MIA-BENCH: TOWARDS BETTER INSTRUCTION FOLLOWING EVALUATION OF MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

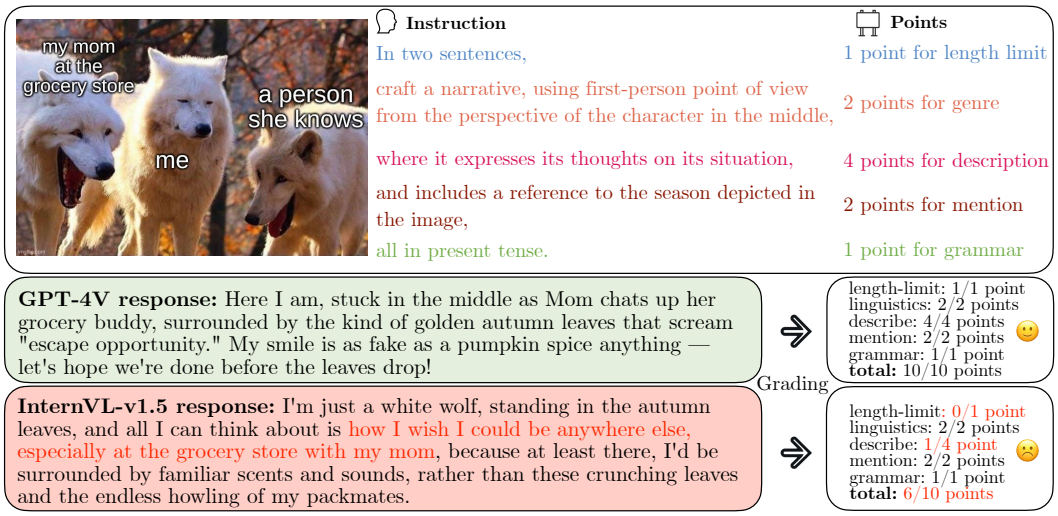

Figure 1: An example from MIA-Bench, featuring an image and a complex instruction to test models' compliance with layered instructions that are compositional in nature. Responses from GPT-4v Achiam et al. (2023) and InternVL-v1.5 Chen et al. (2024b) are evaluated using GPT-4o as the judge.

ABSTRACT

We introduce MIA-Bench, a new benchmark designed to evaluate multimodal large language models (MLLMs) on their ability to strictly adhere to complex instructions. Our benchmark comprises a diverse set of 400 image-prompt pairs, each crafted to challenge the models' compliance with layered instructions in generating accurate responses that satisfy specific requested patterns. Evaluation results from a wide array of state-of-the-art MLLMs reveal significant variations in performance, highlighting areas for improvement in instruction fidelity. Additionally, we create extra training data and explore supervised fine-tuning to enhance the models' ability to strictly follow instructions without compromising performance on other tasks. We hope this benchmark not only serves as a tool for measuring MLLM adherence to instructions, but also guides future developments in MLLM training methods.

1 INTRODUCTION

The rapid advancement of Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Liu et al., 2023c; Dai et al., 2023; Liu et al., 2023b; Bai et al., 2023; Wang et al., 2023a; Lin et al., 2023b; Team, 2023; McKinzie et al., 2024) has been a defining feature of recent AI research, showcasing increased model capabilities to comprehend and respond to visual inputs, often termed as multimodal "instruction following".

To measure the progress of instruction following, many multimodal benchmarks have been developed, which can be roughly divided into two broad categories: (i) fixed-form visual question answering

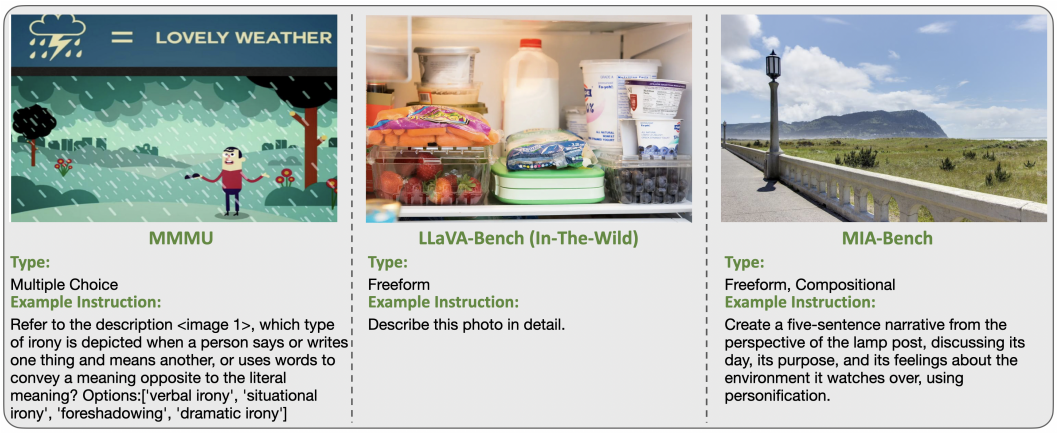

Figure 2: Comparison of various multimodal LLM benchmarks. (Left) Fixed-form visual question answering, often features short answers or multi-choice formats, such as MMMU (Yue et al., 2023). This format is popular due to its ease of evaluation. (Middle) Open-ended responses, such as LLaVA-Bench (in the Wild) (Liu et al., 2023c). (Right) The proposed MIA-Bench, which also uses open-ended responses but focuses on evaluating precise adherence to complex instructions within the prompt.

(VQA), often with short answers or using a multi-choice QA format; and (ii) free-form conversations with open-ended responses. Many current benchmarks have adopted the first format, including VQAv2 (Goyal et al., 2017), TextVQA (Singh et al., 2019b), ScienceQA (Lu et al., 2022), MME (Fu et al., 2023a), MMBench (Liu et al., 2023d), SEED-Bench (Li et al., 2023a), MathVista (Lu et al., 2023), and MMMU (Yue et al., 2023). These benchmarks are popular due to their ease of use in evaluating metrics and presenting model comparisons.

However, as visual assistant models, the ability to engage users in free-form conversations is also crucial. Benchmarks in this format include LLaVA-Bench (Liu et al., 2023c), MM-Vet (Yu et al., 2023b), VisIT-Bench (Bitton et al., 2023), InfiMM-Eval (Han et al., 2023), and the most recent Vibe-Eval (Padlewski et al., 2024) and LLaVA-Bench-Wilder (Li et al., 2024a). Typically, the free-form model responses are evaluated using external models as the judge. These benchmarks are closer to daily-life visual chat scenarios; however, the type of "instruction following" examined in these benchmarks usually gauges a model's ability to perform tasks in a broad, often loosely defined manner. Yet, the precise adherence to complex instructions within prompts – a critical aspect for evaluating LLMs (Chia et al., 2023; Zhou et al., 2023; Qin et al., 2024) – remains less explored in the context of multimodal LLMs.

