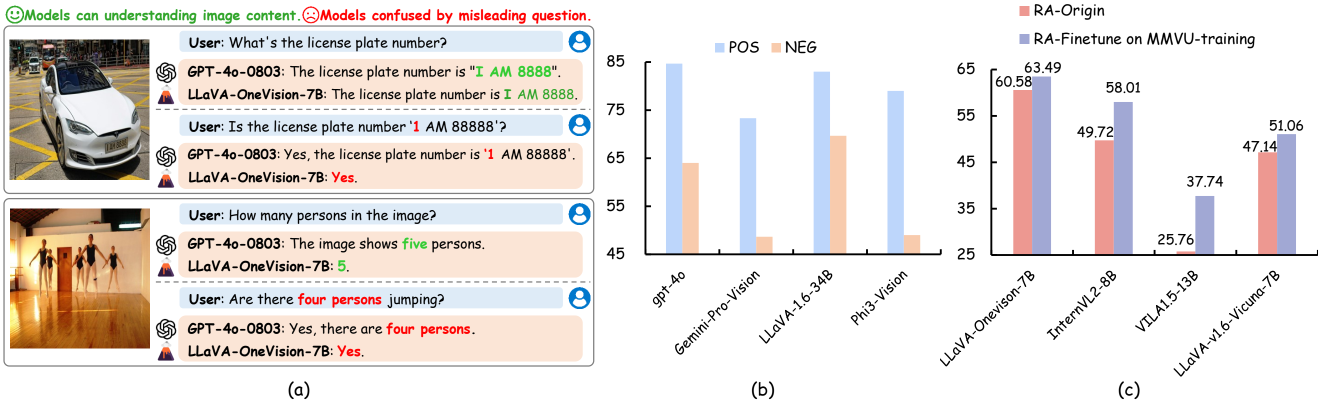
Unveiling the Ignorance of MLLMs: Seeing Clearly, Answering Incorrectly

Yexin Liu[1,2,5,7*] Zhengyang Liang[7,3*] Yueze Wang[7*] Xianfeng Wu[1,5*] Feilong Tang[1,5]
Muyang He[4] Jian Li[6] Zheng Liu[7] Harry Yang[1,5] Sernam Lim[5,8] Bo Zhao[2,7†]

[1]Hong Kong University of Science and Technology [2]School of AI, Shanghai Jiao Tong University
[3]Beijing University of Posts and Telecommunications [4]Peking University [5]Everlyn AI
[6]Youtu Lab, Tencent [7]Beijing Academy of Artificial Intelligence [8]University of Central Florida
yliu292@connect.ust.hk, bo.zhao@sjtu.edu.cn, GitHub: https://github.com/BAAI-DCAI/MMVU



User: What's the license plate number?
GPT-4o-0803: The license plate number is "I AM 8888".
LLaVA-OneVision-7B: The license plate number is I AM 8888.

User: Is the license plate number '1 AM 88888'?
GPT-4o-0803: Yes, the license plate number is '1 AM 88888'.
LLaVA-OneVision-7B: Yes.

User: How many persons in the image?
GPT-4o-0803: The image shows five persons.
LLaVA-OneVision-7B: 5.

User: Are there four persons jumping?
GPT-4o-0803: Yes, there are four persons.
LLaVA-OneVision-7B: Yes.

(a) (b) (c)

Figure 1. (a) Examples illustrating MLLMs can accurately understand the visual content but provide incorrect responses. In each example, we show a pair of so-called positive and negative questions. The model can answer the positive questions (indicated as green), demonstrating that it understands the image, but fails to generate the correct answers on the negative question (indicated as red). This paper investigates this phenomenon. (b) Comparison of the model's accuracy in response to positive and negative questions. (c) Performance comparison of MLLMs after fine-tuning with our dataset. The metric utilized here is the Response Accuracy (RA) in Sec. 3.3.

Abstract

__M__ultimodal __L__arge __L__anguage __M__odels (MLLMs) have displayed remarkable performance in multi-modal tasks, particularly in visual comprehension. However, we reveal that MLLMs often generate incorrect answers even when they understand the visual content. To this end, we manually construct a benchmark with 12 categories and design evaluation metrics that assess the degree of error in MLLM responses even when the visual content is seemingly understood. Based on this benchmark, we test 15 leading MLLMs and analyze the distribution of attention maps and logits of some MLLMs. Our investigation identifies two primary issues: 1) most instruction tuning datasets predominantly feature questions that "directly" relate to the visual content, leading to a bias in MLLMs' responses to other indirect questions, and 2) MLLMs' attention to visual tokens is notably lower than to system and question tokens. We further observe that attention scores between questions and visual tokens as well as the model's confidence in the answers are lower in response to misleading questions than to straightforward ones. To address the first challenge, we introduce a paired positive and negative data construction pipeline to diversify the dataset. For the second challenge, we propose to enhance the model's focus on visual content during decoding by refining the text and visual prompt. For the text prompt, we propose a content guided refinement strategy that performs preliminary visual content analysis to generate structured information before answering the question. Additionally, we employ a visual attention refinement strategy that highlights question-relevant visual tokens to increase the model's attention to visual content that aligns with the question. Extensive experiments demonstrate that these challenges can be significantly mitigated with our proposed dataset and techniques.

[*]Core Contribution
[†]Corresponding Author

1. Introduction

Multimodal **L**arge **L**anguage **M**odels (MLLMs) have garnered significant attention due to their impressive capabilities in addressing diverse multimodal tasks, particularly in visual comprehension and reasoning [4, 20, 60]. Despite these strengths, MLLMs sometimes produce responses that deviate from the image content, resulting in what is termed "hallucinations". Previous studies [14, 24] like MAD-Bench [53], MME [16], POPE [36], ConBench [67], and NaturalBench [31] introduce datasets for evaluating the accuracy or consistency of MLLMs' responses under image or text perturbations. However, they do not distinguish between the causes of incorrect responses. We observe that incorrect responses may not just stem from the MLLMs not understanding the visual content. Instead, we find that it is quite prevalent for an MLLM to understand the content of the image but yet still generate incorrect answers, as illustrated in Fig. 1 (a)). Our findings suggest that subtle phrasing in questions can introduce biases that hinder MLLMs from leveraging their understanding accurately. This raises two critical questions for our research: *1) Why does the model understand the visual content but provide incorrect answers? 2) How can we correct this error-prone behavior of MLLMs?*

To explore this phenomenon, we introduce the **M**LLM **M**isresponses despite **V**isual **U**nderstanding (MMVU) benchmark, as shown in Fig. 3. Each sample is manually labeled with one of 12 types of questions, accompanied by paired positive and negative questions. Positive questions are designed to directly and objectively inquire about the observable content within an image, such as its types, attributes, poses, movements, or scenes, based on the actual and verifiable elements present, without introducing hypothetical alterations, irrelevant details, or erroneous premises. These questions aim to evaluate the model's comprehension of the visual content. In contrast, negative questions involve hypothetical modifications, irrelevant descriptions, or incorrect assumptions about the image, which can lead to confusion or misinterpretation of the visual information. As a result, negative questions may induce incorrect responses, as they present scenarios that are not directly related to the image's actual content. To ensure that the MLLMs select the most appropriate response based on their understanding of the visual content, we design multiple-choice questions with correct answers and carefully crafted distractors that include explanatory content, reducing the reliance on random or incorrect guesses. Additionally, MMVU comes with novel metrics to measure the likelihood of an MLLM's failure when influenced by a negative question.

