MULTI: Multimodal Understanding Leaderboard with Text and Images

Anonymous ACL submission

Abstract

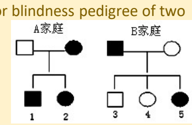
Rapid progress in multimodal large language models (MLLMs) highlights the need to introduce challenging yet realistic benchmarks to the academic community, while existing benchmarks primarily focus on understanding simple natural images and short context. In this paper, we present MULTI, as a cutting-edge benchmark for evaluating MLLMs on understanding complex tables and images, and reasoning with long context. MULTI provides multimodal inputs and requires responses that are either precise or open-ended, reflecting real-life examination styles. MULTI includes over 18,000 questions, and challenges MLLMs with a variety of tasks, ranging from formula derivation to image detail analysis and cross-modality reasoning. We also introduce MULTI-ELITE, a 500-question selected hard subset, and MULTI-EXTEND, with more than 4,500 external knowledge context 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 28.5% and 55.3%. MULTI serves not only as a robust evaluation platform but also paves the way for the development of expert-level AI.

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. They play a crucial role in conveying information, particularly in scientific areas. Therefore, 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., 2023b; 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 may not accurately gauge a model's comprehension and can lead to superficial learning, i.e., the model will not look into other choices if the correct choice is straightforward. A more robust, detailed, and multi-scale dataset is necessary to effectively evaluate MLLMs under diverse conditions and scenarios. Current benchmarks mentioned above are evaluated with English context, while the rapid progression of Chinese MLLMs highlights the need for a Chinese multimodal benchmark with Chinese contents both in text and image and brings new

Question: 下图是A、B两个家庭的色盲遗传系谱图，这两个家庭由于某种原因调换了一个孩子，则调换的两个孩子是[MASK]The following figure shows the color blindness pedigree of two families A and B. For some reason, the two families exchanged a child. The two children who were exchanged are [MASK]
[IMAGE_1]
A. 1 B. 2 C. 3 D. 4 E. 5



Knowledge: 人类红绿色盲症 (human red-green color blindness)
红绿色盲的遗传特点： Genetic characteristics of red-green color blindness:
……
3、女患其父子必患。If a woman is affected, her father and son must be affected too.
红绿色盲的遗传方式： The modes of inheritance of red-green color blindness:
1、 正常女性与色盲男性的婚配图解: 男性的色盲基因只能传给女儿，不能传给儿子。The mating diagram of normal female and color blind male: the male's color blind gene can only be passed to his daughter, not to his son.
……

Ground Truth: BD

Explanation: 色盲属于半X隐性遗传病，其遗传规律是"母病子必病，女病父必病"。由于色盲是伴X隐性遗传病，分析家庭A可知，该家庭的父亲正常，其女儿也应该是正常的，图中显示其女儿患有色盲，因此该女孩不是A家庭中的孩子；B家庭中父亲患病，则女儿可能患病也可能不患病，由于题干信息告诉我们这两个家庭由于某种原因调换了一个孩子，那么肯定是A家庭的2和B家庭的4发生了调换。故选。Color blindness is a sex-linked recessive genetic disease, and its inheritance rule is "mother sick son must be sick, daughter sick father must be sick"……Therefore, choose BD.

Problem Type: 多选 (multiple-choice with multiple answers)

Education: 高中 (senior High)	**Subject:** 生物 (biology)
Difficulty: 5	**Quality:** 5

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

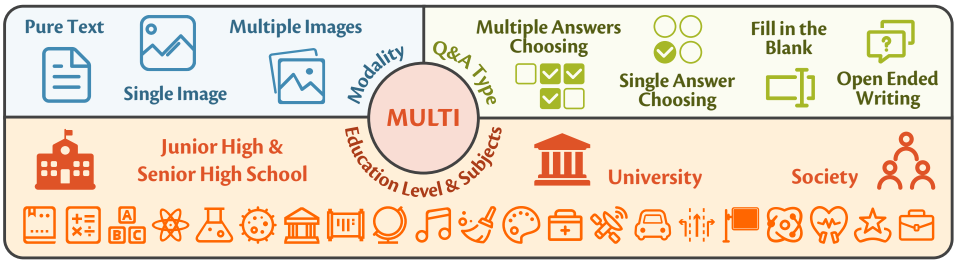


Figure 2: The overview of MULTI.

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 cover a variety of scientific disciplines, including mathematics, physics, computer science, etc., and also pose significant challenges to intricate image reasoning. MULTI serves as the first benchmark incorporating driving tests and administrative aptitude tests in China. The questions are crafted to test understanding and generation in various formats and complexity levels and are categorized into multiple-choice (with single or multiple correct answers), fill-in-the-blank, and open-ended questions.

To further challenge multimodal LLMs, we develop two subsets within MULTI: MULTI-ELITE consists of 500 carefully selected tough questions aiming to probe the limits of these models, and MULTI-EXTEND featuring 4,596 knowledge pieces tests the models' capabilities of learning 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, and more are presented in Appendix G.

We conduct comprehensive experiments on MULTI using leading-edge multimodal and single-modality LLMs. 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 the benchmark is challenging for current models, not to mention the MULTI-ELITE set where GPT-4V only gets a 14.0% score, and most of the other models get a score near random, indicating a large space for improvement.

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

- 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.

2 Related Works

Multimodal Large Language Models (MLLMs). With advancements in aligning features across multiple modalities, like CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021), recent studies have explored projecting vision features into the latent space of LLMs, aiming to enhance their capabilities of comprehending visual information. For example, BLIP-2 (Li et al., 2023c) pioneers this approach by employing Q-Former to translate image features into text representations. Following this, LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), and InstructBLIP (Dai et al., 2023) have introduced visual instruction tuning to bolster the capability of MLLMs of following instructions. Our primary focus is on the proficiency of MLLMs in comprehending instructions in Chinese, which are divided into two main branches: open-source models, which typically build upon existing Chinese LLMs or are fine-tuned on Chinese instruction datasets, examples of which include Chinese-LLaVA (LinkSoul-AI, 2023), VisualGLM (Du et al., 2022), VisCPM (Hu et al., 2023), Qwen-VL (Bai et al., 2023a), InternVL (Zhang et al., 2023a), Yi-VL (01.ai, 2023); and closed-source models, which are often highly powerful, multi-lingual systems such as GPT-4V(ision) (OpenAI, 2023b) and Gemini (Team, 2023). In this paper, we intend to evaluate these models across a range of scientific fields on the MULTI benchmark, offering an extensive assessment and guidance for the onward trajectory of Chinese MLLMs.