To this end, we introduce MIA-Bench,[1] a new benchmark specifically designed for evaluating strict "instruction adherence". Our instruction adherence metric measures the precision with which MLLMs can execute layered and compositional instructions. This involves not only recognizing the content of the instructions, but also meticulously executing the detailed demands without deviation (*e.g.*, answering in a given number of sentences, including specific elements, etc.). By establishing this stricter criterion, our benchmark aims to push the boundaries of model precision and reliability in practical applications, ensuring that outputs not only align with the general intent of the instructions, but also match the exact specifications provided. An example from MIA-Bench is provided in Figure 1, and its comparison with previous MLLM benchmarks is illustrated in Figure 2.

MIA-Bench consists of 400 meticulously created image-prompt pairs, and encompasses diverse image contents including animals, food, landmarks, sport, art, landscape, text, etc. to cover a broad spectrum of real-world scenarios. In constructing this benchmark, we sought not only to evaluate the current capabilities of state-of-the-art MLLMs, but also to push the boundaries of what these models can achieve when rigorously tested against structured and layered instructions. The final prompts are of various complexity levels, and compositional in nature, with five base instruction categories, which

[1] Abbreviation for **M**ultimodal **I**nstruction **A**dherence Benchmark.

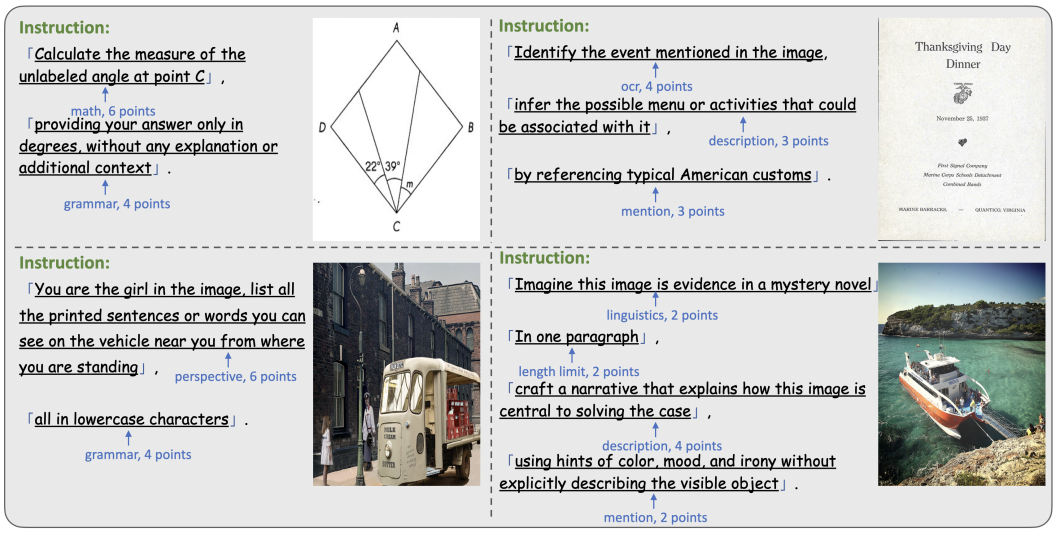

Figure 3: Examples from MIA-Bench, with detailed information on the instruction composition, base instruction weight and type.

are tailored to probe the models' linguistic dexterity, grammatical accuracy, and descriptive fidelity. For example, the prompt in Figure 1 is composed of five base categories, including *description*, *mention*, *grammar*, *length limit*, and *genre*.

We evaluate a wide array of MLLMs on the proposed benchmark, ranging from closed-source models (*e.g.*, GPT-4o (OpenAI, 2024), Gemini Pro (Team, 2023), Claude-3 (Anthropic, 2024), Reka (Team et al., 2024)) to open-source ones (*e.g.*, LLaVA-NeXT (Liu et al., 2024a), Intern-VL-Chat-1.5 (Chen et al., 2024b), CogVLM2 (Wang et al., 2023a), Phi-3-Vision (Abdin et al., 2024)). Our investigations reveal notable variations in model performance, highlighting great opportunities for improvement.

To address these challenges, we further propose to generate training data tailored for supervised fine-tuning (SFT), where we aim to refine the models' abilities to process and comply with multifaceted instructions. Results from our SFT experiments indicate a promising enhancement in the models' performance to strictly adhere to instructions, without hurting performance on other benchmarks.

Our contributions are summarized as follows. (i) We construct MIA-Bench, a new benchmark to comprehensively evaluate MLLMs on their capability to strictly adhere to instructions. (ii) We provide a detailed analysis of popular MLLMs, and suggest training methods for enhanced instruction following. For this purpose, we created training data and conducted experiments for additional supervised fine-tuning. MIA-Bench will be open-sourced, and we hope this benchmark can serve as a useful resource to stimulate further research on multimodal instruction adherence.

2 MIA-BENCH

MIA-Bench consists of 400 image-prompt pairs, with examples shown in Figure 3. The images are collected from diverse sources, including COCO 2017 validation set (Lin et al., 2015), SBU (Ordonez et al., 2011), TextVQA (Singh et al., 2019a), and Flickr. Images in the Flickr subset are photos of a variety of themes, including animals, art, architectures, text, food, math, etc. Images from the other three sources are randomly sampled from each corresponding source. Figure 4 shows the top 15 image content categories and the distribution of the 8 sub-instruction categories in MIA-Bench. The image content is labeled using GPT-4v. For each image, we manually write diverse and challenging instructions that contain multiple sub-instructions.

When constructing the instructions, we follow three principles, detailed below.

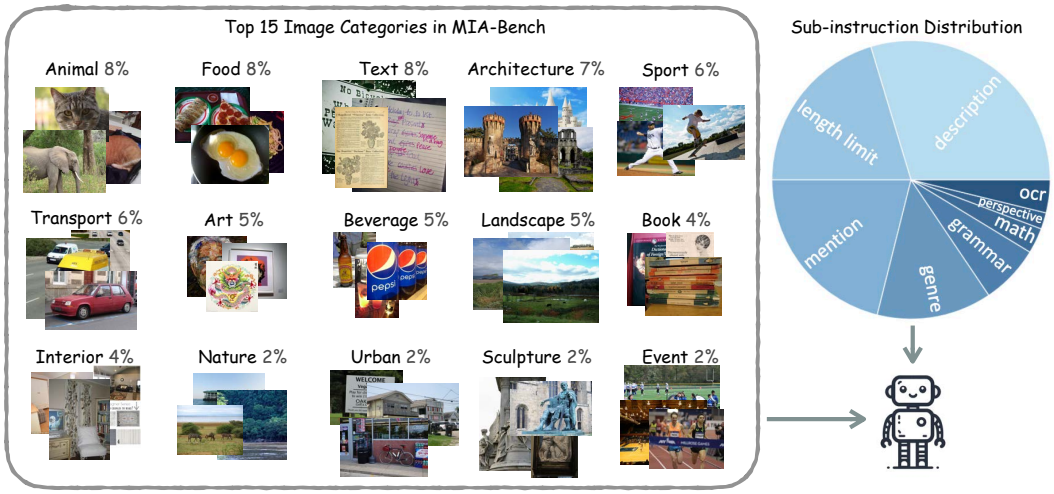

Figure 4: The top-15 image content categories and the distribution of the 8 sub-instruction categories in MIA-Bench.