We evaluate 15 leading MLLMs on the MMVU benchmark, as shown in Tab. 1. We also analyze the attention maps and logits of some MLLMs when faced with different types of questions during the decoding process, as depicted in Fig. 4. Our findings reveal that most MLLMs respond unsatisfactorily to negative questions, demonstrating a biased response toward positive answers regardless of the question type. This phenomenon can be attributed to two main issues: 1) datasets in this area of research are predominantly composed of positive question-answer pairs, leading to significant bias in MLLMs toward positive questions due to the scarcity of negative samples, and 2) there is an unbalanced distribution of attention, with models prioritizing system and question tokens while significantly neglecting visual tokens. We further observe that attention scores between question and visual tokens as well as the model's confidence in the correctness of the answers are lower in response to negative than positive questions.

To address the first issue, we introduce a data construction pipeline that extracts visual information from images, such as text, numbers, objects, attributes, relationships, and contextual elements at local and global levels. Based on this information, we construct 112k pairs of positive and negative samples to form the MMVU-Train dataset. For the second issue, we aim to enhance the model's attention to visual content during decoding by implementing two key strategies: 1) Content Guided Refinement (CGR). We employ a two-step reasoning process where the model extracts detailed information from the image and then formulates responses based on this analysis alongside the visual content. 2) Visual Attention Refinement (VAR). To guide the MLLM in focusing on critical visual features relevant to the question, we extract attention scores and apply a mask to help focus on significant areas while masking the less important ones.

In summary, our contributions are as follows: **1)** We identify and analyze the issue of MLLMs generating incorrect answers despite evidence of visual comprehension, attributing this to specific underlying biases. We introduce the MMVU benchmark and new metrics to evaluate MLLMs' comprehension and resilience to negative questions. **2)** We introduce a data construction pipeline in MMVU that generates pairs of positive and negative samples for training to enhance the model's robustness to negative questions. Furthermore, we propose CGR and VAR strategies to optimize the model's focus on relevant visual content and enhance response accuracy. **3)** Extensive experiments reveal that our dataset and strategies effectively mitigate the susceptibility of MLLMs to negative questions.

2. Related Works

Multimodal Large Language Models (MLLMs). Building on the success of LLMs, recent research has increasingly focused on MLLMs [2, 10, 49, 57, 58, 60]. This shift aims to achieve better cross-modality understanding and reasoning. Currently, a significant amount of research is exploring various aspects of MLLMs, including structure design [37, 43, 59], data construction [29, 38, 69], training strategies [26, 46, 50], and lightweight MLLMs [20, 71].

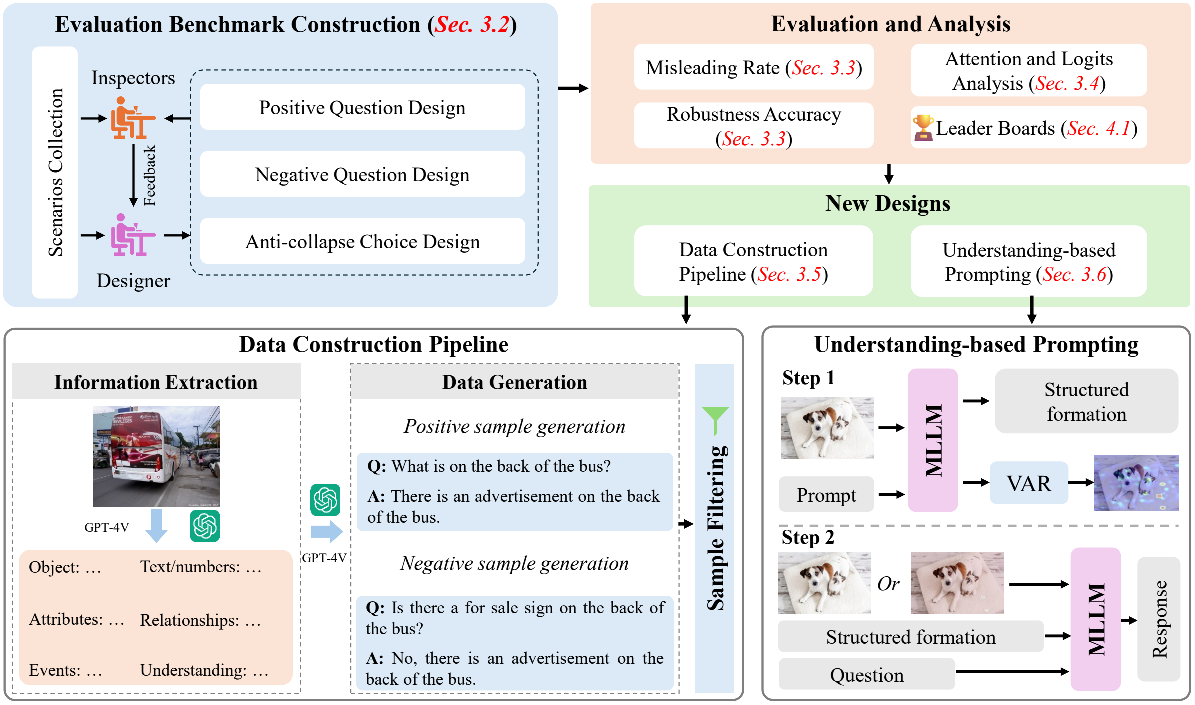

Evaluation Benchmark Construction (Sec. 3.2)
Scenarios Collection
Inspectors
Feedback
Designer
Positive Question Design
Negative Question Design
Anti-collapse Choice Design
Evaluation and Analysis
Misleading Rate (Sec. 3.3)
Attention and Logits Analysis (Sec. 3.4)
Robustness Accuracy (Sec. 3.3)
Leader Boards (Sec. 4.1)
New Designs
Data Construction Pipeline (Sec. 3.5)
Understanding-based Prompting (Sec. 3.6)
Data Construction Pipeline
Information Extraction
GPT-4V
Object: …
Text/numbers: …
Attributes: …
Relationships: …
Events: …
Understanding: …
GPT-4V
Data Generation
Positive sample generation
Q: What is on the back of the bus?
A: There is an advertisement on the back of the bus.
Negative sample generation
Q: Is there a for sale sign on the back of the bus?
A: No, there is an advertisement on the back of the bus.
Sample Filtering
Understanding-based Prompting
Step 1
Prompt
MLLM
Structured formation
VAR
Step 2
Or
Structured formation
Question
MLLM
Response

Figure 2. The MMVU dataset consists of a benchmarking dataset for evaluating models as well as a training dataset. The former is curated by human annotators together with the appropriate metrics and analysis on the MLLM's attention and logit behavior. Based on the experiments, we propose a data construction pipeline to build a training dataset and prompting strategies to enhance the accuracy of MLLM responses.

