Notably, MULTI covers a wide array of subjects and a substantial number of questions (18K), as well as over 10K analysis and 4.6K extensive knowledge content, which is considerably larger than most benchmarks, ensuring a broad and diverse testing environment. MULTI possesses 7.7K images, which is essential for benchmarking MLLMs that require visual understanding alongside textual information. The inclusion of both single and multiple image questions, as well as a variety of answer types, makes MULTI a versatile and challenging benchmark. Meanwhile, the various sources, complex annotation, and processing stages provide sufficient augmentation to alleviate data leakage. MULTI not only encompasses variations of classic questions but also includes newly updated ones, resulting in significantly enhanced diversity.

We list the features of existing benchmarks and make a comparison with MULTI in Table 1. We believe that MULTI assembles the most advantages of the existing benchmarks and is sure to provide a good option for the community to test the capabilities of their Vision LLMs.

[3] https://bing.com/new

[4] https://wikipedia.org

Benchmark	Language	Size					Image		Answer Type				Source
		Sub	Q	Ana	Img	Kn	SI	MI	SA	MA	FB	OP	
VQA (Antol et al., 2015)	en	36	764K	-	265K	-	✓	✗	✗	✗	✓	✗	Repurposed
ScienceQA (Lu et al., 2022)	en	21	21K	19K	10K	0.3K	✓	✗	✓	✗	✗	✗	Textbooks
SciBench (Wang et al., 2023)	en	6	0.8K	-	0.1K	-	✗	✗	✗	✗	✓	✓	Textbooks
M3Exam (Zhang et al., 2023b)	9 lans	4	12K	-	3.1K	-	✓	✗	✓	✗	✗	✗	Exams
AGIEval (Zhong et al., 2023)	zh, en	20	8K	a few	-	-	✓	✗	✓	✓	✓	✗	Exams
MMBench (Liu et al., 2023d)	en	20	3K	-	3K	-	✓	✗	✓	✗	✗	✗	Internet, Repurposed
SEED-Bench (Li et al., 2023c)	en	12	19K	-	19K+	-	✓	✓	✓	✗	✗	✗	Annotated
SEED-Bench-2 (Li et al., 2023b)	en	27	24K	-	22K+	-	✓	✓	✓	✗	✗	✗	Annotated
MLLM-Bench (Ge et al., 2023)	en	42	0.4K	-	0.4K	-	✓	✗	✗	✗	✗	✓	Annotated
Touchstone (Bai et al., 2023b)	en	27	0.9K	-	0.9K	-	✓	✓	✗	✗	✗	✓	Annotated
C-Eval (Huang et al., 2023)	zh	52	14K	a few	-	-	✗	✗	✓	✗	✗	✗	Exams, Internet
SciEval (Sun et al., 2023a)	en	3	18K	-	-	-	✗	✗	✓	✗	✗	✗	Internet, Repurposed
MMMU (Yue et al., 2023)	en	30	12K	2K	11K+	-	✓	✓	✓	✗	✗	✗	Annotated, Internet, Textbooks
MULTI(ours)	zh	23	18K	10K+	7.7K	4.6K	✓	✓	✓	✓	✓	✓	Annotated, Exams, Internet

Table 1: The comparison between MULTI and other existing benchmarks. Sub: Subject or Field, Q: Question, Ana: Analysis or Explanations, Img: Images, Kn: Knowledge or Lecture. SI: the question with a single image, MI: the question with multiple images. SA: multiple-choice question with single correct answer, MA: multiple-choice question with multiple correct answers, FB: fill-in-the-blank question(no more than 10 words), OP: open-ended writing question(more than 10 words).

Creator	Model	# Paras	Form	Modality	Language	Version
Alibaba	Qwen-VL (Bai et al., 2023a)	7B	Weight	Multiple-image	zh, en	`Qwen-VL-Chat`
ModelBest	VisCPM (Hu et al., 2023)	10B	Weight	Single-image[5]	zh, en[6]	`VisCPM-Chat`
THU	VisualGLM (Du et al., 2022)	6B	Weight	Single-image	zh, en	`visualglm-6b`
FDU	MOSS (Sun et al., 2023b)	16B	Weight	Pure-text	zh, en	`moss-moon-003-sft`
SJTU&AISpeech	DFM-2.0 (Chen et al., 2022)	70B	Weight	Pure-text	zh, en	`dfm-2.0-70b-preview`
OpenAI	GPT-4V (OpenAI, 2023b)	-	API	Multiple-image	Multi-lingual	`gpt-4-vision-preview`
OpenAI	GPT-4 (OpenAI, 2023a)	-	API	Pure-text	Multi-lingual	`gpt-4-1106-preview`
OpenAI	ChatGPT(0613) (OpenAI, 2022)	-	API	Pure-text	Multi-lingual	`gpt-3.5-turbo-0613`
OpenAI	ChatGPT(1106) (OpenAI, 2022)	-	API	Pure-text	Multi-lingual	`gpt-3.5-turbo-1106`
Google	Gemini Vision (Team, 2023)	-	API	Multiple-image	Multi-lingual	`gemini-pro-vision`
Google	Gemini (Team, 2023)	-	API	Pure-text	Multi-lingual	`gemini-pro`

Table 2: The list of models evaluated on MULTI. We report Modality as how many images can the model take in one turn. Note that those MLLMs commonly support multiple-image input with chatting in several turns.