- **Correctness.** The instruction needs to be answerable by humans. For example, asking about objects that do not exist in the image makes the prompt unanswerable.
- **No answer leakage.** The instruction should not contain the answer to itself. *'What color is the green object?'* is an example of answer leakage.
- **Image-dependent.** MMStar (Chen et al., 2024a) pointed out that on some multimodal benchmarks, MLLMs can generate correct answers without accessing images half of the time. Multi-modal capabilities are necessary to correctly answer MIA-Bench prompts.

2.1 Instruction Categories

In this paper, we use *instruction* to refer to the entire textual input, which in MIA-Bench can generally be viewed as a composition of multiple individual requests or constraints. We refer to these individual components as *sub-instruction*. Instructions in MIA-Bench are of diverse complexity, and sub-instructions contained are of multiple categories, summarized in Figure 4.

The sub-instructions in MIA-Bench can be categorized into *description*, *length limit*, *mention*, *genre*, *grammar*, *math*, *perspective*, and *OCR*, detailed below.

- *'description'* refers to describing a certain part of the image, with the exception of text-rich parts of the image, which falls under the *'OCR'*;
- *'length limit'* refers to the limitation of response length (*e.g.,* in exactly two sentences, using exactly 60 words);
- *'mention'* refers to mentioning or not mentioning certain objects or entities (*e.g.*, highlighting two similarities and one difference, comparing and contrasting the condition of the buildings with the activity on the street);
- *'genre'* refers to requests for a specific written form (*e.g.*, write a poem, write a narrative, with at least one pun included, all while weaving in a subtle theme of change);
- *'grammar'* refers to grammatical requirements (*e.g.*, use present tense, use capitalized letters, use integers);
- *'math'* refers to requirements to come up with a solution to math problems, or to identify errors in solutions to math problems, or to generate a valid math problem given table, charts, etc.;
- *'perspective'* refers to requirements specifying the viewpoint of an object or person in the image. This requires MLLMs to correctly identify what can or cannot be seen from the specified position,

and understand the spatial relationship of objects in its surrounding with itself (*e.g.*, imagine you are the lady in the image, describe what you can see without turning your head around);

- *'OCR'* refers to requirements related to understanding OCR information in text-rich images such as menus, tickets, bills, etc. For example, given a photo of a ticket, the sub-instruction asking about the price printed on the ticket falls into this category.

Figure 5 shows the most frequently used verbs and co-occurring nouns in MIA-Bench. To guarantee the diversity of prompts, when writing the instructions, we contribute instructions of various levels of complexity: *basic*, *intermediate*, *advanced*, *creative*, and *complex*. The *basic* category is the simplest; the instructions normally only contain one or two sub-instructions, such as "What is the color of the cat?", or "Describe the sofa in two words.". The *intermediate* category consists of instructions that contain three or more sub-instructions, but are in general easy for MLLMs to follow. The *advanced* category contains instructions that are challenging and contain three or more sub-instructions. The *creative* category contains instructions that instruct MLLMs to generate creative pieces of text, such as poems. The *complex* category is a combination of the previous two categories; the instructions in this category are the most complicated as they usually contain multiple challenging sub-instructions. While we found these categories useful to elicit a diverse instruction set, we also found that practical examples were often difficult to categorize objectively. As a result, we only used these categories for data collection, but are not reporting per-category results.

Figure 5: The most frequently used verbs and co-occurring nouns in MIA-Bench.

2.2 Response Evaluation Method

We adopt GPT-4o (OpenAI, 2024) to score MLLMs' responses on each instruction and return a total score using the following prompt:

> Here is the instruction for a multimodal LLM: {instruction} You need to grade if the response from the model follows each component of the instruction. The first component is: {sub-instruction 1}, and the second component is: {sub-instruction 2}, … , and the Nth component is: {sub-instruction N}. The response is: {response} You need to score the response and be strict. The total score ranges from 0 to 10, depending on if the response follows the instruction. The first, second, …, and Nth component is each worth {weight 1}, {weight 2}, …, and {weight N} scores. List scores of each component, and the total score in one sentence in this format: score of component 1: x/2, score of component 2: y/8, total score: z/10. Then explain

Each response is graded by first assessing how well it follows each sub-instruction, then computing the total score. Figure 6 shows an example of how responses from different MLLMs are evaluated and scored. Each sub-instruction in an instruction is assigned a maximum score ranging from 1 to 10; sum of the weight of all sub-instructions in an instruction is 10. The scoring system was carefully designed to reflect the complexity and importance of each task. For example, length limits are often binary in nature—either met or not met—hence the single point allocation. In contrast, a description task may require the model to handle multiple layers of complexity, including accuracy, detail, and relevance, which justified a higher score. For the example in Figure 6, there are 4 sub-instructions (denoted from S1 to S4); the first is worth 4 points and the rest is worth 2 points each. The response from GPT-4o partially follows the first sub-instruction which requires the response to be from the perspective of the dog, as the dog should not be able to see the car behind the man without turning around. The dog should be able to see the guitar, thus GPT-4o gets 2 points out of 4 for the first sub-instruction. It successfully follows the other 3 sub-instructions, achieving full score for them. Thus, the final score GPT-4o reaches is 8 out of 10. We always assign larger weight (6 if there are two

Instruction:

Imagine you are the dog in the image, list 2 objects you can see without moving your head. Respond in bullet points, within 20 words.

Four Sub-instructions (10 points):

S1: Answer from the dog's perspective, without moving its head, 4 points
S2: List 2 objects, 2 points
S3: Respond in bullet points, 2 points
S4: Limit response in 20 words, 2 points

GPT-4o (8 points):

- Guitar with a brown and white finish
- Car with a visible license plate

S1❌the dog can't see the car (2/4); S2✅ (2/2); S3✅ (2/2); S4✅ (2/2)

InternVL-Chat-V1.5 (8 points):

- The person holding me is playing a guitar.
- There's a black case in front of the person.

S1✅ (4/4); S2❌response contains 3 objects instead of 2 (0/2); S3✅ (2/2); S4✅ (2/2)

XComposer2 (8 points):

2 objects: the guitar and person playing it.