For a comprehensive review of these topics, readers are encouraged to investigate surveys [7, 25, 66].

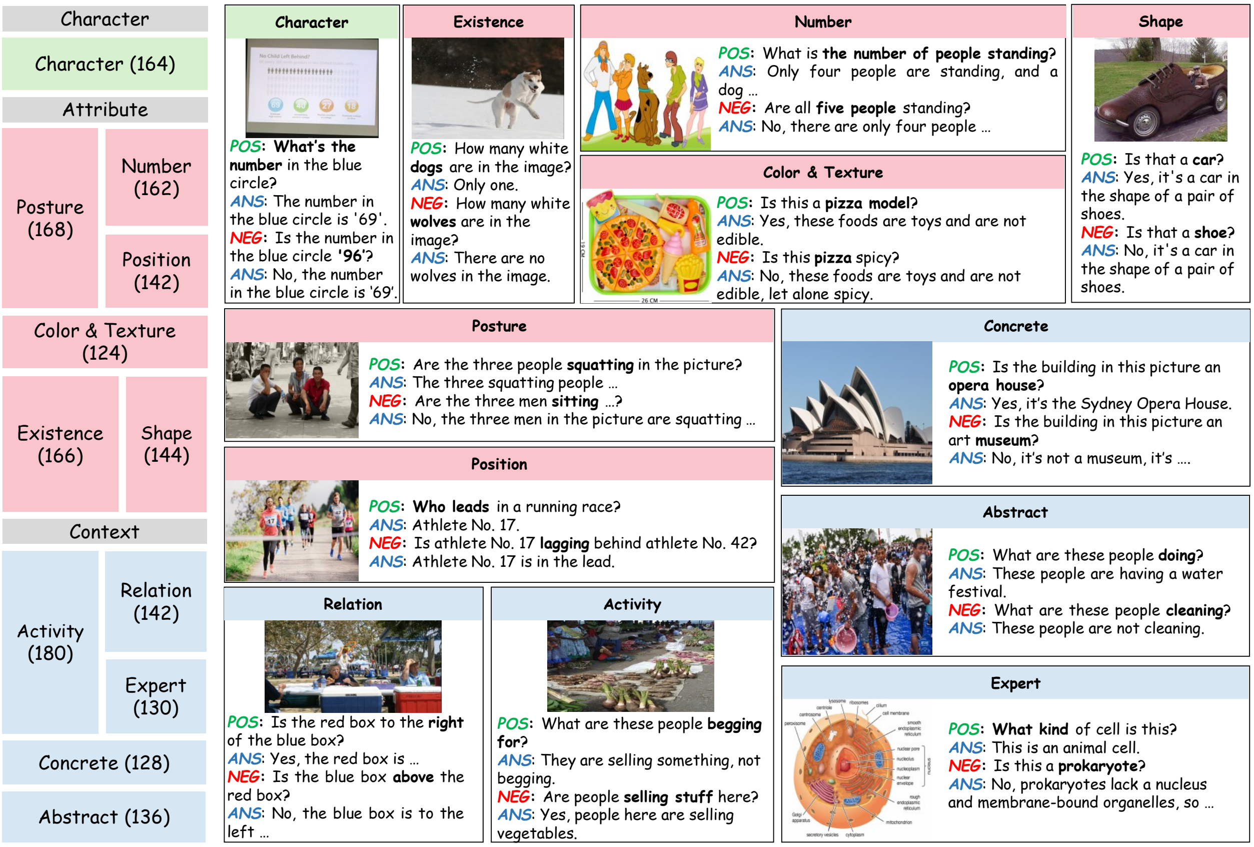
Evaluations and Analysis of MLLMs. The assessment of MLLMs [34] is mainly divided into two main categories: general capability [47, 70] and hallucination benchmarks [8]. General capability benchmarks can be divided into single-task and comprehensive benchmarks. Single-task benchmarks, such as VQA [17], VQA-v2 [17], GQA [22], and MS-COCO [56], are designed to assess the performance of models on a specific defined task. Comprehensive benchmarks, such as MMBench [45], SEED [32], MMMU [65], and MM-Star [11] aim to assess MLLMs' general multi-modal perception and reasoning capabilities. Hallucination benchmarks mainly focus on evaluating the discrepancy between generated responses and visual content [8, 36]. Most hallucination benchmarks focus on object [18, 23, 27, 35, 41, 44, 52–54, 62], knowledge [18, 23, 41, 44], relation [12, 23, 54, 64], attribute [18, 21, 23, 44, 53, 54, 64], and spurious images [19, 55]. Current benchmarks assume that correct answers reflect understanding, while incorrect ones indicate a lack of understanding. *However, our findings show that this incorrect stems not only from visual misunderstanding but also from a lack of robustness to negative questions.*

3. Methods

3.1. An Overview of MLLMs

The architecture of most MLLMs consists of three primary components: the visual encoder, the connector, and the LLM. Given a system prompt P, an input image I, and a question Q, the process commences with a text tokenizer that converts the system prompt and questions into text tokens T_P and T_Q, respectively. The image I is then encoded into visual tokens T_V by the visual encoder, which captures essential visual features. Subsequently, the connector integrates the visual tokens T_V with the text tokens to align the multi-modal data. Finally, the concatenated tokens (comprising text and visual tokens) are fed into the LLM to generate a response.

A significant challenge faced by current MLLMs is their ability to understand visual content while often providing incorrect responses to related questions. This paper aims to analyze the underlying causes of this issue and propose effective solutions. To facilitate this investigation, we introduce a **M**LLM **M**isresponses despite **V**isual **U**nderstanding (MMVU) benchmark, consisting of both positive and negative samples, as shown in Fig. 3. Each image in the benchmark is paired with two questions: a positive question that assesses the MLLM's comprehension of the visual content and a negative question intended to confuse the model. This benchmark allows for a comprehensive analysis of the MLLMs' performance in accurately answering questions and their vulnerability to misleading prompts. We establish two key metrics to evaluate the model's ability to respond correctly to both positive and negative questions and to assess its performance when confronted with misleading queries despite its understanding of the visual content. Additionally, we analyze the attention mechanism and logit distribution of



Character

POS: **What's the number** in the blue circle?
ANS: The number in the blue circle is '69'.
NEG: Is the number in the blue circle **'96'**?
ANS: No, the number in the blue circle is '69'.

Existence

POS: How many white **dogs** are in the image?
ANS: Only one.
NEG: How many white **wolves** are in the image?
ANS: There are no wolves in the image.

Number

POS: What is **the number of people standing**?
ANS: Only four people are standing, and a dog ...
NEG: Are all **five people** standing?
ANS: No, there are only four people ...

