Seeing is Believing? Counting Bananas Helps Multimodal Large Language Models Mitigate Modality Bias

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) often encounter irrelevant or misleading images in real-world applications. To handle such challenges, MLLMs must dynamically adjust their reliance on different modalities based on relevance. However, we find that MLLMs disproportionately favor visual inputs, even when textual cues are equally informative. This modality bias leads to imbalanced reasoning and reduced robustness, especially when irrelevant images are present. In this paper, we systematically investigate modality bias by designing a Banana-Counting dataset, where identical information is embedded in both textual and visual formats, ensuring that models have equal access to both modalities. Our findings reveal that most MLLMs prioritize visual information even when textual cues provide equally informative content. To mitigate this bias, we design a balanced multimodal Banana-Counting training dataset and fine-tune MLLMs using LoRA-based adaptation. Our approach significantly reduces modality bias while maintaining or even improving general reasoning performance on datasets such as ScienceQA, CSQA, and MMLU. Additionally, our fine-tuned models demonstrate enhanced robustness against noisy figures, ensuring more reliable performance in real-world multimodal scenarios. Our study highlights the importance of balanced multimodal training strategies and provides insights into improving MLLMs' ability to integrate information effectively across modalities.

1 Introduction

The rapid evolution of large language models (LLMs) has been a major driving force in the pursuit of Artificial General Intelligence (AGI). To meet the increasing demands of real-world applications, LLMs have advanced beyond single-modality processing, evolving into multimodal LLMs (MLLMs). These models are now capable of integrating and reasoning over multiple data modalities, including text, images, and audio (Liu et al., 2024; Bai et al., 2023; Achiam et al., 2023; Wang et al., 2023; Yao et al., 2024; Wu et al., 2023). With their remarkable performance in tasks such as visual question answering (VQA), image captioning, and multimodal reasoning, MLLMs have become indispensable in a wide range of applications (Yao et al., 2023; Ma et al., 2024; Bianco et al., 2023; Zhang et al., 2023).

Despite these advancements, in this paper, we reveal a concerning phenomenon with current MLLMs: **Modality Bias, which refers to an over-reliance on one modality while neglecting information from the others**. Specifically, our experiments find that many MLLMs tend to prioritize visual cues over textual information, even when text provides critical details. This bias appears to stem from extensive training on VQA and image captioning datasets, which heavily favor visual information. Consequently, this leads to two main issues in real-world applications: **(1) Limited Textual Processing:** MLLMs tend to prioritize visual information over textual content, limiting their ability to process and reason based on text. **(2) Susceptibility to Irrelevant Visual Input:** In real-world scenarios, users may provide irrelevant or misleading images. Ideally, an MLLM should dynamically adjust its reliance on visual inputs based on their relevance, rather than blindly incorporating visual information into its reasoning process.

In this paper, we aim to systematically investigate modality bias in MLLMs, analyze its implications, and propose an effective solution. To explore this issue, we design the Banana-Counting dataset, a multimodal dataset designed to evaluate whether MLLMs can effectively extract numerical information from both textual and visual sources rather than relying on a single modality. As far as we know, our dataset is the first to feature a balanced multimodal design, where identical information is embedded in both text and images, allowing for

a direct comparison of how models balance and integrate the two modalities.

Upon identifying clear evidence of modality bias, we introduce a simple yet effective intervention: a LoRA fine-tuning approach using our custom Banana-Counting training dataset. We train multiple MLLMs on this dataset and conduct a comprehensive evaluation. Our key findings demonstrate that fine-tuning with balanced multimodal data: (1) Effectively reduces modality bias, enabling MLLMs to integrate textual and visual information more equitably. (2) Enhances overall performance on benchmark datasets such as ScienceQA(Lu et al., 2022), CSQA(Talmor et al., 2019), and MMLU(Hendrycks et al., 2021). (3) Improves robustness in scenarios involving noisy or misleading images, making MLLMs more resilient to irrelevant visual input. By addressing modality bias, we provide insights into how MLLMs can be improved for more reliable multimodal reasoning and highlight the importance of balanced dataset construction in MLLM training.

2 Investigating Modality Bias in MLLMs

2.1 Preliminarily Study

Cognitive science research (Holsanova et al., 2009; Sivle and Uppstad, 2018) suggests that effective reading comprehension requires the ability to switch between textual and visual information to construct a coherent understanding. In contrast, many current MLLMs have been trained primarily on VQA-based and Optical Character Recognition (OCR) datasets (Liu et al., 2024) such as OKVQA (Marino et al., 2019), OCRVQA(Mishra et al., 2019), and TextCaps (Sidorov et al., 2020), which heavily emphasize image-dependent question answering. While these datasets effectively train models to interpret images, they may also inadvertently cause modality bias—a tendency for MLLMs to overly rely on visual information while neglecting textual information.