4 Experiments

4.1 Models

We evaluate a wide range of MLLMs that support Chinese, including Qwen-VL (Bai et al., 2023a), VisCPM (Hu et al., 2023), VisualGLM (Du et al., 2022), Gemini Vision (Team, 2023), and GPT-4V (OpenAI, 2023b). We evaluate these models with both multimodal input and text-only input to verify the information gain of input images. We also select several most capable LLMs for comparison with text-only input, including DFM-2 (Chen et al., 2022), MOSS (Sun et al., 2023b), ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023a), and Gemini (Team, 2023). Model specifications are Listed in Table 2. Due to the API request rate limit of Gemini and GPTs, ablation studies are mostly performed on weight-accessible models. We choose the checkpoints with largest model size and latest version, we use FP16 to accelerate interference if provided in official guidelines. We also follow the official guidelines to prompt each model so that the outputs go in the desired way.

4.2 Settings

Prompt We use specialized prompts for each question, an example is shown in Figure 2. The prompts are designed carefully according to the features of each type of question and the answer patterns expected. We also modified input format to fit into official inference guidelines. The complete collection of prompts are presented in Appendix C.

Image MULTI include questions with either none, single, or multiple images. Most MLLMs accept text accompanied by one image as input or a pure-text input. For questions with a single image, the image and text are fed in one turn. We simply drop image information when evaluating LLMs.

For pure-text questions, we use empty question info as input, except for VisCPM which compulsorily demands an image in each turn. In this case, we give the model a blank image, a 24*24 placeholder with color set to (0,0,0), along with the plain text in evaluation. When evaluating pure-text questions on GPT-4V and

你是一名来自中国的考生，你需要运用你所学的{knowledge}知识回答这道{question_type}题。You are a student from China. You need to use your knowledge of {knowledge} to answer this {question_type} question.

这道题目只有唯一的正确选项，请只给出唯一一个大写英文字母作为答案，不包含选项后面的描述，如：A，B，E。This question has only one correct option. Please give only one uppercase letter as the answer, without the description after the option, such as: A, B, E.

这道题目包含图片信息，请基于文字和图片信息，并按照格式给出答案。This question contains image information. Please give your answer directly based on the text and image information.

Question: ……

我们为你提供了一些额外材料，你可以参考这些信息来回答问题，请注意它们并不一定完整，也不一定正确，它们可能有图片输入，也有可能输入图片描述，也有可能只有文字，你需要结合你之前的知识来回答。We provide you with some extra materials. You can refer to these materials to answer the questions. Please note that they are not necessarily complete or correct. You need to combine them with your previous knowledge to answer the questions.

Knowledge: ……

请直接给出你的答案：
Please directly give your answer:

Figure 2: An example of the prompts used when evaluating a multiple-choice question with image context, knowledge piece and single correct answer.

Gemini Vision, we use evaluation results of those on the pure-text model versions, i.e., GPT-4 and Gemini for efficiency.

For questions with multiple images, as the positions of images quite matters in the evaluation result, e.g., a multiple-choice question where each choice consist an image, we use special pattern `[IMAGE_{index}]` to indicate the position and order of images. Qwen-VL, GPT-4V, and Gemini Vision naturally support multiple images as input in one turn, while other models mostly support only one image as input in one turn. We adopt the strategy of splitting the questions into multiple segments after each image and feeding them into the MLLM sequentially as rounds of conversation, where the MLLM receives each segment with a corresponding image respectively. We tune our prompts so that the MLLM may receive all the information but should only give a finalized answer after we show a signal that the question ends. The prompt we use in multi-turn input is shown in Figure 3.

4.3 Metrics

We focus on subjective questions with a certain answer, including multiple-choice and blank-filling questions. We also give a score to each objective open-ended question based on this similarity to the reference answer. The metrics we use are the following:

Multiple-choice with Single Answer (SA) Each question worth one point. We calculate the accuracy of the given answer.

[5]VisCPM only supports a single image as input in one turn and must have one, therefore for some pure text questions, we simply input a blank image of size 256*256 as input.

[6]The model backbone of VisCPM was trained on both Chinese and English corpus, while the visual multimodal module was trained on English text-image pairs.

Multiple-choice with Multiple Answers (MA) We define the total points of an MA question as the number of correct choices. Each correct choice selected will be rewarded one point. If the given answer contain any wrong choice, the score will be counted to zero. We report the score ratio (# points / # total points) as the metric. We also report accuracy as a more rigorous metric.

Fill in the Blank (FB) We define the total points of a blank-filling question as the number of the blanks marked as `[MASK]`. It is required in prompts that each line of given answer correspond to a blank in order. We follow the most strict standard of *exact match*. Therefore, only answers exactly matching the standard answers will be granted points. We report the score ratio as the final metric.

Open-ended Question (OP) The points and counting method is similar to FB, however, we use loose standard and report normalized ROUGE-L (Lin, 2004) score for each point. Please be noted that the reference answer may be concise or in detail, and there could be other possible answers.