S1✅ (4/4); S2✅ (2/2); S3❌not in bullet points (0/2); S4✅ (2/2)

LLaVA-1.6-34b (10 points):

- Microphone stand
- Guitar

S1✅ (4/4); S2✅ (2/2); S3✅ (2/2); S4✅ (2/2)

Figure 6: An example with responses from four MLLMs and their evaluation scores.

sub-instructions, 4 if there are three or more sub-instructions) to the sub-instruction in the *description* category unless this category is absent in some cases, as usually a major part of the response is addressing this sub-instruction. For each MLLM, we compute the average score it gets on all 400 responses, and represent the ratio of the average score divided by 10 using percentage. We also compute the average score for each instruction category.

3 Experiments

In this section, we first present results of different MLLMs on MIA-Bench in Section 3.1, with additional supervised fine-tuning exploration in Section 3.2.

3.1 Benchmark Results

In total, we have evaluated 29 popular MLLMs on MIA-Bench. Results are reported in Table 1. Observations are summarized as follows.

- Overall, the best performance was achieved by GPT-4o (OpenAI, 2024), with a score 88.58, showcasing its superiority across different categories of instruction adherence.
- The ability to describe content accurately was best exhibited by Reka (Team et al., 2024). Other models like Claude-3-Opus (Anthropic, 2024), GPT-4v (Achiam et al., 2023) and GPT-4o also achieved scores higher than 90. This suggests that these models are good at generating coherent and contextually appropriate text.
- In the genre category, the highest proficiency was shown by GPT-4v and GPT-4o with scores above 94, suggesting an exceptional grasp of language nuances. Among open-source models, Phi-3-Vision (Abdin et al., 2024) and LLaVA-NeXT-34b (Liu et al., 2024a) show strong performance with scores of 86.52 and 86.58, respectively. The lowest score on this metric was by Kosmos-2 (Peng et al., 2023a), with a mere 11.55, pointing to difficulties in understanding or generating linguistically complex sentences.
- GPT-4o excelled in grammar with a score of 85.70, which indicates superior ability in syntax correctness and sentence structuring that matches specific requirements in the instruction. Among

Model	Meta-Avg	Description	Len-Limit	Genre	Grammar	Mention	Math	Perspective	OCR
Open Source 1b−8b									
Fuyu-8b (Bavishi et al., 2023)	24.52	52.06	24.52	17.06	17.18	36.43	22.62	66.67	33.09
Kosmos-2 (Peng et al., 2023a)	26.06	50.95	38.52	11.55	19.78	28.70	17.26	50.83	41.88
Sphinx (Lin et al., 2023b)	50.99	75.33	53.51	60.45	48.28	57.75	47.41	70.00	61.04
Idefics-2-8b (Laurençon et al., 2024)	51.42	59.37	62.73	48.07	64.09	46.20	46.51	48.33	61.97
mPLUG-Owl2 (Ye et al., 2023b)	57.86	75.01	65.25	63.39	60.26	57.70	57.22	65.00	62.08
CogVLM-Chat (Wang et al., 2023a)	58.95	60.42	57.86	67.94	60.55	62.92	36.67	60.83	61.87
ShareGPT4V (Chen et al., 2023b)	59.41	81.08	63.49	63.88	58.46	62.49	52.98	82.50	72.29
DeepSeek-VL-7b-chat (Lu et al., 2024)	60.96	86.31	63.26	72.11	54.79	63.75	67.39	74.17	77.85
LLaVA-1.5-7b (Liu et al., 2023c)	62.18	78.00	68.60	63.95	64.18	65.89	47.31	86.67	60.75
LLaVA-NeXT-7b-vicuna (Liu et al., 2024a)	62.27	79.21	68.01	65.63	60.95	63.33	46.67	90.00	65.54
Qwen-VL-Chat (Bai et al., 2023)	63.09	80.51	74.22	66.95	63.11	63.01	45.00	75.83	66.01
XComposer2-7b (Dong et al., 2024a)	67.71	83.47	76.16	73.66	67.69	67.01	48.61	77.50	68.06
CogVLM2 (Wang et al., 2023a)	73.43	87.60	74.52	83.47	71.97	77.01	71.53	90.83	87.16
Phi-3-vision (Abdin et al., 2024)	76.02	84.90	84.46	86.52	67.93	74.70	78.16	74.17	83.96
MiniCPM-Llama3-v2.5 (Hu et al., 2023)	76.27	84.12	79.44	80.33	81.25	76.99	64.08	81.67	76.59
Open Source 8b−13b									
InstructBLIP-13b (Dai et al., 2023)	38.16	50.54	39.57	29.34	38.43	42.28	12.50	50.00	30.42
LLaVA-1.5-13b (Liu et al., 2023c)	63.55	80.98	70.15	64.54	59.30	67.42	45.11	69.17	76.28
LLaVA-NeXT-13b-vicuna (Liu et al., 2024a)	69.16	86.75	69.88	82.07	64.77	74.99	48.56	77.50	75.83
Open Source 13b−110b									
Yi-VL-34b (AI et al., 2024)	53.90	74.89	52.05	59.09	55.91	57.25	54.17	41.85	70.09
InternVL-Chat-v1.5 (Chen et al., 2024b)	75.42	89.13	78.21	79.92	78.16	77.54	76.11	87.50	80.92
LLaVA-NeXT-34b (Liu et al., 2024a)	75.61	88.02	83.50	86.58	71.57	75.83	68.06	87.50	80.26
LLaVA-NeXT-110b (Liu et al., 2024a)	79.84	86.99	84.86	82.49	79.04	80.10	71.94	80.83	75.45
Proprietary									
Gemini-1.0-Pro (Team, 2023)	70.63	82.77	72.83	78.76	76.91	71.67	81.45	89.29	84.11
Reka-Core (Team et al., 2024)	76.95	**91.05**	79.91	85.16	78.98	82.08	82.53	77.50	81.08
Claude-3-Haiku (Anthropic, 2024)	78.25	86.86	77.53	90.27	73.41	82.62	82.22	57.50	86.49
Claude-3-Sonnet (Anthropic, 2024)	79.44	88.06	82.71	90.54	79.60	82.05	82.22	76.67	84.43
Claude-3-Opus (Anthropic, 2024)	84.50	90.50	86.03	91.19	83.82	85.49	85.92	65.00	**86.84**
GPT-4v (Achiam et al., 2023)	86.11	90.03	87.61	**94.59**	80.12	89.37	85.63	59.17	85.26
GPT-4o (OpenAI, 2024)	**88.58**	90.82	**92.73**	94.29	**85.70**	**90.66**	**87.07**	**92.50**	86.54

Table 1: Evaluation results of a wide array of MLLMs on MIA-Bench.

the open-source models, MiniCPM-Llama3-V-2.5 (Hu et al., 2023) is notable with a score of 81.25. Contrastingly, Fuyu-8b (Bavishi et al., 2023) scored the lowest with 17.18, reflecting major challenges in grammar adherence.