To empirically investigate this hypothesis, we evaluate two widely used MLLMs, Qwen2-VL-7B-Instruct (Bai et al., 2023) and Llama3-Llava-next-8b (Liu et al., 2024), on a subset of 1,000 test samples from the ScienceQA dataset (Lu et al., 2022). We design three experimental settings to assess MLLMs' reliance on visual input: (1) Normal: The original dataset with both textual and visual information. (2) No Figure: Images are removed, leaving only textual information. (3) Noisy Figure:

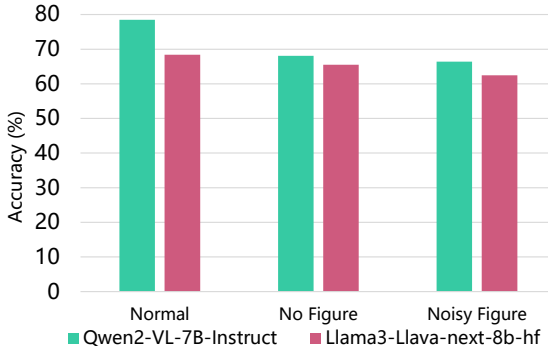

Figure 1: Accuracy of different LLMs on the ScienceQA dataset under various settings. This bar chart presents the accuracy (%) of Qwen2-VL-7B-Instruct and Llama3-Llava-next-8b on the ScienceQA dataset. The evaluation includes three conditions: Normal, No Figure, and Noisy Figure.

Images are replaced with randomly selected, irrelevant images from the ScienceQA dataset. The results are presented in Figure 1. As shown in the Figure, both MLLMs exhibit a strong dependence on visual information. The accuracy further degrades when irrelevant images are introduced (Noisy Figure) compared to No Figure setting, indicating that incorrect visual cues can mislead the model's reasoning process. **Ideally, a robust MLLM should be capable of filtering out misleading or irrelevant images and relying more on textual information when necessary.** However, our findings suggest that current MLLMs fail to do so effectively, as their accuracy in the Noisy Figure setting remains lower than in the No Figure setting. This highlights a critical limitation: **MLLMs may not be assessing the relevance of visual input but instead default to treating images as a primary source of information, even when they introduce noise.** Therefore, based on this findings, we constructed a Banana-Counting dataset to systematically analyze modality bias in MLLMs.

2.2 Banana-Counting Dataset Construction

Banana-Counting dataset is based on SPIQA (Scientific Paper Image Question Answering) (Pramanick et al., 2024), a large-scale QA dataset designed to interpret complex figures and tables within the context of scientific research articles across various domains of computer science. We constructed our dataset using the test-A split from SPIQA, selecting images, their corresponding captions, and associated text. We then inserted a needle phrase into both text and images in the format:

The little monkey counted {number} bananas.

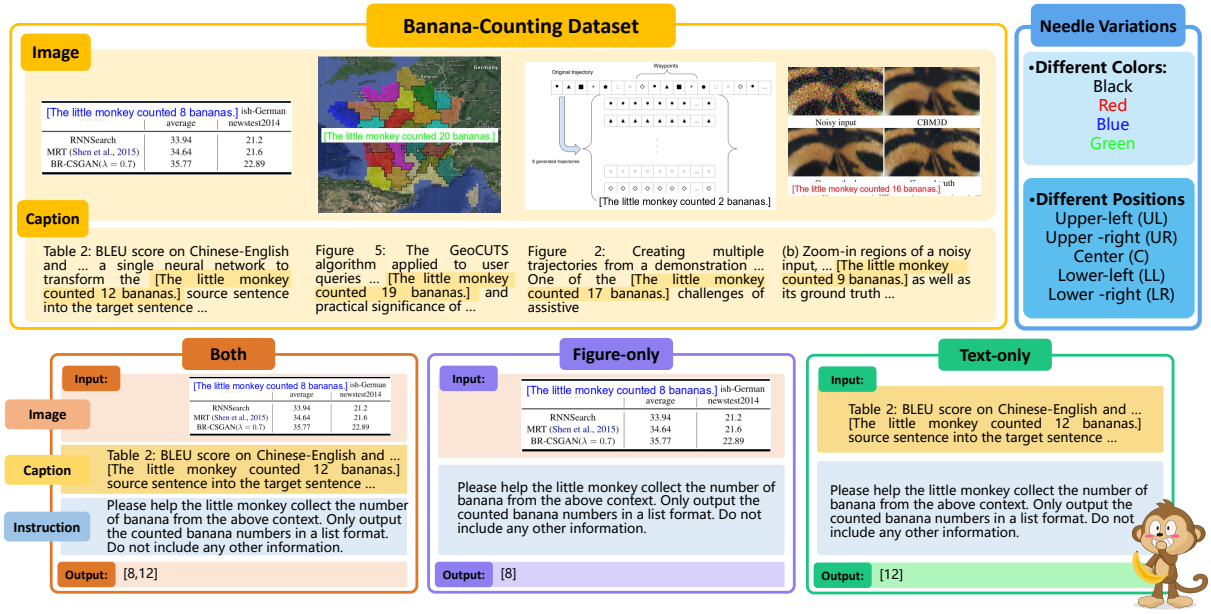

Figure 2: Overview of the Banana-Counting Dataset. We extracted figures and captions from the SPIQA dataset and inserted needle phrases into both modalities. The needle's color and position in figures were randomly assigned while in text, it was inserted at random positions to avoid biases. To evaluate modality bias, we designed three settings: (1) Both: The model receives both figure and caption, (2) Figure-only: The model receives only the figure, and (3) Text-only: The model receives only the caption. This setup examines whether MLLMs effectively integrate multimodal information or favor one modality over the other.

where the number of bananas in each needle was randomly generated within the range of 1 to 20. The Banana-Counting dataset contains a total of 1026 instances for evaluation. The overview of the dataset is illustrated in Figure 2. To eliminate potential biases caused by superficial cues, the **needle's color and position in the figure were randomly assigned**, while in the text, the needle was inserted at random positions. Specifically:

Figure-based Needle Insertion: The needle phrase was placed at random positions within the image, including Upper-left (UL), Upper-right (UR), Center (C), Lower-left (LL), and Lower-right (LR). Additionally, the text color was randomly selected from black, red, blue, and green (as shown in Figure 2).