4.4 Main Experiment Results

We report the overall and field-specific performance of tested models on the whole benchmark in Table 3, 4, and5. The leaderboard will be continually updated in the future.

Model	Overall	NI	SI	MI
Puretext (LLM)				
MOSS	32.6	36.1	27.3	17.1
DFM-2.0	49.7	63.0	28.7	11.3
Gemini	**52.2**	62.5	**36.2**	**18.3**
ChatGPT(0613)	40.1	50.1	24.4	12.4
ChatGPT(1106)	35.9	54.0	6.8	5.1
GPT-4	50.2	**74.5**	11.3	8.8
Text+Image (MLLM)				
VisualGLM	31.1	35.1	25.2	9.7
VisCPM	33.4	36.8	28.4	16.6
Qwen-VL	39.0	43.2	32.7	20.7
Gemini Vision	53.7	62.5	40.0	24.5
GPT-4V	**63.7**	**74.5**	**46.9**	**28.1**

Table 3: The main performance of models evaluated on MULTI. NI: the question with no image, SI: the question with a single image, MI: the question with multiple images.

We report the overall performance in Table 3. Qwen-VL outperform other locally implemented models less than 20 billion parameters, while remaining a noticeable gap between it and the larger-scale models, namingly a 70b DFM-2.0, GPT and Gemini series. However, even the most powerful competitor, GPT-4V, achieves a mere 63.7% score, underscoring the benchmark's complexity and challenge.

We report the performance by the image number in Table 3. It is clear that more image involved in question leads to more difficulty. For MLLMs, a higher score on NI set indicates a higher score on multimodal questions, i.e., SI set and MI set. For LLMs, there is an inverse relationship between the scores on the NI set and those on the SI and MI sets, as we require the model in prompt to tell if visual information is needed to solve the question. A stronger model are more likely to refuse to answer in that case, while a weaker model will simply guess one. A newer version of ChatGPT is also more likely to refuse as its capability has increased through longer time of RLHF (Reinforcement Learning from Human Feedback).

Model	SA	MA	MA Acc.	FB	OP
Puretext (LLM)					
MOSS	38.5	33.1	6.8	2.7	8.7
DFM-2.0	55.8	53.9	29.7	13.3	10.3
Gemini	**58.2**	52.7	22.8	29.1	7.9
ChatGPT(0613)	47.4	38.3	10.2	8.9	**11.0**
ChatGPT(1106)	40.0	39.4	17.9	10.5	7.7
GPT-4	51.3	**60.0**	**53.1**	**32.9**	6.8
Text+Image (MLLM)					
VisualGLM	37.9	30.2	1.9	0.7	3.6
VisCPM	41.7	27.7	0.0	3.8	**14.1**
Qwen-VL	49.8	29.4	2.8	5.8	13.7
Gemini Vision	59.4	54.4	24.3	30.5	12.5
GPT-4V	**67.1**	**70.6**	**58.2**	**42.4**	11.7

Table 4: Performance of models on each type of questions of MULTI. SA: multiple-choice question with single correct answer, MA: multiple-choice question with multiple correct answers, FB: fill-in-the-blank question, OP: open-ended writing question. MA Acc.: Accuracy of MA questions, i.e., the model selects all of the correct choices and none of the wrong choices.

We report the performance by the question type in Table 4. Most of the models exhibit their best performance on the common SA set, with a marginally lower score observed on MA set. Notably, for GPT-4/V, the performance on the MA set surpasses that of the SA set. The accuracy of GPT-4/V on the MA set significantly outperforms that of the Gemini models and others, suggesting a stronger propensity of GPT-4/V to select all correct choices rather than settling for a single answer.

On BF set, the score becomes extremely low with exact matching. This is partly attributed to the failure of one model to adhere to the given instructions, resulting in correct answers delivered in an incorrect format. Additionally, we observe a smaller variance in scores on OP set compared to FB set. This indicates that our dataset is less prone to leakage concerns and presents significant challenges in multimodal content generation.

We report the performance by the education levels and subjects in Table 5. A common pattern across models in terms of performance for each subject at the school level is not observed. Given that a large proportion of the questions in the driving test are judgmental questions (mostly SA with two choices), higher scores in this subject compared to others are expected. Additionally, the Administrative Aptitude Test (AAT) questions, which invariably include at least one image and often require the ability of pattern recognition, can be seen in Figure 7. We observe that the accuracy of locally deployed models is close to or below 25%, which is equivalent to the level of random choice. Those strong LMs perform exceptionally poorly, namingly DFM-2.0 and the text-only versions of GPT, as they reject the majority of AAT questions. Similarly, the multimodal versions of GPT and Gemini also exhibit low performance, highlighting the great challenging nature of MULTI.

