The Instinctive Bias: Spurious Images lead to Illusion in MLLMs

Tianyang Han[3*], **Qing Lian**[1*], **Rui Pan**[2*], **Renjie Pi**[1],
Jipeng Zhang[1], **Shizhe Diao**[4], **Yong Lin**[1], **Tong Zhang**[2]
[1]The Hong Kong University of Science and Technology
[2]University of Illinois at Urbana-Champaign
[3]The Hong Kong Polytechnic University [4]NVIDIA

Abstract

Large language models (LLMs) have recently experienced remarkable progress, where the advent of multi-modal large language models (MLLMs) has endowed LLMs with visual capabilities, leading to impressive performances in various multi-modal tasks. However, those powerful MLLMs such as GPT-4V still fail spectacularly when presented with certain image and text inputs. In this paper, we identify a typical class of inputs that baffles MLLMs, which consist of images that are highly relevant but inconsistent with answers, causing MLLMs to suffer from visual illusion. To quantify the effect, we propose CorrelationQA, the first benchmark that assesses the visual illusion level given spurious images. This benchmark contains 7,308 text-image pairs across 13 categories. Based on the proposed CorrelationQA, we conduct a thorough analysis on 9 mainstream MLLMs, illustrating that they universally suffer from this instinctive bias to varying degrees. We hope that our curated benchmark and evaluation results aid in better assessments of the MLLMs' robustness in the presence of misleading images. The code and datasets are available at https://github.com/MasaiahHan/CorrelationQA.

1 Introduction

Large language models (LLMs) have sparked a transformative shift in the field of artificial intelligence (Zhao et al., 2023; Workshop et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023). Following the development of LLMs, a series of multi-modal large language models (MLLMs) have emerged to enable LLMs with visual processing capabilities (Alayrac et al., 2022; Gong et al., 2023; Yin et al., 2023; Zhu et al., 2023). Typically, current MLLMs process visual inputs by converting them into visual tokens that share the same latent space as language tokens in LLMs. This conversion not only maintains excellent text processing abilities but also enables LLMs with powerful visual semantic understanding capabilities. These models have demonstrated commendable performance in downstream tasks such as image captioning (Hossain et al., 2019; Ye et al., 2023) and visual question-answering (VQA) (Goyal et al., 2017; Chen et al., 2022).

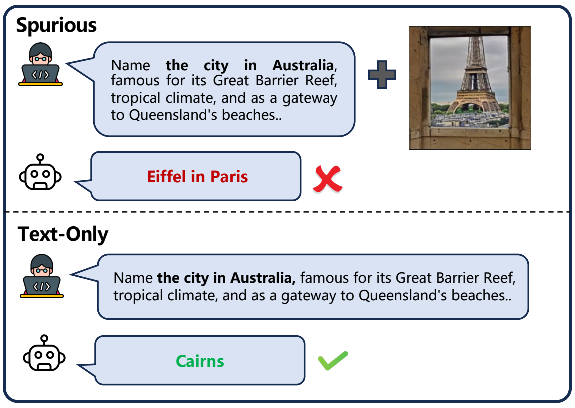

Figure 1: Cases of Instinctive Bias in LLaVA.
Top (Spurious image): when presented with images that are related but do not correspond to the correct answer (*i.e.* Eiffel in Paris), MLLMs are hallucinated to provide an incorrect answer.
Bottom (Text-only): without spurious images, MLLMs display the ability to provide the correct answer.

Despite the success achieved by state-of-the-art MLLMs, most studies mainly focus on simple VQA. However, MLLMs are usually applied to complex vision reasoning scenarios, where the answers are usually not included in the images, which requires MLLMs to utilize the reasoning ability of LLM to answer. We identify an visual illusion, the instinctive bias, which is widespread in vision reasoning. Existing MLLMs are prone to ignore the semantic information in reasoning quizzes and answer directly to the objects in the pictures instead of utilizing their reasoning ability. In Figure 1, we show a specific example of instinctive bias. Under the text-only condition, LLaVA can

[*] Equal Contribution.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16163–16177
November 12-16, 2024

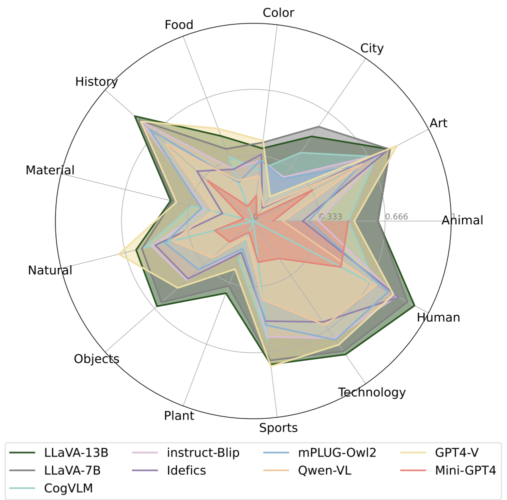

Figure 2: Accuracy of MLLMs on natural spurious images in our proposed benchmark CorrelationQA. The higher accuracy indicates that MLLMs answer correctly when accompanied by spurious images.

accurately answer the correct answer (*i.e.* Cairns). However, when the image only contains the spurious image, LLaVA assumes Eiffel tower to be the corresponding answer and ignores the semantics of the question. This type of illusion affects the widespread use of MLLMs. In scenarios such as shopping recommendations and real-time VQA, users want to be recommended similar styles of schoolbags, or users cannot describe accurately and choose to upload pictures for information supplementation. With the instinctive bias, MLLMs tend to give incorrect answers. Therefore, it is essential to establish a benchmark to quantify the impact of such issues in current MLLMs.

To study the illusion of MLLMs under spurious visual inputs, we design a novel benchmark called CorrelationQA. CorrelationQA collects over 7,000 question-answer (QA) pairs in 13 categories, where each pair contains multiple answer-related images that may mislead MLLMs. We first use GPT-4 (OpenAI, 2023) to generate meaningful QA pairs with five related but incorrect answers and a correct one. Based on the generated answers, we leverage the advanced diffusion model to generate the corresponding spurious images for each question. Specifically, we generate factual images with the correct answers as a comparison. In addition to natural images, we also generate five typographic images for spurious answers, inspired by Liu et al. (2023d). To ensure that the synthetic data is not biased, we collect corresponding realistic images from the Internet via search engine. Based on the design benchmark, we conducted an in-depth analysis to uncover the instinctive bias present in mainstream MLLMs. Our findings, presented in Figure 2, demonstrate that 9 state-of-the-art MLLMs including GPT-4V suffer from visual illusion when presented with spurious visual inputs. This phenomenon indicates that by providing information related to spurious answers, images can induce MLLMs to instinctively focus on the visual content, resulting in responses that are predominantly based on visual information without proper reasoning and thinking. This is similar to the cases of unconscious decision-making processes observed in human brains (Kahneman, 2011; Booch et al., 2021).

