MM-SpuBench: Towards Better Understanding of Spurious Biases in Multimodal LLMs

Wenqian Ye[1] **Guangtao Zheng**[1] **Yunsheng Ma**[2] **Xu Cao**[3] **Bolin Lai**[4]
James M. Rehg[3] **Aidong Zhang**[1]
[1]University of Virginia
[2]Purdue University
[3]University of Illinois Urbana-Champaign
[4]Georgia Institute of Technology
{wenqian,aidong}@virginia.edu

Abstract

Spurious bias, a tendency to use spurious correlations between non-essential input attributes and target variables for predictions, has revealed a severe robustness pitfall in deep learning models trained on single modality data. Multimodal Large Language Models (MLLMs), which integrate both vision and language models, have demonstrated strong capability in joint vision-language understanding. However, whether spurious biases are prevalent in MLLMs remains under-explored. We mitigate this gap by analyzing the spurious biases in a multimodal setting, uncovering the specific test data patterns that can manifest this problem when biases in the vision model cascade into the alignment between visual and text tokens in MLLMs. To better understand this problem, we introduce MM-SpuBench, a comprehensive visual question-answering (VQA) benchmark designed to evaluate MLLMs' reliance on nine distinct categories of spurious correlations from five open-source image datasets. The VQA dataset is built from human-understandable concept information (attributes). Leveraging this benchmark, we conduct a thorough evaluation of current state-of-the-art MLLMs. Our findings illuminate the persistence of the reliance on spurious correlations from these models and underscore the urge for new methodologies to mitigate spurious biases. To support the MLLM robustness research, we release our VQA benchmark at `https://huggingface.co/datasets/mmbench/MM-SpuBench`.

1 Introduction

In recent years, we have witnessed the rise of highly performant Large Language Models (LLMs) [1, 2, 3, 4, 5, 6] and Vision Foundation Models (VFMs) [7, 8] powered by the advancements in language modeling and visual understanding as well as the availability of large-scale training data and substantial computational resources. Building on these advancements, multimodal Large Language Models (MLLMs) [9, 10, 11, 12, 13, 14, 15], which integrate both LLMs and VFMs for joint visual and text understanding, emerge as the new frontier of foundation models. MLLMs have demonstrated significant performance in visual understanding and reasoning tasks, such as image perception [16], visual question answering [17], and instruction following [18], making remarkable strides toward Artificial General Intelligence (AGI).

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Despite the impressive performance of MLLMs, the robustness of MLLMs remains largely under-explored. A well-known robustness issue in deep learning models is the *spurious bias*, a tendency to use spurious correlations between non-essential input attributes and target variables for predictions [19]. For example, image classifiers tend to identify an object by using the image background that frequently co-occurs with the object in the training data [20], and the image background and the target object establish a spurious correlation which is not inherently relevant to the prediction task. Much research [21, 22, 23, 24, 25] has been focusing on single-modality classification tasks. Given the prevalence of spurious biases in deep learning models, it is natural to ask the following question in the multimodal setting:

Are spurious biases prevalent in MLLMs? If so, how much are MLLMs affected?

To answer the above question, it is critical to identify the major cause of spurious biases in MLLMs. A recent finding [26] suggests that the predominant contrastive language-image pre-training (CLIP) [27] objective often leads to vision models overlooking crucial visual details in images. Motivated by this, we reason that in MLLMs, a core visual token representing a class may be spuriously aligned with multiple irrelevant text tokens. Consequently, MLLMs may struggle in answering challenging visual grounding questions which ask MLLMs to identify a target object in an image amongst descriptions of surrounding and spurious objects in the image. To illustrate, given an image of a boot in a bathroom setting and the question “What is the item being held upright on the flat surface next to the hygiene products?" (Fig. 2, Inference Data), an MLLM may not successfully identify the boot in the image where spurious objects, including a mouthwash, a cabinet, a towel, a toilet, and a sink (Fig. 2, Training Data), exist in the background. Indeed, the model incorrectly answers with “Choice A: A container for liquids", utilizing the strong spurious correlation between a spurious text token “container" and the core visual token “boot".