- GPT-4o also showed the best performance with a score of 92.73 in respecting prescribed length limits, which is crucial for tasks requiring concise and precise answers. Among open-source models, LLaVA-NeXT-110b (Liu et al., 2024a) stands out with a score of 84.86.
- Results from LLaVA series also suggest a strong correlation between LLM size and MIA-Bench performance across metrics.

Correlation with other benchmarks. In Table 2, we compare the ranking of 5 state-of-the-art MLLMs on MIA-Bench as well as their meta ranking on MME (Fu et al., 2023a), MMMU (Yue et al., 2023) , MMBench (Liu et al., 2024b), MMVet (Yu et al., 2023b), HallusionBench (Guan et al., 2023), and MathVista (Lu et al., 2023) (meta ranking is computed by averaging rankings across these benchmarks). Our findings reveal a discrepancy between the two sets of rankings. Notably,

Model	MME	MMMU	MMB	MMVet	HallB	Math Vista	Meta Ranking	MIA	MIA Ranking
GPT-4o (OpenAI, 2024)	2328.7	69.1	83.3/82.1	66.5	67.5	63.8	1	88.58	1
GPT-4v (Achiam et al., 2023)	1926.6	56.8	77/74.4	67.6	46.5	49.9	3	86.11	2
Gemini-1.0-Pro (Team, 2023)	1933.4	47.9	73.6/74.3	64.3	45.2	45.2	5*	70.63	6
Claude-3-Opus (Anthropic, 2024)	1586.8	59.4	63.3/59.2	58.1	37.8	50.5	5*	84.50	3
InternVL-Chat-V1-5 (Chen et al., 2024b)	2187.8	45.2	82.2/82	62.8	49.3	53.5	2	75.42	5
LLaVA-NeXT-34b (Li et al., 2024a)	2028	51.1	81.1/79	48.9	47.6	47.7	4	75.61	4

Table 2: Meta ranking of five state-of-the-art MLLMs on existing multimodal benchmarks compared with their ranking on MIA-Bench.

Model	Total Score	Description	Length Limit	Genres	Grammar	Mention	Math	Perspective	OCR
LLaVA-NeXT-13b (Liu et al., 2024a)	69.16	86.75	69.88	82.07	64.77	74.99	48.56	77.50	75.83
+ DIT	78.85	86.90	86.80	88.02	71.34	81.01	60.87	84.17	72.65
+ DIT + LLaVA-Instruct150k	78.90	88.59	74.67	79.95	74.17	66.39	53.70	100.00	80.83

Table 3: Detailed results of LLaVA-NeXT-13b (Liu et al., 2024a) on MIA-Bench before and after supervised fine-tuning on additional constructed diverse instruction-tuning (DIT) data, and the mixture of DIT and LLaVA-Instruct150k. We re-ran the baseline.

InternVL-Chat-V1.5 (Chen et al., 2024b), which holds the highest meta-ranking among the five MLLMs on the other benchmarks, ranks the lowest on MIA-Bench. Conversely, Claude-3-Opus, which has the lowest meta-ranking, secures the second position on the MIA-Bench. This indicates that excelling in tasks evaluated by existing benchmarks does not necessarily translate to superior instruction adherence capability assessed by MIA-Bench.

Correlation with LLM backbone performance. To determine if the performance on MIA-Bench is attributable solely to the underlying LLMs, we also evaluate several MLLMs on IFEval (Qin et al., 2024), a benchmark that assesses the instruction adherence capability of LLMs, and compare their ranking with that on MIA-Bench. This comparison is shown in Appendix, which shows that the instruction adherence capabilities of MLLMs do not consistently align with their LLMs' adherence capability.

Other external models as the judge. Since the evaluation uses GPT-4o as the judge, it is natural to conjecture that GPT-4o may favorably score its own responses. To alleviate this concern, we use Claude-3, a strong performer in Table 1, to evaluate responses from GPT-4o and itself, and compare their scores with each other. The prompt used to grade responses is the same as the one used in GPT-4o grading. We find that even using Claude-3 Opus to score its own and GPT-4o's responses, GPT-4o still achieves a superior score. When scored by Claude-3-Opus, GPT-4o achieves 89.84 score in contrast to Claude-3-Opus' 85.89. Based on this observation, we use GPT-4o for evaluation by default, and observe that results from multiple runs may have around 1% variation.

3.2 Supervised Fine-Tuning (SFT)

The performance of small-scale models such as LLaVA-NeXT-13b is less ideal on MIA-Bench. In this section, we study the use of supervised fine-tuning to enhance model performance.

Additional SFT data construction. First, we randomly sample 1000 images from COCO 2017 training set, and use GPT-4v to generate five instructions for each image, using the prompt below.

> Sometimes multimodal LLM doesn't follow user's instruction. For example, when I ask a model to use three sentences to answer my question, it might answer in four sentences. I want to evaluate models' ability to strictly follow instructions. Help me compose instructions that are of five levels of difficulty to follow, for this image.
> The five levels are 'Basic Instruction, Intermediate Instruction, Advanced Instruction, Creative Instruction, Complex Instruction'. Diversify when composing instructions. After each of your instructions, add ***.

Model	MMBench	TextVQA	VQA2	LLaVA-itw	POPE	VizWiz	MIA-Bench
LLaVA-NeXT-13b (Liu et al., 2024a)	70.6	64.26	82.80	85.8	87.7	60.41	69.16
+ DIT	68.6	63.20	82.58	83.4	86.9	59.72	78.85
+ DIT + LLaVA-Instruct150k	67.27	54.24	77.92	75.8	87.8	58.87	78.90

Table 4: Results of LLaVA-NeXT-13b (Liu et al., 2024a) on MIA-Bench and other major multimodal benchmarks supervised fine-tuning on additional constructed diverse instruction-tuning (DIT) data, and the mixture of DIT and LLaVA-Instruct150k. We re-ran the baseline.