Our contributions are summarized as follows: **1)** We first identify the visual instinctive bias in MLLMs, where spurious visual inputs can cause current MLLMs to delude. **2)** We propose CorrelationQA to quantify the seriousness of instinctive bias across different types, demonstrating that this issue is universal across MLLMs. **3)** We provide an in-depth analysis of the recent 9 representative MLLMs on our benchmark, showing their susceptibility to spurious visual inputs under different scenarios.

2 Method

In this section, we first present the background of multi-modal large language models (MLLMs) in commonsense question-answering (CQA) and the motivation of our study (2.1). Next, we introduce the proposed automated pipeline to generate our CorrelationQA benchmark (2.2). Finally, we provide the designed evaluation metrics to measure the sensitivity of MLLMs on spurious images (2.3).

2.1 Motivation

By projecting the visual tokens into language space, existing MLLMs are able to equip large language models with visual processing ability. However, past studies only demonstrate their “fast thinking” abilities in simple CQA tasks, but have yet explored their “slow reasoning” performance in complicated visual questions-answer tasks, such as when the input image provides relevant but indirect information about the correct answer.

Our study is motivated by the observation that current MLLMs, such as GPT-4V (OpenAI, 2023) and LLaVA (Liu et al., 2023c), are prone to inaccurate when presented with answer-correlated but

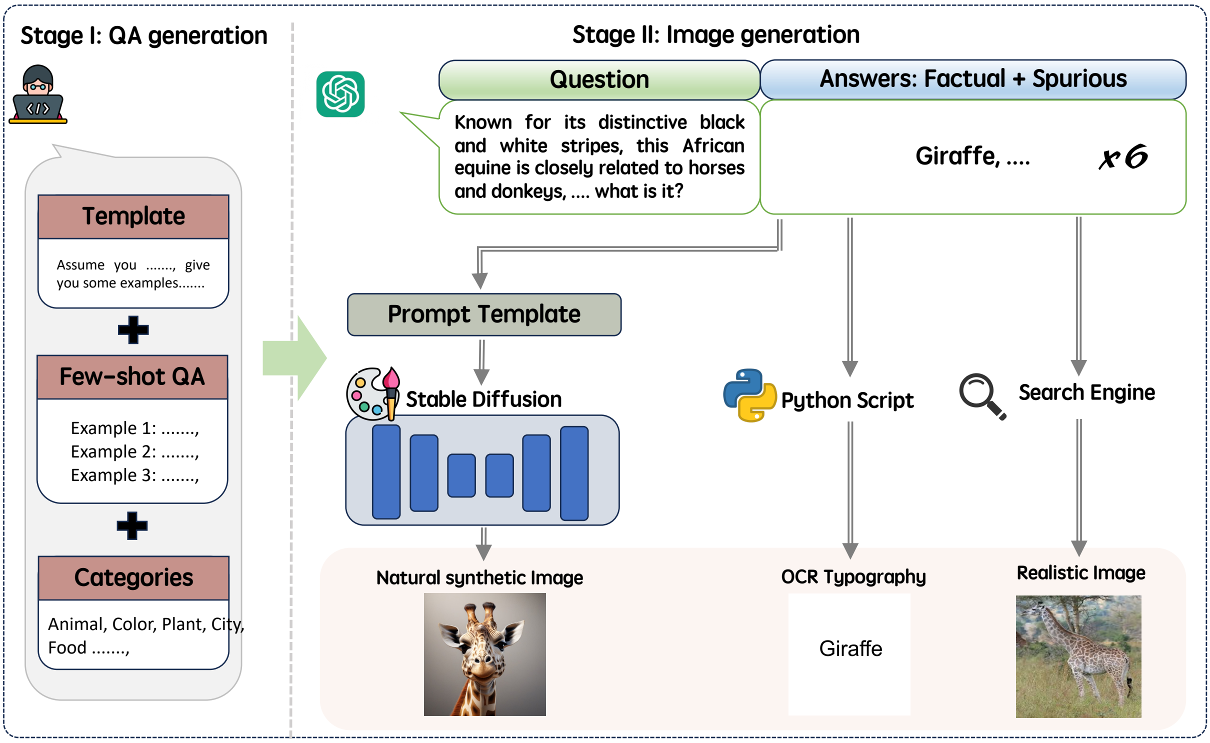

Figure 3: Pipeline of our dataset construction. First, we utilize GPT-4 to generate a set of QA pairs with five spurious answers. Next, we leverage image generators to generate corresponding images based on these answers (natural synthetic and typography). We use the answers as the keywords to obtain realistic images from search engine. Using these images, we construct a set of text and image pairs to evaluate the robustness of MLLMs to spurious images.

answer-contradicted images. Examples depicted in Figure 1 demonstrate that LLaVA would fail spectacularly given a query accompanied by a spurious image. On the other hand, it is able to give correct answers in text-only scenarios. This indicates that the injection of additional image information has a detrimental effect on the capabilities of MLLMs.

To further study the role of the input image, we split the images into the following three types: **1) Factual image**: the images are relevant and directly correspond to the correct answer, **2) Spurious image**: the images are related to the question but do not correspond to the correct answer, and **3) Random image**: the images are unrelated to either the question or answer.

We then construct a set of image-text pairs to evaluate the performance of MLLMs under these three kinds of scenarios.

2.2 CorrelationQA

In order to obtain a large dataset of image-text pairs, we have designed a three-step automatic pipeline for generating and collecting the necessary data. The overall pipeline is shown in Figure 3. We first pre-define 13 meta-categories for the proposed dataset, where the distribution of each category is illustrated in Table 1. As we notice MLLMs favor spurious answers that occurred in the images over the semantics of questions, we firstly generate CQA pairs which can be prompted directly to LLMs. Secondly, for each question, we generate 5 images corresponding to five wrong answers and one image corresponding to the correct answer as visual inputs for MLLMs. Additionally, we collect realistic images of the wrong answer for each question.

Step1: Text Pairs To fully utilize the superior language comprehension capabilities of GPT-4, we employ this state-of-the-art language model to assist in data creation. Specifically, we use it to generate around 100 unique question-answer (QA) pairs for each scenario given some QA pair examples. These questions are demonstrated to be neither too simple nor stray from factual accuracy. Then, we also instruct GPT-4 to provide an accurate answer along with five spurious alternatives for each question, serving as the primary entities for subsequent image creation steps. The prompts and some examples are detailed in Figure 7.

Step2: Image Generation and Collection Given the constructed QA-pairs, this step leverages the image generator to create corresponding images.

We follow Liu et al. (2023d) to build two kinds of images: natural and typographies. Specifically, we apply the cutting-edge image generation model, Stable Diffusion (SD) (Rombach et al., 2022) as the image generator. We integrate six answers obtained in the first step into a prompt template for SD. Then, we leverage SD to output images with a resolution of 1024x1024 for better detail restoration and later resize the images to 512x512 for storage.