Model	JuH	SeH	Uni	Driv	AAT
Puretext (LLM)					
MOSS	21.2	26.7	23.8	44.1	**25.5**
DFM-2.0	42.3	**42.5**	35.7	66.3	3.9
Gemini	47.7	42.3	41.4	66.9	22.5
ChatGPT(0613)	34.7	32.3	36.1	54.1	2.9
ChatGPT(1106)	31.6	23.7	34.9	52.1	1.3
GPT-4	**49.2**	33.7	**55.1**	**69.9**	0.9
Text+Image (MLLM)					
VisualGLM	22.2	25.6	23.6	40.9	24.9
VisCPM	25.2	28.1	23.0	43.4	23.7
Qwen-VL	32.6	32.9	27.2	49.3	26.4
Gemini Vision	48.2	45.2	41.7	67.4	**27.0**
GPT-4V	**58.5**	**52.9**	**59.0**	**80.1**	26.2

Table 5: Performance of models on each subject of MULTI. JuH: level of Junior High school, SeH: level of Senior High school, Uni: level of University, Driv: Chinese driving test, AAT: Administrative Aptitude Test.

4.5 Ablation Study on Image Information Gain

To verify that the images are essential in our dataset, we conduct an evaluation under the setting where a question with images is fed into the MLLM without the images. We adopt BLIP2 (Li et al., 2023d) to extract the image caption for each image, and we also use the Optical Character Recognition (OCR) tool EasyOCR[7] to extract characters in each image. We test the performance of substituting the images with either caption text or OCR content. The results are shown in Table 6.

For questions with exactly one image (as presented in the SI column), this image can provide essential information to answer the question. The gain is extremely large for GPT-4V and Gemini Vision, since they are likely to refuse to answer if images are not provided. For questions containing more than one image (as presented in the MI column), we observe a drop in the final scores when the images are present for several models. This is probably due to the fact that the MLLM fails to utilize the conversation history and remember all the images it has already seen.

The trend is the same across the models when the images are missing but the image information (captions or OCR content) is provided. While most of the models get an improvement with image information, a few are slightly worse when this information is provided. When comparing two types of image information, a model benefits more from a general caption than from several disordered OCR fragments.

The caption mainly focuses on the general figure and generates English content, introducing bilinguals to the models and omitting details. The OCR content provides detailed information but no spatial information, and it is not universal since some images do not have characters. Both types of image information help reduce

[7] https://pypi.org/project/easyocr/

Model	NI	SI				MI			
		w/o. image	w. caption	w. ocr	w. image	w/o. image	w. caption	w. ocr	w. image
Puretext (LLM)									
MOSS	36.1	27.3	27.3 (+0.0)	27.6 (+0.3)	-	**17.1**	**20.7** (+3.6)	**19.0** (+1.9)	-
ChatGPT(1106)	54.0	6.8	9.9 (+3.1)	6.6 (-0.2)	-	5.1	10.7 (+5.6)	5.5 (+0.4)	-
DFM-2.0	**63.0**	**28.7**	**30.2** (+1.5)	**33.4** (+4.7)	-	11.3	15.6 (+4.3)	14.9 (+3.6)	-
Text+Image (MLLM)									
VisualGLM	35.1	20.8	21.4 (+0.6)	20.4 (-0.4)	25.2 (+4.4)	15.3	15.1 (-0.2)	14.5 (-0.8)	9.7 (-5.6)
VisCPM	36.8	27.1	27.6 (+0.5)	27.2 (+0.1)	28.4 (+1.3)	24.8	21.6 (-3.2)	20.9 (-3.9)	16.6 (-8.2)
Qwen-VL	43.2	30.7	30.3 (-0.4)	31.0 (+0.3)	32.7 (+2.0)	**25.5**	**25.0** (-0.5)	**26.2** (+0.7)	20.7 (-4.8)
Gemini/Vision	62.5	**36.2**	**36.9** (+0.7)	**38.4** (+2.2)	40.0 (+3.8)	18.3	23.2 (+4.9)	18.6 (+0.3)	24.5 (+6.2)
GPT-4/V	**74.5**	11.3	9.7 (-1.6)	1.9 (-9.4)	**46.9** (+35.6)	8.8	9.4 (+0.6)	3.1 (-5.7)	**28.1** (+19.3)

Table 6: Performance of models evaluated on the image set of MULTI.

the refuse ratio to answer but may increase the difficulty of reasoning.

4.6 Evaluation on MULTI-ELITE

Model	Overall	SA	MA	MA Acc.	FB	NI	SI	MI
Puretext (LLM)								
MOSS	13.1	14.8	17.0	5.8	0.8	9.4	**14.7**	16.9
DFM-2.0	**18.0**	**19.8**	**21.2**	**9.6**	6.4	**26.9**	12.7	18.6
Gemini	10.5	6.3	16.6	3.9	**8.8**	6.8	9.0	**35.6**
ChatGPT(0613)	6.0	4.7	8.9	1.0	3.2	6.8	4.6	11.9
ChatGPT(1106)	4.7	5.3	4.2	1.9	4.0	8.5	2.2	6.8
GPT-4	5.8	3.8	7.3	5.8	8.0	7.3	2.7	22.0
Text+Image (MLLM)								
VisualGLM	12.8	**14.5**	16.6	0.0	0.8	**16.2**	11.7	6.8
VisCPM	13.0	10.4	22.0	0.0	0.8	10.3	14.2	15.3
Qwen-VL	10.5	7.2	19.3	1.9	0.8	8.5	10.8	16.9
Gemini Vision	12.4	5.3	21.2	5.8	**12.0**	6.8	12.0	**37.3**
GPT-4V	**14.0**	5.3	**25.5**	**15.4**	**12.0**	7.3	**14.9**	33.9

Table 7: Performance of models on MULTI-ELITE.