Text-based Needle Insertion: The needle phrase was embedded into the textual content at random depths. The text primarily comprised the image or table caption. If the caption contained fewer than 100 words, we supplemented it with additional content from the corresponding research paper. If it exceeded 100 words, we truncated it to 100 words.

2.3 Experiments

During the inference phase, we provided input with predefined instructions, as shown in Figure 2, to guide MLLMs in extracting the banana count from the context. We evaluated whether LLMs could **identify and extract the banana count from the given context**. To assess modality bias, we designed three experimental settings: (1) Both: The model receives both the figure and caption as input; (2) Figure-only: The model receives only the figure as input; (3) Text-only: The model receives only the caption as input. In the Both setting, we aim to measure different MLLMs' preferences for specific modalities. In the single-modality settings (Figure-only and Text-only), we further investigated the accuracy of MLLMs when restricted to a single modality. This evaluation serves to demonstrate that MLLMs possess sufficient capability to process each modality independently, ensuring that modality bias is not merely a result of inadequate unimodal processing ability, but rather an inherent preference for one modality over the other.

To evaluate this, we tested ten different MLLMs on the Banana-Counting dataset. The evaluated models include: MiniCPM-V-2_6 (Yao et al., 2024), Qwen2-VL-7B-Instruct, Qwen2-VL-72B (Bai et al., 2023), Cogvlm2-Llama3-chat-19B (Wang et al., 2023), GPT-4o-mini (Achiam et al., 2023), Llava-next-Llama3, Llava-v1.6-vicuna-13B-hf (Liu et al., 2024), MoE-LLaVA-Phi2-2.7B-4e, MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2024), Deepseek-vl2-small (Wu et al., 2024).

Model	Model Size #Parameters (B)	Both		Figure only	Text only
		ACC_{text}	**ACC_{fig}**	**ACC_{fig}**	**ACC_{text}**
MiniCPM-V-2_6 (Yao et al., 2024)	8.10	67.17	**74.93**	92.14	98.48
Qwen2-VL-72B (Bai et al., 2023)	72.00	77.17	**89.96**	99.51	100.00
Qwen2-VL-Instruct (Bai et al., 2023)	7.00	51.66	**72.32**	98.05	99.90
Cogvlm2-Llama3-chat (Wang et al., 2023)	19.00	**67.82**	58.77	96.80	91.29
GPT-4o-mini (Achiam et al., 2023)	-	81.52	**87.15**	99.51	100.00
Llava-next-Llama3 (Liu et al., 2024)	8.00	**47.95**	33.72	42.50	93.86
Llava-v1.6-vicuna (Liu et al., 2024)	13.00	47.37	**54.87**	90.35	97.08
MoE-LLaVA-Phi2-2.7B-4e (Lin et al., 2024)	5.61	22.71	**96.20**	52.05	88.11
MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2024)	5.73	38.15	**82.07**	92.88	78.65
Deepseek-vl2-small (Wu et al., 2024)	16.10	8.00	**78.17**	90.55	98.83
Yi-VL-6B (Young et al., 2024)	6.00	**77.78**	27.17	48.54	90.55

Table 1: Performance of different LLMs on the Banana-Counting dataset. The table reports the accuracy of various models in identifying the number of bananas under three settings: (1) **Both** (2) **Image only**, and (3) **Text only**. **#Parameters** denotes the model size in billions. **ACC_{text}** and **ACC_{fig}** represent the accuracy of extracting the banana count from text and images, respectively. The results highlight significant **modality bias**, where most models favor image-based information (**ACC_{fig}**) while often overlooking text-based cues (**ACC_{text}**).

2.4 Results

The experimental results are presented in Table 1. The table summarizes the performance of different MLLMs in extracting banana counts under the three evaluation settings: Both (where both the figure and caption are provided), Image-only (where only the figure is available), and Text-only (where only the caption is available). The accuracy of retrieving banana counts from text (ACC_{text}) and from images (ACC_{fig}) is reported for each model.

The results reveal clear modality bias across most MLLMs: (1) When both modalities are available, most models prioritize visual information, extracting counts mainly from the figure and neglecting the text, leading to significantly higher ACC_{fig} than ACC_{text}. (2) In the Image-only and Text-only settings, all models demonstrate higher accuracy in extracting banana counts compared to the Both setting. This indicates that when information must be combined across modalities, models struggle to effectively integrate textual and visual cues. (3) Among the models exhibiting an inverse modality bias, Cogvlm, Yi and Llava-next-Llama3 show distinct trends. Cogvlm displays a stronger preference for textual information, leading to lower ACC_{fig} compared to other models. Meanwhile, Yi and Llava-next-Llama3 demonstrates extremely low accuracy in the Image-only setting, suggesting that it has poor image text recognition capabilities, causing it to ignore visual information and rely more heavily on text.

Additionally, to eliminate the influence of the needle phrase's position and color in the figure, we analyzed accuracy across different colors and positions, as shown in Table 2. The results indicate that model performance remains relatively stable regardless of needle position or color, suggesting that **neither factor significantly influences figure accuracy**. This implies that modality bias is primarily driven by an inherent preference for visual or textual information rather than superficial attributes such as text color or placement.