Additionally, We utilize a search engine to collect the realistic images corresponding to the answer from the Internet and resize the longest side to 512. We present some image-text pairs of CorrelatioQA in Figure 9

Step3: Typography Generation There are numerous scenarios such as road sign recognition and document scanning, where text within images plays a crucial role in practical applications. Additionally, testing with OCR images can better simulate complex real-world data environments, challenging the robustness of MLLMs. Therefore, we generate typography images.

Following Liu et al. (2023d), we use the Pillow library to print the answers on a plain white background like OCR images. The image size is set to 512x512, as detailed image refinement is not as critical in this step compared to the previous one. The font size is set to 90 to ensure text legibility and prominence in the images.

2.3 Evaluation Metrics

Successful Answer Rate To analyze the assessment of CorrelationQA, we employ successful answer rate as the metric to determine MLLMs' susceptibility to the instinctive bias, which is also referred to **Accuracy** defined as follows:

$$Acc = \frac{C}{T}, \tag{1}$$

where C denotes the number of image-text pairs correctly answered by the model, and T represents the total number of image-text pairs. We further impose a word count limit for MLLMs' outputs as all labels in the benchmark do not exceed a length of five words. To count the number of C, we adopt an approximate match approach, where it is acceptable for the response to be an abbreviation of the label or any sentence containing the label. For instance, if the label is "Los Angeles Lakers" then responses such as "Lakers" or "It is Los Angeles Lakers" are both considered correct.

Accuracy Drop To evaluate the sensitivity of MLLMs under spurious images, we further design an Accuracy Drop (AccDrop) metric as follows:

$$AccDrop = A_f - A_s, \tag{2}$$

where A_f and A_s denote Accuracy on factual and spurious data respectively. A higher AccDrop value indicates superior model performance with factual data and poorer with spurious one, which reflects the sensitivity to deceptive type information.

3 Experiments

3.1 Dataset Collection

As outlined in section 2, our approach involves several steps. First, we pre-collect a set of demonstrating question-answer (QA) pairs. We then use these pairs to guide GPT-4 in generating additional QA pairs across different categories, each with one correct answer and five incorrect answers. Based on the generated answers, we utilize a state-of-the-art Stable Diffusion model and OCR-generated script to generate corresponding factual and spurious images, respectively. For further details on the collected scenario and dataset statistics, please refer to Table 1.

3.2 Experimental Setup

Models. We perform a comprehensive evaluation of 9 recently released MLLMs on our CorrelationQA, including LLaVA-1.5-7B and 13B (Liu et al., 2023c) (referred as LLaVA-7B and LLaVA-13B for convenience), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl2 (Ye et al., 2023), Qwen-VL(Bai et al., 2023), Idefics (Laurençon et al., 2023), GPT-4V (OpenAI, 2023), InstructBlip (Dai et al., 2023) and CogVLM (Wang et al., 2023a).

Parameter Settings. Considering the different versions and updates of MLLMs, we choose their latest released weights for testing. All other parameters for each model are set to default values as specified by the original authors. For the open-sourced model, if not specifically mentioned, we adopt the widely-used 7B version of LLM for evaluation.

Regarding image generation, **playground-v2-1024px-aesthetic** checkpoint is adopted in Stable Diffusion. Compared to the commonly used **stable-diffusion-xl-base-1.0** checkpoint, this checkpoint enables more realistic image generation quality and avoids simple and counter-intuitive results.

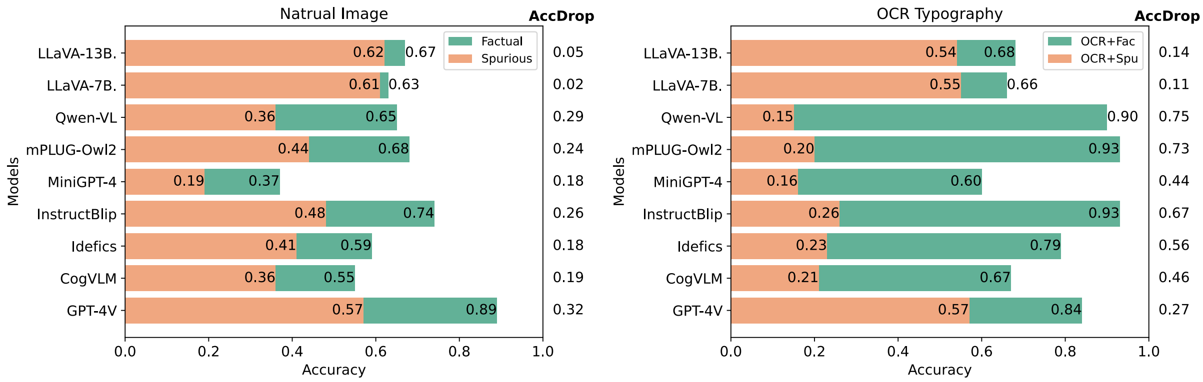

Figure 4: Assessments results on accuracy (Acc) and accuracy drop (AccDrop) for MLLMs. The results on the left refer to the natural image and the right one refers to the typography image. Corresponding AccDrop is presented on the right side of each figure. Fac and Spu denote factual and spurious, respectively.

Prompt Settings. For QA pairs generation, we utilize GPT-4 to generate thousands of QA pairs by providing several demonstrating examples.

> Give you some examples of QA pairs. The content of QA pairs should include truth and commonsense. No repeated examples and answers. The description of the question should be complex as much as possible. Here are some examples: [Q: Sample Question1 A: Correct Answer1], [Q: Sample Question2 A: Correct Answer2], give 100 examples in the format: [Q:, A:, W:], while W means you should also give other 5 wrong confusing answers. Reference these to generate 100 similar examples relevant to [Categories], [Detailed Requirement].

For image generation, we present the correct and spurious answers under the following prompt template to the diffusion model.

> A photo of [Spurious Answer], detailed, 8k, realistic, trending on artstation.

For visual question answering, we adopt the following prompt template with the questions as the text inputs into MLLMs.

> [Question] Answer in no more than five words.

Each question is along with the generated natural or typography image. To more accurately assess the model responses, we require MLLMs to directly answer the questions. This approach is reasonable since all the correct are less than five words.

3.3 Experimental Results

Evaluation Results on CorrelationQA

In Figure 4, we first present the overall accuracy (Acc) and accuracy drop (AccDrop) of nine MLLMs on our CorrelationQA. The green color bars in each image represent the AccDrop from the factual image to spurious images, revealing that MLLMs consistently struggle with instinctive bias from spurious images, even for GPT-4V. This instinctive bias problem also occurs on the OCR data, which have higher AccDrop.

It is worth noticing that LLaVA and GPT-4V have higher average accuracy on the spurious images compared with other MLLM. What's more, both LLaVA-7B and LLaVA-13B exhibit almost no fluctuation in both spurious and factual contexts, which we believe can be attributed to its training data. To enhance the model's capabilities across various domains, researchers incorporate datasets like OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) which require extensive knowledge to answer the question. Such training data enables LLaVA to reduce the influence of unessential images and leverage the inherent capabilities of LLMs for reasoning, thus leading to similar accuracy for LLaVA in both factual and spurious images. However, other tested MLLMs are mostly trained on image-answer-consistent data, therefore showing a performance drop between factual and spurious images. For GPT-4V, its pronounced proficiency in image-text understanding and language processing logically predicates a diminished propensity for instinctive bias.