We report the evaluation results on MULTI-ELITE in Table 7, which consists of 500 questions. These questions were selected based on pre-annotated quality and difficulty scores, and the selection process aimed to maintain a distribution similar to that of MULTI, while also considering the evaluation results on MULTI in § 4.4. The highest two recorded scores on MULTI-ELITE were modest 18.0% achieved by DFM-2.0 and 14.0% achieved by GPT-4V. This result underscores the great challenge of the MULTI-ELITE and highlights a significant room for improvement on the extremely hard questions demanding detailed image comprehension and complex cross-modality reasoning.

4.7 Evaluation with MULTI-EXTEND

Model	window size	w/o. kn	w. kn
DFM-2.0	4,096 tokens	49.7	49.2(-0.5)
Qwen-VL	8,192 tokens	39.0	34.7(-4.3)
ChatGPT(1106)	16,385 tokens	35.9	38.0(+2.1)
Gemini	30,720 tokens	52.2	52.1(-0.1)

Table 8: Performance of models with MULTI-EXTEND.

We further evaluate the In-Context Learning (ICL) capabilities using the MULTI-EXTEND knowledge set, which incorporates relevant concepts of the problem along with some commonly used solutions. The prompt used for adding these knowledge pieces is shown in Figure 3, and the corresponding results are outlined in Table 8. A significant increase in scores is observed for all models upon the inclusion of knowledge background information. It is noteworthy that the average token count per question increase from 65 to 250, and eventually to 850, subsequent to the addition of prompts and the employment of MULTI-EXTEND. This increment poses a challenge for models, as the brief question may get submerged within an extensive context.

5 Conclusion

In this paper, we introduced MULTI, a comprehensive and challenging benchmark designed to rigorously evaluate the performance of multimodal large language models (MLLMs) in detailed cross-modality understanding and scientific reasoning. Our experiments with state-of-the-art models like Qwen-VL, VisCPM, Gemini, and GPT-4 demonstrate that while these models exhibit promising capabilities, there remains a significant gap compared to human performance, particularly in tasks involving cross-modal alignment, logical reasoning, and complex comprehension. This underscores the need for continued research and development in this domain.

The creation of the MULTI-ELITE and MULTI-EXTEND subsets further contributes to the field by providing insights into the strengths and limitations of current MLLMs. These subsets challenge the models' learning and reasoning abilities and encourage the development of more sophisticated and robust multimodal understanding systems.

Looking ahead, the MULTI benchmark opens new avenues for research, particularly in enhancing the MLLMs' ability to integrate and reason over diverse data types, including images, text, and structured data. Future work could focus on expanding the benchmark to include more diverse modalities and question types, further pushing the boundaries of what MLLMs can achieve. By making MULTI publicly available, we hope to foster a collaborative environment where researchers can continually test and improve the capabilities of MLLMs, driving the field toward the development of truly intelligent and versatile AI systems.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Zhi Chen, Jijia Bao, Lu Chen, Yuncong Liu, Da Ma, Bei Chen, Mengyue Wu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022. Dfm: Dialogue foundation model for universal large-scale dialogue-oriented task learning. *arXiv preprint arXiv:2205.12662*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, et al. 2023. Mllm-bench, evaluating multi-modal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023b. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023e. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv*, abs/2306.04387.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023f. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023a. Scieval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023b. Moss: Training conversational language models from synthetic data.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Limitations and Future Work

Multilingual Capabilities MULTI predominantly features simplified Chinese and mainly focuses on subjects taught in Chinese schools, with limited English multimodal content that's relatively straightforward for LLMs. We plan to include translations in future versions. Nonetheless, the presence of Chinese characters in figures poses a significant challenge for MLLMs trained on different linguistic datasets.

Use of Explanations While we annotated explanations in detail, their utilization in subsequent studies was limited. These explanations could potentially serve as valuable training data for model fine-tuning and few-shot learing using methods like CoT(Chain-of-Thoughts), aiding in evaluating reasoning skills.

Metrics for evaluating blank-filling and open-ended writing Our evaluation primarily uses exact match, which might be overly stringent for assessing MLLMs' true capabilities. Assessing open-ended writing tasks that require complex knowledge and reasoning is still a challenge. While only few studies (Wang et al., 2023) involve human evaluation, developing automatic and reliable methods remains an open research area.

Adaptation to various MLLMs Although we have tested several MLLMs, numerous others exist, and new ones are continuously emerging. We encourage the community to evaluate their MLLMs using our benchmark to gauge their cognitive reasoning abilities. We will test more models as soon as the multilingual version is released.

Expansion to more modalities and subjects Our benchmark currently focuses on static images, but incorporating other modalities like audio and video, and subjects like art, music theory, medicine, and sports could present new challenges. Thus, expanding our question set to cover these areas is a promising direction for future research.

B Statistics

We provide a detailed statistic in Table 9. One question may contain more than one scoring points as mentioned in § 4.3.