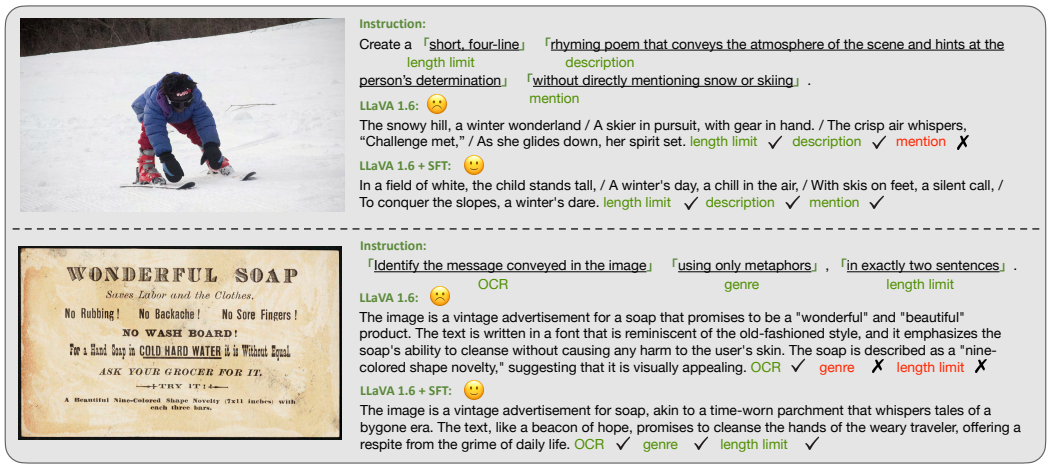

Figure 7: Examples of LLaVA-NeXT-13b responses before and after supervised fine-tuning on additional diverse instruction-tuning data.

We then manually process the generated instructions. The cleaned data for SFT consists of 5000 image-prompt pairs.

Then, we use GPT-4v to generate responses to the constructed prompts. To evaluate the quality of these responses, we sampled 100 responses and manually checked if they adhere to the instructions. We find that 90% of the responses successfully followed all instructions in the prompt, serving as a proper ground-truth response for model training. Examples of this additional training data is provided in the Appendix.

Results. Using LLaVA-NeXT-13b as the backbone, we train the model for 1 epoch on the constructed diverse instruction-tuning (DIT) data. We also performed SFT using the combination of LLaVA Visual Instruct 150K dataset and our diverse instruction-tuning dataset, to examine which data mixture leads to better performance. Results on MIA-Bench and other benchmarks are summarized in Table 4, with detailed results on MIA-Bench reported in Table 3. The performance of the model after finetuning on the DIT data alone on MIA-Bench has been significantly improved by around 10 points, at the cost of minor regressions across other benchmarks. This supports our hypothesis that using the small amount of DIT data, we can enhance the model with better instruction following capability. On the other hand, further adding LLaVA Visual Instruct 150K data did not significantly improve model performance on the tested benchmarks. Examples are shown in Figure 7 to compare responses from LLaVA-NeXT-13b before and after SFT.

4 Related Work

Multimodal LLMs and Benchmarks. Multimodal Large Language Models (MLLMs) have recently emerged as a significant research focus. LLaVA (Liu et al., 2023c) and MiniGPT-4 (Zhu et al., 2023) pioneered visual instruction tuning, and the past year has witnessed a boom of open-source MLLMs based on this concept. Prominent examples include InstructBLIP (Dai et al., 2023), mPLUG-

Owl(-2/Doc) (Ye et al., 2023b;c;a), Qwen-VL (Bai et al., 2023), CogVLM (Wang et al., 2023a), SPHINX(-X) (Lin et al., 2023b; Gao et al., 2024), InternLM-XComposer2-VL (Dong et al., 2024b), InternVL(-1.5) (Chen et al., 2023c; 2024b), VILA (Lin et al., 2023a), MM1 (McKinzie et al., 2024), Mini-Gemini (Li et al., 2024b), Idefics2 (Laurençon et al., 2024), Phi-3-vision (Abdin et al., 2024), to name a few. There is also a rich body of literature on enabling MLLMs for referring and grounding (Peng et al., 2023b; Chen et al., 2023a; You et al., 2024a; Wang et al., 2023b; Lai et al., 2023; Zhang et al., 2023; 2024; You et al., 2024b), image generation and editing (Koh et al., 2023; Sun et al., 2023a; Fu et al., 2023b), *etc*.

Various benchmarks have been proposed to evaluate the performance of MLLMs across different dimensions. Benchmarks like VQAv2 (Goyal et al., 2017), TextVQA (Singh et al., 2019a), ScienceQA (Lu et al., 2022), MME (Fu et al., 2023a), MMbench (Liu et al., 2024b), SEED-Bench (Li et al., 2023a), MathVista (Lu et al., 2023), and MMMU (Yue et al., 2023) aim to assess comprehensive multimodal understanding abilities. Additionally, there are benchmarks that specifically study model hallucination, including POPE (Li et al., 2023b), MHalDetect (Gunjal et al., 2024), GAVIE (Liu et al., 2023a), HallusionBench (Guan et al., 2023), and MAD-Bench (Qian et al., 2024). Many of these benchmarks have gained popularity within the community due to their use of multiple-choice evaluations. However, they do not accurately reflect the common use cases for MLLMs, where user interactions are typically open-ended. To address this, benchmarks like LLaVA-Bench (Liu et al., 2023c), MM-Vet (Yu et al., 2023b), and Vibe-Eval (Padlewski et al., 2024) have been proposed. Our MIA-Bench also falls into this category; however, we focus on studying the exact instruction adherence of MLLMs, a metric that previous benchmarks have only loosely measured.

Instruction Following Benchmarks for LLMs. Several benchmarks have been proposed to measure the instruction adherence ability of LLMs. Instruction-Following Eval (IFEval) (Zhou et al., 2023) is a benchmark for assessing LLMs' adherence ability to the given instructions. Its approach emphasizes verifiable instructions, which enhance objectivity and reproducibility in evaluations. IFEval creates 541 prompts spanning 25 instruction types, revealing a significant performance gap in instruction adherence ability between GPT-4 (Achiam et al., 2023) and PaLM-2 (Anil et al., 2023). This demonstrates the benchmark's ability to effectively differentiate between models in adherence ability. On the other hand, InfoBench (Qin et al., 2024) introduces a new metric called Decomposed Requirements Following Ratio (DRFR) for assessing the instruction-adherence capabilities of LLMs. DRFR dissects complex instructions into simpler sub-instructions, allowing for a granular evaluation of compliance with various task aspects. InfoBench contains 500 diverse instructions consisting of 2,250 decomposed questions in multiple constraint categories. The evaluation of advanced LLMs using this framework highlights their strengths and areas for improvement, especially in complex instruction adherence scenarios. Compared with these previous work, we are the first known effort that specifically focuses on benchmarking the instruction adherence ability of multimodal LLMs.

5 Conclusion

This paper introduces MIA-Bench, a benchmark designed to evaluate the ability of MLLMs to strictly adhere to complex instructions within prompts. Through the analysis of 400 image-prompt pairs from diverse sources, our findings highlight variability in model performance and much room for improvement, underscoring a critical need for enhanced training methods to improve instruction compliance. We further explored supervised fine-tuning (SFT) using LLaVA-NeXT as the backbone, which yielded promising results. Going forward, future research can expand on both SFT and alignment methods such as RLHF (Ouyang et al., 2022; Sun et al., 2023b) and DPO (Rafailov et al., 2024; Yu et al., 2023a), enhancing MLLMs to achieve higher accuracy and reliability in practical applications across diverse instructional contexts.

