MIBench: Evaluating Multimodal Large Language Models over Multiple Images

Haowei Liu[1,2], **Xi Zhang**[3], **Haiyang Xu**[3†], **Yaya Shi**[4], **Chaoya Jiang**[5], **Ming Yan**[3], **Ji Zhang**[3], **Fei Huang**[3], **Chunfeng Yuan**[1,2†], **Bing Li**[1,2], **Weiming Hu**[1,2,6]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Alibaba Group [4]University of Science and Technology of China [5]Peking University
[6]School of Information Science and Technology, ShanghaiTech University, China
liuhaowei2019@ia.ac.cn, cfyuan@nlpr.ia.ac.cn
{shuofeng.xhy, ym119608}@alibaba-inc.com

Abstract

Built on the power of LLMs, numerous multimodal large language models (MLLMs) have recently achieved remarkable performance on various vision-language tasks. However, most existing MLLMs and benchmarks primarily focus on single-image input scenarios, leaving the performance of MLLMs when handling realistic multiple images underexplored. Although a few benchmarks consider multiple images, their evaluation dimensions and samples are very limited. In this paper, we propose a new benchmark **MIBench**, to comprehensively evaluate fine-grained abilities of MLLMs in multi-image scenarios. Specifically, MIBench categorizes the multi-image abilities into three scenarios: multi-image instruction (MII), multimodal knowledge-seeking (MKS) and multimodal in-context learning (MIC), and constructs 13 tasks with a total of 13K annotated samples. During data construction, for MII and MKS, we extract correct options from manual annotations and create challenging distractors to obtain multiple-choice questions. For MIC, to enable an in-depth evaluation, we set four sub-tasks and transform the original datasets into in-context learning formats. We evaluate several open-source and closed-source MLLMs on the proposed MIBench. The results reveal that although current models excel in single-image tasks, they exhibit significant shortcomings when faced with multi-image inputs, such as limited fine-grained perception, multi-image reasoning and in-context learning abilities. The annotated data of MIBench is available at https://huggingface.co/datasets/StarBottle/MIBench.

1 Introduction

Recently, leveraging the powerful comprehension and reasoning abilities of LLMs, many MLLMs such as LLaVA-1.5 (Liu et al., 2024) and mPLUG-Owl2 (Ye et al., 2024b) have demonstrated outstanding performance across various vision-language tasks (*e.g.* image captioning, VQA and visual grounding). Concurrently, numerous benchmarks like MME (Fu et al., 2023), MMBench (Liu et al., 2023) and SEED-Bench (Li et al., 2024)) have been proposed to evaluate the abilities of MLLMs in terms of different perspectives such as recognition, localization and reasoning.

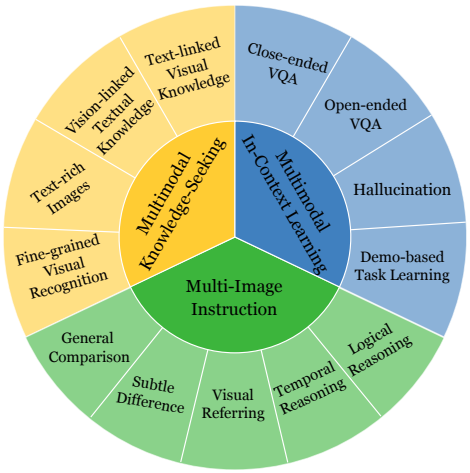

Figure 1: Overview of our MIBench, which covers three multi-image scenarios and a total of 13 tasks.

However, most existing MLLMs focus on single-image scenarios. Accordingly, previous benchmarks primarily evaluate MLLMs based on single-image inputs. In contrast, real-world multimedia information, such as web pages and social media, generally contains multiple images and corresponding text in interleaved forms. Therefore, multi-image scenarios have greater practical value than single-image scenarios, which makes it worth exploring whether existing single-image MLLMs possess emergent abilities for multi-image inputs. Moreover, some methods like Sparkles (Huang et al., 2023) and Mantis (Jiang et al., 2024) explore multi-image scenarios but have not comprehensively evaluated the models' multi-image abili-

[†]Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22417–22428
November 12-16, 2024

Benchmark	Scenario	#Multi-Image Task	#Multi-Image Sample	Answer Type	Evaluator
MME	Single-Image	0	0	Yes/No	Metrics
MMBench	Single-Image	0	0	Multi-choice	GPT
SEED-Bench	Single-Image	4	829	Multi-choice	Metrics
Sparkles-Eval	Multi-Image Dialogue	1	150	Open-ended	GPT-4
Mantis-Eval	Multi-Image Reasoning	1	217	Multi-choice & Short Answer	Metrics
MIBench	Comprehensive Multi-Image	13	13K	Multi-choice & Short Answer	Metrics

Table 1: Comparison of the proposed MIBench with recent MLLM benchmarks.

ties. As shown in Table 1, Sparkles evaluates the model solely on a small-scale multi-image chat dataset, and the assessment relies entirely on scoring by GPT-4. Mantis-Eval focuses on multi-image reasoning and has a limited scale of 217 samples.