Benchmarks for MLLMs. 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

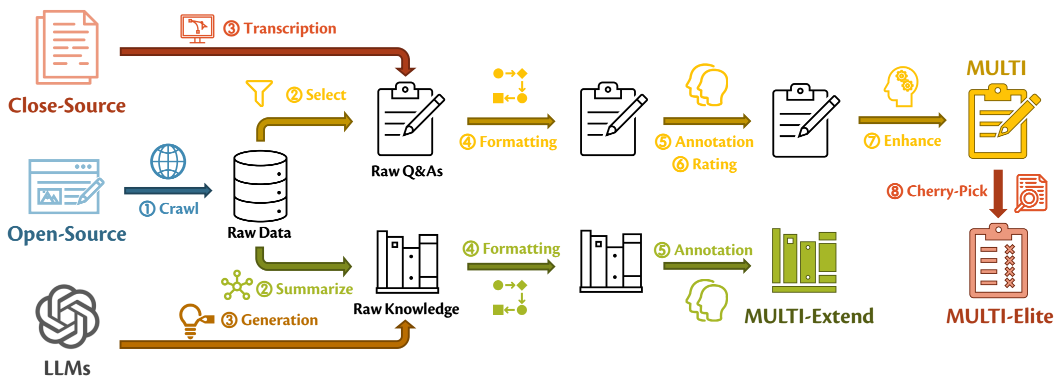


Figure 3: The construction pipeline of MULTI.

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 MLLMs on dealing with various tasks in real world. Benchmarks like LLaVA-Bench (Liu et al., 2023b), MMBench (Liu et al., 2023c), MM-VET (Yu et al., 2023), TouchStone (Bai et al., 2023b), MLLM-bench (Ge et al., 2023), and SEED-Bench(Li et al., 2023b,a), for instance, leverage GPT to evaluate the relevance and helpfulness of human-like long responses in the reality. POPE (Li et al., 2023d) 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 example, 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

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.

3.1 Data Construction Process

The data construction pipeline is shown in Figure 3. To develop MULTI, we follow several key steps to ensure high-quality and precise annotation. Firstly, we crawl open-source raw question data from the Internet and transcript close-source exams from paper documents. Secondly, we format each question and knowledge piece into markdown and LaTeX formula format to maintain precision and quality. Thirdly, we revise and refine each question multiple times to prevent data leakage and increase difficulty. Lastly, We rate every question based on its difficulty and content richness.

Data Source We collect 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 E. We also collect questions from internal exams and practices of several top universities. After the selection, we obtain over 18K questions as the raw data.

Data Process and Annotation The data process and annotation for our dataset involve a comprehensive series of steps to ensure high-quality, diverse content.

In the **Data Pre-process** stage, raw data with formats like HTML, photocopy, hand script, or plain text are refined by removing irrelevant HTML tags, converting text styles into markdown format, and transcribing math functions and chemical structures into LaTeX format, with complex tables saved as screenshot images after HTML rendering. OCR tools are utilized for text conversion from photocopies and hand scripts.

During the **Data Annotation** stage, an online platform facilitates annotators, mostly skilled undergraduates (involved in the work as authors), in tasks across format, content, labelling, and semantic levels. This includes converting content into markdown and LaTeX, splitting sub-questions into individual ones, evaluating the difficulty and quality, and correcting errors for factual accuracy.

The **Data Post-process** stage employs strategies like formation, disambiguration, distillation, and transformation to enhance question difficulty and diversity, including modifying question formats and reducing assistance information.

Throughout these stages, we process 2.7 million questions in total and pick out 18,430, incorporating 23,320 scoring points, 7,658 images, and 4,595 knowledge pieces. MULTI highlights a broad diversity in question types, including multiple-choice questions with both single and multiple answers, along with fill-in-the-blank and open-ended writing questions enriching the testing scenarios. [1] The stages during data processing

[1]For the sake of simplifying writing, in the following

and annotation significantly increase the diversity and difficulty of the dataset. For details of data processing and annotation, please refer to Appendix F.

3.2 The MULTI-ELITE Set

We select 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 are also referred to in the selection, where the results of GPT-4V(OpenAI, 2023b) are given the most consideration.

3.3 The MULTI-EXTEND Background Knowledge Dataset

External knowledge is crucial to provide critical information that assists 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 outer knowledge source (e.g. New Bing[2] and Wikipedia[3]). 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, 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.

3.4 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. Furthermore, the questions without images also test the MLLMs' ability on dealing with plain text information. 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 recently updated questions, which significantly enhances its 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.

4 Experiments

4.1 Models

We evaluate a wide range of MLLMs that support Chinese, including Chinese-LLaVA (LinkSoul-AI, 2023), Qwen-VL (Bai et al., 2023a), VisCPM (Hu et al., 2023), VisualGLM (Du et al., 2022), InternVL (Chen et al., 2023), Yi-VL (01.ai, 2023), 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), and the performance of these models on questions with images will reflect their abilities on finding loss of information. Model specifications are listed in Table 10. 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, and use FP16 or INT4 quantization to accelerate inference if officially provided. We 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 shown in Figure 4. The prompts are designed carefully according to the features of each type of question and the answer patterns expected. We also modify the input format to fit into official inference guidelines. The complete collection of prompts are presented in Appendix D.