Limitation

In designing the instructions for our benchmark, we incorporated a wide range of categories to enhance the diversity of sub-instructions. Nonetheless, the real world presents an infinite variety of instructions, many of which may pose significant challenges for MLLMs.

REPRODUCIBILITY STATEMENT

For reproducibility purpose, we release our evaluation code at https://anonymous.4open.science/r/anonymous-mia-bench-D0F4/evaluation.ipynb and benchmark at: https://anonymous.4open.science/r/anonymous-mia-bench-D0F4/instruction_benchmark_all.json.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023c.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024b.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*, 2023.

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024a.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024b.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023a.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023b.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, et al. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. *arXiv e-prints*, pp. arXiv–2311, 2023.

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. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

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

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

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024a. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

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

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024b.

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

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2015.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023b.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

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

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*, 2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023a.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023b.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2024.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019b.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023b.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023a.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023b.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023a.

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

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023c.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024a.

Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms. *arXiv preprint arXiv:2404.05719*, 2024b.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023a.

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

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

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

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

A Appendix

A APPENDIX

A.1 COMPARISON BETWEEN MIA-BENCH AND IFEVAL

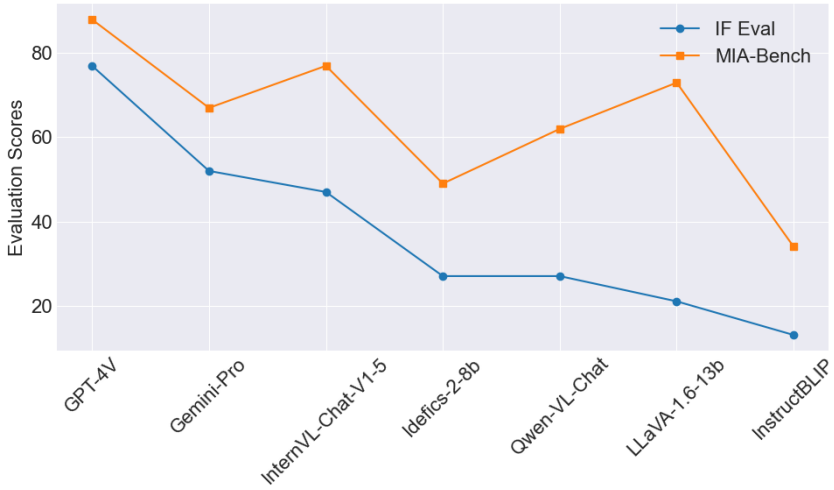

Figure 8: Comparison between MIA-Bench and IFEval.

A.2 EXAMPLES OF DIVERSE INSTRUCTION-TUNING DATA USED FOR SFT

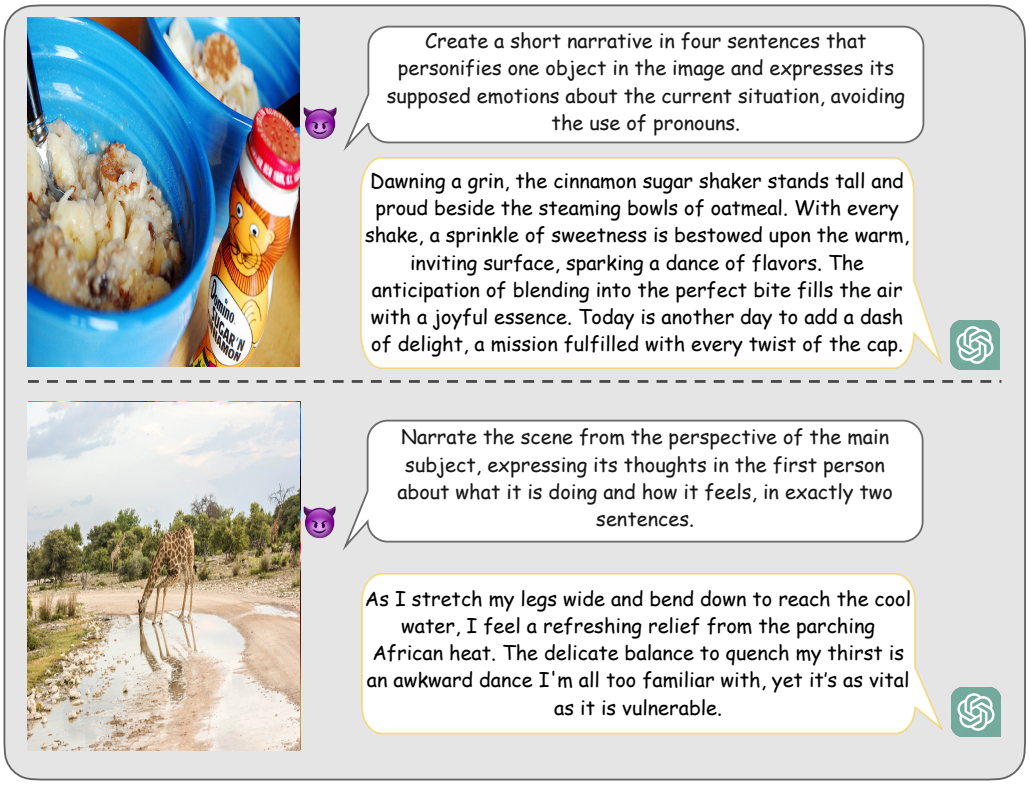

Figure 9: Examples of diverse instruction-tuning data used for SFT.

A.3 EXAMPLES OF HOW MLLMS RESPOND TO INSTRUCTIONS IN MIA-BENCH

Instruction:
Imagine a short story of two sentences about why the animal is in its current location, but you must include a weather condition, a time of day, and an emotion. components: ["Imagine a short story about why the animal is in its current location", "of two sentences", "you must include a weather condition, a time of day, and an emotions"], component weight: [4, 3, 3]

XComposer2-7b:
On a sunny afternoon, the polar bear, feeling relaxed, lounges on a rock in its enclosure, enjoying the warmth of the sun.
Score of component 1: 4/4, score of component 2: 0/3, score of component 3: 3/3, total score: 7/10.
Reasons: Component 1: The story explains why the polar bear is in its current location (lounging on a rock in its enclosure enjoying the warmth of the sun). Full score awarded. Component 2: The response is not two sentences long. Component 3: The story includes a weather condition (sunny), a time of day (afternoon), and an emotion (relaxed). Full score awarded.