In this paper, to comprehensively evaluate the multi-image ability of MLLMs, we propose a large-scale multi-image benchmark **MIBench**, which covers 13 different tasks with a total of 13K high-quality samples. As shown in Figure 1, MIBench contains three multi-image scenarios, *i.e.* **Multi-Image Instruction (MII)**, **Multimodal Knowledge-Seeking (MKS)** and **Multimodal In-Context Learning (MIC)**. MII is a basic multi-image scenario, where the instructions involve perception, comparison and reasoning across multiple images. MKS presents a different scenario, in which models are provided with interleaved image-text data as external knowledge, while the question itself is about a single image or even independent of any image. MIC is another scenario where MLLMs respond to queries (*e.g.* image & question) by conditioning on a series of multimodal demos. The three scenarios are further divided into 13 different tasks, and examples are shown in Figure 2.

The MII and MKS scenarios comprise 9K multiple-choice questions. To get these questions, we first sample images from nine existing datasets, and convert the original annotations into questions and ground truth options according to the tasks. To obtain challenging distractors and mitigate inherent biases of options, we devise task-specific strategies to sample from annotations or generate distractors using GPT-4. For MKS, we also devise corresponding strategies to sample images and associated texts from the datasets as external knowledge. The MIC scenario contains 4K short-answer questions, covering close-ended VQA, open-ended VQA, object hallucination, and demo-based task learning. We convert the data sampled from four datasets into the VQA format, and retrieve samples of the same task to construct demos. To ensure high quality, we combine automated filtering and manual verification to remove samples with ambiguous or duplicate options. For multiple-choice questions, we use accuracy as the metric and employ circular evaluation (Liu et al., 2023) to mitigate the position bias of LLMs. For short-answer questions, we use exact matching as the metric.

We evaluate several existing MLLMs on the proposed MIBench, including both closed-source (*e.g.* GPT-4o) and open-source models (*e.g.* LLaVA-1.5, Idefics2 and mPLUG-Owl3). The evaluation results reveal that current MLLMs especially open-source models have major flaws in multi-image scenarios. The annotated data of our MIBench is publicly available to spur progress in improving the multi-image abilities of MLLMs.

Our contributions are summarized as follows:

- We propose the first large-scale and comprehensive benchmark MIBench for evaluating the multi-image abilities of MLLMs, covering three scenarios and 13 tasks in total.
- The evaluation on MIBench reveals that existing MLLMs especially open-source models face significant challenges in **fine-grained perception** and **multi-image reasoning**.
- Current MLLMs perform poorly in the **multimodal knowledge-seeking** scenario. And there still exists considerable room for improvement in the **multimodal in-context learning** abilities.

2 Related Work

2.1 Multimodal Large Language Models

Recent research (Zhu et al., 2023; Liu et al., 2024; Dai et al., 2024; Ye et al., 2023) has expanded LLMs (*e.g.* LLaMA Touvron et al., 2023) into multimodal scenarios, enabling them to process both visual and textual information. Some studies (Jiang et al., 2024; Huang et al., 2023; Laurençon et al.,

Figure 2: Examples of the multi-image scenarios with a total of 13 tasks. The correct answers are marked in blue.

2024; Ye et al., 2024a) have further explored augmenting MLLMs with multi-image understanding abilities. However, there lacks a comprehensive benchmark for evaluating the multi-image abilities of MLLMs, which limits the full exploration of these models' potential and hinders the development of this field.