These findings highlight a fundamental challenge in MLLM design—many models overly rely on visual cues when both modalities are available, which can lead to suboptimal decision-making when textual information is equally or more informative. The next section explores potential approaches to mitigating this modality bias and improving multimodal reasoning in MLLMs.

Further Exploration To eliminate the influence of instruction phrasing, we conducted an additional experiment using Explicit Instruction. Specifically, we modified the instruction in Figure 2, explicitly directing the model to retrieve banana counts from both the image and text. The detailed experimental setup and results are provided in Appendix A. Our results indicate that even when explicitly instructed to integrate information from both modalities, MLLMs still predominantly rely on image-based cues. This further reinforces the presence of modality bias and highlights the necessity for improved training strategies to encourage balanced multimodal reasoning.

3 Mitigating Modality Bias in MLLMs

Given the findings from the previous section, we propose a simple yet effective approach to mitigate

Model	Figure Needle Color				Figure Needle Position				
	Blue	Green	Red	Black	M	UR	LR	UL	LL
MiniCPM-V-2_6 (Yao et al., 2024)	75.94	71.47	75.06	**77.63**	75.25	77.72	69.51	**79.12**	72.14
Qwen2-VL-72B (Bai et al., 2023)	89.52	**90.59**	89.96	89.75	89.69	89.41	90.10	90.18	**90.52**
Qwen2-VL-Instruct (Bai et al., 2023)	71.96	72.43	70.85	**74.15**	**78.92**	68.81	76.24	60.62	78.61
Cogvlm2-Llama3-chat (Wang et al., 2023)	59.63	58.82	56.36	**60.25**	57.94	55.35	57.23	**62.65**	60.58
GPT-4o-mini (Achiam et al., 2023)	87.45	**88.60**	87.45	84.83	**91.57**	84.06	84.26	85.31	90.87
Llava-next-Llama3 (Liu et al., 2024)	31.73	29.41	35.22	**39.41**	33.63	27.23	39.11	28.32	**42.20**
Llava-v1.6-vicuna (Liu et al., 2024)	**56.83**	52.57	53.85	56.36	53.36	46.04	67.33	43.36	**67.63**
MoE-LLaVA-Phi2-2.7B-4e (Lin et al., 2024)	33.95	32.72	**35.22**	31.78	34.98	29.70	36.14	29.20	**38.15**
MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2024)	22.51	**24.26**	21.46	22.46	**36.77**	15.84	13.86	23.01	22.54
Deepseek-vl2-small (Wu et al., 2024)	**80.07**	79.78	78.95	73.31	76.23	78.22	**83.17**	72.57	82.08
Yi-VL-6B (Young et al., 2024)	39.48	42.65	40.89	**43.22**	**51.12**	37.62	37.62	42.92	36.42

Table 2: Accuracy of different LLMs in identifying the figure banana needle across various colors and positions. The **Figure Needle Color** columns represent different text colors in the figure: Blue, Green, Red, and Black. The **Figure Needle Position** columns correspond to different placements within the image: **M** (Middle), **UR** (Upper Right), **LR** (Lower Right), **UL** (Upper Left), and **LL** (Lower Left). The results suggest that neither color nor position significantly affects the ability of LLMs to locate the figure banana needle.

Needle Type	Multimodal Text-only	Multimodal Both	Unimodal Text-only
# Instances	2101	1680	2101

Table 3: Statistics of the Banana-Counting training dataset. # Instances represents the number of samples in each category.

modality bias in MLLMs. Our method focuses on targeted training using a structured Banana-Counting training dataset, designed to balance the model's reliance on both text and images.

3.1 Dataset Construction and Training Strategy

To address modality bias, we constructed a specialized Banana-Counting training dataset. The overall dataset construction follows a similar process as described in Section 2.2. However, unlike the test set, the training dataset is designed to provide a more balanced exposure to different modality configurations. Specifically, the training set consists of three distinct formats of Banana-Counting tasks: (1) Multimodal-Text-Only: Both text and images are provided, but the needle phrase appears only in the text. (2) Multimodal - Both: Both text and images are provided, and the needle phrase appears both in the image and the text. (3) Unimodal - Text-Only: Only text is provided, with the needle phrase embedded solely within the text. The detailed statistics of the Banana-Counting training dataset are shown in Table 3.

To fine-tune the models, we leveraged the Llama-Factory framework (Zheng et al., 2024) and applied LoRA (Hu et al., 2021) for parameter-efficient adaptation. The detailed hyperparameter settings for the training process are provided in Table 8 in Appendix B. We selected five different models to investigate the impact of LoRA fine-tuning on the Banana-Counting training dataset, including: Qwen2-VL-7B-Instruct (Bai et al., 2023), Llama3-Llava-next-8b (Liu et al., 2024), Llava-v1.6-vicuna-13b (Liu et al., 2024), MiniCPM-V-2_6 (Yao et al., 2024) and Yi-VL-6B (Young et al., 2024).

