Difficult Task Yes but Simple Task No: Unveiling the Laziness in Multimodal LLMs

Sihang Zhao[1], Youliang Yuan[2], Xiaoying Tang[1], and Pinjia He[†2]

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[2]School of Data Science, The Chinese University of Hong Kong, Shenzhen
{sihangzhao, youliangyuan}@link.cuhk.edu.cn
{tangxiaoying, hepinjia}@cuhk.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) demonstrate a strong understanding of the real world and can even handle complex tasks. However, they still fail on some straightforward visual question-answering (VQA) problems. This paper dives deeper into this issue, revealing that models tend to err when answering easy questions (e.g., Yes/No questions) about an image, even though they can correctly describe it. We refer to this model behavior discrepancy between difficult and simple questions as model *laziness*. To systematically investigate model laziness, we manually construct *LazyBench*, a benchmark that includes Yes/No, multiple choice, short answer questions, and image description tasks that are related to the same subjects in the images. Based on *LazyBench*, we observe that laziness widely exists in current advanced MLLMs (e.g., GPT-4o, Gemini-1.5-pro, Claude 3, LLaVA-1.5, LLaVA-1.6, and QWen-VL). We also analyzed the failure cases of LLaVA-1.5-13B on the VQA-v2 benchmark and discovered that about half of these failures are due to the model's laziness. This further highlights the importance of ensuring that the model fully utilizes its capability. To this end, we conduct a preliminary exploration of how to mitigate laziness and find that chain of thought can effectively avoid this issue. The data can be accessed at https://github.com/Akutagawa1998/LazyBench.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2023c) integrate multimodal content such as images into large language models (LLMs) (Touvron et al., 2023). Represented by OpenAI's GPT-4 (OpenAI, 2023b), MLLMs have demonstrated impressive capabilities across various complex multimodal tasks (OpenAI, 2023a; Yang et al., 2023). However, existing research indicates that even state-of-the-art MLLMs still suffer from some straightforward visual questions (e.g., "Is the door of the truck cab open?" for an image of a truck.) (Tong et al., 2024). A natural question arises:

> *Why do MLLMs struggle with these easy questions?*

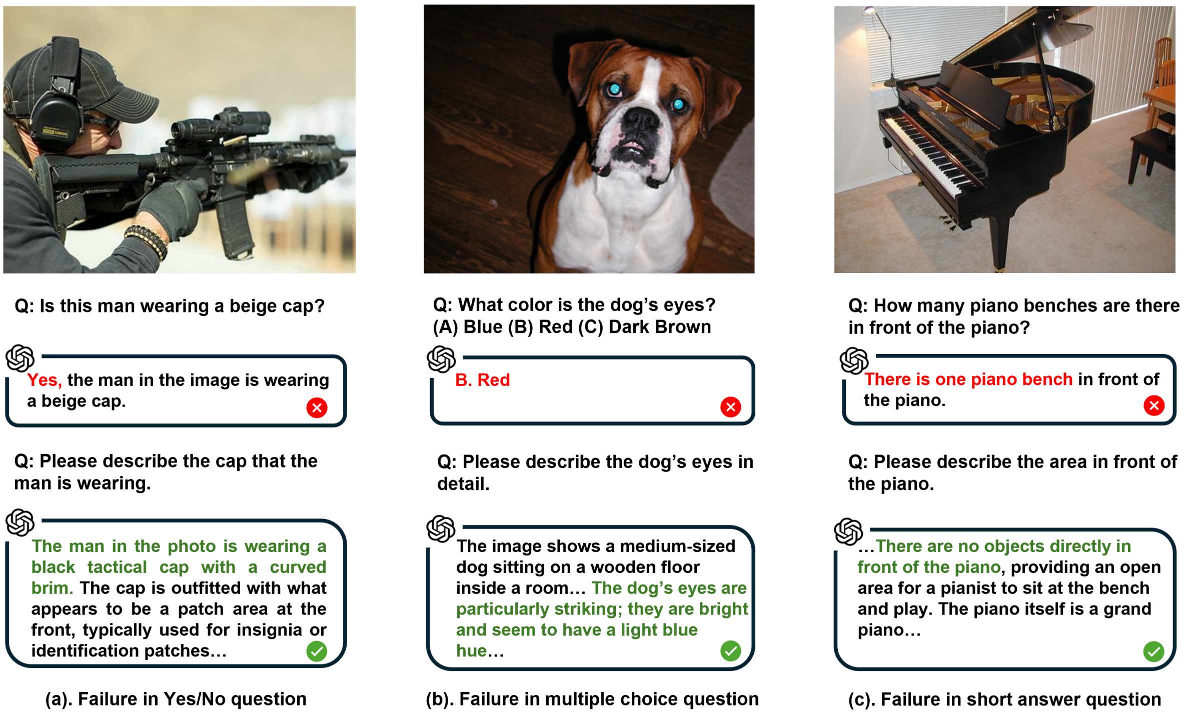
In this work, we dive deeper to explore this question and find that MLLMs often struggle with simple questions (like Yes/No questions) about an image, even though they can accurately describe the image itself. For example, as present in Figure 1, when we asked GPT-4V, "`Is this man wearing a beige cap?`" GPT-4V answered "`Yes`", which is incorrect. In contrast, when we asked it a similar but more difficult question, "`Please describe the cap that the man is wearing`", GPT-4V correctly described its color. In this paper, we describe this phenomenon where MLLMs perform well on the description tasks but make mistakes on simpler tasks as model **laziness**[1].

To systematically study model laziness, we manually construct a benchmark called **LazyBench**. We found image pairs encoded as "similar images" by the pretrained Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) model and designed simple Yes/No questions on their visual differences. We collected images from where GPT-4V (website version) fails in the above-mentioned questions. Then for each image, we handcraft three different types of questions about the same subject of the Yes/No question: multiple choice, short answer question, and a description task. We use LazyBench to evaluate advanced closed-source models like GPT-4o, GPT-4V (OpenAI, 2023a), Gemini-1.5-pro (Reid et al., 2024), and Claude 3 (Anthropic, 2024), and open-source models like

[†]Pinjia He is the corresponding author.

[1]The explanation of why we define description tasks as the harder tasks can be found in Appendix A.1

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7535–7548
November 12-16, 2024 ©2024 Association for Computational Linguistics



Q: Is this man wearing a beige cap?

Yes, the man in the image is wearing a beige cap. ✗

Q: Please describe the cap that the man is wearing.

The man in the photo is wearing a black tactical cap with a curved brim. The cap is outfitted with what appears to be a patch area at the front, typically used for insignia or identification patches... ✓

(a). Failure in Yes/No question

Q: What color is the dog's eyes? (A) Blue (B) Red (C) Dark Brown

B. Red ✗

Q: Please describe the dog's eyes in detail.

The image shows a medium-sized dog sitting on a wooden floor inside a room... The dog's eyes are particularly striking; they are bright and seem to have a light blue hue... ✓

(b). Failure in multiple choice question

Q: How many piano benches are there in front of the piano?

There is one piano bench in front of the piano. ✗

Q: Please describe the area in front of the piano.

...There are no objects directly in front of the piano, providing an open area for a pianist to sit at the bench and play. The piano itself is a grand piano... ✓

(c). Failure in short answer question

Figure 1: MLLMs sometimes fail to correctly answer straightforward Yes/No or multiple-choice questions based on images. However, they often manage to avoid these errors when describing the images. We refer to this phenomenon as "model laziness."

LLaVA-1.5 (Liu et al., 2023c), LLaVA-1.6 (Liu et al., 2024) and QWen-VL (Bai et al., 2023). The results show that these state-of-the-art MLLMs significantly suffer from laziness: they show a low accuracy on Yes/No questions (e.g., GPT-4V: 28.72%, Claude 3: 34.66%), and multiple choice questions (e.g., GPT-4V: 54.45%, Claude 3: 55.45%), while performing significantly better on the corresponding description tasks (e.g., GPT-4V: 71.28%, Claude 3: 57.43%).

We further explore to what extent MLLM laziness is prevalent in the widely used visual question-answering (VQA) benchmarks. To this end, we propose a simple LLM-based framework that automatically evaluates the extent of laziness in their failure cases.