2.2 MLLM Benchmarks

The rapid development of MLLMs has led to the emergence of a series of benchmarks, such as LVLM-eHub (Xu et al., 2023), MMBench (Liu et al., 2023), MM-Vet (Yu et al., 2023) and SEED-Bench (Li et al., 2023a). However, these benchmarks primarily focus on single-image evaluation, and often overlook multi-image perception and reasoning abilities, which hold even greater practical value. Some recent studies develop benchmarks for assessing multi-image capabilities. Sparkles-Eval aims to establish a benchmark for multi-turn dialogues and multi-image scenarios. However, it exclusively focuses on the dialogue scenario, and relies entirely on GPT-4 for evaluation. Besides, it has a small data scale. Other datasets such as Mantis-Eval (Jiang et al., 2024) and SEED-Bench2 (Li et al., 2024) also cover a small number of multi-image tasks, with a limited scale due to reliance on manual annotation.

In this paper, we propose a large-scale benchmark covering three multi-image scenarios and 13 tasks, to comprehensively evaluate the multi-image capabilities of MLLMs.

3 MIBench

3.1 Evaluation Taxonomy

We categorize multi-image inputs into three scenarios: Multi-Image Instruction (MII), Multimodal Knowledge-Seeking (MKS) and Multimodal In-Context Learning (MIC). As Figure 2 shows, MII refers to cases where instructions involve perception, comparison and reasoning across multiple images. For instance, "Do the two images show the same number of cats?" MKS examines the ability of MLLMs to acquire relevant information from external knowledge, which is provided in an interleaved image-text format. Compared to MII, the questions in the MKS scenario can be about a single image or even independent of any visual content. MIC is another popular scenario, in which MLLMs respond to visual questions while being provided with a series of multimodal demonstrations (*i.e.*, demos).

3.1.1 Multi-Image Instruction

According to the semantic types of the instructions, we further categorize the Multi-Image Instruction scenario into the following five tasks.

General Comparison (GC) task examines the model's general understanding of each image (*e.g.* scene, attribute and location), and comparison across different images. GC represents the most fundamental aspect of multi-image abilities. We use the image-pair description dataset NLVR2 (Suhr et al., 2018) for data construction.

Subtle Difference (SD) task examines the model's ability to perceive subtle differences between similar images. Compared to general comparison, the SD task requires more fine-grained perception ability. The image editing dataset MagicBrush (Zhang et al., 2024) is adopted in this task.

Visual Referring (VR) task evaluates whether the model can utilize the referring information provided by input images to comprehend the relationships between different objects. Figure 2(c) shows an example of the VR task, whose data is from the visual relation dataset VrR-VG (Liang et al., 2019).

Temporal Reasoning (TR) task assesses the model's understanding of the temporal relationships among a series of consecutive images, and its comprehension of the overall content conveyed by these images. We employ the video understanding dataset Something-Something V2 (Goyal et al., 2017a) for this task.

Logical Reasoning (LR) task requires the model to perform logical reasoning and analyze the causal relationships between objects or events shown in the input images. The video QA dataset NExT-QA (Xiao et al., 2021) is used for data construction.

3.1.2 Multimodal Knowledge-Seeking

Based on the forms of external knowledge, we categorize the Multimodal Knowledge-Seeking scenario into the following four tasks.

Fine-grained Visual Recognition (FVR) task examines the model's ability to recognize the object in the query image when given multiple reference images. It requires the model to understand the image-label correspondence in the reference images, as well as link similar images. A combination of several fine-grained recognition datasets (Khosla et al., 2011, Wah et al., 2011 and Nilsback and Zisserman, 2008) is used for this task.

Text-Rich Images (TRI) VQA task evaluates the model's ability to understand text-rich images and extract information relevant to the question, which is very common in real-world scenarios (*e.g.* reading slides or documents). We adopt the SlideVQA (Tanaka et al., 2023) dataset for data construction.

Vision-linked Textual Knowledge (VTK) task corresponds to a very practical scenario where the question is beyond the visual content of the query image, such as querying background knowledge. The provided external knowledge encompasses images and corresponding text which are possibly retrieved from a knowledge base (*e.g.*, Wikipedia). The model is required to link the query image to the relevant image, and extract useful information from the corresponding text. Figure 2(h) shows an example, whose data is from the InfoSeek dataset (Chen et al., 2023).

Text-linked Visual Knowledge (TVK) task refers to cases where the text-only question is about the visual attributes of a specific object. For instance, "Is the China National Stadium round or square?" When provided with external knowledge in an interleaved image-text form, the model needs to link the question to the relevant text, and extract visual information from the corresponding image. This task is very common in real life such as browsing web pages. Figure 2(i) shows an example, whose data is from the WebQA dataset (Chang et al., 2022).