Statistics	Number	Points
Total Problems	17251	-
Total Questions	18430	-
Total Points	23320	-
Total Images	7658	-
Total Knowledge	4595	-
Multiple Choices[8]	16100(87.36%)	19904(85.35%)
- Single Answer	13963(75.76%)	13963(59.88%)
- Multiple Answers	2137(11.60%)	5941(25.48%)
Fill in the Blank	1432(7.77%)	2211(9.48%)
Open Ended Writing	798(4.33%)	1205(5.17%)
Others	100(0.54%)	-
Question with Images	7489(40.63%)	9042(38.77%)
- Single Image	7265(39.42%)	8767(37.59%)
- Choices within Image	1179(6.40%)	1181(5.06%)
- Multiple Images	224(1.22%)	275(1.18%)
Question with Explanations	10565(57.33%)	13186(56.54%)
Question with Knowledge	7987(43.34%)	11858(50.85%)

Table 9: The statistic overview of MULTI.

Data Distribution We provide the distribution of choices in multiple-choice questions as shown in Table 10. Each row corresponds to a different total number of options available in the questions. The columns represent the frequency of each specific choice option. The table showcases a well-balanced distribution of choices. Notably, the distribution reveals a higher frequency of questions with four choices and a single correct answer, indicating a common format in multiple-choice questions.

Type	# choices	# A	# B	# C	# D	# E,F,G...
SA	2	1819	1376	0	0	0
	3	272	287	262	0	0
	4	2193	2638	2708	2379	0
	5	0	2	7	9	0
MA	3-13	1467	1568	1510	1303	91
Total	2-13	5751	5871	4487	3691	91

Table 10: The choice distribution for multiple-choice questions.

C Prompts

The complete collection of prompts designed for evaluation on MULTI is shown in Figure 3. One of the prompt pieces in each row are selected according to the evaluation setting and data format. Please note that some prompt will not take effect, for instance, if the knowledge is not given.

D Data Selection Algorithm

We mostly pick questions based on its content length L_q, calculated with function

$$L_q = \left(a \times \begin{bmatrix} \mathcal{H}(L_{q,\#\text{characters in question}}) \\ \mathcal{H}(L_{q,\#\text{characters in answer}}) \\ \mathcal{H}(L_{q,\#\text{characters in analysis}}) \end{bmatrix} + b \times \begin{bmatrix} \mathcal{H}(L_{q,\#\text{images in question}}) \\ \mathcal{H}(L_{q,\#\text{images in answer}}) \\ \mathcal{H}(L_{q,\#\text{images in analysis}}) \end{bmatrix} \right)^\top \begin{bmatrix} 1.0 \\ 0.1 \\ 0.5 \end{bmatrix}$$

where $q = 1, b = 1$ are customized weights.

In the formula above, we use a harmonic mean function $\mathcal{H}$ to normalize content length $L_{q,i}$ of each target value i within each knowledge piece k.[9]

$$\mathcal{H}(L_{q,i}) = \frac{1}{\frac{1}{L_{q,i}} + \frac{1}{\overline{L_{q,i}}}} = \frac{2L_{q,i}\overline{L_{q,i}}}{L_{q,i}^2 + \overline{L_{q,i}}^2}$$

where $\overline{L_{q,i}}$ is the arithmetic average of $L_{q,i}$ for all questions with k.

[9]Note that for those questions without knowledge information, we simply use a "null" string as a keyword.

Figure 3: The prompts for evaluation on MULTI.

Then we pick N_k questions within each knowledge piece k.

$$N_k = \lceil \alpha \times \lg(\#\text{questions of } k) \rceil$$

where $\alpha = 3$ is a customized parameter.

Now we sort $L_{q,k} = L_q : q \in k$ in descendent order.

Then we assign a pick-up probability to select these questions

$$Pr[\text{pick up } q] = \begin{cases} 1 \text{ , for } q \text{ s.t. } L_{q,k}[0] \\ p \text{ , if } q = 1 \text{ , for } q \text{ of } L_{q,k}[1:m] \\ \qquad\qquad\qquad\quad \text{or } L_{q,k}[-m:] \\ p\frac{N_k - 2m}{\#\text{questions of } k} \text{ , otherwise} \end{cases}$$

E Annotation Platform

Here we show our online platform for annotation. The platform consists of text boxes for editing contents and regions for rendering the text to see the final appearance of the data. Details can be found in Figure 4.

F More Examples

In Figure 5 and Figure 6, we show several examples of complex formation and modification during data annotation and post-process stage. In Figure 7, we show more examples for annotated questions. English translations of Chinese text are shown for better readability, except Figure 5 for better format clarity. The markdown, LaTeXand html format codes are remained as it is.