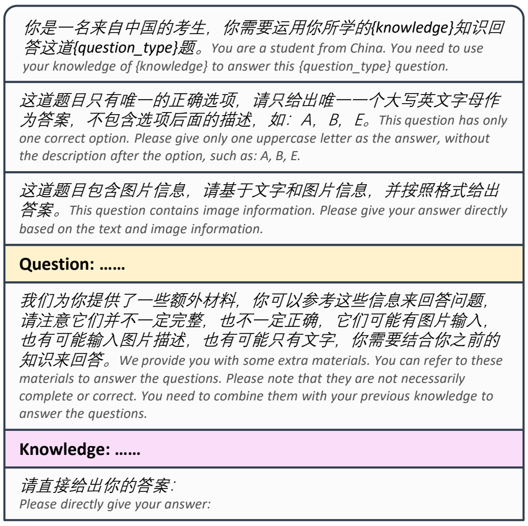


Figure 4: An example of the prompts used when evaluating a multiple-choice question with image context, knowledge piece and single correct answer.

paragraphs we may use abbreviations. We denote multiple choices questions with a single answer as **SA** or *Single Answer Choosing* and those with multiple answers as **MA** or *Multi Answer Choosing*. We use **FB** for fill-in-the-blank questions and **OP** for open-ended writing questions.

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

[3] https://wikipedia.org

Benchmark	Lang	Size					Image			Answer Type				Source
		Sub	Q	Ana	Img	Kn	NI	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 langs	4	12K	-	3.1K	-	✓	✓	✗	✓	✗	✗	✗	Exams
AGIEval (Zhong et al., 2023)	zh, en	20	8K	a few	-	-	✓	✗	✗	✓	✓	✓	✗	Exams
MMBench (Liu et al., 2023c)	en	20	3K	-	3K	-	✗	✓	✗	✓	✗	✗	✗	Web, Repurposed
SEED-Bench (Li et al., 2023b)	en	12	19K	-	19K+	-	✗	✓	✓	✓	✗	✗	✗	Anno.
SEED-Bench-2 (Li et al., 2023a)	en	27	24K	-	22K+	-	✗	✓	✓	✓	✗	✗	✗	Anno.
MLLM-Bench (Ge et al., 2023)	en	42	0.4K	-	0.4K	-	✗	✓	✗	✗	✗	✗	✓	Anno.
Touchstone (Bai et al., 2023b)	en	27	0.9K	-	0.9K	-	✗	✓	✓	✗	✗	✗	✓	Anno.
C-Eval (Huang et al., 2023)	zh	52	14K	a few	-	-	✗	✗	✗	✓	✗	✗	✗	Exams, Web
SciEval (Sun et al., 2023a)	en	3	18K	-	-	-	✓	✗	✗	✓	✗	✗	✗	Web, Repurposed
MMMU (Yue et al., 2023)	en	30	12K	2K	11K+	-	✗	✓	✓	✓	✗	✗	✗	Anno., Web, Textbooks
MULTI (ours)	zh	23	18K	10K+	7.7K	4.6K	✓	✓	✓	✓	✓	✓	✓	Anno., Exams, Web

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. NI: the question with pure text, 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). Anno.: Annotation

Image MULTI includes 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 the text as input. For some models like VisCPM, InternVL and Yi-VL which compulsorily demand an image in each turn, we feed the model a blank image with color set to RGB(0,0,0) along with plain text in evaluation. For efficiency, results of GPT-4 and Gemini on pure-text questions are directly used as the results of GPT-4V and Gemini Vision respectively.

For questions with multiple images, as the positions of images matter a lot, e.g., a multiple-choice question where each choice is an image, special patterns with `[IMAGE_{index}]` are used 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 VisCPM and VisualGLM support only one image as input in one turn. We adopt the strategy of splitting the content into multiple segments divided by each image and feeding them into the MLLM sequentially as rounds of conversation, where the MLLM receives each segment along with the corresponding image. 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 5. As the released versions of Chinese-LLaVA, InternVL and Yi-VL do not support multiple images as input, currently only the first image is used for evaluating each question.

4.3 Metrics

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

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

Multiple-choice with Multiple Answers (MA) We define the total points of an MA question as the number of correct choices, and each correct choice selected is 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, where only correctly giving all the choices without wrong ones will be granted points. [4]

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, but we use a 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 2, 3,

[4] For example, a question with correct answer `ACE` worth 3 points, and answer `AC` will be granted 2 points and answer `BC` or `ABCE` will be granted 0 points. However, on calculating accuracy none will be counted, and only `ACE` will be calculated as correct

and 4.

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	35.9	54.0	6.8	5.1
GPT-4	50.2	**74.5**	11.3	8.8
Text+Image (MLLM)				
Chinese-LLaVA	28.5	32.3	22.6	17.8
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
InternVL	44.9	50.9	35.5	25.1
Yi-VL	55.3	63.8	42.0	24.5
Gemini Vision	53.7	62.5	40.0	24.5
GPT-4V	**63.7**	**74.5**	**46.9**	**28.1**

Table 2: 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.

Overall comparison. We report the overall performance in Table 2. **The most powerful competitor, GPT-4V, achieves a mere 63.7% score**, underscoring the benchmark's complexity and challenge. Yi-VL outperforms other open-source models, but there still remains a notable gap with GPT-4V, and those smaller models do not get as much as half of the scores.

Comparison by number of images. In Table 2, we also present the performance categorized by image number. For MLLMs, a higher score on the Non-Image (NI) set suggests improved performance on multimodal questions, including the Single Image (SI) set and Multiple Image (MI) set. **It is evident that questions requiring more images are more challenging.** A significant drop in performance is observed when answering questions with more than one image. Only GPT-4V (28.1%) manages to exceed the average baseline set by random guessing.

Conversely, for LLMs, there exists a reverse correlation between scores on the NI set and those on the SI and MI sets. This is because we prompt the model to determine whether visual information is necessary for answering a question and if so the model needs to refuse to answer. Less capable models may simply make a guess, but more sophisticated models tend more to withhold an answer, resulting in lower but more reliable overall scores. The results on SI and MI sets for LLMs indicate a long way before mitigating hallucination.