To comprehensively assess the effectiveness of LoRA fine-tuning in reducing modality bias, we evaluated the pre-trained and fine-tuned models on four datasets: Banana-Counting, ScienceQA (Lu et al., 2022), CSQA (Commonsense QA) (Talmor et al., 2019), and MMLU (Hendrycks et al., 2021) where: (1) Banana-Counting dataset is used to directly measure the model's ability to extract numerical information from both text and images. It serves as the primary benchmark for assessing modality bias reduction. (2) ScienceQA is a multimodal multiple-choice science question dataset collected from elementary and high school curricula which requires both textual reasoning and visual interpretation to answer correctly. (3) CSQA is designed for commonsense reasoning, containing multiple-choice questions with five possible answers. Unlike ScienceQA, CSQA emphasizes world knowledge and logical inference, making it an ideal benchmark for evaluating a model's ability to apply general reasoning skills. (4) MMLU is a large-scale benchmark consisting of multiple-choice questions from a wide range of disciplines, including humanities, social sciences, hard sciences, and mathematics. Covering 57 different

Model	Banana				ScienceQA	CSQA	MMLU
	Both		Text-only	Figure-only			
	$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$	$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$			
Qwen2-VL-7B-Instruct	72.32	51.66	99.90	98.05	78.50	78.71	69.20
Qwen2-VL-7B-Instruct_lora	**100.00**	**100.00**	**100.00**	**100.00**	**79.50**	**78.79**	**69.38**
Llama3-Llava-next-8b	33.72	47.95	93.86	42.50	**68.40**	66.83	61.32
Llama3-Llava-next-8b_lora	**99.31**	**100.00**	**100.00**	**99.51**	67.80	**68.63**	**61.64**
Llava-v1.6-vicuna-13b	54.87	47.37	97.08	90.35	**56.90**	48.89	56.79
Llava-v1.6-vicuna-13b_lora	**99.42**	**98.93**	**100.00**	**99.61**	**56.90**	**51.27**	**56.94**
MiniCPM-V-2_6	75.24	67.64	98.83	92.20	84.80	**79.77**	63.07
MiniCPM-V-2_6_lora	**99.81**	**100.00**	**100.00**	**99.81**	**85.20**	**79.77**	**63.20**
Yi-VL-6B	20.66	77.78	83.04	48.54	**63.10**	75.59	**61.22**
Yi-VL-6B_lora	**55.26**	**99.61**	**100.00**	**53.22**	62.20	**76.16**	61.17

Table 4: Performance comparison of different models before and after LoRA fine-tuning (highlighted in yellow) across various datasets, with accuracy (%) averaged over ten runs. The table includes results on the Banana-Counting dataset, with accuracy measured for text-based and image-based banana counting in both multimodal (both text and image), figure-only and text-only settings. Additionally, performance on ScienceQA, CSQA, and MMLU benchmarks is provided to assess whether LoRA fine-tuning improves Banana-Counting accuracy without negatively impacting broader multimodal and reasoning tasks.

subjects, MMLU is designed to evaluate models on broad world knowledge and complex problem-solving capabilities. Details about the dataset statistics and examples can be found in Appendix C.

Since the LoRA fine-tuning process was conducted solely on the Banana-Counting training dataset, testing on unseen datasets (ScienceQA, CSQA, and MMLU) allows us to determine whether **reducing modality bias in one task affects the model's performance on broader multimodal and reasoning tasks.**

3.2 Results

Table 4 presents the evaluation results of different models and their LoRA fine-tuned versions (highlighted in yellow) across multiple datasets. As shown in the table, fine-tuned models exhibit perfect or near-perfect accuracy across all Banana-Counting scenario. This confirms that targeted training effectively enables models to extract numerical information from both text and images without favoring one modality. Furthermore, **LoRA fine-tuning significantly improves Banana-Counting accuracy while maintaining or even enhancing general benchmark performance**. The fine-tuned models achieve comparable or better results on ScienceQA, CSQA, and MMLU, demonstrating that reducing modality bias does not degrade broader multimodal reasoning capabilities. Notably, LoRA fine-tuning yields substantial performance gains on the CSQA dataset for Llava models, where Llama3-Llava-next-8b improves by nearly 2 points after fine-tuning and Llava-v1.6-vicuna-13b sees an improvement of approximately 3 points. This aligns with our hypothesis that forcing the model to process information from both text and images during training strengthens its contextual reasoning abilities. Additionally, we observe that while most models experience only minor variations in MMLU performance, fine-tuning does not degrade their ability to generalize across diverse knowledge domains.

These findings reinforce the importance of balanced multimodal training, demonstrating that **reducing modality bias can enhance general performance across a range of multimodal and knowledge-intensive tasks**.

3.3 Robustness to Noisy Figures

In Figure 1, we observed that modality bias can cause models to overly rely on a single modality, particularly the image modality for most MLLMs. This over-reliance makes them highly susceptible to performance degradation when presented with irrelevant images, a critical issue in real-world applications. To address this concern, this section investigates whether fine-tuning on the Banana-Counting training dataset using LoRA can improve model robustness against noisy figures. To evaluate this, we constructed noisy versions of the ScienceQA, CSQA, and MMLU datasets, where we randomly inserted unrelated images from ScienceQA into different instances. The experimental results are summarized in Table 5. From Table 5, we observe that across all datasets, models fine-tuned with LoRA show higher accuracy in noisy

Model	Noisy		
	ScienceQA	CSQA	MMLU
Qwen2-VL-7B-Instruct	66.38	77.58	67.89
Qwen2-VL-7B-Instruct_lora	**66.68**	**78.54**	**67.97**
Llama3-Llava-next-8b	62.44	66.39	**57.82**
Llama3-Llava-next-8b_lora	**64.12**	**68.12**	57.72
Lava-v1.6-vicuna-13b	53.96	60.38	42.22
Llava-v1.6-vicuna-13b_lora	**55.02**	**63.04**	**43.00**
MiniCPM-V-2_6	69.38	77.39	62.14
MiniCPM-V-2_6_lora	**69.86**	**78.18**	**62.32**
Yi-VL-6B	**59.06**	72.17	59.74
Yi-VL-6B_lora	58.70	**72.92**	**59.83**