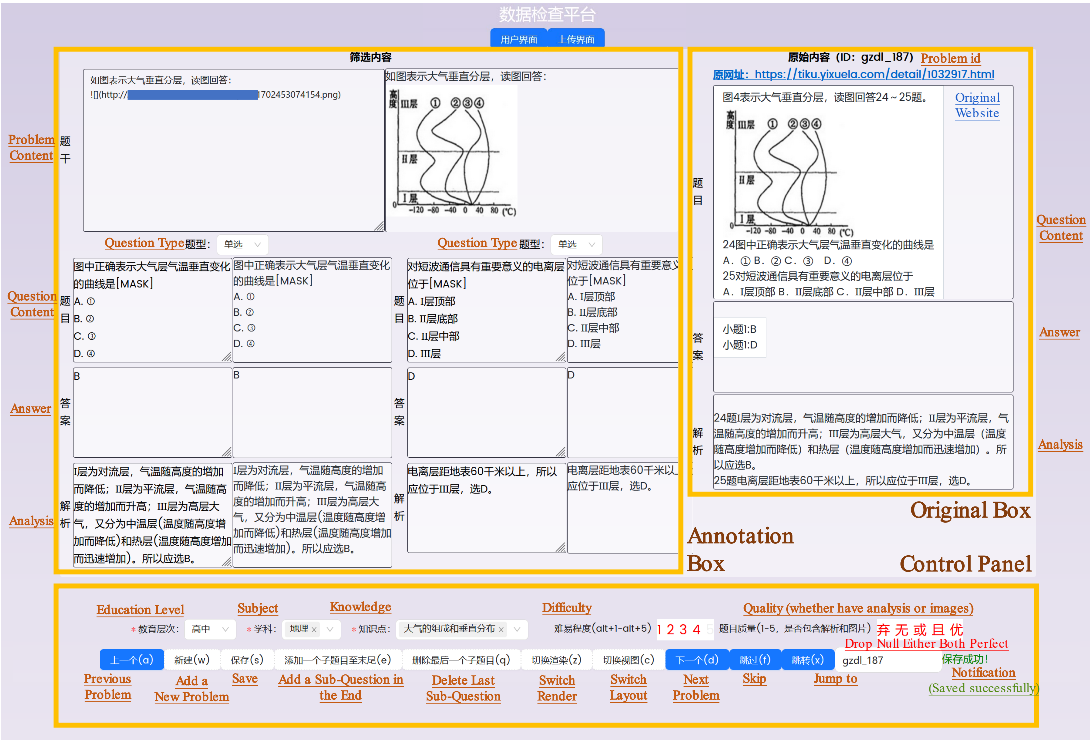

Figure 4: A screenshot for the main page of the data annotation platform.

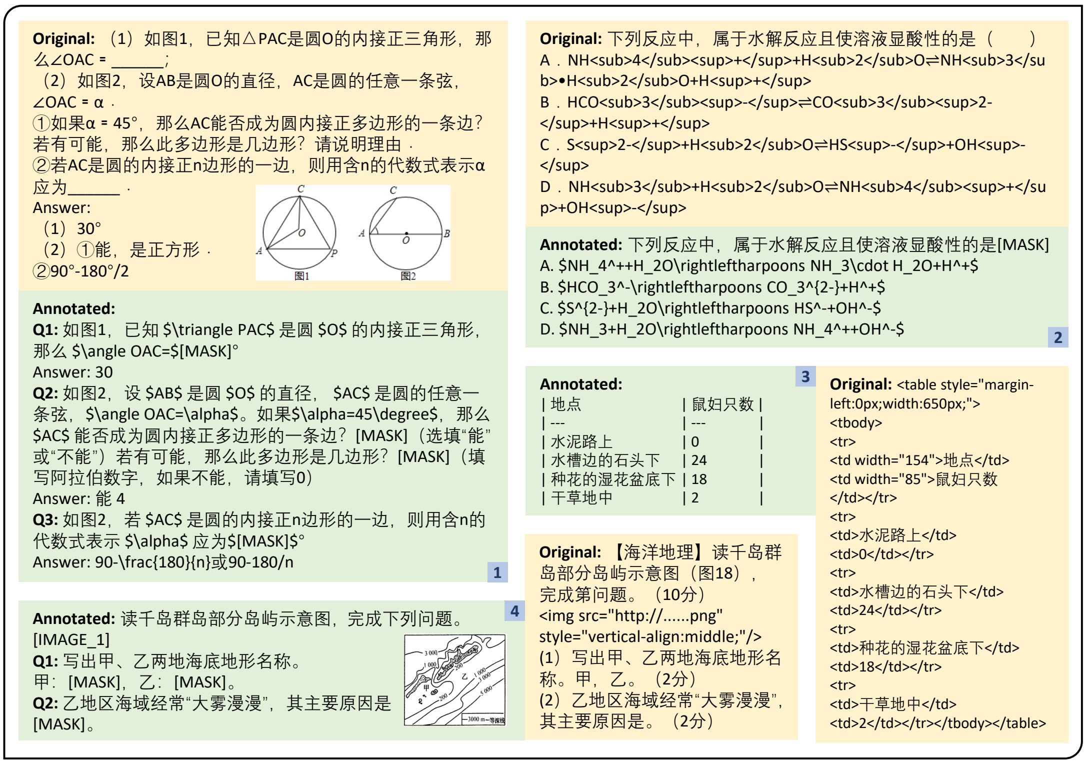

Figure 5: Several data annotation examples when constructing MULTI.