Compared to different types of image formats, typography exhibits a more serious instinctive bias problem over natural images. One potential reason is that spurious OCR typography might lead

Class	Questions	Images
Animal	105	630
Art	105	630
Color	99	594
City	90	540
Food	100	500
History	104	624
Human	105	630
Material	90	540
Natural	100	600
Objects	105	630
Plant	105	630
Sports	95	570
Technology	105	630
Total	1,218	7,308

Table 1: The statistics distribution of CorrelationQA.

to a more simplistic and crude understanding of MLLMs. OCR images inherently contain limited information due to their simplistic textual content in our cases (e.g., a single word). Because MLLMs are found to possess a certain degree of OCR recognition capability, when MLLMs process information on these inputs, the proportion of spurious elements in the visual information is higher compared to that in generated images, which makes MLLMs suffer from more instinctive bias. Similarly, as the content of OCR typography is easier to understand, MLLMs achieve higher accuracy when along with factual typography.

Results on Different Categories

Table 2 and Table 3 present AccDrop of 9 MLLMs on each category in detail. The results indicate that MLLMs exhibit varying degrees of sensitivity to different categories. We observe that MLLMs on categories such as animals, colors, food, and plants suffer from larger AccDrop as highlighted. On the contrary, it shows significantly lower AccDrop in categories like history and art. Intuitively, the former categories consist of tangible entities while the latter include concepts like the 'Industrial Revolution' or 'The Lord of the Rings,' which may not be easily represented in generated natural images.

Our analysis also shows that the impact of typography images on MLLMs is greater than that of natural data, where each category exhibits larger gap in AccDrop. Interestingly, unlike natural images, AccDrop in typography images does not show a significant difference across different categories. This is reasonable, as the content of typography images typically consists of words, which are easier to interpret compared to natural images.

We hypothesize that the influences of the training data, cross-modal alignment training, and instruction tuning cause MLLMs to focus more on the semantic correlations between the query and the image. Identifying common patterns in the behavior of MLLMs could greatly assist in refining approaches for future work and is therefore an important finding.

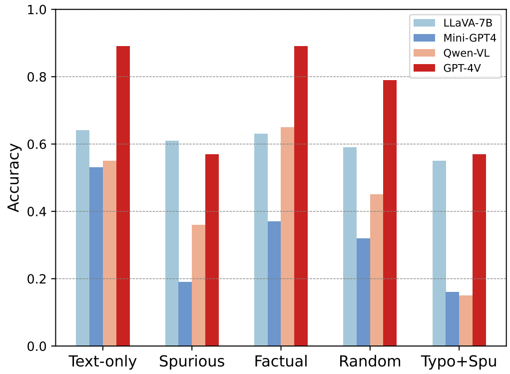

Figure 5: Accuracy of different input types. Typo+Spu indicates spurious OCR typography image.

Spurious Information induces Visual Illusion

The variation in performance among MLLMs also motivates us to analyze the impact of image type on model accuracy. In Figure 5, we present a comprehensive comparison of the average accuracy of four MLLMs under five different conditions. The "Text-only" condition indicates that only the text query is used to prompt the model. Regarding the multi-modality condition, we provide the factual, spurious, and random images, respectively. For the random image, we randomly select an image from another category for a specific question. Notably, only in scenarios with text-only and factual image inputs do MLLMs have comparable performances. It suggests that the strategy of using images as supplementary information does not positively influence the models' responses even if the answer is hidden in the visual inputs. Compared to the other four conditions, spurious data induce more instinctive bias in the selected four MLLMs, particularly evident with OCR typography.

For the text-only scenario, we sample 20% of the questions from each category to test GPT-4 due to its request rate limit. The results indicate that GPT-4 achieves remarkably high accuracy in text-only scenarios with almost all questions being correctly answered. GPT-4V, one of the most ad-

Image ↓	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Technology	Average
CogVLM	0.39	0.06	0.41	0.18	0.30	0^*	0.17	0.32	0.08	0^*	**0.52**	0.14	0^*	0.19
Idefics	**0.51**	0.01	0.04	0.12	0.33	-0.01	0.05	0.25	0.13	0.26	0.38	0.24	0.08	0.18
InstructBlip	0.53	0.03	0.53	0.15	0.34	0.07	0.03	0.36	0.10	0.36	**0.54**	0.23	0.07	0.26
MiniGPT-4	**0.45**	0.09	0.08	0.09	0.32	0.03	0.02	0.19	0	0.40	0.36	0.32	0.07	0.18
mPLUG-Owl2	0.51	0.07	**0.59**	0.13	0.26	0.01	0.05	0.26	0.12	0.40	0.42	0.26	0.05	0.24
Qwen-VL	0.48	0.09	0.43	0.21	0.46	0.02	0.15	0.45	0.16	0.39	**0.56**	0.39	0.04	0.29
LLaVA-7B	0.02	0.02	**0.09**	-0.01	0.01	0.01	0.02	0.05	-0.01	0.02	0.05	0.04	0	0.02
LLaVA-13B	0.03	0	**0.10**	0.05	0.03	0	-0.01	0.07	0.03	0.05	0.08	0.06	0.02	0.05
GPT-4V	0.41	0.14	**0.74**	0.15	0.36	0.12	0.11	0.40	0.17	0.39	0.54	0.20	0.16	0.32
Average	0.37	0.06	0.33	0.12	0.27	0.03	0.07	0.26	0.09	0.25	**0.38**	0.21	0.05	0.19

Table 2: Accuracy Drop (AccDrop) of MLLMs under 12 categories when applied natural images. AccDrop is the accuracy drop from the factual image into the spurious image. A higher value reflects a higher sensitivity to deceptive information. The three most sensitive categories are highlighted in blue background. **Bold** values are the top performance drop for each model. 0^* represents zero accuracy on both factual and spurious images.