Comparison by question type. In Table 3, we present the performance categorized by question type. **A majority of the models achieve their highest scores on the Single Answer Choosing (SA) set, with a lower performance on the Multiple Answers Choosing (MA) set**. A notable discrepancy is observed between the scores for the MA set and its accuracy, highlighting the smaller models' inability to identify all correct options accurately.

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	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)					
Chinese-LLaVA	34.5	26.9	3.9	2.4	8.4
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
InternVL	56.4	33.4	2.1	14.2	13.1
Yi-VL	61.3	42.0	36.4	14.6	8.9
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 3: Performance of models on each type of questions of MULTI. MA Acc.: Accuracy of MA questions.

For the Fill-in-the-Blank (FB) set, which requires short but exact matches, **the scores further decline**. This is partially due to failure to follow the specified instructions, often leading to correct responses being presented in an unacceptable format.

Furthermore, **we note significantly lower scores on the Open-ended Writing (OP) set in comparison to the FB set**. VisCPM stands out but only with the best score of 14.1% on the OP set, suggesting that our dataset minimizes the risk of data leakage and poses considerable challenges for models in generation across modalities.

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	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)					
Chinese-LLaVA	21.1	25.4	20.7	35.8	21.8
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
InternVL	39.3	36.5	30.6	57.7	24.8
Yi-VL	46.6	46.0	45.4	71.1	26.5
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 4: 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.

Comparison by education level and subjects. In Table 4, we present the performance categorized by educational levels and subjects. **The performance**

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	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)									
Chinese-LLaVA	32.3	26.1	26.3 (+0.2)	25.5 (-0.6)	22.6 (-3.5)	17.6	19.9 (+2.3)	19.6 (+2.0)	17.8 (+0.2)
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)
InternVL	50.9	33.4	33.3 (-0.1)	33.1 (-0.3)	35.5 (+2.1)	24.8	21.9 (-2.9)	22.9 (-1.9)	25.1 (+0.3)
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)
Yi-VL	63.8	**39.9**	**38.7** (-1.2)	**39.4** (-0.5)	42.0 (+2.1)	24.1	**26.5** (+2.4)	24.2(+0.1)	24.5 (+0.4)
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)
average			+0.30	-0.35	+5.98		+1.54	-0.30	+0.98

Table 5: Performance of models evaluated on the image set of MULTI.

trends for high school and university level questions remain consistent with the overall results observed. For questions at the society level, we anticipate higher scores on the Driving Test. This may be caused by a larger percentage of judgmental questions (in the format of SA with two options), as well as its nature with knowledge of regulations.

Furthermore, questions from the Administrative Aptitude Test (AAT), which typically include at least one image and often examine skills on image pattern recognition (illustrated in the first two examples in the left column of Figure 9), tend to have scores around or below randomly choosing baseline. **Even the strongest competitor, GPT-4V, shows limited success, with a performance of only 27.0% on these questions** as detailed in the study cited in the paper (OpenAI, 2023b). This underscores the significant challenge posed by multimodal questions. Notably, the stronger LLMs, specifically DFM-2.0 and the text-only versions of GPT, perform poorly on AAT questions as expected, as they often reject answering the majority of them.

4.5 Ablation Study on Image Information Gain

To assess the necessity of images in MULTI for solving problems, we conduct an ablation study where we either remove images from the SI and MI sets or substitute them with textual descriptions, such as captions and OCR-derived text. We utilize BLIP2 (Li et al., 2023c) for generating image captions and EasyOCR[5] to extract text from images. The results are shown in Table 5.

For questions that incorporate a single image (as indicated in the SI column), the presence of images significantly aids in answering the questions, with an average performance boost of 5.98%. Notably, GPT-4V experiences a substantial increase of 35.6% in performance, primarily due to its tendency to abstain from answering in the absence of images.

In settings where images are omitted and replaced by their textual descriptions (captions or OCR text), there's a marginal improvement of 0.30% observed with captions, but a minor reduction of -0.35% with OCR text. Captions, which generally summarize the images, introduce bilingual elements to the models and usually miss details. OCR text, while detailed, lacks spatial information and is not universally applicable, as some images contain no text at all. Both forms of textual information lower the models' refusal rate, and LLMs benefit more from these image information than MLLMs. However, they potentially complicate reasoning processes. Nevertheless, a generic caption is found to be more beneficial than scattered OCR fragments.

For questions that involve multiple images (as discussed in the MI column), we categorize models into three groups: 1) Close-source models, specifically GPT-4V and Gemini Pro, which leverage all images and achieve significant improvement. 2) Open-source models capable of handling multiple images within a dialogue or at a time, namely VisualGLM, VisCPM, and Qwen-VL, all of which exhibit a notable performance decline. 3) Open-source models without multi-image support, like Chinese-LLaVA, InternVL, and Yi-VL, show slight improvements. The second group's decline could be attributed to their inability to utilize conversation history effectively and remember previously seen images. The third group's limitation likely stems from providing only the first image, insufficient for comprehending all necessary information to answer the question, but to some degree avoiding distraction.

4.6 Evaluation on MULTI-ELITE

We conduct evaluations on MULTI-ELITE, as outlined in Table 6, which includes 500 specifically chosen questions. These questions are selected based on pre-annotated quality and difficulty scores, in addition to the evaluation results on MULTI discussed in § 4.4. The selection aims to ensure a distribution that mirrors MULTI's but also bring challenge to strong MLLMs. Yi-VL achieves the highest score on MULTI-ELITE with 26.2%, while scores for other models vary between 10.5% and 20.7%. This highlights the substantial challenge presented by MULTI-ELITE, indicating significant potential for improvement in tackling extremely difficult questions that require in-depth image understanding and

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

Model	Overall	SA	MA	MA Acc.	FB	NI	SI	MI
Chinese-LLaVA	12.3	15.7	13.1	1.0	1.6	13.7	11.0	15.3
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
InternVL	20.7	24.8	23.2	0.0	4.8	17.9	21.0	28.8
Yi-VL	**26.2**	**33.0**	**29.0**	8.7	3.2	**32.5**	**22.7**	25.4
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 6: Performance of models on MULTI-ELITE.

intricate reasoning across modalities. It is important to highlight the accuracy of multiple answers choosing (MA Acc.) as the most demanding task for MLLMs, necessitating a thorough grasp of the relationships between the choices and the questions, and reflecting model reliability of selecting all answers correctly.