We find that 41.15% failure cases of LLaVA-1.5-13B on VQA-v2 are caused by model laziness. We believe this provides valuable insights into the way to improve the capability of MLLMs: in addition to allowing MLLMs to learn more knowledge, it is equally important to ensure that MLLMs are fully utilizing the knowledge learned.

To mitigate the influence caused by model laziness in simple tasks, we implemented a chain of thought (CoT) (Wei et al., 2022) based method to make the task "harder". We require MLLMs to handle the description task first before answering a Yes/No or a multiple-choice question. The results show that our method fixed around 40% cases of laziness and effectively improved MLLMs' performance in those tasks.

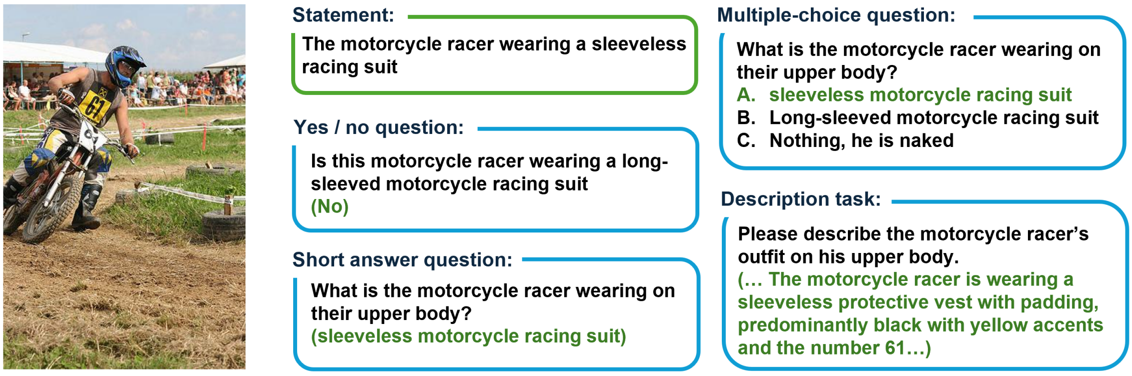
In summary, our contributions are listed below:

- We conduct an in-depth study on the phenomenon of MLLMs making errors on easy questions, discovering that current advanced MLLMs exhibit significant laziness.
- We manually construct a dataset called LazyBench to investigate the laziness phenomenon in MLLMs.
- We provide a CoT-based method that can effectively prevent models from being lazy.

2 Related Work

2.1 Visual Question and Answering

With the success of LLMs, increasing attention has been given to integrating visual embeddings into language models. Initially, researchers applied transformers to connect visual encoders with



Statement:

The motorcycle racer wearing a sleeveless racing suit

Yes / no question:

Is this motorcycle racer wearing a long-sleeved motorcycle racing suit
(No)

Short answer question:

What is the motorcycle racer wearing on their upper body?
(sleeveless motorcycle racing suit)

Multiple-choice question:

What is the motorcycle racer wearing on their upper body?
A. sleeveless motorcycle racing suit
B. Long-sleeved motorcycle racing suit
C. Nothing, he is naked

Description task:

Please describe the motorcycle racer's outfit on his upper body.
(... The motorcycle racer is wearing a sleeveless protective vest with padding, predominantly black with yellow accents and the number 61...)

Figure 2: The green box represents a correct, brief statement about the "question subject" in the image. The blue box contains four different types of questions about this subject (Yes/No, multiple-choice, short-answer questions, and descriptive requests). They are used to evaluate the model's laziness, and the construction of these questions is described in Section 3.2.

LLMs, pretraining them on image-text matching datasets (Lin et al., 2014; Krishna et al., 2017; Changpinyo et al., 2021) and fine-tuning them on specific datasets (e.g., VQA (Antol et al., 2015), VQA-v2 (Goyal et al., 2017)). Then, to improve MLLMs' performances and generalization abilities, researchers began using VQA format data for instruction tuning (Liu et al., 2023c). Despite MLLMs showing considerable capabilities in some complex VQA tasks (Fu et al., 2022; Hu et al., 2022, 2023b; Fu et al., 2023a,b), these studies seem to focus primarily on the textual reasoning abilities of MLLMs (Wei et al., 2022), rather than on whether MLLMs are truly extracting information from the images. Our work bridges this gap by studying the model laziness.

2.2 Benchmarks for Visual Perceptions

Increasing attention is being given to the evaluation of MLLMs' visual perception. Tong et al., 2024 suggest that due to encoding flaws in the CLIP pretrained model, CLIP-based MLLMs might make mistakes on some simple questions. POPE (Li et al., 2023) and NOPE (Lovenia et al., 2023) designed questions about the presence or absence of objects in images to measure MLLM hallucination; however, these consist solely of Yes/No questions. Hallusibench (Liu et al., 2023a) provides a benchmark for evaluating MLLMs' hallucinations across different tasks. MathVerse (Zhang et al., 2024) is a benchmark for visual problems in mathematical domains such as tables and charts. It reveals that MLLMs may not be thoroughly reading these charts, but they lack analysis of simpler and more straightforward VQA tasks. LazyBench is the first benchmark to focus on the consistency of MLLMs' answers to the same question about the same subject in the same image when asked in different forms.

3 MLLMs Are Being Lazy

To thoroughly understand and analyze the lazy phenomenon, where MLLMs perform well on descriptive tasks but fail on simple tasks, we construct the **LazyBench** benchmark. Therefore, in this section, we first introduce the methods and steps for constructing LazyBench. Subsequently, we measure the extent of the lazy phenomenon of current state-of-the-art MLLMs on LazyBench. Finally, we use a CoT-based method to mitigate the MLLMs laziness.

3.1 Samples of LazyBench

Each item in **LazyBench** consists of an image, a ground truth statement, and 4 different questions (i.e., Yes/No, multiple-choice, short answer, description) together with their ground truth answers. For instance, in Figure 2, for the image, there are:

- One Yes/No question: "Is this motorcycle racer wearing a long-sleeved motorcycle racing suit?" and its ground truth answer is "No".
- One multiple-choice question has 3 options: "sleeveless motorcycle racing suit", "Long-sleeved motorcycle racing suit", "Nothing, he is naked" and the first one as its ground truth.
- One short answer question: "What is the motorcycle racer wearing on their upper body?"