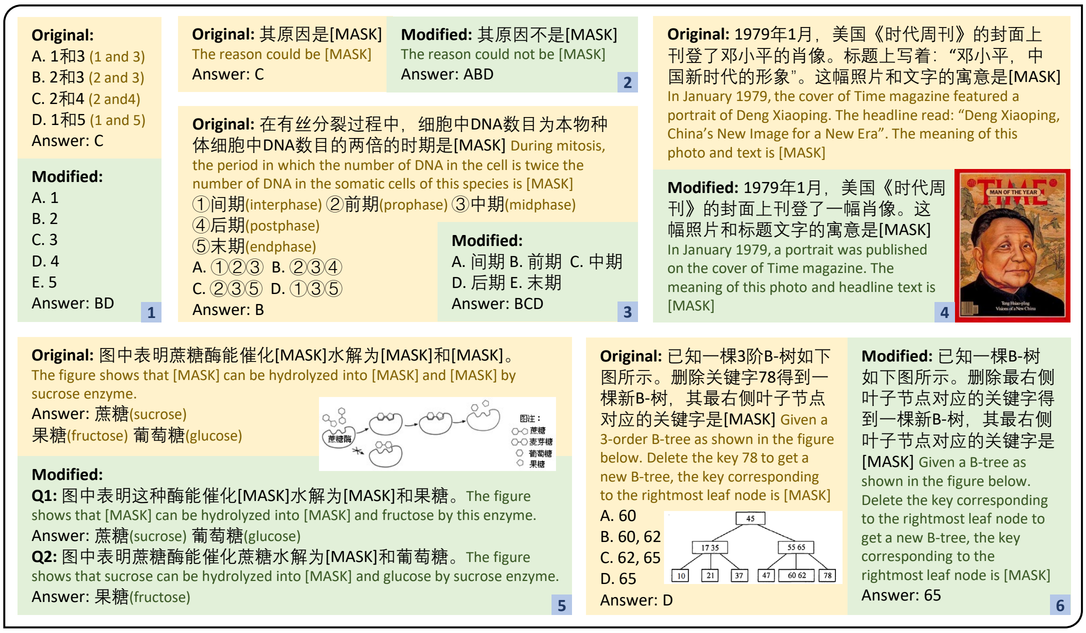

Figure 6: Several data augmentation examples when constructing MULTI.

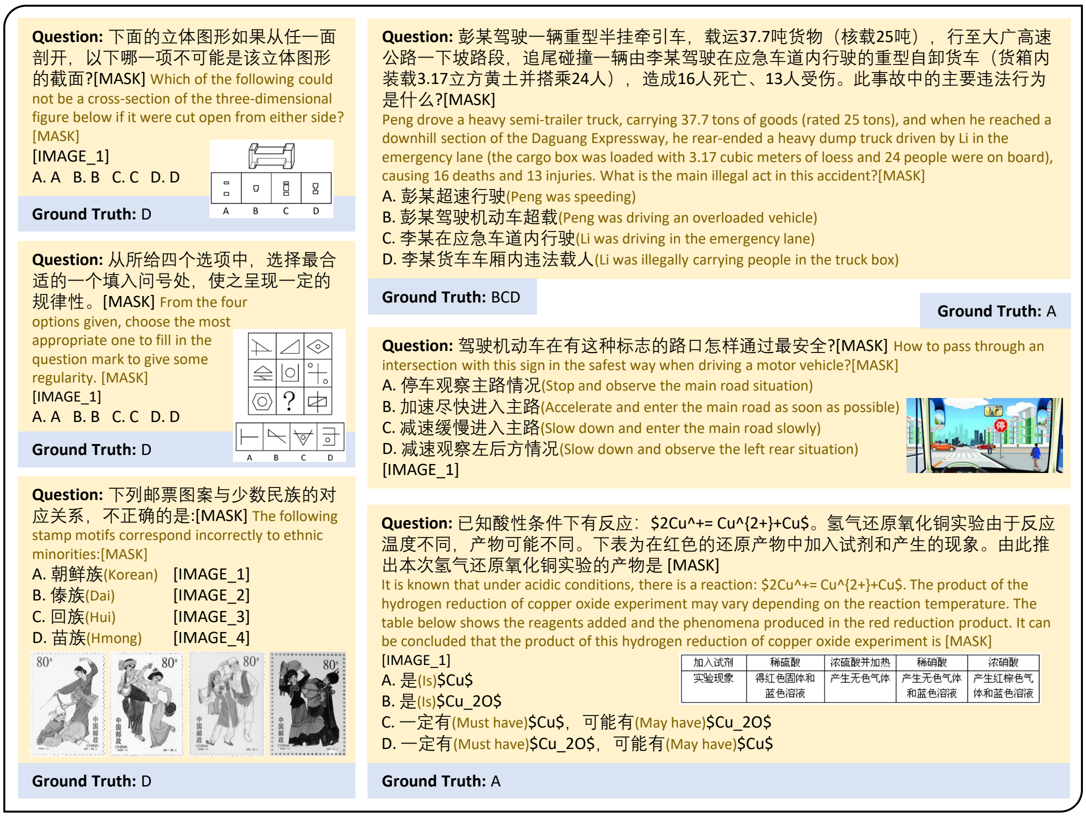

Figure 7: More example of MULTI.