4.7 Evaluation on MULTI-ELITE with MULTI-EXTEND

Model	window size	w/o. kn	w. kn
InternVL	768 tokens	20.7	19.9 (-0.8)
Yi-VL	4,096 tokens	26.2	21.4 (-4.8)
Qwen-VL	8,192 tokens	10.5	13.0 (+2.5)
Gemini Vision	30,720 tokens	12.4	17.0 (+4.6)
GPT-4V	128,000 tokens	14.0	21.3 (+7.3)

Table 7: Performance of models with MULTI-EXTEND on MULTI-ELITE.

The significant challenges posed by MULTI-ELITE prompt further investigation into the In-Context Learning (ICL) capabilities of MLLMs through the utilization of the MULTI-EXTEND knowledge set. This set is designed to include relevant concepts and frequently utilized solutions related to the problems. The study is conducted on several MLLMs, with the prompts for incorporating these knowledge pieces shown in Figure 5, and the results are listed in Table 7. Notably, the average number of tokens per question escalates from 65 to 250, and further to 850, following the integration of prompts and the adoption of MULTI-EXTEND, with the most extensive examples surpassing 10,000 tokens. MULTI-EXTEND poses a significant challenge in terms of the necessary window size and the capacity to handle lengthy contexts. It is observed that models equipped with larger window sizes, i.e. Gemini Vision and GPT-4V, benefit more from MULTI-EXTEND, whereas there is a notable decline in performance for MLLMs with smaller window sizes. The increase in tokens may also presents a hurdle for models, as the concise question may become overshadowed by the extensive context.

4.8 Takeaways

- GPT-4V demonstrates the highest performance with a 63.7% score, indicating significant challenge of MULTI, while Yi-VL leads among open-source models.
- MLLMs show a performance drop in questions requiring more images, with only GPT-4V exceeding a basic guessing baseline in multi-image scenarios.
- LLMs show a reverse correlation in performance between non-image and single/multiple image sets, highlighting the challenge of avoiding hallucination in visual questions.
- Models generally perform better on questions requiring shorter answers, i.e. SA > MA > FB > OP. The results of MA Acc. emphasize the importance of balancing recall and precision.
- Performance trends are consistent across educational levels, with lower scores on AAT questions due to their multimodal complexity.
- The inclusion of images significantly boosts question-answering performance, with captions offering a slight improvement and OCR text potentially complicating reasoning processes.
- In the MULTI-ELITE evaluation, Yi-VL achieves the highest 26.2% score, illustrating the difficulty of MULTI-ELITE and the need for advanced image understanding and reasoning across modalities.
- The aid of MULTI-EXTEND help improve performance on models with long window sizes, yet it may yield adverse effects on less capable models.

5 Conclusion

In this paper, we introduce MULTI, a comprehensive and challenging benchmark designed to rigorously evaluate the performance of MLLMs in detailed cross-modality understanding and scientific reasoning. Our experiments with state-of-the-art models like Qwen-VL, InternVL, Yi-VL, 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 continuous 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.

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 may 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 continuously test and improve the capabilities of MLLMs, driving the field toward the development of truly intelligent and versatile AI systems.

References

01.ai. 2023. Yi-vl. https://github.com/01-ai/Yi.

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*.

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

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*.

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.

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

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

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*.

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. 2023d. 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.

LinkSoul-AI. 2023. Chinese llava. https://github.com/LinkSoul-AI/Chinese-LLaVA.

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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. 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. 2023c. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

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*.

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 have annotated explanations in detail, the utilization in subsequent studies remains limited. These explanations could potentially serve as valuable training data for model fine-tuning and few-shot learning using methods like CoT (Chain-of-Thoughts) or RAG (Retrieval Augmented Generation) and may aid in evaluating reasoning skills.

Metrics for evaluating blank-filling, open-ended writing and others 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. We also have 100 questions that do not belong to traditional categories, such as questions requiring geographic drawing, and the evaluation on them will be even more challenging. Now that 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 topics. Thus, expanding our question set to cover these areas is a promising direction for future research.

B Statistics

We providedetailed statistics in Table 8. One question may contain more than one scoring points as mentioned in § 4.3.

B.1 Data Distribution on Question Types

Our benchmark showcases a remarkable diversity in the choice setting 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 unique correct option to those with multiple correct choices. We provide the distribution of choices in multiple-choice questions as shown in Table 9. 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.

Statistics	Number	Points
Total Problems	17251	-
Total Questions	18430	-
Total Points	23320	-
Total Images	7658	-
Total Knowledge	4595	-
Multiple Choices[6]	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	9048(49.09%)	12919(55.40%)

Table 8: The statistic overview of MULTI.

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 9: The choice distribution for multiple-choice questions.

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 scenarios.

C Models

The model specifications are listed in Table 10.