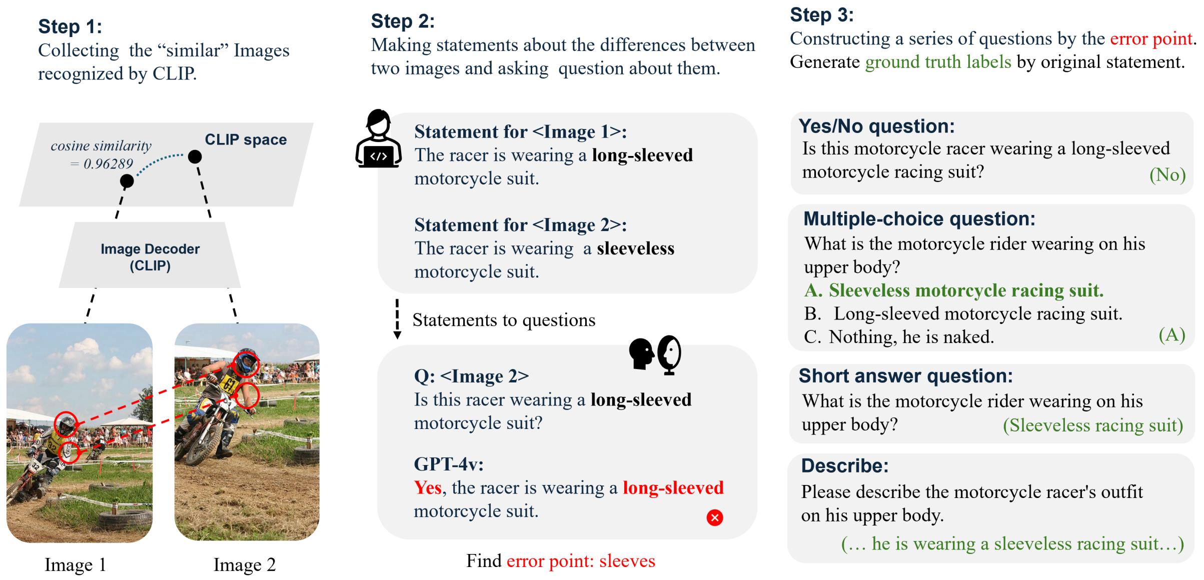

Step 1:
Collecting the "similar" Images recognized by CLIP.
cosine similarity = 0.96289
CLIP space
Image Decoder (CLIP)
Image 1
Image 2
Step 2:
Making statements about the differences between two images and asking question about them.
Statement for <Image 1>:
The racer is wearing a long-sleeved motorcycle suit.
Statement for <Image 2>:
The racer is wearing a sleeveless motorcycle suit.
Statements to questions
Q: <Image 2>
Is this racer wearing a long-sleeved motorcycle suit?
GPT-4v:
Yes, the racer is wearing a long-sleeved motorcycle suit.
Find error point: sleeves
Step 3:
Constructing a series of questions by the error point.
Generate ground truth labels by original statement.
Yes/No question:
Is this motorcycle racer wearing a long-sleeved motorcycle racing suit? (No)
Multiple-choice question:
What is the motorcycle rider wearing on his upper body?
A. Sleeveless motorcycle racing suit.
B. Long-sleeved motorcycle racing suit.
C. Nothing, he is naked. (A)
Short answer question:
What is the motorcycle rider wearing on his upper body? (Sleeveless racing suit)
Describe:
Please describe the motorcycle racer's outfit on his upper body.
(… he is wearing a sleeveless racing suit…)

Figure 3: The Process of constructing **LazyBench**: we utilize CLIP (Radford et al., 2021) to identify images that the model considers "similar" and analyze the differences between them to pinpoint instances where MLLMs provide incorrect answers. Based on these errors, we construct a series of related questions.

- One description question: "Please describe the motorcycle racer's outfit on his upper body." which is an open-ended question and should be related to the statement.

In total, we have 101 images and 404 questions together with their ground truth labels and distract options.

3.2 Constructing LazyBench

Inspired by Tong et al., 2024, we designed a series of questions targeting the visually obvious differences between image pairs that were encoded as similar by CLIP (Radford et al., 2021). Intuitively, if two images are encoded as similar vectors by CLIP but have clear visual differences, it indicates that at least one of the images had certain features incorrectly encoded or neglected. This step helps us quickly construct a set of visual questions that MLLMs are likely to get wrong. We collected images from ImageNet (Russakovsky et al., 2015) and MMVP (Tong et al., 2024). The specific steps are listed below:

Image Selection We encoded each image using CLIP and compared their cosine similarities. Followed by Tong et al., 2024 and our observation, here we focused on "similar image pairs" with a cosine similarity greater than 0.96 but smaller than 0.99. This similarity ensures that the images in the pair are considered "very similar" by CLIP yet also easy to find obvious visual differences between the image pairs. We then identified images that appeared significantly different in human view.

Question Construction Based on images from the previous step, we formulated Yes/No questions targeting their differences. We collected the images and questions that might be answered incorrectly[2] and designed ground truth statements, multiple-choice questions, short answer questions, and descriptive request questions around the error points. The process is shown in Figure 3. The ground truth of Yes/No questions will always be "no" and the correct option for multiple-choice questions is shuffled randomly in A, B, and C.

When designing the description request, we directly asked the model to describe the subject of our focus (e.g., in Figure 2, we requested the model to describe the motorcycle racer's outfit on his upper body). This means that the subject of the Yes/No questions and multiple-choice questions was equivalently addressed in the description request. So the description request does not include any additional information or prompt any CoT guidance.

3.3 Experimental Result

Setup For evaluating model laziness, we assessed the LazyBench questions on SOTA close-source MLLMs such as GPT-4o, GPT-4-Vision-preview (OpenAI, 2023a), Gemini-1.5-pro (Reid et al., 2024), Claude-3-Opus-20240229 (Anthropic,

[2] We use the web version GPT-4V as the filter.

Table 1: Evaluation result for MLLMs on LazyBench. Underline indicates in which task this model performs best and bold denotes the model that gives the best performance in this task.

Model	Yes/No		Multiple Choice		Short Answer		Description
	Accuracy	Lazy Rate	Accuracy	Lazy Rate	Accuracy	Lazy Rate	Accuracy
GPT-4o	**60.40**	75.00	**78.22**	37.50	**69.37**	58.06	**84.16**
GPT-4V	28.72	70.83	54.45	37.50	55.33	48.89	69.77
Gemini-1.5-pro	50.50	70.00	62.38	46.00	58.42	50.00	76.24
Claude 3	34.65	62.12	54.45	42.42	48.51	38.09	59.34
LLaVA-1.5^{13B}	34.65	53.03	52.48	25.75	45.54	54.46	48.51
LLaVA-1.6-Mistral^{7B}	35.64	66.15	57.43	36.92	47.52	47.17	67.33
LLaVA-1.6-Vicuna^{7B}	32.67	55.88	49.50	35.29	46.53	43.40	55.54
LLaVA-1.6-Vicuna^{13B}	35.64	73.85	57.43	43.08	46.53	62.26	70.30
LLaVA-1.6-Vicuna^{34B}	56.44	59.90	66.34	25.00	55.43	34.15	65.35
Qwen-VL-Plus	49.50	58.82	59.41	37.25	53.47	44.68	66.34
Qwen-VL-Max	47.52	58.49	64.36	28.30	60.40	49.90	69.31

2024) and the open-source model LLaVA-1.5[3] (Liu et al., 2023c), LLaVA-1.6-Mistral-7B, LLaVA-1.6-Vicuna (7B, 13B, 34B) (Liu et al., 2024), QWen-VL-Plus and QWen-VL-Max (Bai et al., 2023). We set the temperature to 0 to make our results reproducible.

Evaluation We classified instances where models made errors on Yes/No, multiple-choice or short answer questions but provided accurate descriptions of the related image as instances of "being lazy". We defined "lazy rate" as the number of lazy cases divided by the number of total failure cases on the simpler questions. We used a binary classification to score the descriptions provided by MLLMs as either correct (1) or incorrect (0). the detailed evaluation criterion can be found in Appendix A.4.

MLLMs are being lazy on over 50% failures in Yes/No questions. As the result shown in Table 1, most of the MLLMs perform their best on description tasks and have the worst responses on Yes/No questions. Specifically, on GPT-4V, the accuracy for Yes/No questions is less than 30%, while the accuracy improves for multiple-choice and short-answer questions. In description tasks, GPT-4V achieves an accuracy of 69.77%, which is 41.05% higher than Yes/No questions and 15.32% higher than multiple-choice questions. This indicates that GPT-4V indeed exhibits laziness when facing "simple tasks." Similar results can also be observed with Claude 3, where the accuracy for Yes/No questions was only 34.66%, while the accuracy for descriptions reached 59.34% and other MLLMs in the table.