3.1.3 Multimodal In-Context Learning

The in-context learning ability enables LLMs to gain performance boost when provided with a series of demos. Recent studies (Alayrac et al., 2022; Awadalla et al., 2023; Laurençon et al., 2024) have also explored multimodal in-context learning (MIC). For the evaluation of the MIC ability, existing methods solely assess the model's performance via a holistic metric, such as accuracy on the VQAv2 (Goyal et al., 2017b) dataset. To evaluate the model's MIC ability in a fine-grained manner, we categorize the MIC scenario into the following four distinct tasks.

Close-ended VQA task requires the model to answer from a predefined set of responses, which is provided via multimodal demos. This task examines the model's ability to learn the image-label mapping relationships from the demos. We use the Mini-ImageNet dataset (Vinyals et al., 2016) for data construction.

Open-ended VQA task has an open range of possible answers which cannot be fully covered by the provided demos. The task evaluates the model's ability to learn task patterns through demos. We conduct a balanced sampling of different knowledge types from the OK-VQA dataset (Marino

et al., 2019) for this task.

Hallucination phenomenon is a significant challenge faced by MLLMs. In this task, we convert the hallucination dataset POPE (Li et al., 2023b) into in-context learning format, and study the impact of the model's MIC ability on the hallucination phenomenon.

Demo-based Task Learning is a core aspect of in-context learning, which enables the model to rapidly adapt to new tasks given a few demos. To investigate existing MLLMs' demo-based task learning ability, we select several visual tasks from the VQAv2 dataset and remove the task instructions. Instead, we present the task demos in the form like "rabbit: 3". Figure 2(m) shows an example.

3.2 Data Generation

In Section 3.1, we introduced the evaluation tasks and the corresponding data source of the proposed MIBench. However, the generation of test samples using the original datasets is nontrivial. We meticulously devise a data generation pipeline, including various strategies of question generation, distractor generation and external knowledge sampling for different tasks.

Question Generation. To enhance the diversity of questions, we devise corresponding prompts for the tasks, and employ GPT-4 to generate a variety of question forms. We then randomly sample from the question pool to construct the test samples. For instance, for the General Comparison (GC) task, the questions such as "Is the subsequent sentence an accurate portrayal of the two images?" and "Can the given sentence accurately illustrate what's in these two images?" are utilized.

Distractor Generation. For different tasks, we adopt two methods of distractor generation. One way is to sample from the original annotations following certain strategies. For instance, for the Temporal Reasoning (TR) task, we utilize the Something-something V2 dataset for data construction. To prevent the model from taking shortcuts by identifying objects to choose the correct options, we sample different temporal relationships of the same object from the annotations as distractors. In this way, the constructed test samples can more accurately reflect the model's understanding of temporal relationships. The second method is to generate distractors with the help of GPT-4. For instance, in the Text-Rich Images (TRI) VQA task, we prompt GPT-4 to generate distractors according to the question and the correct answer.

External Knowledge Sampling. For the Multimodal Knowledge-Seeking (MKS) scenario, reasonably sampling interleaved image-text data as external knowledge is very important to the quality of test samples. For instance, in the Vision-linked Textual Knowledge (VTK) task, we select text and images from the original annotations which have the same question as the current query but with different answers as external knowledge. This approach avoids selecting text and images unrelated to the current query, and can thus generate more challenging distractors. Additionally, some datasets require more complex information extraction. For instance, we use GPT-4 to extract question-related segments from the original wiki entries of the InfoSeek dataset, which can be as long as several thousand words.

3.3 Quality Control

To mitigate data contamination, our construction of test data exclusively utilizes the validation or test sets from existing datasets. Furthermore, we combine automated filtering and manual verification to ensure the quality and reliability of the test data.

Specifically, after the data generation process, we perform two automated filtering strategies on the obtained data. **1)** We remove images from the input samples, and test multiple advanced MLLMs on them. Then we discard samples which can still be answered correctly without visual input. This avoids the overestimation of model performance due to the textual bias of the questions and options. **2)** For the Multimodal Knowledge-Seeking scenario, we eliminate external knowledge from the samples and test them using multiple MLLMs. Then we remove samples which the models can answer correctly without external knowledge. This mitigates the impact of internal knowledge of the model, and provides a more accurate assessment of the model's ability of utilizing external knowledge.

As stated in Section 3.2, for some tasks such as Visual Referring, we employ GPT-4 to generate distractors. To ensure the high quality of the generated samples, we apply manual verification after automated filtering. The process is conducted by three trained annotators who possess relevant professional backgrounds. Specifically, a sample is discarded if there are duplicate options or more than one correct option.