D Prompts

The complete collection of prompts designed for evaluation on MULTI is shown in Figure 5. One of the

Creator	Model	# Paras	Form	Modality	Lang	Version
FDU	MOSS (Sun et al., 2023b)	16B	Weight	T	zh, en	`moss-moon-003-sft`
SJTU&AISpeech	DFM-2.0 (Chen et al., 2022)	70B	Weight	T	zh, en	`dfm-2.0-70b-preview`
LinkSoul-AI	Chinese-LLaVA (LinkSoul-AI, 2023)	7B	Weight	One	zh, en	`Chinese-LLaVA-Cllama2`
THU	VisualGLM (Du et al., 2022)	6B	Weight	SI	zh, en	`visualglm-6b`
ModelBest	VisCPM (Hu et al., 2023)	10B	Weight	SI	zh, en	`VisCPM-Chat`
Alibaba	Qwen-VL (Bai et al., 2023a)	7B	Weight	MI	zh, en	`Qwen-VL-Chat`
OpenGVLab	InternVL (Chen et al., 2023)	19B	Weight	One	zh, en	`InternVL-Chat-Chinese-V1.1`
01-ai	Yi-VL (01.ai, 2023)	34B	Weight	One	zh, en	`Yi-34B-Chat`
Google	Gemini (Team, 2023)	-	API	T	ML	`gemini-pro`
Google	Gemini Vision (Team, 2023)	-	API	MI	ML	`gemini-pro-vision`
OpenAI	ChatGPT (OpenAI, 2022)	-	API	T	ML	`gpt-3.5-turbo-1106`
OpenAI	GPT-4 (OpenAI, 2023a)	-	API	T	ML	`gpt-4-1106-preview`
OpenAI	GPT-4V (OpenAI, 2023b)	-	API	MI	ML	`gpt-4-vision-preview`

Table 10: 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. W: accessible through weight. T: pure text LLM, One: only one image in the beginning, SI: single image in each turn, MI: multiple images in one turn. The underline means the model must have an image as input. ML: Multi-lingual.

prompt pieces in each row are selected according to the evaluation setting and data format. Please note that some prompt will not take effect under certain cases, for instance, the prompt related to knowledge will be ommitted if the knowledge is not given.

E 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.[7]

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

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

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

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

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

F Data Process and Annotation

Initially, we extract 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 conduct 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 filter out and modify questions based on predefined criteria. The second round of data annotation focus more on semantic analysis and data enhancement. This post-processing stage significantly increases 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 remove over 800 similar questions.

F.1 Data Pre-process

The raw data range from HTML, photocopy, hand script, and plain text, and we conduct pre-processing 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 underlines (<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.

F.2 Data Annotation

We develop an online platform for data annotation stage. The platform consists of text boxes for editing contents and regions for rendering the text to see the final appearance of the data as shown in Figure 6. We employ skilled human annotators annotators and involve them as authors, primarily undergraduate students from top universities in China familiar with exam quizzes 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.

这道题目有不小于两个可行的答案，请选出所有的正确选项，格式为连续的多个大写英文字母，不包含选项后面的描述，如：AC，BDE。
This question has no less than two possible answers. Please choose all the correct options, in the format of consecutive uppercase letters, without the description after the options, such as: AC, BDE.

每一个'[MASK]' 对应一个最简且确定的答案，多个'[MASK]' 的答案之间换行隔开，如：文艺复兴\n0.5\n①。
Each '[MASK]' corresponds to a simple and definite answer. The answers for multiple '[MASK]'s are separated by line breaks, such as: Renaissance\n0.5\n①.

这道题需要你对问题进行详细的分析作答，请以'我的分析如下：'作为开头。
This question requires you to analyze the problem in detail. Please start with 'My analysis is as follows:'.

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.

这道题目不包含图片信息{，我们会输入一张纯黑图片}，请基于文字信息，给出你对这道题的思考。
This question does not contain image information. {We will input a pure black image. } Please give your answer directly based on the text information in the question.

这道题目包含图片信息，{但我们不会提供这部分信息，请基于题目中的文字信息回答问题/我们使用生成的图片描述来代替图片，你可以参考这些描述来回答问题}。如果你认为题目中的文字信息不足以确定正确答案，请回答'缺少图片信息'而非随便猜测一个答案，否则请按照格式给出答案。
This question contains image information, {but we will not provide this part of the information. Please give your answer directly based on the text information in the question/We use the generated image description to replace the image. You can refer to these descriptions to answer the questions}. If you think that the text information in the question is not enough to determine the correct answer, please answer 'Lack of image information' instead of guessing an answer at will. Otherwise, please give your answer directly based on the text and image information.

这道题目包含多张图片信息，你将通过多轮问答的方式接收到所有的图片。请注意，直到你被要求开始作答之前，题目均未加载完成，你可以在每一轮对话的过程中给出你对当前信息的理解与思考，但我们只会采纳你最后一轮得出的答案作为最终结果。请基于全部的文字和图片信息，给出你对这道题的思考。
This question contains multiple images. You will receive all the images through multiple rounds of dialogue. Please note that until you are asked to start answering, the question has not been loaded completely. You can give your understanding and thoughts on the current information during each round of dialogue, but we will only adopt the answer you obtained in the last round as the final result. Please give your answer directly based on all 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:

请先在此处，逐步给出你对所给问题的思考过程、推理：Please show your reasoning process step by step:

根据以上思考过程，你的最终答案是：According to the reasoning you have given above, the final answer should be:

Figure 5: The prompts for evaluation on MULTI.

markdown rules, to 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 7). This includes transforming complex mathematical and chemical equations, usually in HTML, into LaTeX. For this purpose, Mathpix [8] 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.

In Figure 7, we show several examples of complex formation and modification during data annotation stage. The markdown, LaTeXand HTML format code is remained for better format clarity.

F.3 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 format the questions in a render-friendly manner, and meanwhile reduce the similarity to contents that the MLLMs are trained on. During this stage, we assess if there are any omissions or missing elements.