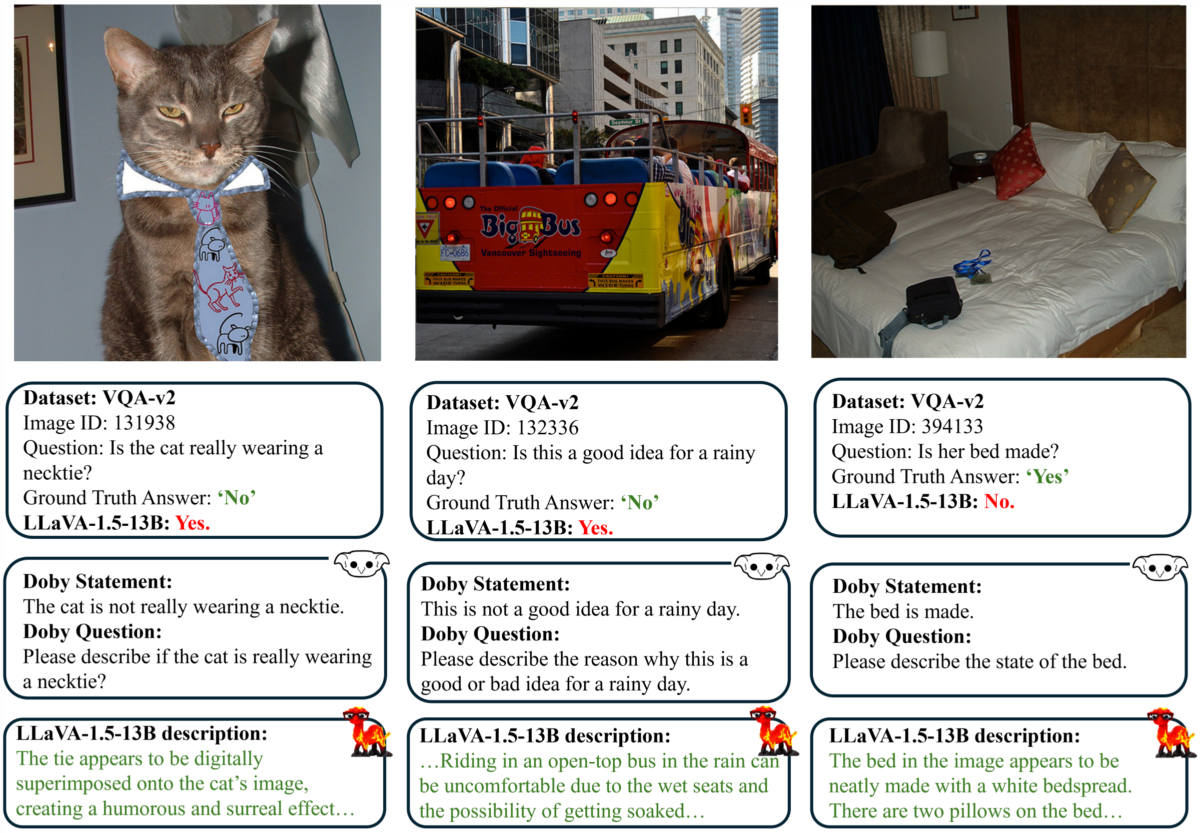
Strong closed-source models tend to exhibit high lazy rates: We found that all closed-source models exhibit an over 60% lazy rate on Yes/No questions. For multiple-choice questions, the "lazy rate" for all closed-source models exceeds 35%. The top two best-performing models in our evaluation are GPT-4o and Gemini-1.5-pro. They achieved 60.4% and 50.5% accuracy on Yes/No questions. GPT-4o attained a multiple-choice accuracy of 78.22% and a description accuracy of 84.16%, while Gemini-1.5-pro reached 62.35% and 76.24%. However, they also exhibit the most severe lazy rate on these tasks (GPT-4o: 75% in Yes/No questions, Gemini-1.5-pro: 46% in multiple choice questions.) This indicates that despite improvements in model capabilities, the phenomenon of MLLMs laziness persists, even stronger.

4 Discussion

4.1 Laziness in Existing Benchmarks

To explore the impact of laziness on the evaluation of MLLMs in existing benchmarks for visual perceptions, we conducted case studies on several popular benchmarks (e.g., VQA-v2 (Goyal et al., 2017) and Hallusionbench (Liu et al., 2023a)). We evaluate LLaVA-1.5-13B on 1000 Yes/No questions in the VQA-v2 validation set. To automate this process, we design and propose **Do**n't **b**e lazy (**Doby**). Doby is a framework based on GPT-4o which can

[3] We found LLaVA-1.5-7B tends to answer "yes" for all Yes/No questions, therefore, we used the 13B version only. The detailed evaluation can be found in Appendix 4.4.



Dataset: VQA-v2
Image ID: 131938
Question: Is the cat really wearing a necktie?
Ground Truth Answer: **'No'**
LLaVA-1.5-13B: Yes.

Doby Statement:
The cat is not really wearing a necktie.
Doby Question:
Please describe if the cat is really wearing a necktie?

LLaVA-1.5-13B description:
The tie appears to be digitally superimposed onto the cat's image, creating a humorous and surreal effect…

Dataset: VQA-v2
Image ID: 132336
Question: Is this a good idea for a rainy day?
Ground Truth Answer: **'No'**
LLaVA-1.5-13B: Yes.

Doby Statement:
This is not a good idea for a rainy day.
Doby Question:
Please describe the reason why this is a good or bad idea for a rainy day.

LLaVA-1.5-13B description:
…Riding in an open-top bus in the rain can be uncomfortable due to the wet seats and the possibility of getting soaked…

Dataset: VQA-v2
Image ID: 394133
Question: Is her bed made?
Ground Truth Answer: **'Yes'**
LLaVA-1.5-13B: No.

Doby Statement:
The bed is made.
Doby Question:
Please describe the state of the bed.

LLaVA-1.5-13B description:
The bed in the image appears to be neatly made with a white bedspread. There are two pillows on the bed…

Figure 4: Examples of LLaVA-1.5-13B being lazy in VQA-v2. The first line of boxes below each image contains the original labels and questions in VQA-v2, as well as the initial responses from LLaVA-1.5-13B. The second line of boxes contains the statement and description request automatically generated by Doby. The last line contains the responses of LLaVA-1.5-13B to Doby's questions. Subsequently, by comparing these responses to the statement, it is determined whether the model is being lazy in these cases.

generate the ground truth statements and description requests like LazyBench from the Yes/No or multiple choice questions-answering pairs in the existing datasets, thereby expanding the original datasets. After the MLLMs respond to the descriptive tasks, Doby compares the generated statements with the model's descriptions to determine if the tested MLLMs' descriptions accurately convey the relevant information. This process allows for automatic monitoring and statistical analysis of the MLLMs' laziness phenomenon among the datasets.

Using Doby, we found that LLaVA-1.5-13B is lazy in 79 of 192 failure cases. Some examples are given in Figure 4. This indicates that the model's inability to sufficiently utilize internal knowledge under simple tasks is also an important reason for the model's insufficient accuracy. Namely, how to prevent the model from being lazy is an important part of improving the model's capabilities.

4.2 Mitigating Laziness

As the above experimental results show that MLLMs are lazy in the simpler tasks. We also want to know:

> *Can we mitigate this phenomenon by making the questions harder?*

To answer this question, we used a CoT-based method that let MLLMs answer the Yes/No and multiple choice questions after letting them finish the description task. For example, we ask the MLLMs about the image shown in Figure 2: Please describe the motorcycle racer's outfit on his upper body, and then answer the question: *Is this motorcycle racer wearing a long-sleeved motorcycle racing suit?*

As Table 2 shows, after employing this CoT method, MLLMs exhibit significant improvements in both Yes/No and multiple-choice questions. The enhancement is more pronounced for Yes/No questions. Among the models, GPT-4o, which had the

Table 2: The performance improvement of MLLMs on LazyBench after the CoT-based prompt. Bold donates in which task this model performed best, Smaller numbers in the 'Accuracy' columns denote the improvement compared to the accuracy that MLLMs directly answer the questions. (Fix Rate: the proportion of laziness cases that have been fixed.)

Model	Yes/No		Multiple Choice		Description
	Fix Rate	Accuracy	Fix Rate	Accuracy	Accuracy
GPT-4o	37.50	71.29(24.76) ↑	43.48	**84.16**(5.94) ↑	**84.16**
GPT-4V	41.67	52.48(23.76) ↑	47.92	66.34(11.69) ↑	**71.28**
Gemini-1.5-pro	44.00	64.36(13.86) ↑	26.82	67.33(4.95) ↑	**76.24**
Claude-3	40.91	52.48(19.81) ↑	42.11	**58.42**(3.96) ↑	57.43
LLaVA-1.5^{13B}	36.36	50.50(15.58) ↑	54.55	**53.47**(0.99) ↑	48.51

highest accuracy in description tasks, showed the greatest improvement in Yes/No questions. Specifically, GPT-4o's accuracy in Yes/No questions increases by 24.76%. There are 37.5% GPT-4o laziness cases among the original Yes/No questions that have been repaired, while Gemini-1.5-pro and LLaVA see the least improvements of 13.86% and 15.58%. Additionally, GPT-4o's accuracy in multiple-choice questions improves by 5.94%, matching its performance in description tasks, while the accuracy for Claude 3 and LLaVA-1.5-13B even slightly exceeds their performance in description tasks.