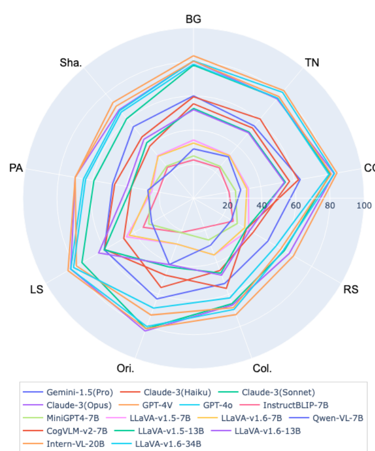


Figure 1: Comparative performance of different MLLMs across 9 types of spurious biases in MM-SpuBench.

Revealing and benchmarking spurious biases in MLLMs require dedicated evaluation data and methodologies that specifically target robustness pitfalls in MLLMs. However, there is a scarce of works that systematically evaluate spurious biases in MLLMs. To this end, we propose an automatic attribute-based Visual Question Answering (VQA) construction method based on our theoretical analysis on spurious biases in MLLMs. The idea is to test whether MLLMs produce wrong answers when spurious correlations are shifted in both vision and language modalities. We consider nine categories of spurious correlations when constructing VQA questions, creating a challenging evaluation scenario that exposes MLLMs' reliance on spurious correlations between vision and language modalities. To facilitate future research, we propose MM-SpuBench and a Visual Question Answering (VQA) benchmark specifically designed to evaluate the reliance of MLLMs on instance-level spurious correlations in training data. By investigating the reliance on spurious correlations of state-of-the-art vision encoders in MLLMs we carefully select 10,773 image data from five open-sourced datasets and design 2,400 VQA questions containing derived core/spurious attributes and types of spurious biases. Our experiments highlight the urge for better modality alignment techniques and how the information from the benchmark can help to improve the performance of current MLLMs as shown in Fig. 1.

Our contributions are summarized as follows:

- We formally define multimodal spurious bias in MLLMs, highlighting how spurious correlations can propagate from vision encoders and lead to failures in current MLLMs.
- We propose MM-SpuBench, a comprehensive benchmark featuring 10,773 realistic images with concept-based attribute information, paired with a subset with 2,400 VQA data,

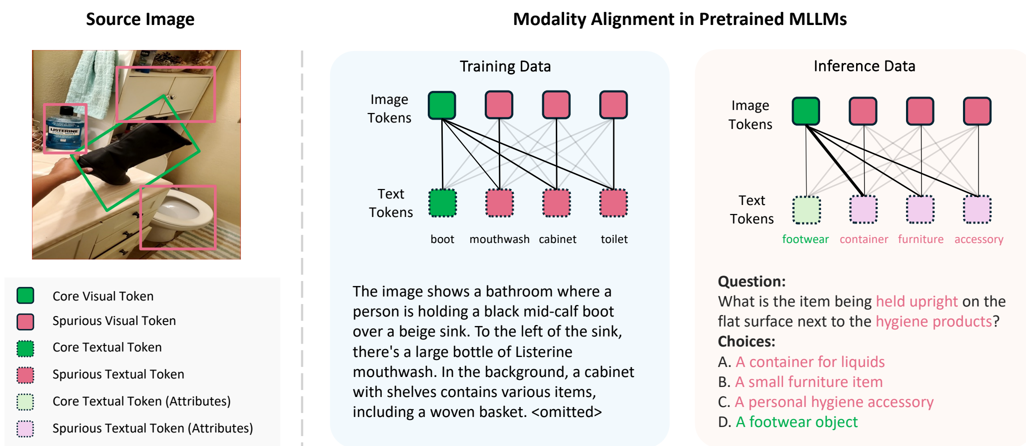


Figure 2: **Illustration of multimodal spurious bias:** From the training data of MLLMs, these models aim to learn visual grounding through instance-level correlations between visual objects and text descriptions. During inference, these correlations can be influenced by other attributes which refer to the same object attributes. In this case, we break the previous correlations and reveal the underlying spurious correlations learned by the model. It can be observed by the failures to accurately interpret the objects in the vision modality.

designed to systematically evaluate current MLLMs across 9 distinct categories of spurious biases.

- We conduct an in-depth analysis of current representative MLLMs, including 5 close-sourced and 10 open-sourced models with different parameter sizes, revealing the existing limitations on achieving effective alignment between vision and language modalities.