Table 5: Performance of MLLMs before and after LoRA fine-tuning (highlighted in yellow) on noisy datasets. The table shows accuracy (%) on ScienceQA, CSQA, and MMLU after injecting irrelevant images to assess model robustness against noisy figures. Each result is averaged over ten runs. The findings demonstrate that fine-tuned models are more resilient to noisy figure interference, maintaining higher accuracy than their pre-trained counterparts.

settings compared to their original versions, indicating a **stronger ability to filter out irrelevant visual information**. Llama3-Llava-next-8b and Llava-v1.6-vicuna-13b benefit significantly from fine-tuning. Llava-v1.6-vicuna-13b's accuracy improves by 1.06 percentage points in ScienceQA and 2.66 in CSQA, highlighting that fine-tuning strengthens multimodal alignment and robustness.

These findings confirm that **fine-tuning with a balanced multimodal dataset enhances a model's ability to focus on relevant context while ignoring misleading visual inputs**, making it more reliable for real-world applications where noisy or irrelevant figures are prevalent.

3.4 MMLU Performance Analysis

In this section, we examine how LoRA fine-tuning influences model performance across different domains in the MMLU dataset. Following the approach of Llama-Factory (Zheng et al., 2024), we categorize MMLU instances into four broad groups: STEM, Social Sciences, Humanities, and Other. Table 6 presents the results for each model before and after LoRA fine-tuning. From the table, we observe that fine-tuning on the Banana-Counting dataset has varying effects across different categories. The most consistent improvements appear in the STEM and Social Sciences categories, where models such as Qwen2-VL-7B-Instruct and Llama3-Llava-next-8b show noticeable gains. For example, Qwen2-VL-7B-Instruct-lora improves by 0.69 points in STEM, while Llama3-Llava-next-8b-lora gains 0.20 points in Social Science. The Humanities and Other category shows diverse trends. For example, in Other category, while Llava-v1.6-vicuna-13b-lora sees a notable improvement of +0.65 points, other models like MiniCPM-V-2-6 and Yi-VL-6B experience minimal changes ($+0.15$ and $+0.13$, respectively). This suggests that the impact of fine-tuning on general knowledge-based tasks is model-dependent.

One key takeaway is that fine-tuning with a multimodal dataset does not universally enhance performance across all MMLU categories. **The improvements are more pronounced in domains requiring structured reasoning and numerical understanding** (STEM and Social Sciences). These findings highlight the need for a more balanced fine-tuning approach when adapting multimodal LLMs to diverse real-world tasks.

4 Related Work

Our work mainly relates to two areas of research: multimodal needle retrieval and modality bias in multimodal learning. Wang et al. (2024) introduce MM-NIAH, a benchmark for evaluating MLLMs' ability to retrieve information from long multimodal documents, testing whether models can locate and process dispersed information across tens of thousands of tokens. In contrast, our study focuses on modality bias, ensuring that models have equal access to identical needle information in both text and image modalities within short-context instances. Unlike MM-NIAH, which does not explicitly analyze how models balance text vs. image information when both provide the same answer, our work is the first to systematically evaluate whether MLLMs truly integrate multimodal inputs or inherently favor one modality over the other.

Beyond needle retrieval, prior research has extensively examined dataset-induced modality bias in multimodal learning. Park et al. (2024) investigate modality bias in Video Question Answering (VidQA) datasets, introducing the Modality Importance Score (MIS) to quantify the contribution of each modality. Their findings reveal that many VidQA benchmarks are inherently skewed toward a single modality, often allowing models to answer questions unimodally. While their work primarily identifies dataset-level biases, our study shifts the focus to model-induced modality bias, analyzing whether MLLMs exhibit inherent modality preferences even when provided with fully balanced

Model	STEM	Social Sciences	Humanities	Other
Qwen2-VL-7B-Instruct	63.82	79.07	62.95	73.90
Qwen2-VL-7B-Instruct_lora	64.51 (+0.69)	79.10 (+0.03)	63.02 (+0.07)	73.90 (+0.00)
Llama3-Llava-next-8b	52.42	70.91	56.45	67.58
Llama3-Llava-next-8b_lora	52.49 (+0.07)	71.11 (+0.20)	57.19 (+0.74)	67.61 (+0.03)
Llava-v1.6-vicuna-13b	46.26	65.91	52.99	63.45
Llava-v1.6-vicuna-13b_lora	46.36 (+0.10)	66.17 (+0.26)	52.77 (-0.22)	64.10 (+0.65)
MiniCPM-V-2_6	54.17	73.64	58.64	67.74
MiniCPM-V-2_6_lora	54.14 (-0.03)	73.55 (-0.09)	59.00 (+0.36)	67.89 (+0.15)
Yi-VL-6B	51.49	72.90	54.88	68.38
Yi-VL-6B_lora	51.89 (+0.40)	72.86 (-0.04)	54.43 (-0.45)	68.51 (+0.13)