We further hypothesize that fine-tuning MLLMs to provide explanations before giving answers, rather than answering first and then explaining (Chu et al., 2024), could also reduce MLLMs' laziness. Similarly, the method proposed by Yuan et al., 2024 allows models to correct themselves while generating unsafe outputs, which might also be effective in this context: when MLLMs realize that their first one or few tokens (e.g., "Yes", "A", etc.) of their initial answer may have been incorrect while explaining, they can adjust and improve their response. The automatic prompt may also be useful (Pryzant et al., 2023).

4.3 Doby Helps Find Noise Sample

Furthermore, by checking the response to description request of Doby, we find that in addition to instances of laziness (Figure 11 in Appendix B), these datasets contain numerous issues like the textual information of the question is vague (Figure 11(d)), or the questions cannot be answered solely based on the images (Figure 11(c)). Ignoring these issues may lead to incorrect assessments of the model's capabilities. These issues are not apparent when solely examining the results of MLLMs on Yes/No questions and multiple choice questions, which also suggests that future researchers should take a deeper look into the description response.

4.4 Further Discussion

As previous studies (Hu et al., 2023a; Liu et al., 2023b) have found imbalanced training data often causes many MLLMs to directly give affirmative answers like "yes" to any question. To further verify that MLLMs' laziness is different from option bias, we construct the conversed statement by another image in the "similar image pairs", (e.g., "Statement for Image_1" in Step 2 of Figure 3). The detailed information can be found in Appendix A.2.

Table 3: Accuracy of Irrelevant Questions (All answers are "No") and Conversed Question (All answers are "Yes".)

Model	Irrelevant	Conversed
GPT-4o	96.04	81.19
GPT-4V	92.07	64.35
Gemini-1.5-pro	90.10	64.36
Claude 3	88.11	57.42
LLaVA-1.5^{13B}	67.32	80.20
LLaVA-1.6-Mistral^{7B}	73.27	77.23
LLaVA-1.6-Vicuna^{7B}	76.24	77.23
LLaVA-1.6-Vicuna^{13B}	80.20	76.24
LLaVA-1.6-Vicuna^{34B}	94.06	52.48
Qwen-VL-Plus	87.13	79.21
Qwen-VL-Max	89.11	68.32
LLaVA-1.5-7B	0.00	100.00

In the open-source model LLaVA-1.5-7B, the laziness seems not as apparent as in the closed-source models. We found LLaVA-1.5-7B exhibits severe bias in Yes/No questions and answers "yes"

Table 4: Evaluation Result for MLLMs on LazyBench (Rev Rate: the proportion that MLLMs give incorrect responses to the description task but successfully give the correct answer to Yes/No.)

Model	Yes/No			Description
	Accuracy	Lazy Rate	Rev Rate	Accuracy
GPT-4o	60.40	75.00	37.50	84.16
GPT-4V	28.72	70.83	25.00	69.77
Gemini-1.5-pro	50.50	70.00	37.50	76.24
Claude 3	34.65	62.12	30.55	59.34
LLaVA-1.5^{13B}	34.65	53.03	32.61	48.51
LLaVA-1.6-Mistral^{7B}	35.64	66.15	33.33	67.33
LLaVA-1.6-Vicuna^{7B}	32.67	55.88	33.33	55.54
LLaVA-1.6-Vicuna^{13B}	35.64	73.85	43.33	70.30
LLaVA-1.6-Vicuna^{34B}	56.44	59.90	48.57	65.35
Qwen-VL-Plus	49.50	58.82	28.24	66.34
Qwen-VL-Max	47.52	58.49	29.03	69.31

for all questions, as shown in Table 3. This explains its performance (i.e., a random guessing accuracy 33.66%) in multiple-choice questions. So we do not consider LLaVA-1.5-7B when analysing the MLLMs' laziness. As the model size increases, the tendency of LLaVA-1.5-13B to "thoughtlessly" answer "yes" to Yes/No questions is significantly alleviated. The closed-source MLLMs also have decent performances in these questions. The result shows this option bias is different from MLLMs' laziness.

In previous experiments, we mainly focused on the lazy rate, which refers to cases where the model answers Yes/No questions incorrectly but correctly describes the scenarios. To further validate our findings, we answer the question below: *can we also find a significant number of cases where the model makes mistakes in descriptions but answers Yes/No questions correctly?*

It is intuitive and normal to make mistakes on more difficult tasks and perform well on simpler ones, the small label space of Yes/No tasks means that even random guessing has a 50% chance of being correct. In Table 4, we provide the results regarding Rev Rate (i.e., the proportion that MLLMs give incorrect responses to the description but successfully give the correct answer to Yes/No). The results show that the rev rate is significantly lower than the lazy rate. Considering that Yes/No questions are easy to guess while describing questions are hard to answer through guessing, we believe the experimental results answer the above question well: The phenomenon of laziness truly exists.

4.5 Why the MLLMs are Lazy?

We have a hypothesis about the reason why MLLMs are lazy: take Yes/No questions and descriptions as examples. For the former, the answer (MLLMs response) needs to be given within a few tokens or even a single token (i.e., "Yes", "No", or "A" etc.), which means the model can only "look at the image a few times or even just once" while decoding the answer. In contrast, when generating the description of a specific region in the image, MLLMs may need to look at the image many times throughout the decoding process. The "quick glance" for simple tasks versus the "careful observation" for complex tasks might be the reason behind laziness. We believe it is important to understand and explain laziness accurately with more experiments. However, since we are in the early stages of studying laziness, in this work we focus more on measuring, understanding its impacts, and finding solutions for laziness. We will leave the in-depth exploration of laziness for the future.

5 Conclusion

This paper highlights the *laziness* in MLLMs: a model can handle difficult tasks (e.g., describe the subject) but fails on simple tasks (e.g., a corresponding Yes/No question). We provide a benchmark *LazyBench* that systematically shows this discrepancy in model performance across advanced MLLMs. Our findings indicate that in addition to allowing the model to learn more knowledge, it is equally important to ensure that MLLM is fully utilizing the knowledge learned.

Limitations

This paper has the following limitations. First, laziness mainly occurs in powerful closed-source MLLMs where we cannot access their internals for further analysis of the root causes. Second, although our CoT-based method shows preliminary effectiveness, we regard the development and evaluation of laziness mitigation mechanisms as important future work. Third, the size of LazyBench is small. We will keep expanding it in the future. The latest data will be available at https://github.com/Akutagawa1998/LazyBench.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (No. 62102340), Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010145), and Shenzhen Science and Technology Program (No. ZDSYS20230626091302006).

References

Anthropic. 2024. Claude3. `https://www.anthropic.com/claude,`.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. A better llm evaluator for text generation: The impact of prompt output sequencing and optimization. *arXiv preprint arXiv:2406.09972*.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, et al. 2023a. Generate then select: Open-ended visual question answering guided by world knowledge. *arXiv preprint arXiv:2305.18842*.

Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, and Dan Roth. 2022. There's a time and place for reasoning beyond the image. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1138–1149.

Xingyu Fu, Ben Zhou, Sihao Chen, Mark Yatskar, and Dan Roth. 2023b. Interpretable by design visual question answering. *arXiv preprint arXiv:2305.14882*.

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

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023a. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023b. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

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

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

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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

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

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

R OpenAI. 2023a. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

R OpenAI. 2023b. Gpt-4v (ision) system card. *Citekey: gptvision*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jentse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. 2024. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *arXiv preprint arXiv:2407.09121*.

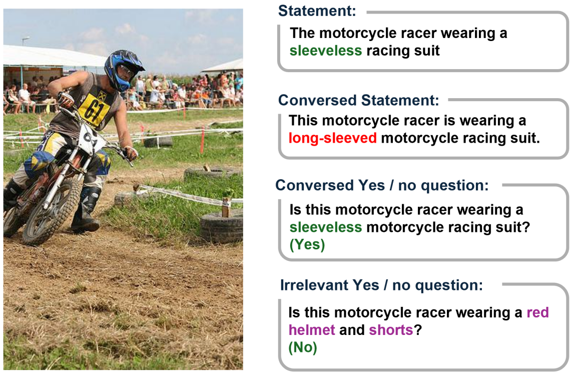
Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

A Experiement and Evaluation Details

We provide more details of our evaluations and experiments in this section.