LLaVA-NeXT 110b:
The polar bear is lounging on the rocks by the pool, enjoying the warmth of the sun on its fur. It's a beautiful day at the zoo, and the bear seems content in its enclosure.
Score of component 1: 3/4, score of component 2: 3/3, score of component 3: 2/3, total score: 8/10.
Reasons: Component 1 (short story about why the animal is in its current location): The response provides context about why the polar bear is in its current location (lounging by the pool on a beautiful day at the zoo). However, it could be more detailed in explaining why it's there specifically. Component 2 (two sentences): The response is effectively two sentences. Component 3 (weather condition, time of day, and emotion): The response includes a weather condition (warmth of the sun, beautiful day), a time of day (implicitly daytime implied by the sun), and an emotion (content). However, the time of day could be made more explicit.

Qwen-VL-Chat:
The polar bear is resting on the rock in the sun because it's a warm summer day. It feels relaxed and comfortable in its natural habitat.
Score of component 1: 4/4, score of component 2: 3/3, score of component 3: 2/3, total score: 9/10.
Reasons: Component 1: The story does explain why the polar bear is in its current location. Component 2: The response is exactly two sentences in length, fully meeting this requirement. Component 3: The response includes a weather condition ("warm summer day") and an emotion ("relaxed and comfortable"). However, it does not clearly specify a time of day, losing 1 point.

GPT-4v:
Under the harsh midday sun, the polar bear lounged by the water's edge, seeking relief from the sweltering heat. Despite the warmth, a glimmer of contentment shone in its eyes as it enjoyed a moment of quiet repose.
Score of component 1: 4/4, score of component 2: 3/3, score of component 3: 3/3, total score: 10/10.
Reasons: Component 1: The response explains why the polar bear is at the water's edge, seeking relief from the heat. Full marks. Component 2: The response has exactly two sentences. Full marks. Component 3: The response includes a weather condition (harsh midday sun), a time of day (midday), and an emotion (contentment). Full marks.

Figure 10: An example of how MLLMs respond to instructions in MIA-Bench.

Instruction:
Explain the message conveyed by the text on the sign in the image, using exactly two sentences.
components: ["Explain the message conveyed by the text on the sign in the image", "using exactly two sentences"], component weight: [6, 4]

InternVL-Chat-V1-5:
The text on the sign in the image conveys a message of stopping and being cautious, as it reads "STOP" and "VOLDEMICKY." The word "VOLDEMICKY" is a play on the word "volde-mekky," which is a Dutch term for a person who is easily influenced or manipulated. The message suggests that one should be careful not to be easily swayed or manipulated by others.
Score of component 1: 3/6, score of component 2: 0/4, total score: 3/10.
Reason: Component 1: The response explains the message of stopping and being cautious, but it misinterprets "VOLDEMICKY" as "volde-mekky" and provides an incorrect definition. Additionally, there seems to be a misunderstanding as "VOLDEMICKY" might be intended as a playful reference to a combination of Voldemort and Mickey Mouse, neither of which is explained.\n- Component 2: The response uses three sentences instead of the required two, hence it does not follow the instruction of using exactly two sentences.

Claude Opus:
The text on the illuminated stop sign conveys an emphatic command to halt or cease movement. The unusual octagonal shape and red color of the sign reinforce the urgent, attention-grabbing nature of the message to stop immediately.
Score of component 1: 4/6, score of component 2: 4/4, total score: 8/10.
Reason: The response explains the message conveyed by the text on the stop sign effectively, noting both the command to stop and the reinforcement by the shape and color of the sign. However, it doesn't address the text "Voldemort" beneath the word "STOP," which is a crucial part of the message. The response uses exactly two sentences, fulfilling the second component perfectly.

Gemini:
The sign says "Stop Voldemort!" It is a reference to the Harry Potter series, in which Voldemort is the main antagonist.
Score of component 1: 6/6, score of component 2: 4/4, total score: 10/10.
Reason: The response successfully explains the message conveyed by the text on the sign, which is to "Stop Voldemort" and relates it to the Harry Potter series where Voldemort is the main antagonist. Additionally, the response adheres to the requirement of using exactly two sentences. Therefore, it fulfills both components of the instruction fully.

GPT-4v:
The sign is a modification of the traditional stop sign, introducing the words "Stop Violence." This is a call to action to end violence and can be interpreted as a public message advocating for peace and safety.
Score of component 1: 2/6, score of component 2: 4/4, total score: 6/10.
Reason: The response does follow the second component perfectly by using exactly two sentences. However, it does not correctly explain the message on the sign. The sign in the image has the text "Stop Voldemort," not "Stop Violence," indicating a message related to a fictional character rather than a general call to end violence. Hence, the response accurately addresses the number of sentences but fails significantly in conveying the correct message from the sign.

Figure 11: Another example of how MLLMs respond to instructions in MIA-Bench.

Instruction:
Describe the scene shown in the image using exactly ten words.
components: ["Describe the scene shown in the image", "using exactly ten words"], component weight: [6, 4]

InternVL-Chat-V1-5:
Elephants drinking water in a field.
Score of component 1: 6/6, score of component 2: 0/4, total score: 6/10.
Reason: The response accurately describes the scene shown in the image with elephants drinking water in a field, fulfilling the first component. However, it does not use exactly ten words, thus failing the second component.

XComposer2-7b:
A herd of elephants walking through a field.
Score of component 1: 6/6, score of component 2: 0/4, total score: 6/10.
Reason: The response correctly describes the scene depicted in the image. The only issue is that the response does not use exactly ten words, hence it's not fully compliant with the instruction.

LLaVA-NeXT 110b:
Elephant, water, trees, grass, sky, dirt, hill, clouds, reflection, wildlife, natural.
Score of component 1: 4/6, score of component 2: 0/4, total score: 4/10.
Reason: The response describes key elements present in the image, such as "Elephant, water, trees, grass, sky, dirt, hill, clouds, reflection, wildlife," but it leaves out some critical relational or scene-setting words that fully describe the image scene, resulting in a score of 4/6 for component 1. The response uses eleven words instead of ten ("water, trees, grass, sky, dirt, hill, clouds, reflection, wildlife, and natural"), resulting in a score of 0/4 for component 2.

Claude Opus:
Elephant drinking water from shallow pond in grassy African savanna.
Score of component 1: 6/6, score of component 2: 4/4, total score: 10/10.
Reason: The description "Elephant drinking water from shallow pond in grassy African savanna." accurately describes the scene shown in the image, fulfilling the first component of the instruction. Additionally, the response is exactly ten words, satisfying the second component. Therefore, the response fully follows both instructions and earns a perfect score.

GPT-4v:
Elephant drinks at waterhole; herd in African savanna background.
Score of component 1: 6/6, score of component 2: 0/4, total score: 6/10.
Reason: The response accurately describes the scene shown in the image (an elephant drinking at a waterhole with a herd in the African savanna background). However, the response uses nine words instead of the required ten words. Therefore, full points are awarded for the first component, but no points are given for the second component.

Figure 12: The third example of how MLLMs respond to instructions in MIA-Bench.