Table 6: Performance of different MLLMs before and after LoRA fine-tuning (highlighted in yellow) on the MMLU dataset across four major categories: **STEM, Social Sciences, Humanities, and Other**. Accuracy values are reported as percentages, with improvements (or declines) from LoRA fine-tuning shown in parentheses. The results illustrate how fine-tuning on the Banana-Counting dataset impacts broader knowledge-based reasoning tasks.

multimodal inputs. Relatively few studies have investigated modality bias at the model level. Gat et al. (2021) introduced the Perceptual Score, measuring prediction stability under modality perturbations, revealing that early small-scale VQA models over-relied on textual cues. However, rather than perturbing modalities, our approach injects identical information into both text and image modalities, explicitly testing whether modern large-scale multimodal LLMs (MLLMs) can integrate both modalities or favor one. Our findings indicate a reversal in modality bias trends—whereas early VQA models primarily relied on text, modern MLLMs tend to prioritize visual information over text. Further research has examined modality robustness and preference in multimodal learning. Yang et al. (2024) propose a two-step training framework that regulates uni-modal representation margins and adjusts modality integration factors to enhance robustness against adversarial perturbations. Similarly, concurrent work (Park et al., 2025) explores modality imbalance in vision-language models, where models perform significantly better on text-based tasks than their visual counterparts. While modality imbalance refers to performance gaps across modalities (text vs. image modalities) in complex reasoning tasks, our study focuses on modality bias, where models systematically favor one modality even when both provide equally informative content. Additionally, while their work evaluates tasks presented in either textual (e.g., LaTeX) or image format, we explicitly inject identical information into both modalities to assess whether MLLMs genuinely integrate multimodal inputs or exhibit a preference.

5 Conclusion

In this work, we conducted a systematic investigation into modality bias in MLLMs, revealing that many models inherently favor visual cues over textual ones. To quantify this bias, we designed a Banana-Counting dataset, where numerical information is embedded identically in both text and images, ensuring that models have an equal opportunity to utilize both sources. Our experiments demonstrate that most MLLMs prioritize visual information, leading to biased decision-making and reduced robustness when textual information is more reliable. To address this issue, we introduced a balanced multimodal Banana-Counting training dataset and fine-tuned MLLMs using LoRA. Our results show that fine-tuning significantly reduces modality bias while maintaining or even improving performance on general reasoning benchmarks such as ScienceQA, CSQA, and MMLU. Furthermore, fine-tuning enhanced model robustness against noisy images, ensuring that MLLMs do not blindly rely on visual inputs but instead make decisions based on the most relevant modality.

Our findings suggest that MLLMs often rely too heavily on vision, even when textual cues provide equally valid answers. This highlights the need for balanced multimodal training strategies that teach MLLMs that *"maybe seeing is not believing"* and true multimodal intelligence requires context-aware, adaptive reasoning across both text and images. Future work could explore extending this methodology to other multimodal tasks, investigating whether similar biases exist in real-world multimodal applications, and developing more sophisticated training strategies to further enhance multimodal reasoning capabilities.

Limitations

While Banana-Counting performance reached near 100%, variations in broader benchmark scores suggest that fine-tuning strategies could be further optimized to generalize across diverse multimodal tasks. Future work should explore more sophisticated training paradigms that explicitly encourage dynamic modality selection based on task relevance. We hope that our study will inspire further research into developing more robust and unbiased MLLMs, capable of true multimodal reasoning in real-world applications.

References

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

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

Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. 2023. Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593*.

Itai Gat, Idan Schwartz, and Alexander G. Schwing. 2021. Perceptual score: What data modalities does your model perceive? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21630–21643.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jana Holsanova, Nils Holmberg, and Kenneth Holmqvist. 2009. Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(9):1215–1226.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *Preprint*, arXiv:2401.15947.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2024. CoCo-agent: A comprehensive cognitive MLLM agent for smartphone GUI automation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9097–9110, Bangkok, Thailand. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. 2024. Assessing modality bias in video question answering benchmarks with multimodal large language models. *CoRR*, abs/2408.12763.

Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2025. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms? *arXiv preprint arXiv:2501.02669*.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. *CoRR*, abs/2407.09413.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 742–758. Springer.

Anders D Sivle and Per H Uppstad. 2018. Reasons for relating representations when reading digital multimodal science information. *Visual Communication*, 17(3):313–336.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024. Needle in A multimodal haystack. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. 2024. Quantifying and enhancing multi-modal robustness with modality preference. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond chain-of-thought, effective graph-of-thought reasoning in language models. *arXiv preprint arXiv:2305.16582*.

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

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Model	Model Size (# Parameters(B))	Both	
		$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$
MiniCPM-V-2_6	8.10	63.26	**82.07**
Qwen2-VL-Instruct	7.00	48.34	**83.24**
Cogvlm2-Llama3-chat	19.00	51.27	**59.55**
Llava-next-Llama3	8.00	37.91	**38.89**
Llava-v1.6-vicuna	13.00	35.87	**53.12**
Deepseek-vl2-small	16.10	7.60	**83.24**

Table 7: Effect of explicit instruction on Banana-Counting dataset. $\mathbf{ACC_{text}}$ and $\mathbf{ACC_{fig}}$ represent the accuracy of extracting the banana count from text and images, respectively. Despite explicitly instructing models to extract information from both image and text, strong modality bias persists.

Hyper-parameter	Value
finetuning_type	lora
lora_target	all
per_device_train_batch_size	1
gradient_accumulation_steps	8
learning_rate	1.00E-04
num_train_epochs	5
lr_scheduler_type	cosine
warmup_ratio	0.1
val_size	0.1

Table 8: Detailed hyper-parameter settings for LoRA fine-tuning.