2 Related Works

Robustness in multimodal LLMs. Recent close-sourced MLLMs, such as GPT-4V [10], Claude [11], and Gemini [12], have demonstrated notable robustness to various distribution shifts. These models showcase the potential of MLLMs in handling diverse and challenging real-world scenarios. On the other hand, open-source methods like InstructBLIP [13], MiniGPT-4 [14], and LLaVA [15] emphasize the importance of high-quality visual instruction tuning data in improving the robustness of MLLMs [26]. However, MLLMs still face challenges in handling visually complex images due to limitations in visual search mechanisms [28] and visual grounding capabilities [26]. Moreover, MLLMs are susceptible to spurious correlations that can lead to hallucinations and non-trustworthy behaviors [29, 30]. Our paper focuses on the spurious bias issue in the multimodal setting, as it covers a broad family of biases prevalent in current MLLMs.

Spurious attribute detection. Spurious attributes can negatively impact a model's generalization capabilities during training [20]. Detecting these spurious attributes often requires domain knowledge [31, 32] and human annotations [33, 34]. Previous studies have identified object backgrounds [35] and image texture [36] as spurious attributes that can pose biases to the predictions of deep learning models. Recent research [37, 38] has employed explainable methods to automatically detect spurious attributes and their corresponding features through the neural networks. Additionally, [39] utilizes a pre-defined concept bank as an auxiliary knowledge base for spurious feature detection. In our work, we aim to automatically build human-understandable concept information based on both ground truth and incorrectly predicted labels with state-of-the-art vision encoders, and use this information for building more challenging tasks for MLLMs.

Benchmarks on multimodal LLMs. Previous benchmarks such as TextVQA [40] and GQA [41] have focused on traditional VQA queries. More recently, works like MM-Vet [42], POPE [43], and MM-Bench [44] have been developed to specifically evaluate multimodal LLMs in terms of hallucination, reasoning, and robustness. These evaluations have highlighted that multimodal LLMs can suffer from hallucination [45, 46], catastrophic forgetting [47], and a lack of robustness [48]. Unlike previous VQA benchmarks, which only include question-answer data, our benchmark also incorporates concept-based information on both core and spurious attributes. This addition helps future researchers distinguish between core and spurious information, thereby facilitating the development of spurious bias mitigation methods.

3 Spurious Biases in Multimodal LLMs

3.1 Problem Setting

In this study, we consider a common multimodal setting with the vision modality $\mathcal{X}$ and the language modality $\mathcal{Y}$. Given the image input $\mathbf{x} \in \mathcal{X}$ and text input (prior) $\mathbf{y} \in \mathcal{Y}$, an MLLM algorithm learns the mapping $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{C}$ such that $c = \phi(\mathbf{x}, \mathbf{y})$, where $c \in \mathcal{C} \subset \mathcal{Y}$ denotes the response (generated autoregressively) to $\mathbf{y}$ conditioned on $\mathbf{x}$. For example, the $\mathbf{x}$ could be an image showing a boot in the middle, and the $\mathbf{y}$ could be a question starting with “What is the object in the middle of the image?", then the output c of ϕ could be “a boot". To elucidate spurious biases in MLLMs, without loss of generality, we consider each input from any data modalities to have a spurious feature, a core feature, and a noise feature [21]. Specifically, we denote $\mathbf{x} = [x_{\text{core}}, x_{\text{spu}}, x_{\text{noise}}]$, representing the core, spurious, and noise features of $\mathbf{x}$. Similarly, we denote $\mathbf{y} = [y_{\text{core}}, y_{\text{spu}}, y_{\text{noise}}]$. In any of the two modalities, the core features are essential to generating the desired response c, spurious features are non-essential to c, and noise features storing sample-specific information.

In the multimodal setting, inputs from different modalities can have the same attribute. For example, both a text description and an image can contain the attribute “footwear" (Fig. 2). We use a latent feature vector $\mathbf{z} \in \mathcal{Z}$ [49] to model a modality-agnostic attribute, which is obtained by mapping modality-specific features to the latent feature vector space $\mathcal{Z}$. To analyze spurious biases in the multimodal setting, given a multimodal data tuple $(\mathbf{x}, \mathbf{y}, c)$, we restrict $\mathbf{z}$ to only representing *a spurious attribute* that is shared by $\mathbf{x}$ and $\mathbf{y}$ but not by c. For example, as illustrated in Fig. 2, $\mathbf{x}$ is an image showing a boot in a bathroom, $\mathbf{y}$ is a question regarding the boot, c is “A footwear object", and $\mathbf{z}$ could represent “A container for liquids".