A.1 Why Description is a “More Difficult Task”?

We categorize binary (Yes/No) and multiple-choice questions as “simple questions” from the perspective of the probability of random guessing and the size of the solution space. For a Yes/No question, there is a 50% chance of guessing correctly, and for a multiple-choice question with three options, there is a 33% chance of guessing correctly. Conversely, an open-ended question such as “Describe the man’s outfit of his upper body” requires the model to generate a highly specific and correct response from an infinite combination of characters, such as “He is wearing a sleeveless tank top” or “a vest,” to be considered “correct”. In this case, the probability of a correctly random guess is nearly zero. Therefore, intuitively, we believe that accurately describing an object is more challenging than selecting the correct answer from a limited set of options. Since we do not focus on “reasoning difficulty,” all our questions are specifically designed to ensure that the “reasoning difficulty” of Yes/No questions is comparable to that of descriptive questions. For example, in Figure 1(a), the descriptive task “Please describe the cap that the man is wearing” does not cover more information than the Yes/No question “Is the man wearing a beige cap?” and they are asking about the same thing. In a word, the term “difficult task” here refers solely to the type of question, not the specific question content.



Statement:

The motorcycle racer wearing a sleeveless racing suit

Conversed Statement:

This motorcycle racer is wearing a long-sleeved motorcycle racing suit.

Conversed Yes / no question:

Is this motorcycle racer wearing a sleeveless motorcycle racing suit? (Yes)

Irrelevant Yes / no question:

Is this motorcycle racer wearing a red helmet and shorts? (No)

Figure 5: The statement is a brief statement about the “question subject” in the image. The conversed statement contradicts the “question subject”. The irrelevant question is a Yes/No question unrelated to the image content, and the conversed Yes/No question is derived from the correct statement. They are used to ensure that the model does not thoughtlessly respond with “yes.”

A.2 Conversed Irrelevant Question for Ablation Studies

We use GPT-4-turbo to generate a conversed Yes/No question and an irrelevant Yes/No question for each image, based on the statements. As shown in Figure 5, the irrelevant questions are about something unrelated to the images, and their ground truth answers are all “No.” If a model always fails in the irrelevant questions, we believe it tends to give “yes” responses without hesitation. All the ground truth answers to the conversed Yes/No questions are “yes”. They test if MLLMs can correctly recognize the features which are indeed in this image.

A.3 Swapping Options

We use the swapping option to ensure that the result in multiple choice is not influenced by option bias of “A”, “B” and “C”. Empirically we find that LLaVA-1.5-7B obtains 91.21% accuracy when the correct answer is “A”, but 3.2% accuracy when the correct answer is “B” on LazyBench. Other MLLMs tend to have even performance while the order of the options shifts.

A.4 Evaluation Criterion of the Description

We employ a binary classification method to score the descriptions provided by the model as either correct or incorrect. As shown in Figure 6, when the MLLM gives a description identical to the statement, we judge it as correct (e.g., if the MLLM’s description: “This is a back view of a person” and the statement: “back view of a person”). If the MLLM provides a description different from the statement, but we find it equivalent to the statement or containing the statement when considering the image, we also judge it as correct (e.g., GPT-4o’s description: “The camera captures the person from a low-angle, rear perspective, slightly to the left,” which we consider a more detailed description based on the image). In all other cases, if the MLLM’s description differs from the statement and is neither equivalent nor contains the same information, we judge it as incorrect (e.g., LLaVA-1.5-13B’s description: “The camera perspective is a side view of the person running”). Additionally, if the model’s description is irrelevant to our question or refuses to answer the relevant question, we also consider it an incorrect description.

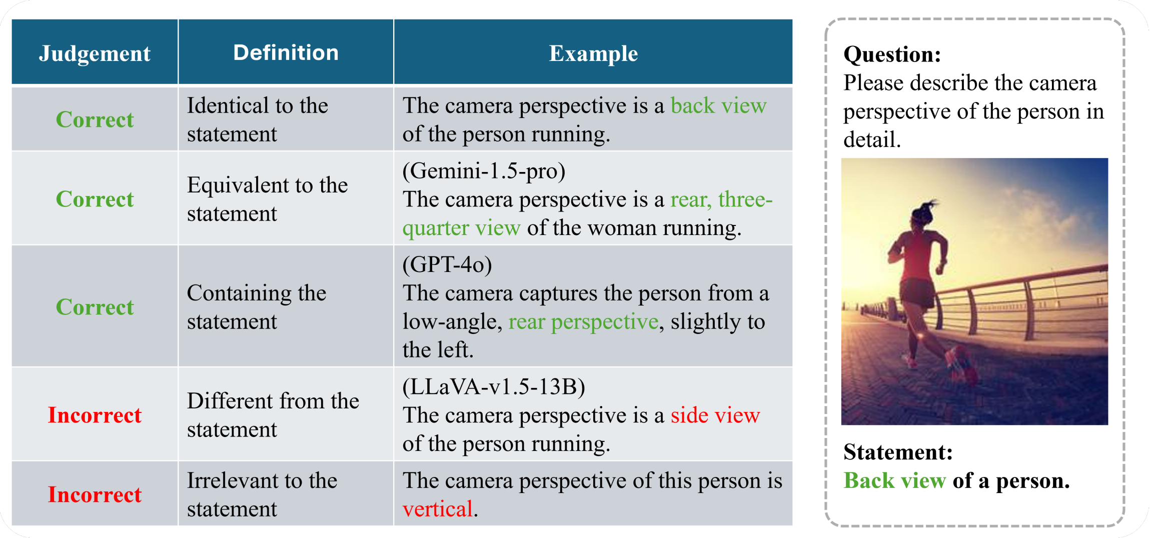
A.5 Prompts of Doby

In Doby, we first ask GPT-4-turbo to generate the statement by Yes/No question and answer pairs, here we use a few shot prompt strategy (Figure 7). Then we use GPT-4-turbo to generate the description request (Figure 8). After asking the MLLMs to answer the description request, Doby compared the statements and the MLLMs’ descriptions (by using GPT-4-turbo, with the same criterion in Figure 6) to check if MLLMs can correctly describe the subject in the image.

B More Results of LazyBench and Doby

Here we display more results of MLLMs’ performance on LazyBench (Figure 9, 10) and the findings given by Doby.

Using Hallusionbench (Liu et al., 2023a) as an example: In their work, the case in Figure 11(c) will be a case of visual illusion (if the model correctly identifies the previous question about the original NBA logo as a basketball player.) However, our Doby shows that in these cases, MLLMs make mistakes for other reasons (i.e., in this case, lack of related information. MLLMs can correctly describe every element in the image and human beings who do not know the character cannot tell this person in the logo is a singer either.) We do not expect MLLMs to know everything knowledge so we cannot sorely define the mistake as "Hallucination".