	ScienceQA	CSQA	MMLU
# Instances	1000	1221	14042

Table 9: Detailed statistics of the datasets used in our experiments. #Instances refers to the number of samples in each dataset.

A Results for Explicit Instruction

To further investigate the impact of instruction phrasing on modality bias, we conducted an Explicit Instruction experiment. In our original setting (implicit instruction), the model was guided with the following prompt:

> Please help the little monkey collect the number of bananas from the above context. Only output the counted banana numbers in a list format. Do not include any other information.

For the explicit instruction setting, we modified the prompt to explicitly instruct the model to extract information from both image and text:

> Please help the little monkey collect the number of bananas from the above image and text. Only output the counted banana numbers in a list format. Do not include any other information.

Apart from this change in instruction, all other experimental settings remained identical to those described in Section 2. The results are presented in Table 7.

From Table 7, it is evident that even with explicit instructions directing the model to extract information from both modalities, a significant modality bias remains prevalent. Across all tested models, the accuracy for extracting banana counts from images remains consistently higher than that from text, indicating a persistent tendency to prioritize visual information over textual input.

These findings suggest that modality bias is not merely an artifact of instruction phrasing but rather an inherent characteristic of current MLLMs. Even when explicitly prompted to integrate information from both modalities, the models still predominantly rely on image-based cues, further reinforcing the need for improved training strategies to mitigate this bias.

B Hyper-parameter Settings

The detailed hyper-parameter settings for LoRA fine-tuning can be found in Table 8

C Dataset Statistics and Examples

To evaluate the effectiveness of our approach, we conduct experiments on ScienceQA, CSQA, and MMLU, three widely-used multimodal and commonsense reasoning benchmarks. ScienceQA (Lu et al., 2022) consists of multiple-choice science questions, often accompanied by images, requiring models to integrate textual and visual information for reasoning. CSQA (Talmor et al., 2019) is a commonsense question-answering dataset that evaluates a model's ability to infer everyday knowledge from text. MMLU (Hendrycks et al., 2021) is a large-scale benchmark containing questions from diverse disciplines, including STEM, social sciences, humanities, and other general knowledge areas. These datasets collectively provide a comprehensive evaluation framework, testing models on multimodal reasoning, commonsense inference, and factual knowledge retrieval.

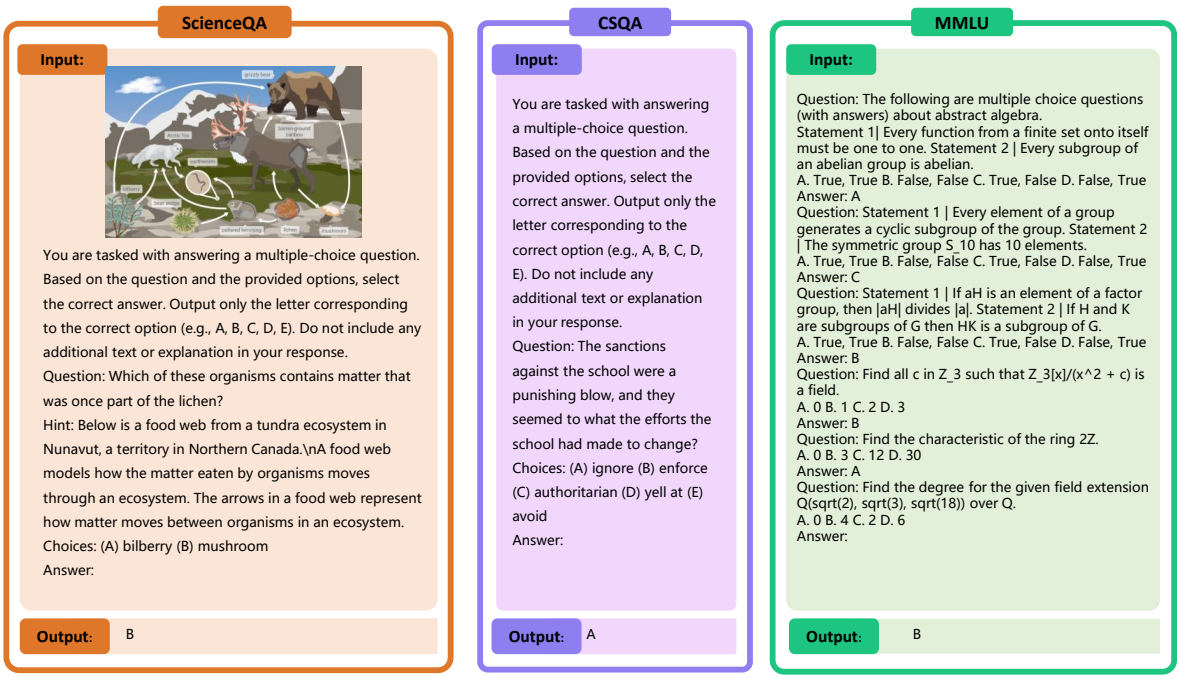

Figure 3: Example instances from the ScienceQA, CSQA, and MMLU datasets. ScienceQA contains multimodal science-related multiple-choice questions, CSQA evaluates commonsense reasoning, and MMLU covers a broad range of subjects requiring advanced reasoning skills.

The detailed dataset statistics are presented in Table 9, while Figure 3 provides representative examples from each dataset.