Model	Multi-Image Instruction					Multimodal Knowledge-Seeking			
	GC	SD	VR	TR	LR	FVR	TRI	VTK	TVK
Closed-source MLLMs									
GPT-4o	80.7	90.5	46.8	68.0	69.8	98.3	74.8	54.7	63.3
GPT-4V	72.8	79.2	45.8	61.8	66.3	90.2	71.0	52.0	56.0
Open-source MLLMs									
mPLUG-Owl3	86.4	70.1	33.0	46.8	67.2	76.4	50.1	31.1	48.8
Mantis	83.0	54.1	37.6	45.5	63.4	16.4	37.7	26.4	41.7
Idefics2-I	83.1	49.7	32.6	44.8	56.4	42.4	43.9	25.6	39.0
MMICL	53.7	46.4	41.1	47.0	59.6	56.6	27.6	22.1	35.9
mPLUG-Owl2	64.2	40.1	35.6	30.7	41.3	13.3	39.0	17.0	25.6
Qwen-VL	45.9	22.5	16.3	27.5	36.8	58.8	35.9	22.9	18.1
LLaVA-1.5	40.6	14.9	24.1	30.1	44.8	18.2	25.8	16.7	26.3
mPLUG-Owl	19.1	4.0	21.7	8.0	29.2	17.3	12.1	14.9	20.6

Table 2: Evaluation results on the multi-image instruction and multimodal knowledge-seeking scenarios of MIBench.

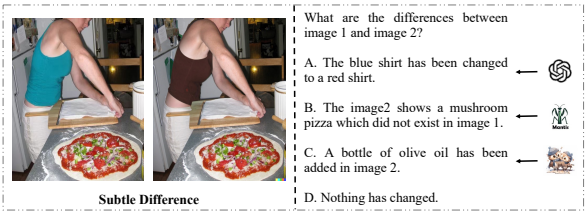

Figure 3: A qualitative case of the Subtle Difference task, where open-source MLLMs show inferior performance due to limited fine-grained perception ability.

3.4 Evaluation

For the multiple-choice questions, we employ the accuracy of the predicted options as the evaluation metric. Notably, early MLLMs such as mPLUG-Owl tend to produce longer responses rather than directly outputting the option. For these models, we use GPT-4 to determine which option matches the predicted content. In addition, similar to the observation of MMBench, we find that different MLLMs show preferences for specific options (*i.e.* position bias). Therefore, we set the correct option sequentially to "A", "B", "C" and "D". A model is considered to have correctly answered a sample only if it consistently provides the correct response across multiple tests. In this way, the impact of position bias on the evaluation results is mitigated.

4 Experiments

4.1 Models

In this section, we evaluate MLLMs using the constructed MIBench dataset. We first evaluate MLLMs on the Multi-Image Instruction and Multi-modal Knowledge-Seeking scenarios. These models can be categorized into three distinct groups: (1) closed-source models, including GPT-4V and GPT-4o; (2) open-source single-image MLLMs, including mPLUG-Owl (Ye et al., 2023), LLaVA-1.5 (Liu et al., 2024), Qwen-VL (Bai et al., 2023) and mPLUG-Owl2 (Ye et al., 2024b); (3) open-source models natively supporting multi-image input, including Mantis (Jiang et al., 2024), Idefics2 (Laurençon et al., 2024) and mPLUG-Owl3 (Ye et al., 2024a). For the open-source models, we employ greedy decoding for prediction generation.

Then we evaluate open-source MLLMs on the Multimodal In-Context Learning (MIC) scenario. However, as most of these models have neither been pre-trained on large-scale interleaved image-text data nor fine-tuned on ICL format data, they do not exhibit MIC capabilities. In the tests across the four MIC tasks, they consistently exhibit a negative ICL effect, *i.e.*, their performance decreases as the number of demos increases. Therefore, we only present the evaluation results of models that possess multimodal ICL abilities, *i.e.* OpenFlamingo, MMICL, IDEFICS and IDEFICS-I.