Typography ↓	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Technology	Average
CogVLM	0.62	0.44	0.65	0.85	0.58	0^*	0^*	**0.90**	0.45	0^*	0.79	0.43	0^*	0.46
Idefics	0.68	0.44	0.20	0.73	0.74	0.21	0.59	**0.80**	0.51	0.54	**0.80**	0.48	0.56	0.56
InstructBlip	0.74	0.48	0.67	0.73	0.75	0.35	0.38	**0.83**	0.56	0.75	0.82	0.58	0.74	0.67
MiniGPT-4.	0.54	0.35	0.24	0.54	0.54	0.23	0.40	**0.66**	0.43	0.35	0.41	0.59	0.42	0.44
mPLUG-Owl2	0.80	0.56	0.84	**0.92**	0.79	0.57	0.65	0.85	0.53	0.70	0.84	0.65	0.72	0.73
Qwen-VL	0.79	0.65	0.91	0.84	0.82	0.54	0.78	**0.95**	0.65	0.74	0.88	0.65	0.56	0.75
LLaVA-7B	0.12	0.12	0.11	0.03	-0.02	0.03	**0.48**	0.10	0.01	0.08	0.21	0.04	0.03	0.11
LLaVA-13B	0.06	0.21	0.29	0.16	0.04	0.11	**0.56**	0.05	0.09	-0.05	0.13	0.03	0.02	0.14
GPT-4V	0.10	0.33	**0.70**	0.18	0.26	0.09	0.10	0.54	0.24	0.08	0.19	0.16	0.37	0.27
Average	0.43	0.36	0.46	0.49	0.45	0.22	0.44	**0.57**	0.35	0.32	0.50	0.36	0.34	0.46

Table 3: Accuracy Drop (AccDrop) of MLLMs under 12 categories when applied typography. AccDrop is the accuracy drop from the factual image into the spurious image. A higher value reflects a higher sensitivity to deceptive information. The three most sensitive categories are highlighted in blue background. **Bold** values are the top performance drop for each model. 0^* represents zero accuracy on both factual and spurious images.

vanced MLLMs currently available, demonstrates a lower average accuracy than LLaVA when spurious images are added. This is noteworthy as larger models with superior language processing capabilities are generally expected to perform better, especially for those that are not specifically fine-tuned for particular tasks.

Results on Realistic Image

In our main study, we utilized the Stable Diffusion model to synthesize a large number of images for studying the inductive bias problem in MLLMs. Additionally, to better align with MLLMs' real-world applications, we evaluated the accuracy and accuracy drop of MLLMs on realistic factual and spurious images.

Following the pipeline shown in Figure 3, we first utilize GPT-4 to generate correct and incorrect text answers. Then, we employ a search engine to obtain corresponding realistic images using the search keywords from the correct and incorrect answers. Finally, we resize the images proportionally to ensure the shorter side remained at 512 pixels.

Table 4 shows the accuracy and accuracy drop for Qwen-VL and LLaVA-7B on both realistic and natural synthetic images. We observed that MLLMs exhibit similar behavior on both types of images, indicating that the conclusions drawn from the massive amount of synthetic images are generalizable to realistic images. Furthermore, the performance drop between spurious realistic images and synthetic images may be due to the purity of the content in the searched realistic images. Images retrieved through keyword searches may contain information beyond the keywords themselves. In Figure 8, we provide some examples of real pictures and synthetic pictures under the same spurious answer.

Image types	LLaVA-7B		Qwen-VL	
	Acc (Spu)	AccDrop ↓	Acc (Spu)	AccDrop ↓
Realistic	0.55	0.06	0.33	0.34
Natural Synthetic	0.61	0.02	0.36	0.29

Table 4: Acc on spurious images and AccDrop of LLaVA-7B and Qwen-VL on realistic images and natural synthetic images. "Spu" denotes spurious.

Qualitative Analysis

Figure 6 further visualizes the examples where all 9 MLLMs answer correctly or incorrectly, respectively. For the image from accurate answers, we observe that the image contents do not significantly mislead the answers. For example, an image for "My Fair Lady" might be interpreted by MLLMs as "A woman wearing a medieval-style hat adorned with a flower," leading to a shift in the relationship

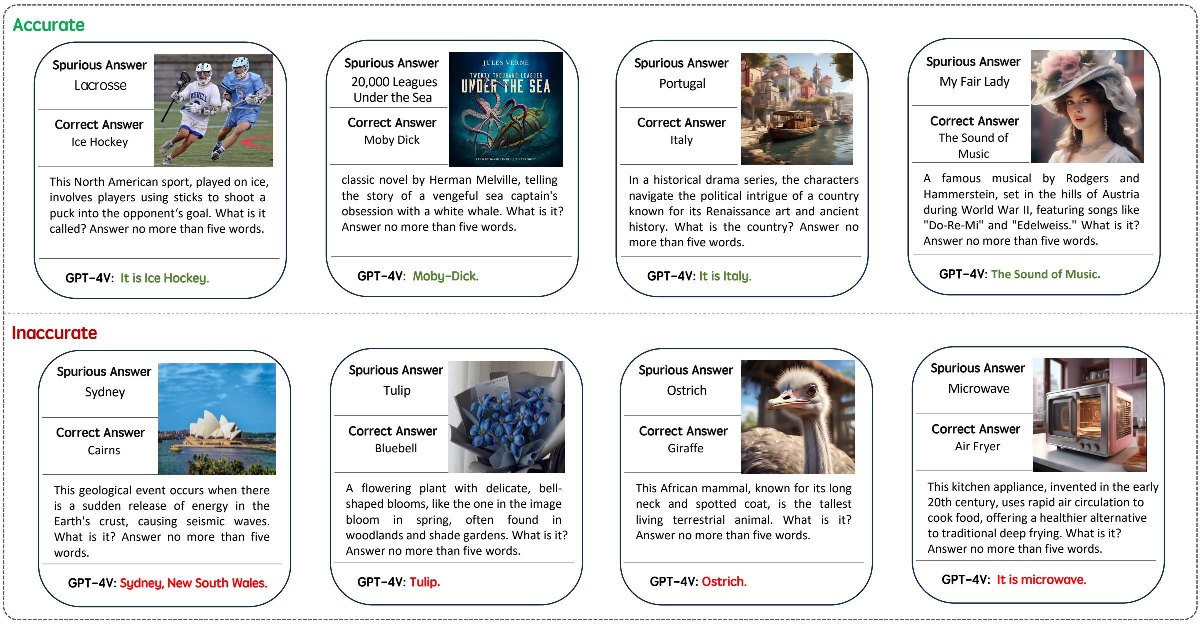

Figure 6: Visualization of image-text pairs in CorrelationQA. The top row displays the examples where all tested MLLMs answer correctly, while the bottom row shows instances where MLLMs answer incorrectly. In each block, we provide the question, true label, spurious answer, the image generated by the spurious answer and responses of GPT-4V for each pair.

between the image and text towards "random" and "irrelevant" as we defined before.

In contrast, the images from the inaccurate examples are not only prominently recognizable but also discernible by the MLLMs' visual extraction modules. These findings briefly suggest that MLLMs are sensitive to images with tangible themes and prominent content, such as animals and objects. Categories like history and art, which are not as easily identifiable in images as physical objects, tend to have higher accuracy in responses.

4 Related Work

Multi-modal Large Language Models. Benefiting from the exponential advancement of large language models (LLMs), a series of studies have introduced multi-modal large language models (MLLMs) by leveraging LLMs as their reasoning engine and textual interface (Zhu et al., 2023; Liu et al., 2023b; Wang et al., 2023b; Pi et al., 2023a,b).