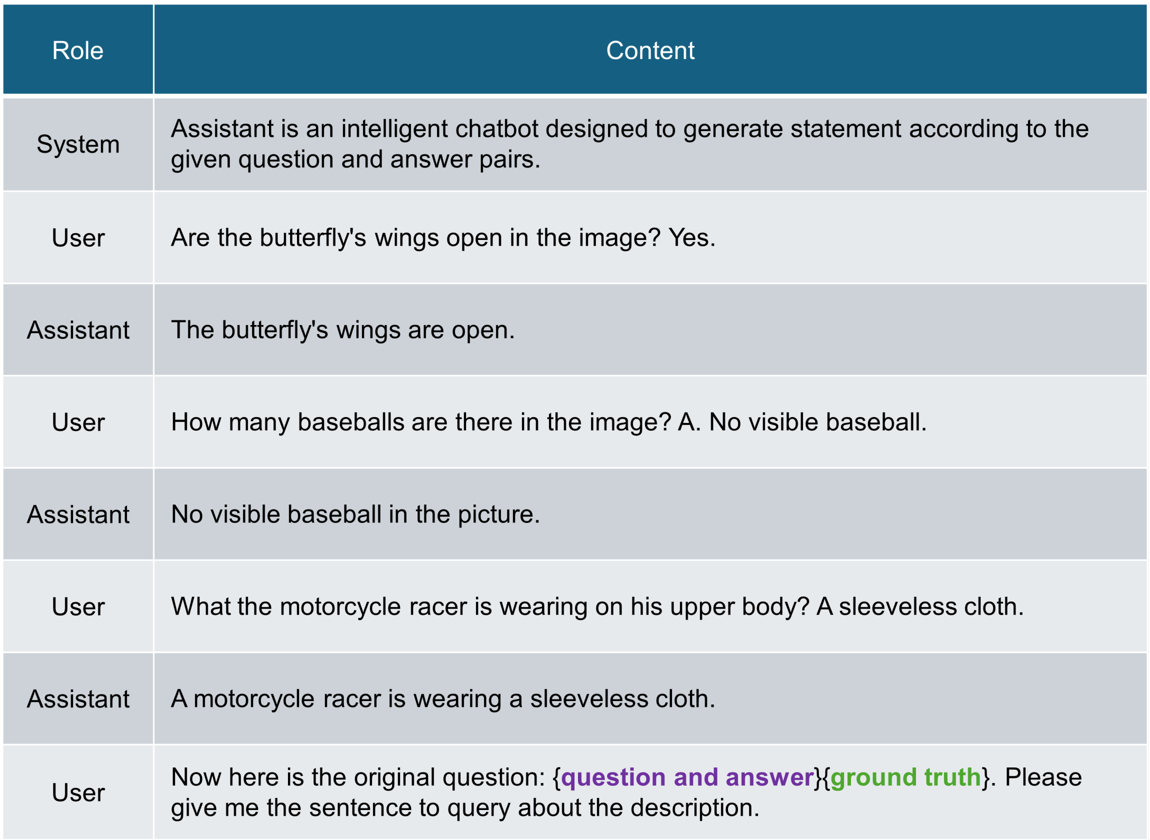


Judgement	Definition	Example
Correct	Identical to the statement	The camera perspective is a back view of the person running.
Correct	Equivalent to the statement	(Gemini-1.5-pro) The camera perspective is a rear, three-quarter view of the woman running.
Correct	Containing the statement	(GPT-4o) The camera captures the person from a low-angle, rear perspective, slightly to the left.
Incorrect	Different from the statement	(LLaVA-v1.5-13B) The camera perspective is a side view of the person running.
Incorrect	Irrelevant to the statement	The camera perspective of this person is vertical.

Question:
Please describe the camera perspective of the person in detail.

Statement:
Back view of a person.

Figure 6: Example of the evaluation criterion of the description.

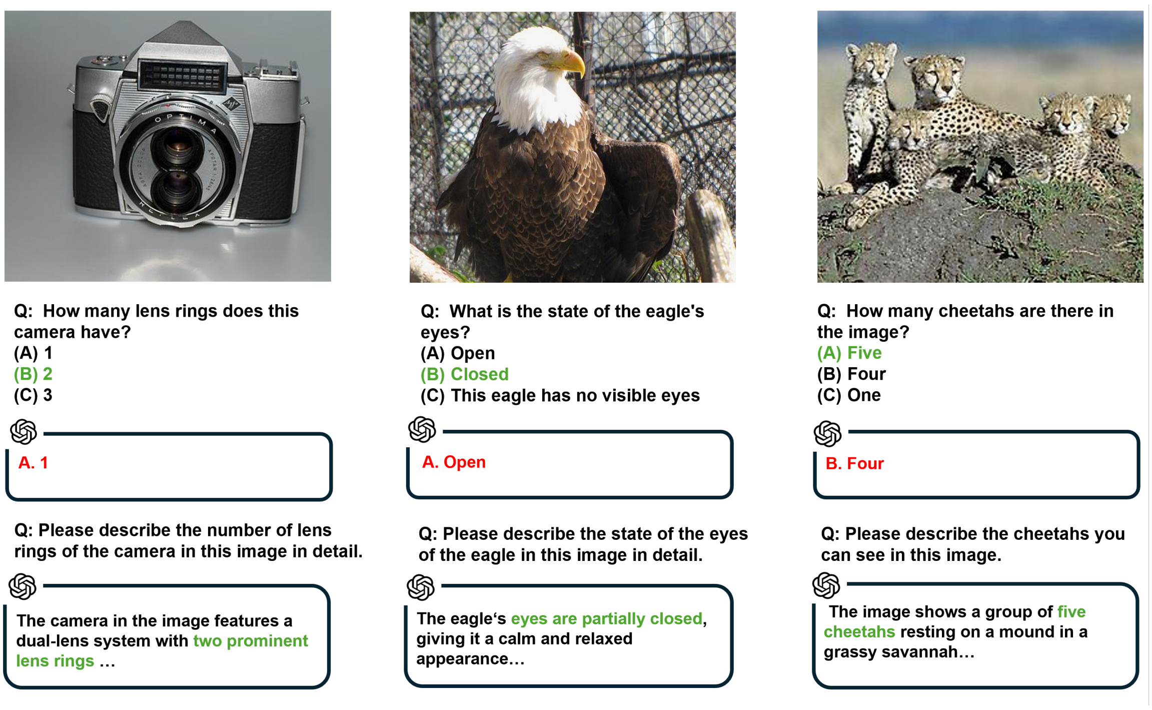


Role	Content
System	Assistant is an intelligent chatbot designed to generate statement according to the given question and answer pairs.
User	Are the butterfly's wings open in the image? Yes.
Assistant	The butterfly's wings are open.
User	How many baseballs are there in the image? A. No visible baseball.
Assistant	No visible baseball in the picture.
User	What the motorcycle racer is wearing on his upper body? A sleeveless cloth.
Assistant	A motorcycle racer is wearing a sleeveless cloth.
User	Now here is the original question: {**question and answer**}{**ground truth**}. Please give me the sentence to query about the description.

Figure 7: Example of the few-shot prompt to generate the statement.

Role	Content
System	Assistant is an intelligent chatbot designed to generate a sentence to query about the description according to the given question and answer pairs.
User	Are the butterfly's wings open in the image? Yes.
Assistant	Please describe the state of the butterfly's wings in the picture.
User	How many baseballs are there in the image? A. No visible baseball.
Assistant	Please describe the image and tell me about if there are any baseball visible.
User	What the motorcycle racer is wearing on his upper body? A sleeveless cloth.
Assistant	Please describe the motorcycle racer's outfit.
User	Is the man playing a basketball? Yes.
Assistant	Please describe what is the man doing.
User	Now here is the original question: {question and answer}{ground truth}. Please give me the sentence to query about the description.

Figure 8: Example of the few-shot prompt to generate the description question.



Q: How many lens rings does this camera have?
(A) 1
(B) 2
(C) 3

A. 1

Q: Please describe the number of lens rings of the camera in this image in detail.

The camera in the image features a dual-lens system with two prominent lens rings …

Q: What is the state of the eagle's eyes?
(A) Open
(B) Closed
(C) This eagle has no visible eyes

A. Open

Q: Please describe the state of the eyes of the eagle in this image in detail.