4.2 Evaluation Results

4.2.1 Multi-Image Instruction & Multimodal Knowledge-Seeking

Table 2 shows the main results of the Multi-Image Instruction (MII) and Multimodal Knowledge-Seeking (MKS) scenarios. Through these results, we have several valuable observations:

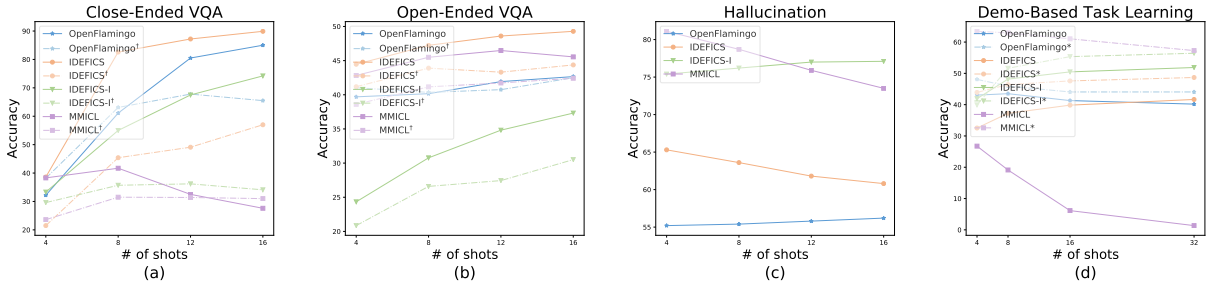

Figure 4: Evaluation results on the Multimodal In-Context Learning scenario.

Closed-source MLLMs exhibit superior performance than open-source models. In most MII and MKS tasks, the performance of open-source models lags significantly behind that of proprietary models. For instance, on the Temporal Reasoning (TR) task, the best-performing open-source model MMICL achieves an accuracy of 47.0%, falling behind GPT-4o by 21.0%.

Open-source MLLMs are inadequate in fine-grained perception tasks. Although many open-source MLLMs have decent performance on the General Comparison (GC) task, their performance is significantly worse on the Subtle Difference (SD) and Text-Rich Images (TRI) VQA tasks. For instance, Idefics2-I achieves 83.1%, 49.7% and 43.9% on the three tasks respectively. In contrast, GPT-4V and GPT-4o largely outperform open-source models, due to their high-resolution input strategy. Figure 3 provides a qualitative case supporting the above point.

Multi-image MLLMs perform better than single-image models on most tasks. This verifies that pre-training on interleaved image-text data (*e.g.* Idefics2-I) and instruction tuning on multi-image data (*e.g.* Mantis) are both beneficial for improving multi-image abilities. Combining multi-image pre-training and instruction tuning, mPLUG-Owl3 achieves better performance than other open-source MLLMs on most tasks.

The Visual Referring (VR) task is particularly challenging for existing MLLMs. As it requires integration of fine-grained perception, spatial correspondence and relation reasoning, most models have not achieved satisfactory performance on the VR task. Even the best-performing model GPT-4o has not exceeded a 50% accuracy rate.

Most existing open-source MLLMs perform poorly in the Multimodal Knowledge-Seeking (MKS) scenario. For instance, the accuracy rates of mPLUG-Owl2 on both the Vision-linked Textual Knowledge (VTK) and Text-linked Visual Knowledge (TVK) tasks are below 30%. In the Fine-grained Visual Recognition (FVR) task, which requires the combination of fine-grained perception and comparison abilities, mPLUG-Owl2's performance is merely over 10%. Compared to single-image MLLMs, multi-image models such as Idefics2-I exhibit better capabilities in utilizing multimodal external knowledge. However, there is still significant room for improvement, as the performance of Idefics2-I on both the VTK and TVK tasks is under 40%.

4.2.2 Multimodal In-Context Learning

Figure 4 shows the performances of OpenFlamingo, MMICL, and IDEFICS on multimodal ICL scenarios. The horizontal axis represents different shots (*i.e.*, the number of demos), and the vertical axis represents accuracy. To investigate the impact of images on multimodal ICL, the models that remove the images from demos († in Figure 4) are evaluated on close-ended VQA and open-ended VQA.

The current models exhibit multimodal ICL abilities on close-ended VQA. As shown in Figure 4(a), after removing the images in the demos, the performance of most models declines, and the extent of this decline increases with the number of shots. This indicates that these models have learned the image-label mapping relationships in the demos, exhibiting a certain degree of multimodal ICL ability.