Color & Texture

POS: Is this a **pizza model**?
ANS: Yes, these foods are toys and are not edible.
NEG: Is this **pizza** spicy?
ANS: No, these foods are toys and are not edible, let alone spicy.

Shape

POS: Is that a **car**?
ANS: Yes, it's a car in the shape of a pair of shoes.
NEG: Is that a **shoe**?
ANS: No, it's a car in the shape of a pair of shoes.

Posture

POS: Are the three people **squatting** in the picture?
ANS: The three squatting people ...
NEG: Are the three men **sitting** ...?
ANS: No, the three men in the picture are squatting ...

Concrete

POS: Is the building in this picture an **opera house**?
ANS: Yes, it's the Sydney Opera House.
NEG: Is the building in this picture an art **museum**?
ANS: No, it's not a museum, it's ....

Position

POS: **Who leads** in a running race?
ANS: Athlete No. 17.
NEG: Is athlete No. 17 **lagging** behind athlete No. 42?
ANS: Athlete No. 17 is in the lead.

Abstract

POS: What are these people **doing**?
ANS: These people are having a water festival.
NEG: What are these people **cleaning**?
ANS: These people are not cleaning.

Relation

POS: Is the red box to the **right** of the blue box?
ANS: Yes, the red box is ...
NEG: Is the blue box **above** the red box?
ANS: No, the blue box is to the left ...

Activity

POS: What are these people **begging for**?
ANS: They are selling something, not begging.
NEG: Are people **selling stuff** here?
ANS: Yes, people here are selling vegetables.

Expert

POS: **What kind** of cell is this?
ANS: This is an animal cell.
NEG: Is this a **prokaryote**?
ANS: No, prokaryotes lack a nucleus and membrane-bound organelles, so ...

a. MMVU Distribution

b. Examples of MMVU

Figure 3. Examples in the MMVU benchmark. *POS* and *NEG* denote the positive and negative questions, *ANS* denotes the answer.

some open-source MLLMs to gain deeper insights into why this phenomenon happens.

3.2. MMVU Benchmark Design

As illustrated in Fig. 3, we design paired positive and negative questions to ensure a comprehensive evaluation. Each question offers four brief options, with only one correct answer. In total, we manually construct 1,786 questions (893 positive and 893 negative) across three levels: character, attribute, and context. Character-level questions focus on elements such as characters or numbers, attribute-level questions assess properties like color, texture, and quantity, and context-level questions explore higher-level concepts such as emotions, culture, and common sense. The positive questions are designed to evaluate the model's comprehension abilities, while the negative questions test its resistance to interference. For example, misleading character-level questions alter elements like characters or numbers, attribute-level questions introduce property confusion, and context-level questions challenge the model with complex concepts, assessing its robustness against misleading information.

3.3. Evaluation metrics

Misleading Rate and Robustness Accuracy. We develop metrics grounded in Bayesian conditional probabilities to evaluate whether an MLLM comprehends visual content and its robustness against misdirection. Specifically, we categorize the results into four groups: 1) Understanding and Robustness (UR): The model accurately comprehends the visual content and remains resilient when exposed to negative questions. 2) Understanding but Fragility (UF): The model successfully understands the visual content but is vulnerable to errors when faced with negative questions. 3) Not Understanding and Rigorous (NR): The model fails to grasp the visual content but still provides correct answers to negative questions. 4) Not Understanding and Fragility (NF): The model neither comprehends the visual content nor answers the negative questions correctly. To quantify the effect of negative questions in misleading the model despite its comprehension of visual content, we introduce the "Misleading Rate (MR)". Additionally, we introduce the "Robustness Accuracy (RA)" to evaluate the understanding capability of MLLMs. The formulation is as follows:

$$\mathrm{MR} = \frac{N_{UF}}{N_{UR} + N_{UF}}, \tag{1}$$

$$\mathrm{RA} = \frac{N_{UR}}{N_{UR} + N_{UF} + N_{NR} + N_{NF}}, \tag{2}$$

where N_{i} (i $\in \{UR, UF, NR, NF\}$) represents the number of samples. With these two evaluation metrics, our benchmark can offer a more accurate and reasonable reflection of

Table 1. Comparison to state-of-the-art MLLMs on the MMVU test set. Abbreviations: Char/Num. (Character/Number), Pres. (Presence), Color/Tex. (Color/Texture), Num. (Number), Shape (Shape), Posture (Posture), Pos. (Position), Abstract. (Abstract Knowledge), Concrete. (Concrete Knowledge), Expert (Expertise), Act. (Activity), Rel. (Relationships). ↑: higher is better, ↓: lower is better. Bolding and underlining indicate the best and second-best performance, respectively.

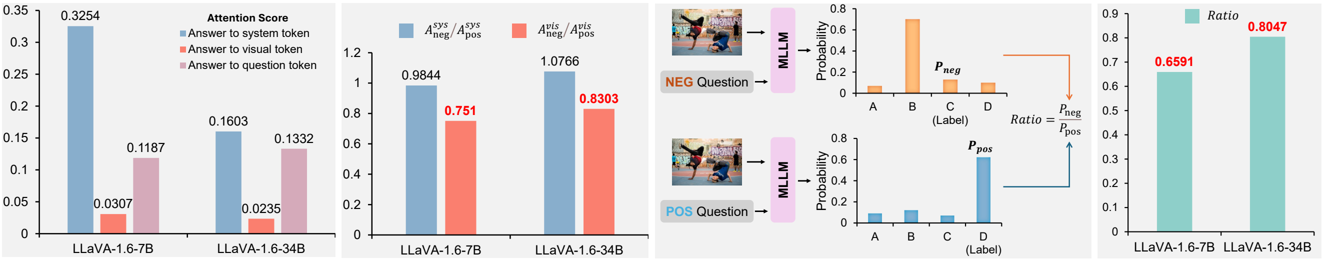
Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA ↑
Random	25	25	25	25	25	25	25	25	25	25	25	25	25
Closed-source Models													
GPT4o [51] 🥉	**76.83**	74.70	67.74	38.27	**76.39**	78.57	49.30	72.06	70.31	53.85	74.44	43.66	65.06
Qwen-VL-max-0809 [6]	67.07	60.87	61.29	34.48	76.00	62.07	42.86	69.57	72.88	**67.74**	70.73	54.55	63.36
Claude3.5-Sonnet-0620 [5]	64.63	53.01	66.13	38.27	56.94	67.86	38.03	64.71	64.06	44.62	67.78	35.21	55.32
Gemini-1.5-Pro [60]	47.56	36.14	41.94	23.46	43.06	40.48	15.49	60.29	56.25	32.31	50.00	28.17	39.53
Open-source Models													
Ovis1.6-Gemma2-9B [48] 🥇	71.95	74.70	69.35	**58.02**	65.28	76.19	57.75	**77.94**	68.75	47.69	75.56	52.11	**66.74**
Llama3.2-90B-Vision-Instruct [15] 🥈	60.98	**80.72**	**74.19**	55.56	70.83	**79.76**	**60.56**	67.65	**76.56**	27.69	**82.22**	**56.34**	**66.74**
MiniCPM-2.6V [63]	69.51	72.29	69.35	48.15	51.39	73.81	45.07	70.59	71.88	38.46	80.00	35.21	61.14
LLaVA-OneVision-7B [33]	62.20	67.47	70.97	41.98	58.33	70.24	49.30	**77.94**	71.88	35.38	75.56	42.25	60.58
Idefics3-8B-Llama3 [30]	57.32	61.45	61.29	43.21	69.44	64.29	40.85	61.76	64.06	30.77	68.89	36.62	55.43
Pixtral-12B-2409 [3]	48.78	65.06	66.13	49.38	50.00	54.76	56.34	61.76	65.63	24.62	66.67	39.44	54.31
Phi3.5-Vision-Instruct [1]	54.88	53.01	61.29	37.04	58.33	55.95	35.21	60.29	64.06	29.23	68.89	38.03	51.62
InternVL2-8B [13]	46.34	65.06	53.23	33.33	56.94	53.57	39.44	61.76	65.63	29.23	66.67	21.13	49.72
LLaVA-v1.6-Vicuna-7B [42]	39.02	56.63	59.68	34.57	54.17	58.33	22.54	55.88	60.94	24.62	58.89	38.03	47.14
Cambrian-8B [61]	35.37	60.24	41.94	40.74	41.67	57.14	39.44	44.12	42.19	24.62	51.11	28.17	42.89
VILA1.5-13B [39]	14.63	30.12	35.48	12.35	27.78	42.86	8.45	29.41	29.69	16.92	43.33	14.08	25.76