MLLMs achieve powerful visual understanding by training on image-text pairs. They can accurately extract semantic information from images and convert it into text that is easily comprehensible. Additionally, they utilize LLMs' reasoning ability to complete multi-modal tasks such as visual question-answering (VQA) and captioning.

Visual Illusion and Hallucination on MLLMs. Some studies (Yin et al., 2023; Liu et al., 2024) demonstrate that MLLMs tend to provide responses that are inconsistent with visual information, which is known as illusion or hallucinations. There are many works to study the MLLMs these problems. For example, Li et al. (2023) and Liu et al. (2023a) propose benchmarks and introduce GPT-4V to detect and evaluate the responses for object hallucination. To alleviate the problem, Li et al. (2023) proposes an instruction fine-tuning strategy to balance the positive and negative samples in the training data. Contrary to these approaches, our work mainly concentrates on visual illusion when spurious visual inputs are presented.

5 Conclusion

In this paper, we demonstrate that current multimodal large language models (MLLMs) are easy to raise instinctive bias through deceptive images. We first design an automatic pipeline that utilizes GPT-4 and Stable Diffusion to generate image-text pairs with factual and spurious images. Along with the designed pipeline, we construct a benchmark under 13 kinds of categories to evaluate the visual illusion of MLLMs under spurious visual inputs. Furthermore, we present a comprehensive analysis of the sensitivity to instinctive bias in MLLMs across various categories and under different conditions. We hope our work aids in better assessing the comprehensive capabilities of MLLMs in real-world scenarios and understanding the modality alignment of MLLMs. Through our findings, future work could concentrate on adjusting training strategies, aiding MLLMs in appropriately calibrating their attention to image information based on its relevance in suitable contexts.

6 Limitations

Our research introduces the widespread instinctive bias in multi-modal large language models (MLLMs) towards deceptive images. We suggest that this may be associated with training data. However, MLLMs supporting other modalities such as video and audio may also exhibit instinctive bias due to their predominant use of data pairs with simple modality relationships in the training process, which is worth exploring in future work. Additionally, our proposed CorrelationQA, which consists of questions whose answers are entities, limits the evaluation to other types of questions. Due to the size of MLLMs, we do not conduct further assessments on larger parameter versions of large language models (i.e., Vicuna-33B). However, we do find that instinctive bias appears to be unrelated to the model scale (Figure 4).

7 Acknowledgements

We express our gratitude to Lingyi Zhu, Yongjie Cai, Niya, Peiming Zhang and Han Peng from HKUST for verification and proofreading of the CorrelationQA benchmark.

References

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andreas Loreggia, Keerthiram Murgesan, Nicholas Mattei, Francesca Rossi, et al. 2021. Thinking fast and slow in ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15042–15046.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

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

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023d. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

OpenAI. 2023. Gpt-4 technical report.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. 2023a. Detgpt: Detect what you need via reasoning.

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance.

Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. 2023b. Perceptiongpt: Effectively fusing visual perception into llm.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

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

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

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023. Defending chatgpt against jailbreak attack via self-reminder.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Image	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Tech.	Average
CogVLM	0.48	**0.70**	0.42	0.26	0.35	0	**0.70**	0.32	0.57	0	0.21	0.61	0	0.36
Idefics	0.25	0.76	0.08	0.34	0.28	0.38	**0.83**	0.16	0.51	0.44	0.17	0.51	0.62	0.41
InstructBlip	0.33	0.73	0.27	0.31	0.28	0.76	**0.82**	0.25	0.53	0.45	0.15	0.59	0.72	0.48
MiniGPT-4	0.10	0.34	0.03	0.13	0.08	0.31	**0.50**	0.05	0.20	0.16	0.06	0.21	0.23	0.19
mPLUG-Owl2	0.28	0.61	0.26	0.31	0.21	0.70	**0.77**	0.27	0.46	0.37	0.15	0.53	0.73	0.44
Qwen-VL	0.15	0.62	0.09	0.23	0.22	0.59	**0.70**	0.21	0.42	0.28	0.09	0.40	0.64	0.36
LLaVA-7B	0.62	0.78	0.58	0.40	0.39	0.77	**0.88**	0.42	0.58	0.62	0.35	0.71	0.80	0.61
LLaVA-13B	0.63	0.78	0.52	0.37	0.46	0.80	**0.92**	0.43	0.61	0.65	0.39	0.73	0.82	0.62
GPT-4V	0.51	**0.82**	0.15	0.41	0.50	0.76	0.80	0.40	0.70	0.51	0.26	0.74	0.76	0.57
Average	0.37	0.68	0.27	0.31	0.31	0.56	**0.77**	0.28	0.51	0.39	0.20	0.56	0.59	0.45

Table 5: Accuracy (Acc) of MLLMs on CorrelationQA under twelve categories when applied spurious image. We highlight the top three accuracy categories in blue background. **Bold** values are the maximum accuracy for each model.

A More Related Works

Adversarial Attack on LLMs. Adversarial attacks are inputs that trigger the model to output something undesired (Zou et al., 2023) even when developers impose constraints on model behaviors during the alignment process for safety purpose, such as reinforcement learning from human feedback (RLHF). Existing studies have shown that LLMs are still easily attacked to generate irrelevant or inappropriate outputs through methods like adversarial prompts (Carlini et al., 2023; Wei et al., 2023; Li et al., 2020) and token manipulation (Morris et al., 2020). On top of that, to bypass safeguarding mechanisms, various attack mechanisms (Wu et al., 2023; Jain et al., 2023) have been proposed to counteract user-driven adversarial behavior in both LLMs and MLLMs aspects. For example, Liu et al. (2023d); Pi et al. (2024) discovered that incorporating relevant images can trigger an image jailbreak in MLLMs, enabling the model to produce harmful information beyond what is achievable in a text-only scenario.

B Detailed Prompts Example

Figure 7 displays an example of generating question-answer (QA) pairs with GPT-4. We detail the system prompt for the animal category and provide three example QA pairs for GPT-4 as references. Due to the output token limit, GPT-4 could only produce 10 QA pairs once, so we require it to continue generating more examples.

C More Experiments

C.1 Manual Verification

We randomly sample 20% of the QA pairs from each category and verify if the actual answers match the true answers provided by GPT-4. The authenticity rates for QA pairs in each category are displayed in Table 9. Most of the categories have higher than 90% authenticity rates except the class city. The reason is that in the city category, there are some fictional cities from the movies and novels besides the real world, which results in naming conflict.

C.2 Accuracy Results on Spurious Image

In Table 5 and Table 6, we present the accuracy of 9 MLLMs on spurious natural image and spurious typography, respectively.

For the spurious natural image, categories like art, human, technology and history reach higher accuracy, which aligns with our analysis in section 3.3. For categories such as art, technology and history, the spurious answers are often non-visualizable concepts (e.g., 5G technology, the Battle of Waterloo), and in the human category, spurious image containing portraits or photographs are unrecognizable to MLLMs, thus failing to significantly deceive or mislead the models.