Multimodal ICL abilities of different models appears to be driven by different modalities. As shown in Figure 4(b), when the number of demos increases, all models show consistent performance improvement. However, for OpenFlamingo, removing images from the demos does not cause a significant performance change, indicating that OpenFlamingo's ICL on this task is primarily driven by text. In contrast, the absence of images leads

	Text-rich Images VQA		Text-linked Visual Knowledge		Vision-linked Textual Knowledge	
	w/ Dis.	w/o Dis.	w/ Dis.	w/o Dis.	w/ Dis.	w/o Dis.
mPLUG-Owl2	39.0	42.1	25.6	29.6	17.0	90.1
Mantis	37.7	42.6	41.7	47.7	26.4	88.1
Idefics2-I	43.9	46.8 (59.5)	39.0	45.2	25.6	91.0

Table 3: Ablation study of the impact of distractors on various tasks on the multimodal knowledge-seeking scenario.

	Confusion		Reasoning	
	Conf. A	Conf. B	Tem.	Obj.
mPLUG–Owl2	87.0	25.0	30.7	56.6
Qwen–VL	89.2	26.8	27.5	60.9
LLaVA–1.5	91.8	31.6	30.1	59.3
Mantis	91.2	83.6	45.5	75.7

Table 4: Ablation study on the multi-image confusion phenomenon and the temporal reasoning task.

to a significant performance decline for IDEFICS and MMICL, indicating that they possess a certain degree of multimodal ICL ability.

Multimodal ICL abilities of current models do not alleviate the hallucination phenomenon. As shown in Figure 4(c), on object hallucination task, only IDEFICS-I and Idefics2-I exhibit slight accuracy improvements with an increasing number of shots, while other models show negative effects. It indicates that ICL provides very limited help in mitigating hallucinations and may even exacerbate them. Comparing the base and instruction-tuned versions of IDEFICS, it is evident that instruction tuning can help alleviate hallucinations.

Most models possess some capacity of demo-based task learning, but the capacity is relatively limited. Figure 4(d) shows the model performance under different shots in counting and color tasks demonstrated only through examples. It is evident that with an increasing number of demos, IDEFICS shows significant gains, OpenFlamingo quickly reaches saturation, and MMICL even fails to follow the task format presented in the demos. In fact, except for MMICL, these models can effectively follow the output format with just 4 shots, and their performance improves with more shots. It reflects that OpenFlamingo and IDEFICS possess a certain degree of demo-based task learning ability. In addition, compared to the experimental results with explicit task instructions (*e.g.*, 'How many people are in the room?'), there remains a significant performance gap, indicating that the demo-based task learning abilities of current models still have substantial room for improvement.

4.3 Analysis

4.3.1 Multi-image Confusion Phenomenon

When evaluating MLLMs on the MIBench benchmark, we observe that open-source MLLMs, particularly single-image models, exhibit confusion when handling multiple images. To validate this issue, we derive two confusion subsets with 500 samples respectively from the POPE dataset used in the hallucination task. In subset A, each sample consists of one image and one question. The question asks whether a specific object is present in the image, which actually is not contained. In subset B, an extra image containing the object in the question is added to each sample in subset A as a distractor. As shown in Table 4, it can be observed that the performance of the three single-image models significantly decline after the addition of the extra image, while the multi-image model Mantis also has a slight performance drop. It confirms that current open-source MLLMs, especially single-image models, suffer from severe confusion, thereby affecting their performance in multi-image instructions and multimodal knowledge-seeking.

4.3.2 Limited Reasoning Ability

In the construction of temporal reasoning, we utilize the ground truth of videos as the correct option and sample different actions of the same object as distractors. Under this setting, the majority of MLLMs demonstrate poor performance. To further study these results, we replace the same objects in the distractors with different objects and test several representative models. As indicated in the table, under the setting where distractors contain different objects, the performance of mPLUG-Owl2, LLaVA-1.5 and Mantis models significantly improves, as the models can take shortcuts by distinguishing between objects. The results indicate that for current MLLMs, the reasoning ability across multiple images is significantly inferior to their spatial perception ability.

4.3.3 Bottlenecks of the MKS task

Compared to multi-image instruction, multimodal knowledge-seeking requires the model to extract relevant information from external image-text knowledge sources and is thus more challenging. To investigate the bottlenecks of multimodal knowledge-seeking tasks, we compare the impact of distracting content.

As shown in Table 3, for text-linked visual knowledge, removing distracting content(*i.e.*, only retaining the information relevant to the question) results in a certain performance improvement. It indicates that the model's ability to identify relevant information (*i.e.*, link by text) still can be improved. On the other hand, even after the removal of distracting content, the performance remains poor. It suggests that the primary bottleneck for this task is the deficiencies of MLLMs in perceiving and reasoning with visual information.

In contrast, for the task of vision-linked textual knowledge, the removal of distracting content leads to a significant performance improvement. It suggests that the bottleneck for this task lies in the MLLMs' ability to mine effective messages through image comparison (*i.e.*, link by image).