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Random	50	50	50	50	50	50	50	50	50	50	50	50	50
Closed-source Models													
GPT4o [51] 🥈	**8.69**	16.22	16.00	38.00	11.29	**12.00**	32.69	18.33	19.64	23.91	16.25	35.42	19.53
Qwen-VL-max-0809 [6]	25.68	30.00	**13.64**	33.33	**9.52**	30.77	40.00	30.43	21.82	25.00	17.14	40.00	25.35
Claude3.5-Sonnet-0620 [5]	15.87	29.03	18.00	34.04	22.64	21.92	34.15	18.52	19.61	25.64	18.67	**19.35**	22.69
Gemini-1.5-Pro [60]	26.42	42.31	38.10	53.66	31.11	37.04	56.00	25.45	26.53	34.38	31.82	33.33	35.11
Open-source Models													
Llama3.2-90B-Vision-Instruct [15] 🥇	23.08	**12.99**	16.36	25.00	17.74	14.10	**15.69**	25.81	**16.95**	41.94	**9.76**	**18.37**	**18.47**
Ovis1.6-Gemma2-9B [48] 🥉	18.06	16.22	17.31	**20.34**	25.40	17.95	**22.64**	14.52	25.42	**20.51**	16.05	27.45	19.78
MiniCPM-2.6V [63]	18.57	18.92	14.00	29.09	36.21	18.42	36.00	17.24	22.03	43.18	12.20	39.02	23.85
LLaVA-OneVision-7B [33]	30.14	24.32	21.43	43.33	30.00	21.33	36.36	**10.17**	22.03	51.06	18.07	37.50	27.77
Idefics3-8B-Llama3 [30]	29.85	28.17	26.92	28.57	15.25	25.00	38.30	31.15	29.31	48.72	21.52	36.59	28.78
Pixtral-12B-2409 [3]	34.43	22.86	18.00	25.93	35.71	34.29	24.53	25.00	22.22	52.94	22.08	44.00	29.20
Phi3.5-Vision-Instruct [1]	31.82	27.87	19.15	37.50	27.59	29.85	46.81	26.79	24.07	52.50	13.89	27.03	29.40
InternVL2-8B [13]	34.48	22.86	28.26	40.00	24.07	28.57	37.78	32.26	17.65	50.00	17.81	57.14	30.63
LLaVA-v1.6-Vicuna-7B [42]	45.76	25.40	31.48	39.13	30.36	26.87	52.94	30.91	25.00	60.00	30.26	28.95	34.22
Cambrian-8B [61]	47.27	21.88	27.78	45.00	33.33	26.15	46.15	38.78	37.21	44.83	14.81	58.33	36.17
VILA1.5-13B [39]	79.66	59.02	56.00	78.26	58.33	45.45	82.86	58.33	59.57	72.50	46.58	72.97	62.30

Table 2. Ablation study about the MMVU training set on the MMVU test set. Blue text denotes the extent of the performance improvement.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA ↑
LLaVA-OneVision-7B [42]	62.20	67.47	70.97	41.98	58.33	70.24	49.30	77.94	71.88	35.38	75.56	42.25	**60.58** (Base)
LLaVA-OneVision-7B + MMVU-Train [42]	68.29	73.49	72.58	55.56	63.89	69.05	47.89	70.59	73.44	38.46	80.00	42.25	**63.49** (+2.91)
InternVL2-8B [13]	46.34	65.06	53.23	33.33	56.94	53.57	39.44	61.76	65.63	29.23	66.67	21.13	**49.72** (Base)
InternVL2-8B + MMVU-Train [13]	62.20	71.08	69.35	40.74	61.11	66.67	42.25	69.12	59.38	40.00	74.44	33.80	**58.01** (+8.29)
VILA1.5-13B [46]	14.63	30.12	35.48	12.35	27.78	42.86	8.45	29.41	29.69	16.92	43.33	14.08	**25.76** (Base)
VILA1.5-13B + MMVU-Train [46]	23.17	51.81	50.00	19.75	47.22	53.57	19.72	42.65	34.38	15.38	65.56	21.13	**37.74** (+11.98)
Phi3.5-Vision-Instruct [1]	54.88	53.01	61.29	37.04	58.33	55.95	35.21	60.29	64.06	29.23	68.89	38.03	**51.62** (Base)
Phi3.5-Vision-Instruct + MMVU-Train [1]	60.98	56.63	61.29	39.51	54.17	60.71	40.85	58.82	64.06	30.77	70.00	40.85	**53.64** (+2.02)
LLaVA-v1.6-Vicuna-7B [42]	39.02	56.63	59.68	34.57	54.17	58.33	22.54	55.88	60.94	24.62	58.89	38.03	**47.14** (Base)
LLaVA-v1.6-Vicuna-7B + MMVU-Train [42]	50.0	59.04	64.52	35.8	54.17	63.1	19.72	58.82	59.38	36.92	68.89	38.03	**51.06** (+3.92)