3.2 From Single Modality to Multi-modality

To define spurious biases in the multimodal setting, we start with an analysis on a single modality scenario. Without loss of generality, we consider the vision modality $\mathcal{X}$ as an example. Given a data pair $(\mathbf{x}, c)$ from a training dataset, the target c typically represents a class label, and $\mathbf{x}$ has a spurious attribute $\mathbf{z}$. When spurious attributes and class labels in the training dataset have strong spurious correlations, the conditional probability distributions regarding $\mathbf{z}$ have the following relation: $p_{\text{train}}(\mathbf{z}|c, x_{\text{core}}) \gg p_{\text{train}}(\mathbf{z}|x_{\text{core}})$ [50], which describes strong correlations between a spurious attribute $\mathbf{z}$ and a class label c in the existence of the core input feature x_{core}. Spurious biases describe the tendency of a model using the spurious correlations described above for predictions.

Following the analysis on a single modality scenario, we extend our analysis to the multimodal setting. We define multimodal spurious bias as follows.

Definition 3.1 (Multimodal Spurious Bias)**.** Given an input image $\mathbf{x} = [x_{\text{core}}, x_{\text{spu}}, x_{\text{noise}}]$, a text input $\mathbf{y} = [y_{\text{core}}, y_{\text{spu}}, y_{\text{noise}}]$, the desired response c to the joint inputs $\mathbf{x}$ and $\mathbf{y}$, and a spurious attribute $\mathbf{z}$ shared by $\mathbf{x}$ and $\mathbf{y}$, the spurious correlations in the multimodal setting are expressed as follows.

$$p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}, c) \gg p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}). \tag{1}$$

The multimodal spurious bias is the tendency to use the spurious correlations between spurious attributes $\mathbf{z}$ and the desired responses c to generate responses given the core features in both modalities.

The inequality in Eq. (1) is derived from our assumptions on the varied degrees of spurious correlations in the vision and language modalities and on the weak correlation between the two modalities. Formally, we have the following proposition.

Proposition 3.1. Given that the vision and the language modalities are weakly correlated and that conditional distributions in the vision and language modalities have the following relations:

$$\textbf{Vision modality: } p(\mathbf{z}|c, x_{\text{core}}) \gg p(\mathbf{z}|x_{\text{core}}); \tag{2}$$

$$\textbf{Language modality: } p(\mathbf{z}|c, y_{\text{core}}) \approx p(\mathbf{z}|y_{\text{core}}), \tag{3}$$

Type	Description
Background (BG)	Occurs when the model relies on background context instead of the subject, e.g., identifying animals by natural backgrounds and failing in urban settings.
Texture and Noise (TN)	Arises when the model focuses on textures or noise patterns instead of shapes. E.g., misclassifying fruits due to changes in surface texture.
Co-occurring Objects (CO)	Happens when the model associates frequently appearing objects together. E.g., labeling any scene with a microwave as a kitchen.
Relative Size (RS)	Occurs when the model uses the relative size of objects as a cue. E.g., misclassifying a toy car as a real car due to a close-up perspective.
Colorization (Col.)	Related to reliance on specific colors for predictions. E.g., failing to recognize bananas that are green or brown.
Orientation (Ori.)	Arises when the model depends on the orientation of objects. E.g., struggling with faces not shown upright or from side profiles.
Lighting and Shadows (LS)	Occurs when predictions are influenced by lighting conditions or shadows. E.g., misclassifying objects in images with different lighting conditions.
Perspective and Angle (PA)	Emerges when the model relies on the viewing angle of objects. E.g., car recognition failing with top-down or oblique views.
Shape (Sha.)	Arises when an object has an unusual shape resembling another object. E.g., misidentifying a deformed fruit as a different type due to shape similarity.

Table 1: Types of spurious correlations categorized in MM-SpuBench.

the inequality in Eq. (1) holds.

Typically, in the vision modality, images are not balanced in terms of spurious attributes across different classes, leading to Eq. (2). In contrast, due to the flexibility of language and the massive amount of text data, a spurious attribute $\mathbf{z}$ often exhibits a weak correlation with a specific response c given a core text feature y_{core}, which leads to Eq. (3). Under a mild condition on the correlation between the vision and language modalities, we can prove Prop. 3.1. The details of the derivation are provided in the Appendix.