[8] `https://mathpix.com/snipping-tool`

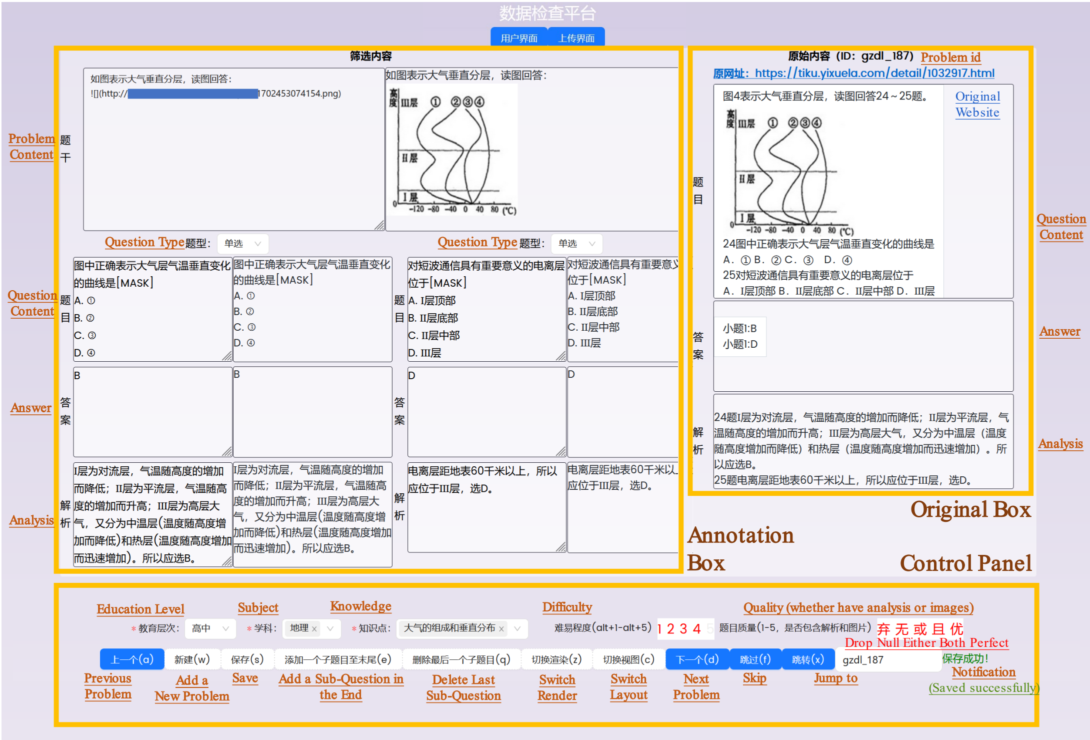


Figure 6: A screenshot for the main page of the data annotation platform.

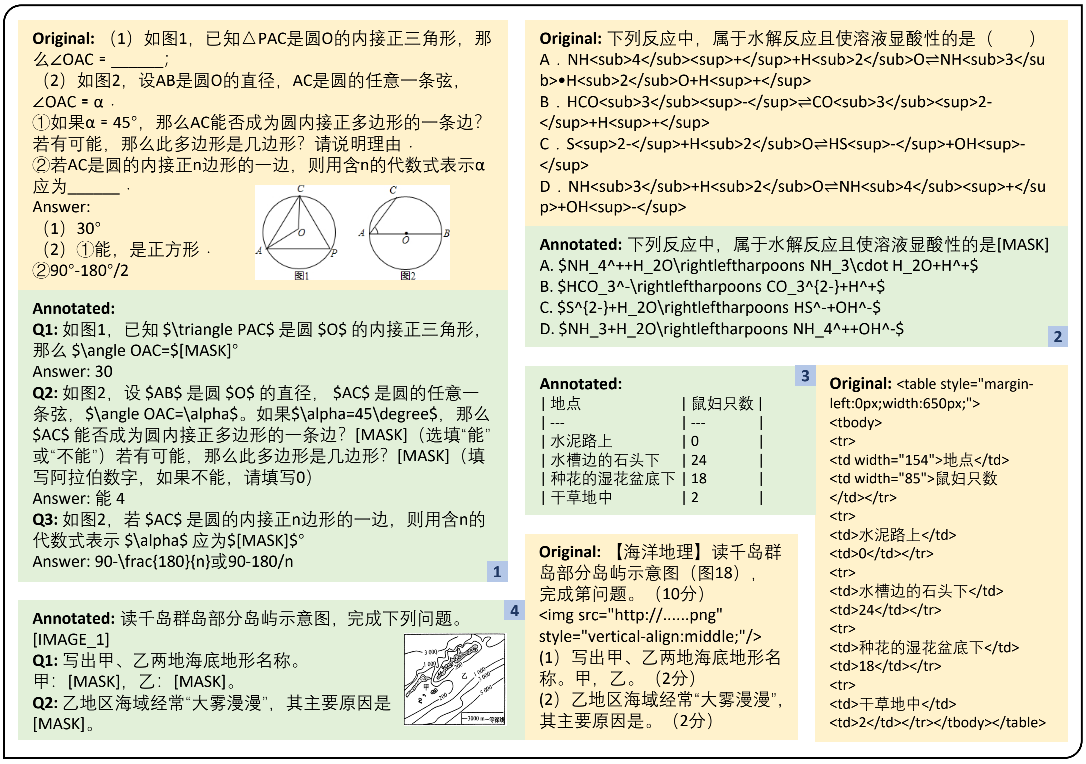


Figure 7: Several data annotation examples when constructing MULTI.

- **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 8) in order to determine a unique fixed answer for each question.