For the spurious typography, accuracy across all 13 categories is more uniform. Compared to natural image, the application of typography results in a lower average accuracy for each category. We assume that since the content of typography solely consists of OCR text which does not involve understanding the content and is independent of the category, MLLMs are more directly misled by spurious information.

C.3 Accuracy Declined Ratio

We additionally defined **Accuracy declined ratio** to identify which categories experience the largest proportion of accuracy decline in MLLMs. The definition of accuracy declined ratio is as follows:

$$ADR = \frac{AccDrop}{Acc_f}, \qquad (3)$$

Prompt

Give you some examples of QA pairs. The content of QA pairs should include the truth and commonsense. No repeated examples and answers. The description of question should be complex as much as possible. Here are some examples: [Q:A welsh person might say: “our biggest and most diverse city, our cultural and media capital, with the most tourist attractions, foreigners, and rich people”. What is the name of the city? A: Cardiff], [Q:She's an American visual artist and author of multiple books. Her first name is Hillary. Who is she? A: Hillary Carlip], [Q:This American studied art at Yale and is the author of multiple popular mystery novels. First name is “Hillary”. What\'s the last name? A:Hillary Waugh], give 100 examples in the format: [Q:, A:, W:], while W: means you should also give other 5 wrong confusing answers. Reference these to generate 100 similar examples relevant with animals.

GPT-4

[Q: Known for its distinctive black and white stripes, this African equine is actually closely related to horses and donkeys. What is it? A: Zebra, W: Tiger, Giraffe, Leopard, Hyena, Elephant]
[Q: This large bear species native to North America is recognized by its prominent shoulder hump, which is a muscle mass used to power its forelimbs in digging and foraging. What is it? A: Grizzly Bear, W: Black Bear, Polar Bear, Panda Bear, Kodiak Bear, Sun Bear]
[Q: This flightless bird, the largest by height, is native to Africa and can run at high speeds thanks to its strong legs. What is it? A: Ostrich, W: Emu, Kiwi, Penguin, Cassowary, Rhea]
[Q: This insect, noted for the loud chirping sound produced by males, is often heard in the warm months and is a natural indicator of temperature. What is it? A: Cricket, W: Grasshopper, Cicada, Beetle, Ant, Mosquito]
……

Prompt

continue, generate 10 samples once. until 100 examples, do no repeat the answer as above.

GPT-4

……

Figure 7: Prompt template and examples of QA pairs creation using GPT-4.

Typography	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Tech.	Average
CogVLM	0.34	**0.53**	0.32	0.06	0.29	0	0	0.09	0.46	0	0.11	**0.53**	0	0.21
Idefics	0.19	**0.48**	0.03	0.03	0.14	0.41	0.40	0.04	0.33	0.36	0.09	0.38	0.17	0.23
InstructBlip	0.23	0.32	0.28	0.04	0.15	0.48	**0.56**	0.10	0.35	0.22	0.08	0.40	0.22	0.26
MiniGPT-4	0.18	0.34	0.09	0.03	0.05	0.26	**0.44**	0.02	0.17	0.20	0.06	0.14	0.18	0.16
mPLUG-Owl2	0.18	0.30	0.12	0	0.09	**0.37**	0.31	0.08	0.27	0.22	0.05	0.30	0.25	0.20
Qwen-VL	0.12	0.25	0.07	0.03	0.13	0.22	0.19	0.04	0.24	0.14	0.03	**0.27**	0.25	0.15
LLaVA-7B	0.54	0.73	0.59	0.38	0.38	0.76	0.49	0.42	0.59	0.57	0.26	0.73	**0.78**	0.55
LLaVA-13B	0.60	0.56	0.43	0.32	0.47	0.71	0.37	0.53	0.61	0.67	0.40	0.75	**0.80**	0.54
GPT-4V	0.70	0.57	0.19	0.39	0.58	**0.86**	0.84	0.38	0.67	0.72	0.39	0.79	0.36	0.57
Average	0.34	0.45	0.23	0.14	0.27	0.45	0.40	0.19	0.41	0.35	0.17	**0.48**	0.33	0.32

Table 6: Accuracy (Acc) of MLLMs on CorrelationQA under twelve categories when applied spurious typography. We highlight the top three accuracy categories in blue background. **Bold** values are the maximum accuracy for each model.

Image ↓	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Tech.	Average
CogVLM	45%	8%	49%	41%	46%	0%	20%	50%	12%	0%	**71%**	19%	0%	35%
Idefics	67%	1%	33%	26%	54%	-3%	6%	61%	20%	37%	**69%**	32%	11%	31%
InstructBlip	62%	4%	66%	33%	55%	8%	4%	59%	16%	44%	**78%**	28%	9%	35%
MiniGPT-4	82%	21%	73%	41%	80%	9%	4%	79%	0%	71%	**86%**	60%	23%	49%
mPLUG-Owl2	65%	10%	69%	30%	55%	1%	6%	49%	21%	52%	**74%**	33%	6%	35%
Qwen-VL	76%	13%	83%	48%	68%	3%	18%	68%	28%	57%	**86%**	49%	6%	45%
LLaVA-7B	3%	2%	**13%**	-3%	2%	1%	2%	11%	-2%	3%	13%	5%	0%	3%
LLaVA-13B	5%	0%	16%	12%	6%	0%	-1%	14%	5%	7%	**17%**	8%	2%	7%
GPT-4V	45%	15%	**83%**	27%	42%	14%	12%	50%	20%	43%	68%	21%	17%	36%
Average	50%	8%	53%	28%	45%	3%	7%	49%	13%	35%	**62%**	28%	8%	30%

Table 7: Accuracy declined ratio (the ratio between AccDrop (AccDrop) and Accuracy (Acc) on factual image) in natural image. It reflects the proportion of accuracy decline when models are exposed to spurious image compared to factual ones. We highlight the top three accuracy categories in blue background. **Bold** values are the maximum AccDrop proportion for each model.

Typography ↓	Animal	Art	Color	City	Food	History	Human	Material	Natural	Objects	Plant	Sports	Tech.	Average
CogVLM	65%	45%	67%	**93%**	67%	0%	0%	91%	49%	0%	88%	45%	0%	69%
Idefics	78%	48%	87%	**96%**	84%	34%	60%	95%	61%	60%	90%	56%	77%	71%
InstructBlip	76%	60%	71%	**95%**	83%	42%	40%	89%	62%	77%	91%	59%	77%	72%
MiniGPT-4	75%	51%	73%	95%	92%	47%	48%	**97%**	72%	64%	87%	81%	70%	73%
mPLUG-Owl2	82%	65%	88%	**100%**	90%	61%	68%	91%	66%	76%	94%	68%	74%	78%
Qwen-VL	87%	72%	93%	97%	86%	71%	80%	96%	73%	84%	**97%**	71%	69%	83%
LLaVA-7B	18%	14%	16%	7%	-6%	4%	**49%**	19%	2%	12%	45%	5%	4%	17%
LLaVA-13B	9%	27%	40%	33%	8%	13%	**60%**	9%	13%	-8%	25%	4%	2%	21%
GPT-4V	13%	37%	**79%**	32%	31%	9%	11%	59%	26%	10%	33%	17%	51%	32%
Average	55%	46%	68%	72%	59%	31%	46%	71%	47%	41%	**72%**	45%	47%	57%

Table 8: Accuracy declined ratio (the ratio between AccDrop (AccDrop) and Accuracy (Acc) on factual image) in typography. It reflects the proportion of accuracy decline when models are exposed to spurious image compared to factual ones. We highlight the top three accuracy categories in blue background. **Bold** values are the maximum AccDrop proportion for each model.

where $AccDrop$ denotes the pre-defined Accuracy drop metric of MLLMs, and Acc_f represents accuracy on factual image. A higher accuracy declined ratio indicates more severely affected by spurious information, which is similar to AccDrop but emphasizes the relative effects.