On text-rich images VQA, removing distracting content brings some performance boost. Based on this, Idefics2-I further boosts from 46.8% to 59.5% by employing image splitting for higher resolution inputs. The significant performance gain indicates that the bottleneck of this task is more related to information loss caused by low resolution.

From the above comparisons, it can be concluded that the current MLLMs' abilities in perceiving, contrasting, and reasoning with visual information are remarkably inferior to their abilities with text, and face substantial challenges in understanding rich-text images due to resolution issues.

5 Conclusion

While MLLMs have shown strong performance in various vision-language tasks, their abilities with multi-image inputs remain underexplored. To address this, we introduce MIBench in this paper, a benchmark that evaluates MLLMs across three multi-image scenarios: multi-image instruction, multimodal knowledge-seeking and multimodal in-context learning, covering 13 tasks with 13K annotated samples. The evaluation results reveal that existing models, despite excelling in single-image tasks, face significant challenges with multi-image inputs. The annotated data is publicly available to facilitate further research. We hope this work can spur progress in improving the multi-image abilities of MLLMs.

Limitations

Due to the input length limitation of current MLLMs, the Multi-Image Instruction and Multimodal Knowledge-Seeking scenarios of our benchmark include 2 to 8 input images in each sample. However, real-world scenarios may involve a large number of images. We'll investigate the evaluation of MLLMs over more images in future work.

Acknowledgement

This work is supported by Beijing Natural Science Foundation (JQ21017, L243015, L223003), the National Key Research and Development Program of China (No. 2020AAA0105802), the Natural Science Foundation of China (No. 62036011, 62192782), and the Project of Beijing Science and Technology Committee (No. Z231100005923046).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. arxiv 2306.13394 (2023).

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017a. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. 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.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

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

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

Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10403–10412.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang,

Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024a. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

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*.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36.

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*.

Task	Image Source	Question Source	Distractor Source
General Comparison	NLVR2	GPT-4 generated	Original annotations
Subtle Difference	MagicBrush	GPT-4 generated	Sampled from annotations
Visual Referring	VrR-VG	Manual	GPT-4 generated
Temporal Reasoning	Something-Something V2	Manual	Sampled from annotations
Logical Reasoning	NeXT-QA	Original annotations	Original annotations
Fine-grained Visual Recognition	Dogs / Birds / Flowers / Cars	GPT-4 generated	Sampled from annotations
Text-Rich Images	SlideVQA	Original annotations	GPT-4 generated
Vision-linked Textual Knowledge	InfoSeek	Extracted from annotations	Sampled from annotations
Text-linked Visual Knowledge	WebQA	Sampled from annotations	GPT-4 generated
Close-ended VQA	Mini-ImageNet	Sampled from annotations	-
Open-ended VQA	OKVQA	Sampled from annotations	-
Hallucination	POPE	Sampled from annotations	-
Demo-based Task Learning	VQAv2	Converted from annotations	-

Table 5: More details of the data generation process.

Task	Image Number Per Sample	Average Question Length	Average Answer Length
General Comparison	2	33.3	1.0
Subtle Difference	2	19.0	10.0
Visual Referring	3	27.0	6.9
Temporal Reasoning	8	39.0	6.2
Logical Reasoning	8	44.7	3.1
Fine-grained Visual Recognition	5	35.4	2.6
Text-Rich Images	4	25.9	2.9
Vision-linked Textual Knowledge	5	562.7	1.7
Text-linked Visual Knowledge	4	76.7	3.6
Close-ended VQA	5-17	5.0	1.4
Open-ended VQA	5-17	8.1	1.2
Hallucination	5-17	7.2	1.0
Demo-based Task Learning	5-33	3.2	1.1
Overall	125K (in total)	68.2	4.1

Table 6: Image number, average question/answer length of each task.

A More Details of MIBench

Table 5 presents the detailed data generation information of each task. Note that "sampled from annotations" isn't simple random sampling from the original annotations. Instead, as stated in Section 3.2, it involves designing specific sampling strategies tailored to the task.

Table 6 shows the detailed statistics of each task, including image number per sample, average question length and average answer length. Note that "Image Number Per Sample" for the Multimodal In-Context (MIC) learning scenario is a range determined by the number of demos. And the whole benchmark has 125K images in total. "Average Answer Length" refers to the average length of options for multiple-choice questions and the average length of answers for short-answer questions.