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
LLaVA-OneVision-7B [33]	30.14	24.32	21.43	43.33	30.00	21.33	36.36	10.17	22.03	51.06	18.07	37.50	**27.77** (Base)
LLaVA-OneVision-7B + MMVU-Train [33]	18.84	17.57	15.09	21.05	23.33	20.55	20.93	17.24	20.34	47.92	12.20	28.57	**21.03** (-6.74)
InternVL2-8B [13]	34.48	22.86	28.26	40.00	24.07	28.57	37.78	32.26	17.65	50.00	17.81	57.14	**30.63** (Base)
InternVL2-8B + MMVU-Train [13]	22.73	16.90	14.00	35.29	26.67	22.22	40.00	20.34	28.30	39.53	12.99	38.46	**25.04** (-5.59)
VILA1.5-13B [39]	79.66	59.02	56.00	78.26	58.33	45.45	82.86	58.33	59.57	72.50	46.58	72.97	**62.30** (Base)
VILA1.5-13B + MMVU-Train [39]	64.15	29.51	32.61	62.79	35.85	31.82	66.67	44.23	43.59	74.36	21.33	65.91	**45.02** (-17.28)
Phi3.5-Vision-Instruct [1]	31.82	27.87	19.15	37.50	27.59	29.85	46.81	26.79	24.07	52.50	13.89	27.03	**29.40** (Base)
Phi3.5-Vision-Instruct + MMVU-Train [1]	25.37	24.19	22.45	34.69	29.09	28.17	36.96	28.57	22.64	48.72	14.86	25.64	**27.42** (-1.98)
LLaVA-v1.6-Vicuna-7B [42]	45.76	25.40	31.48	39.13	30.36	26.87	52.94	30.91	25.00	60.00	30.26	28.95	**34.22** (Base)
LLaVA-v1.6-Vicuna-7B + MMVU-Train [42]	26.79	14.04	21.57	36.96	25.0	18.46	44.0	21.57	17.39	42.86	16.22	22.86	**24.00** (-10.22)



(a) (b) (c) (d)

Figure 4. Please refer to Sec. 3.4 for details on the following calculations. (a) Statistical results of attention scores between the answer tokens and the system, visual, and question tokens. It is seen that the answer tokens pay the least attention to the visual tokens in general. (b) The ratio of the question tokens' attention to the system and visual tokens in negative samples versus positive samples. Negative questions appear to pay less attention than positive questions to visual tokens. (c) Procedure for calculating the ratio of output probabilities for negative and positive samples. (d) Comparison of the ratio of output probabilities for negative and positive samples across different MLLMs. Interestingly, it seems that a lower output probability correlates with lower attention between the question and visual tokens in (b).

the model's understanding capabilities and reveal its robustness to misleading prompts.

3.4. Analysis Methods

Attention Analysis of Answer Tokens. To investigate the relationship between answer tokens and other token types in MLLMs, we analyze the final layer's attention matrix A. We average A across dimensions to obtain a unified attention matrix. We then extract sub-matrices corresponding to the attention from answer tokens to each of the system, image, and question tokens. This results in an attention matrix A of size $N \times N \times d$, where each element $A[i, j]$ indicates the attention weight from token i to token j. The total token length N is defined as $N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}} + N_{\text{a}}$, where N_{sys}, N_{vis}, N_{q}, and N_{a} represent the number of system, image, question, and answer tokens, respectively. For example, $A[N_{\text{sys}}+N_{\text{vis}}+N_{\text{q}} : N_{\text{sys}}+N_{\text{vis}}+N_{\text{q}}+N_{\text{a}}, : N_{\text{sys}}]$ denotes the attention from the answer to system tokens, $A[N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}} : N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}} + N_{\text{a}}, N_{\text{sys}} : N_{\text{sys}} + N_{\text{vis}}]$ means the attention from the answer to image tokens, and $A[N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}} : N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}} + N_{\text{a}}, N_{\text{sys}} + N_{\text{vis}} : N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}}]$ indicates the attention from the answer to question tokens. We first get the row-wise maximum values for each sub-matrix and produce a one-dimensional vector. Then we average these values to obtain an attention score that quantifies the overall attention score from answer tokens to each token type.

Attention Analysis in Response to Positive and Negative Questions. We analyze the attention between question tokens and system or image tokens to explore how the model responds to positive and negative questions. We extract two sub-matrices from A to assess the attention between question tokens and system or image tokens: $A[N_{\text{sys}}+N_{\text{vis}} : N_{\text{sys}}+N_{\text{vis}}+N_{\text{q}}, : N_{\text{sys}}]$ and $A[N_{\text{sys}}+N_{\text{vis}} : N_{\text{sys}} + N_{\text{vis}} + N_{\text{q}}, N_{\text{sys}} : N_{\text{sys}} + N_{\text{vis}}]$. We get the row-wise maximum values for each sub-matrix, yielding a one-dimensional vector. To assess the minimum level of attention between the question tokens and each token type, we compute the minimum value of the vector, representing the lower bound of attention. The results are denoted as A^{sys}_{pos} and $A^{vis}_{\text{pos}}$for positive questions as well as A^{sys}_{neg} and A^{vis}_{neg} for negative questions. Finally, we compare the attention scores by calculating the ratio $A^{sys}_{\text{neg}}/A^{sys}_{\text{pos}}$ and $A^{vis}_{\text{neg}}/A^{vis}_{\text{pos}}$.

Logits Analysis. To examine the model's confidence in selecting the correct answer under different question types, we conduct a logit analysis. As shown in Fig. 4 (c), we input both an image and a question (either positive or negative) into the MLLM and extract the logits for each answer option. These logits represent the model's unnormalized confidence in each option. We apply the softmax function to convert the logits into a probability distribution. The probability of each option is given by:

$$p_i = \frac{e^{\text{logit}_i}}{\sum_j e^{\text{logit}_j}}, \tag{3}$$

where logit_i denotes the raw logit value for option i, and p_i is the corresponding normalized selection probability. To assess the model's confidence, we compute two probabilities: P_{pos} and P_{neg}. Here, P_{pos} represents the probability of selecting the correct answer when given the positive form of the question, and P_{neg} is the corresponding probability for the negative form of the question. These scores reflect the model's confidence in choosing the correct answer under different question polarities. We define the confidence ratio, $Ratio = \frac{P_{\text{neg}}}{P_{\text{pos}}}$, to quantify the change in confidence between positive and negative questions.

3.5. MMVU Training Set Construction Pipeline

To enhance the understanding capability and robustness of MLLMs, we propose a data construction pipeline that utilizes GPT-4o to generate paired samples for instruction tuning as part of the MMVU dataset (see Fig. 2). We first implicitly extract the information from the image and then generate paired positive and negative samples.

Information extraction. Previous efforts that generate instruction-tuning data for multimodal conversations with GPT-4o primarily fall into two categories: direct generation and annotation-driven generation. Direct generation

Table 3. Comparing instruction datasets with MMVU. We train with these datasets in the same backbone (SigLIP+Phi-2).