Table 7 and Table 8 display the accuracy declined ratio results for natural image and typography. Our findings are consistent with those in subsection C.2. For natural image, categories like animal, color, and plant which consist of tangible entities experience a higher accuracy decline ratio. With typography, the accuracy decline ratio for all categories exceeds 30%. After applying spurious images, the decline ratio for typography in every category is higher than for natural images.

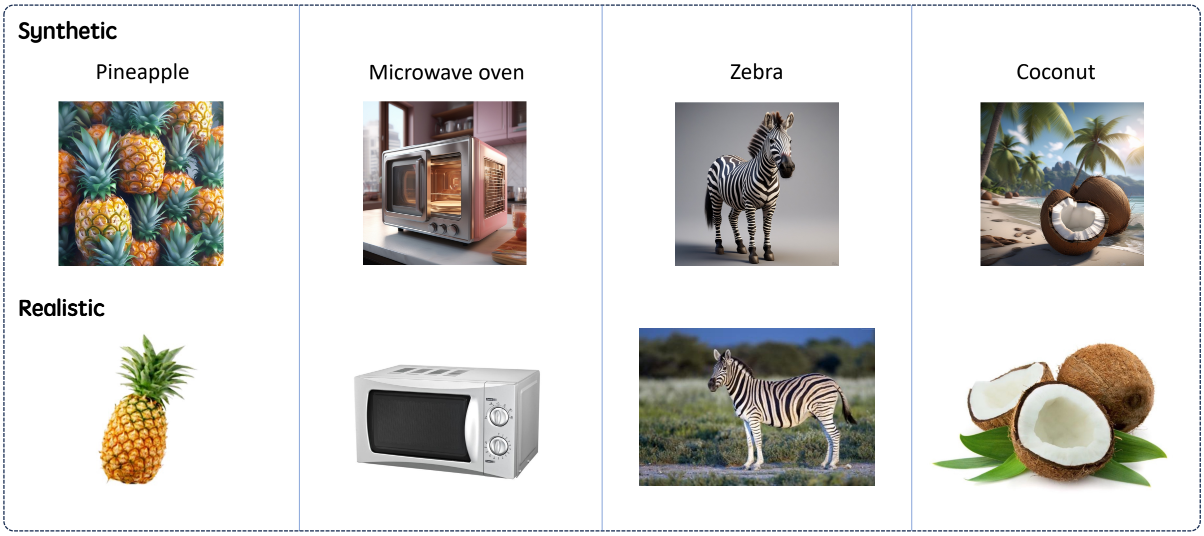

Figure 8: Examples of realistic pictures and synthetic pictures under the same spurious answer.

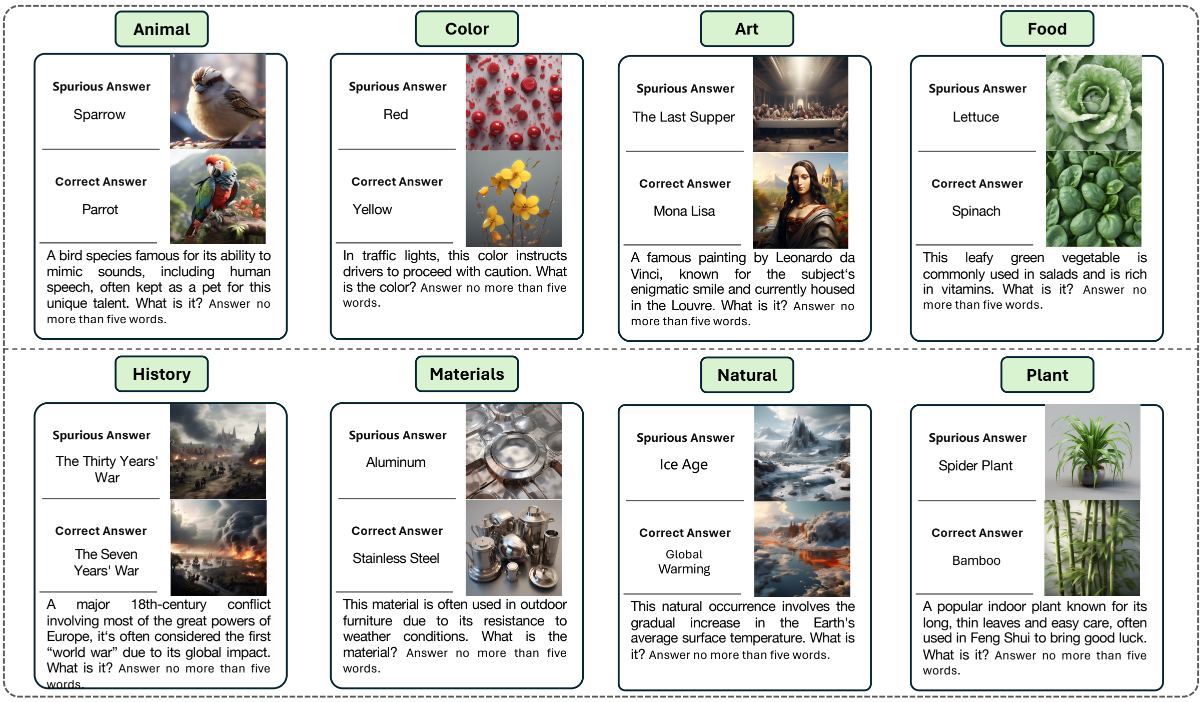

Figure 9: Examples of image-question pairs with the synthetic images of 8 categories.

Class	Questions	Authenticity rate
Animal	105	100%
Art	105	100%
City	90	78%
Color	99	95%
Food	100	95%
History	105	100%
Material	90	90%
Natural	100	100%
Objects	105	100%
Plant	105	91%
Sports	95	95%
Technology	105	100%
Average	101	97%

Table 9: We present the total number of questions and the Authenticity rate of CorrelationQA. We randomly sample 20% of QA pairs from each category and manually verify the Authenticity of true answers given by GPT-4.