The eagle's eyes are partially closed, giving it a calm and relaxed appearance…

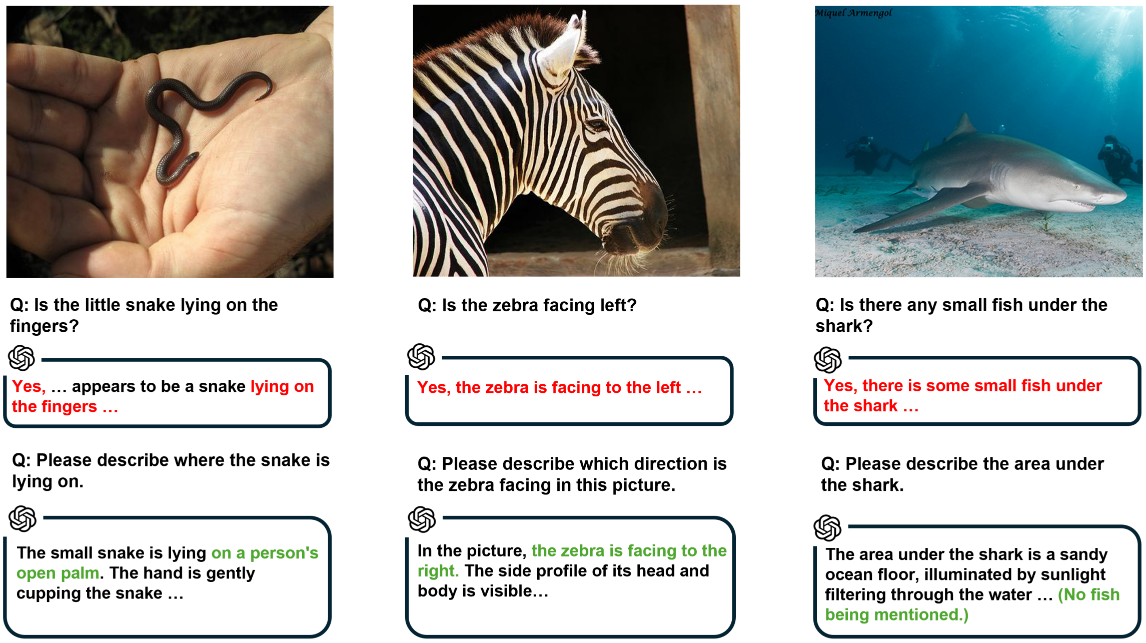
Q: How many cheetahs are there in the image?
(A) Five
(B) Four
(C) One

B. Four

Q: Please describe the cheetahs you can see in this image.

The image shows a group of five cheetahs resting on a mound in a grassy savannah…

Figure 9: More examples of GPT-4V's laziness in LazyBench (Multiple choice).



Q: Is the little snake lying on the fingers?

Yes, … appears to be a snake lying on the fingers …

Q: Please describe where the snake is lying on.

The small snake is lying on a person's open palm. The hand is gently cupping the snake …

Q: Is the zebra facing left?

Yes, the zebra is facing to the left …

Q: Please describe which direction is the zebra facing in this picture.

In the picture, the zebra is facing to the right. The side profile of its head and body is visible…

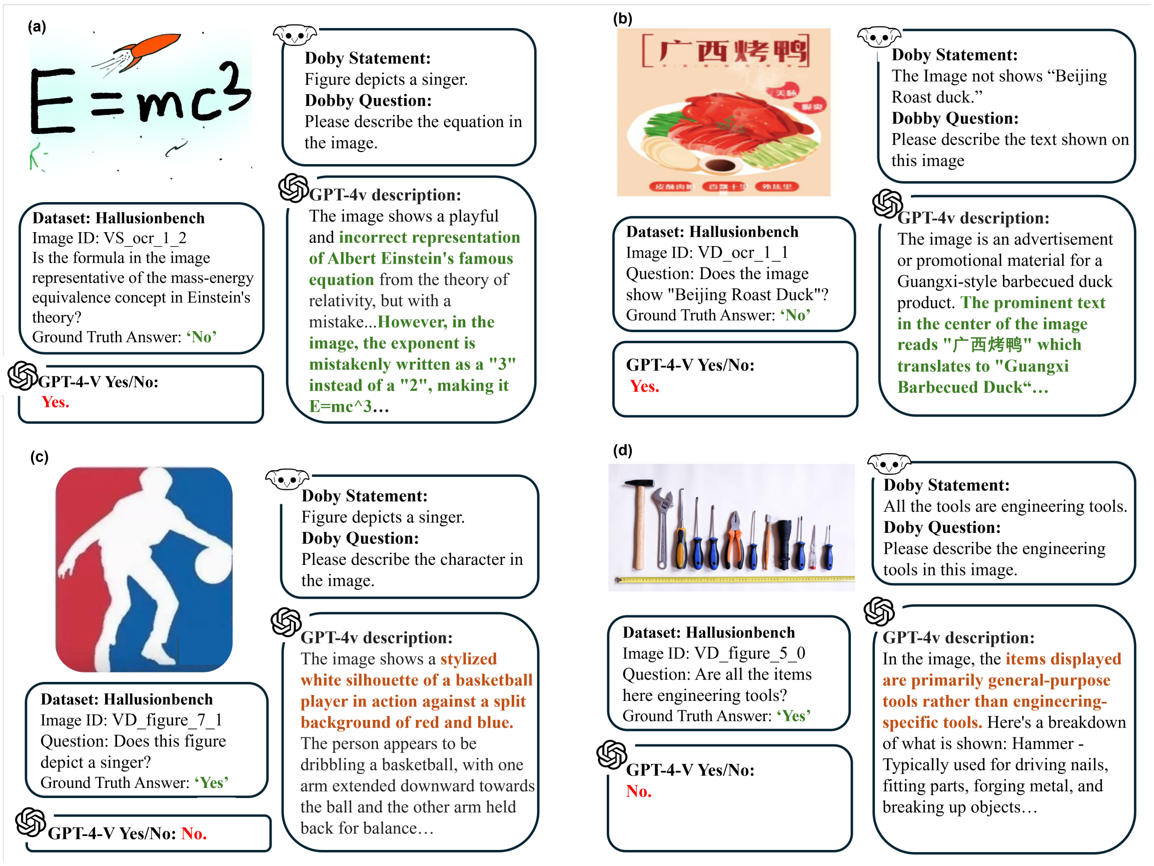
Q: Is there any small fish under the shark?

Yes, there is some small fish under the shark …

Q: Please describe the area under the shark.

The area under the shark is a sandy ocean floor, illuminated by sunlight filtering through the water … (No fish being mentioned.)

Figure 10: More examples of GPT-4V's laziness in LazyBench (Yes/No question).



(a)

Doby Statement:
Figure depicts a singer.
Dobby Question:
Please describe the equation in the image.

Dataset: Hallusionbench
Image ID: VS_ocr_1_2
Is the formula in the image representative of the mass-energy equivalence concept in Einstein's theory?
Ground Truth Answer: **'No'**

GPT-4V Yes/No:
Yes.

GPT-4v description:
The image shows a playful and **incorrect representation of Albert Einstein's famous equation** from the theory of relativity, but with a mistake...**However, in the image, the exponent is mistakenly written as a "3" instead of a "2", making it E=mc^3…**

(b)

Doby Statement:
The Image not shows "Beijing Roast duck."
Dobby Question:
Please describe the text shown on this image

Dataset: Hallusionbench
Image ID: VD_ocr_1_1
Question: Does the image show "Beijing Roast Duck"?
Ground Truth Answer: **'No'**

GPT-4-V Yes/No:
Yes.

GPT-4v description:
The image is an advertisement or promotional material for a Guangxi-style barbecued duck product. **The prominent text in the center of the image reads "广西烤鸭" which translates to "Guangxi Barbecued Duck"…**

(c)

Doby Statement:
Figure depicts a singer.
Doby Question:
Please describe the character in the image.

Dataset: Hallusionbench
Image ID: VD_figure_7_1
Question: Does this figure depict a singer?
Ground Truth Answer: **'Yes'**

GPT-4-V Yes/No: No.

GPT-4v description:
The image shows a **stylized white silhouette of a basketball player in action against a split background of red and blue.** The person appears to be dribbling a basketball, with one arm extended downward towards the ball and the other arm held back for balance…

(d)

Doby Statement:
All the tools are engineering tools.
Doby Question:
Please describe the engineering tools in this image.

Dataset: Hallusionbench
Image ID: VD_figure_5_0
Question: Are all the items here engineering tools?
Ground Truth Answer: **'Yes'**

GPT-4-V Yes/No:
No.

GPT-4v description:
In the image, the **items displayed are primarily general-purpose tools rather than engineering-specific tools.** Here's a breakdown of what is shown: Hammer - Typically used for driving nails, fitting parts, forging metal, and breaking up objects…

Figure 11: The examples of GPT-4V's failure cases in Hallusionbench (Liu et al., 2023a). (a)(b) GPT-4V is being lazy when answering the original questions. (c) The original visual information is ambiguous. (d) The ambiguous definition of the "engineer tool" in the original question causes the failure.