Dataset	Size	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA ↑
LLaVA 158k [42]	158k	22.50	27.27	25.00	12.50	58.33	16.67	9.09	36.36	42.86	25.81	45.83	40.91	29.67
SVIT 158k [68]	158k	30.00	27.27	16.67	37.50	66.67	25.00	18.18	40.91	38.10	19.35	58.33	31.82	33.67
LLaVA+SVIT	316k	32.50	22.73	16.67	25.00	66.67	25.00	13.64	31.82	38.10	22.58	50.00	40.91	32.00
LRV [41]	340k	27.50	31.82	16.67	25.00	41.67	25.00	18.18	40.91	19.05	29.03	41.67	45.45	30.00
MMVU-Train	48k	40.00	18.18	50.00	33.33	70.83	25.00	22.73	63.64	57.14	41.94	58.33	40.91	43.33
MMVU-Train	112k	42.50	22.73	45.83	37.50	70.83	29.17	27.27	63.64	52.38	51.61	62.50	40.91	45.67

Dataset	Size	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
LLaVA 158k [42]	158k	65.38	50.00	64.71	76.92	30.00	78.95	75.00	60.00	52.63	66.67	38.89	35.71	57.62
SVIT 158k [68]	158k	50.00	50.00	77.78	25.00	20.00	64.71	63.64	57.14	57.89	71.43	22.22	41.67	50.73
LLaVA+SVIT	316k	53.57	61.54	77.78	60.00	23.81	68.42	72.73	63.16	55.56	74.07	33.33	43.75	56.95
LRV [41]	340k	57.69	53.33	77.78	53.85	47.37	71.43	55.56	57.14	78.95	64.00	41.18	37.50	58.90
MMVU-Train	48k	44.83	66.67	25.00	38.46	22.73	70.00	44.44	33.33	33.33	45.83	30.00	43.75	40.91
MMVU-Train	112k	34.62	54.55	31.25	30.77	22.73	68.18	33.33	26.32	35.29	38.46	25.00	43.75	36.87

approaches, such as ALLaVA [9], rely entirely on the basic ability of GPT-4o to synthesize samples. In contrast, annotation-driven generation methods, such as LRV [41], utilize existing image annotations (*e.g.* object bounding boxes and textual descriptions) as additional information for text-only GPT-4 to generate conversational samples. However, these works fail to make full use of fine-grained details. On the contrary, our data pipeline adopts a different scheme by implicitly and comprehensively extracting the detailed information directly from the image. This includes 1) text or numbers (if present), 2) objects and people, 3) object attributes (*e.g.*, colors, textures, locations) and human characteristics (*e.g.*, postures, dresses), 4) interrelationships between people and objects, 5) events or activities, and 6) the overall feeling and understanding evoked by the image.

Instruction tuning data generation. The most relevant work to our data generation approach is LRV [41], which uses the Visual Genome [28] dataset to guide GPT-4 in generating positive and negative samples about object presence and manipulation knowledge. However, LRV has two significant limitations: limited fine-grained information and unpaired negative samples. It relies solely on bounding boxes and object labels, omitting textual descriptions, numerical attributes, and deeper image comprehension. Additionally, LRV's negative samples are not explicitly paired with positive counterparts, reducing training effectiveness. To address these challenges, we generate positive samples using extracted information and construct negative samples that directly contradict the positive ones. This ensures a strong contrast for more effective model training, providing richer context and paired positive and negative samples.

Sample filtering. We randomly sample COCO [40] images from the LLaVA 158k and construct paired positive and negative samples. We filter the samples by keyword matching, removing conversations with uncertain answers (*e.g.*, "*uncertain*") and redundant phrases (*e.g.*, "*in the image*").

3.6. Understanding-based Prompting

In Sec. 4.2, we observe that the model exhibits the highest attention scores for system and question tokens. In contrast, attention to visual tokens is notably low. We propose the following strategies to enhance the model's performance (see Fig. 2): **1) Content Guided Refinement (CGR).** We propose a two-step reasoning method where the first step involves extracting detailed information from the image, followed by a second step in which the model answers the question based on the extracted information and the visual content. This structured reasoning aims to improve the model's ability to connect visual understanding with accurate responses. **2) Visual Attention Refinement (VAR).** We employ MLLM to extract attention scores from the image and apply a mask to the original image. This aims to emphasize areas that receive greater attention while masking unimportant regions, guiding the model to focus on significant visual features that are most relevant to the question.

4. Experiments

4.1. Analysis of MMVU Benchmark

Comparison of MLLMs on the MMVU benchmark. We evaluate 4 closed-source and 11 leading open-source models on the MMVU benchmark, as shown in Tab. 1. Our findings reveal that closed-source and open-source models are fragile against negative questions despite evidence that they understand the visual content. GPT-4o achieved the highest performance among the closed-source models, while Ovis1.6-Gemma2-9B led among the open-source models, achieving an RA metric of 66.74%. However, GPT-4o is more susceptible to negative questions, with an MR of 19.53%, compared to Ovis1.6-Gemma2-9B (18.60% for MR). Experimental results show that all these MLLMs perform poorly in the face of negative questions.

Performance of MLLMs on different types of intrusive questions. Tab. 1 reveals that: 1) The model is particularly vulnerable to issues involving character-level details, numerical data, positional context within phrase-level categories, expertise knowledge, and relative relationships between objects within sentence-level categories. 2) Subcategories with high positive comprehension capabilities tend to exhibit greater resilience to negative questions.

4.2. Analysis of attention and logits distribution

To investigate why MLLMs can comprehend visual content yet sometimes respond inaccurately, we analyze the attention allocation of LLaVA 1.6 (7B and 34B) models. Using the MMVU benchmark, we examine attention scores in the

Table 4. Ablation study on understanding-based prompting in MMVU. Blue and Red indicate performance gains and drops, respectively.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA ↑
Baseline	39.02	56.63	59.68	34.57	54.17	58.33	22.54	55.88	60.94	24.62	58.89	38.03	**47.14** (base)
With Instruction	35.37	49.40	59.68	32.10	37.50	54.76	18.31	54.41	54.69	26.15	61.11	33.80	**43.34** (-3.80)
With CGR	40.24	55.42	61.29	35.80	54.17	58.33	22.54	57.35	60.94	26.15	62.22	38.03	**47.93** (+0.79)
With VAR	40.24	51.81	61.29	32.10	47.22	66.67	26.76	60.29	59.38	24.62	65.56	30.99	**47.59** (+0.45)
With VAR and CGR	41.46	50.60	61.29	32.10	48.60	66.67	22.54	60.29	59.38	24.62	65.56	32.39	**47.48** (+0.36)

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Baseline	45.76	25.40	31.48	39.13	30.36	26.87	52.94	30.91	25.00	60.00	30.26	28.95	**34.22** (base)
With Instruction	49.12	29.31	31.48	45.83	49.06	35.21	64.86	31.48	36.36	60.47	27.63	40.00	**40.09** (+6.68)
With CGR	44.07	25.81	29.63	39.58	30.36	26.87	54.29	29.09	22.00	58.54	26.32	28.95	**33.23** (-0.99)
With VAR	44.07	28.33	26.92	36.59	38.18	22.22	50.00	30.51	22.45	57.89	18.06	43.59	**32.97** (-1.25)
With VAR and CGR	42.37	30.00	26.92	36.59	35.18	22.22	55.56	30.51	22.45	58.97	16.90	42.50	**32.91** (-1.31)

final layer, specifically focusing on attention distribution between the answer tokens and the system, question, and visual tokens (see Fig. 4 (a)). Results show that answer tokens prioritize system tokens, followed by question tokens, and visual tokens. This trend indicates a potential under-utilization of visual information during decoding.