- **Distillation**. This is completed during our annotation process. We reduce assistance information so that the answer must depend on more detailed

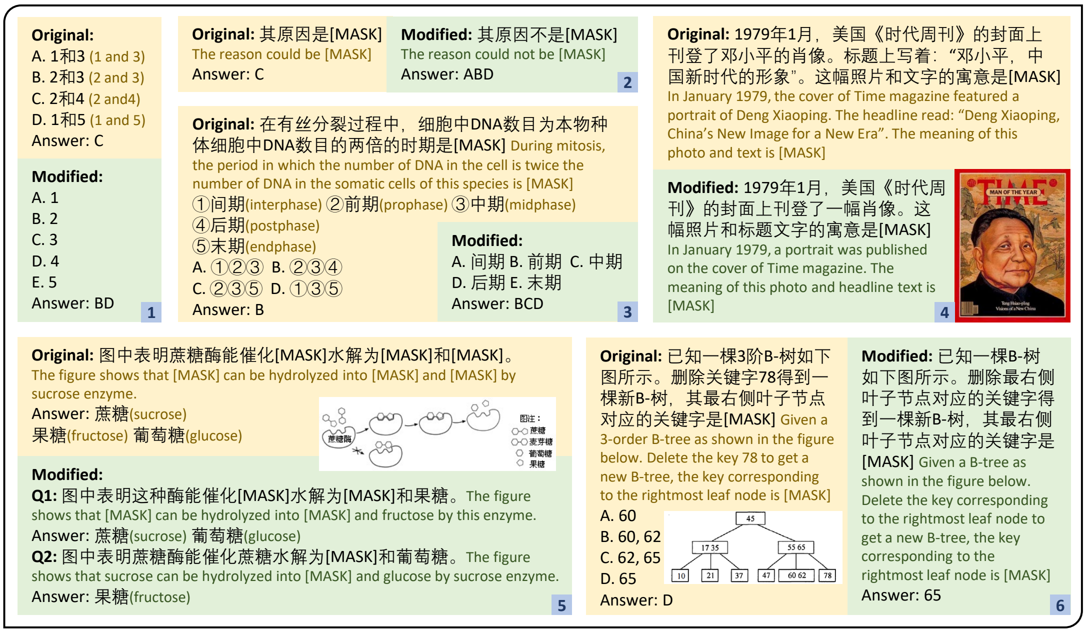


Figure 8: Several data augmentation examples when constructing MULTI.

analysis (refer to example (4) in Figure 8). In this way, we greatly enhance question difficulty.

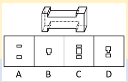
- **Transformation**. We randomly modify the questions such as from single-choice to blank-filling (refer to example (2) in Figure 8), or convert certain kinds of single-choice questions into multiple-choice ones (refer to example (1) and (5) in Figure 8). Lots of single-choice questions have a list 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.

In Figure 8, we show several examples of complex formation and modification during data postprocess stage. English translations of Chinese text are shown for better readability.

G More Examples

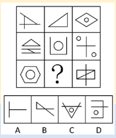
In Figure 9, we show more examples for annotated questions. English translations of Chinese text are shown for better readability.

Question: 下面的立体图形如果从任一面剖开，以下哪一项不可能是该立体图形的截面?[MASK] Which of the following could not be a cross-section of the three-dimensional figure below if it were cut open from either side?[MASK]
[IMAGE_1]
A. A B. B C. C D. D



Ground Truth: D

Question: 从所给四个选项中，选择最合适的一个填入问号处，使之呈现一定的规律性。[MASK] From the four options given, choose the most appropriate one to fill in the question mark to give some regularity. [MASK]
[IMAGE_1]
A. A B. B C. C D. D



Ground Truth: D

Question: 下列邮票图案与少数民族的对应关系，不正确的是:[MASK] The following stamp motifs correspond incorrectly to ethnic minorities:[MASK]
A. 朝鲜族(Korean) [IMAGE_1]
B. 傣族(Dai) [IMAGE_2]
C. 回族(Hui) [IMAGE_3]
D. 苗族(Hmong) [IMAGE_4]

Ground Truth: D

Question: 彭某驾驶一辆重型半挂牵引车，载运37.7吨货物（核载25吨），行至大广高速公路一下坡路段，追尾碰撞一辆由李某驾驶在应急车道内行驶的重型自卸货车（货箱内装载3.17立方米黄土并搭乘24人），造成16人死亡、13人受伤。此事故中的主要违法行为是什么?[MASK]
Peng drove a heavy semi-trailer truck, carrying 37.7 tons of goods (rated 25 tons), and when he reached a downhill section of the Daguang Expressway, he rear-ended a heavy dump truck driven by Li in the emergency lane (the cargo box was loaded with 3.17 cubic meters of loess and 24 people were on board), causing 16 deaths and 13 injuries. What is the main illegal act in this accident?[MASK]
A. 彭某超速行驶(Peng was speeding)
B. 彭某驾驶机动车超载(Peng was driving an overloaded vehicle)
C. 李某在应急车道内行驶(Li was driving in the emergency lane)
D. 李某货车车厢内违法载人(Li was illegally carrying people in the truck box)

Ground Truth: BCD

Ground Truth: A

Question: 驾驶机动车在有这种标志的路口怎样通过最安全?[MASK] How to pass through an intersection with this sign in the safest way when driving a motor vehicle?[MASK]
A. 停车观察主路情况(Stop and observe the main road situation)
B. 加速尽快进入主路(Accelerate and enter the main road as soon as possible)
C. 减速缓慢进入主路(Slow down and enter the main road slowly)
D. 减速观察左后方情况(Slow down and observe the left rear situation)
[IMAGE_1]

Question: 已知酸性条件下有反应：$2Cu^+= Cu^{2+}+Cu$。氢气还原氧化铜实验由于反应温度不同，产物可能不同。下表为在红色的还原产物中加入试剂和产生的现象。由此推出本次氢气还原氧化铜实验的产物是 [MASK]
It is known that under acidic conditions, there is a reaction: $2Cu^+= Cu^{2+}+Cu$. The product of the hydrogen reduction of copper oxide experiment may vary depending on the reaction temperature. The table below shows the reagents added and the phenomena produced in the red reduction product. It can be concluded that the product of this hydrogen reduction of copper oxide experiment is [MASK]
[IMAGE_1]

加入试剂	稀硫酸	浓硫酸并加热	稀硝酸	浓硝酸
实验现象	得红色固体和蓝色溶液	产生无色气体	产生无色气体和蓝色溶液	产生红棕色气体和蓝色溶液

A. 是(Is)Cu
B. 是(Is)Cu_2O
C. 一定有(Must have)Cu，可能有(May have)Cu_2O
D. 一定有(Must have)Cu_2O，可能有(May have)Cu

Ground Truth: A

Figure 9: More example of MULTI.