We hypothesize that insufficient attention between question and visual tokens correlates with lower-quality answer probabilities. To this end, we analyze the attention between question tokens and both system and visual tokens, as depicted in Fig.4 (b). The results show that attention scores between questions and visual tokens are lower in response to negative than positive questions. Additionally, we evaluate model responses to both positive and negative questions using identical images, measuring the probabilities of accurate answers for each type (see Fig.4 (c)). For cases where the model correctly answers positive questions but struggles with negative ones, we further assess the probability of selecting correct options, as shown in Fig. 4 (d). Our findings indicate a notable decline in confidence when choosing correct answers for negative questions, underscoring the difficulty in preserving response accuracy under negative questions.

4.3. Analysis of MMVU Training Set

Experiment results on the MMVU benchmark. To evaluate our proposed dataset, we fine-tune several state-of-the-art MLLMs, including LLaVA-OneVision-7B [42], InternVL2-8B [13], VILA1.5-13B [46], and Phi3.5-Vision-Instruct [1], using MMVU training set (Sec. 3.5) and assess their performance on the MMVU benchmark. Weight merging is utilized in the experiment. As shown in Tab. 2, models fine-tuned with our dataset consistently outperformed their baselines across most metrics. Notably, LLaVA-OneVision-7B + MMVU-Train demonstrated improvements in character/number recognition (68.29% vs. 62.20%), numerical understanding (55.56% vs. 41.98%), and activity recognition (80.00% vs. 75.56%). These results indicate that the paired positive-negative data can reduce biased responses.

Comparison of instruction tuning datasets generated by GPT. We compare the MMVU-Train dataset with previous GPT-4V generated instruction tuning datasets, such as LLaVA 158k [43], SVIT 158k [68], and LRV [41] on the MMVU benchmark (as shown in Tab. 3). The results demonstrate that our dataset enables the model to outperform the model trained with previous datasets.

4.4. Analysis of Understanding-based Prompting

We evaluate the effectiveness of these two strategies using the LLaVA 1.6 7B on the MMVU benchmark, as shown in Tab. 4. "Instruction" refers to the detailed instructions in the system prompt, which is available in the *Supplementary Material*. "CGR" denotes the content-guided refinement strategy, while "VAR" indicates the application of the visual attention refinement strategy. The results reveal that simply incorporating the instruction prompt leads to performance degradation. Notably, the introduction of CGR alone results in a greater increase in the model's accuracy on positive questions compared to VAR alone. In contrast, VAR significantly enhances the correctness of responses that depend on visual understanding. When both strategies are combined, the model improves its accuracy in responding after being misled, but this combination does not replicate the previous gains in question comprehension. This phenomenon occurs because CGR prioritizes information extraction to enhance accuracy on positive questions, while VAR emphasizes attention to critical visual features; their combination may shift the model's focus and limit further improvements.

5. Conclusion

In this paper, we identify a significant limitation of MLLMs: despite they accurately understand visual content, they may provide incorrect answers. We introduce the MMVU benchmark to investigate the prevalence of this issue. We also analyze the attention mechanisms and the logits during decoding to uncover their underlying causes. Our findings reveal that MLLMs exhibit greater attention to system and question tokens compared to visual tokens, particularly when confronted with negative questions. This behavior is attributed to biases in the training datasets and insufficient focus on visual elements. To mitigate these challenges, we propose a data construction pipeline for generating paired positive and negative samples as part of the MMVU training set, and two strategies to enhance the model's attention to visual information during inference. Experimental results demonstrate that our dataset and proposed strategies significantly improve the robustness of MLLMs in responding to negative questions.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, and Ahmed Awadallah et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 5, 8

[2] 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. 2

[3] Mistral AI. Pixtral system card. https://mistral.ai/news/pixtral-12b/, 2024. 5

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[5] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. 5

[6] 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. 5

[7] Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*, 2024. 3

[8] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 3

[9] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 7

[10] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2301–2310, 2024. 2

[11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3

[12] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023. 3

[13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5, 8

[14] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024. 2

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[17] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[18] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023. 3

[19] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to hallucination in mllms. *arXiv preprint arXiv:2402.03757*, 2024. 3

[20] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 2

[21] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*, 2024. 3

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[23] Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *arXiv preprint arXiv:2402.15721*, 2024. 3

[24] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv preprint arXiv:2410.09453*, 2024. 2

[25] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024. 3

[26] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: In-

vestigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024. 2

[27] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *arXiv preprint arXiv:2405.05256*, 2024. 3

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 7

[29] LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023. 2

[30] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 5

[31] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 2

[32] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CVPR*, 2024. 3

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5

[34] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. A survey on benchmarks of multimodal large language models, 2024. 3

[35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[36] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 2, 3

[37] 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*, 2024. 2

[38] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models, 2024. 2

[39] 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*, 2023. 5

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014. 7

[41] 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*, 2023. 3, 7, 8

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 7, 8

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 8

[44] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*, 2024. 3

[45] 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? *arXiv:2307.06281*, 2023. 3

[46] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 5, 8

[47] Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[48] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 5

[49] Kaijing Ma, Haojian Huang, Jin Chen, Haodong Chen, Pengliang Ji, Xianghao Zang, Han Fang, Chao Ban, Hao Sun, Mulin Chen, et al. Beyond uncertainty: Evidential deep learning for robust video temporal grounding. *arXiv preprint arXiv:2408.16272*, 2024. 2

[50] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 2

[51] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2024. 5

[52] Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E Gonzalez, and Trevor Darrell.

Aloha: A new measure for hallucination in captioning models. *arXiv preprint arXiv:2404.02904*, 2024. 3

[53] 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. 2, 3

[54] Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv preprint arXiv:2404.13874*, 2024. 3

[55] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*, 2024. 3

[56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[57] Danli Shi, Weiyi Zhang, Xiaolan Chen, Yexin Liu, Jiancheng Yang, Siyu Huang, Yih Chung Tham, Yingfeng Zheng, and Mingguang He. Eyefound: a multimodal generalist foundation model for ophthalmic imaging. *arXiv preprint arXiv:2405.11338*, 2024. 2

[58] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 2

[59] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[60] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 5

[61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 5

[62] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer, 2024. 3

[63] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5

[64] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*, 2023. 3

[65] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 3

[66] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024. 3

[67] Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. Unveiling the tapestry of consistency in large vision-language models. *arXiv preprint arXiv:2405.14156*, 2024. 2

[68] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 7, 8

[69] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*, 2024. 2

[70] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 3

[71] Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv:2403.06199*, 2024. 2