Prop. 3.1 shows that spurious correlations in the vision modality can propagate to the joint distribution of the visual and text data, posing a great challenge to the alignment between vision and language modalities. Considering that the predominant CLIP objective [27] for training vision encoders may overlook crucial visual details in images [26], a vision encoder in an MLLM may exploit the spurious correlations in the vision modality and develop spurious biases, which can be propagated to the MLLM affecting its alignment between visual and text tokens.

3.3 How to Reveal Multimodal Spurious Bias

In principle, to reveal spurious biases in models, we aim to create a set of test data with spurious correlations different from those in the training data. For example, in the vision modality, a common approach [21] is to curate a test set so that the spurious correlation between a spurious attribute $\mathbf{z}$ and a target c in it becomes $p_{\text{test}}(\mathbf{z}|c, x_{\text{core}}) = p_{\text{test}}(\mathbf{z}|x_{\text{core}})$ [50]. This shows a significant distribution shift from the training distributions, where $p_{\text{train}}(\mathbf{z}|c, x_{\text{core}}) \gg p_{\text{train}}(\mathbf{z}|x_{\text{core}})$, such that the strong correlation between a spurious attribute $\mathbf{z}$ and a target c that holds in the training data no longer holds in the test data.

However, obtaining such a test set requires knowing $\mathbf{z}$ a priori and controlling over groups of test samples, which is challenging in the multimodal scenario where a massive amount of multimodal data is available. Therefore, we propose an instance-level method that creates distribution shifts in both the vision and language modalities. We first select challenging images aiming to approximate the relation $p_{\text{test}}(\mathbf{z}|c, x_{\text{core}}) \approx p_{\text{test}}(\mathbf{z}|x_{\text{core}})$ in the vision modality. Based on these images, we create individual VQA tasks with *generic* derived textual attributes. In this way, c will have reduced reliance on $\mathbf{z}$ given the core features x_{core} and y_{core}, and we can create a shifted test data distribution by bringing $p_{\text{test}}(\mathbf{z}|x_{\text{core}}, y_{\text{core}}, c)$ closer to $p_{\text{test}}(\mathbf{z}|x_{\text{core}}, y_{\text{core}})$. We realize this idea with a comprehensive VQA benchmark in the following section.

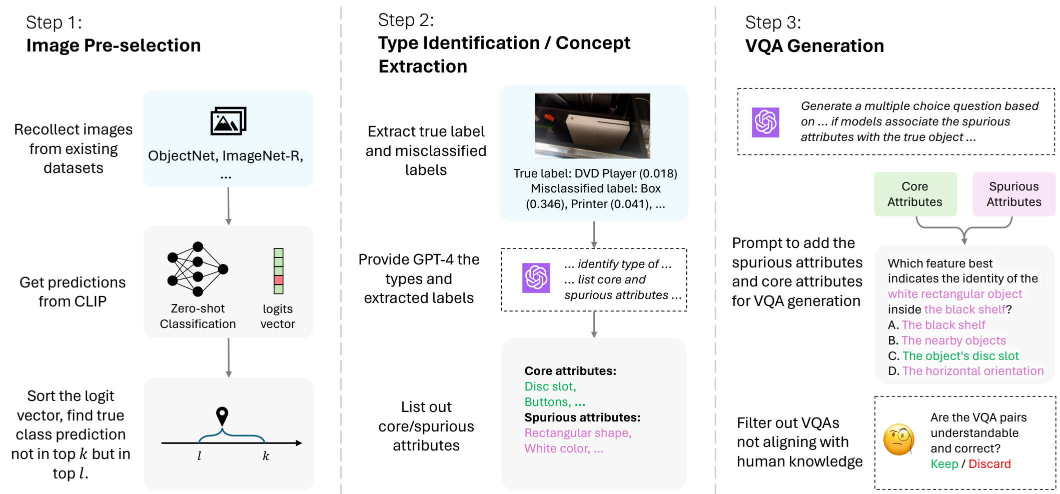


Figure 3: Construction of the MM-SpuBench. **Left:** Pre-select images where CLIP's true class prediction is not in top k but is in top l. **Middle:** Use GPT-4V to identify spurious correlations and lists core/spurious attributes. **Right:** Generate multiple-choice questions based on the spurious bias type and core/spurious attributes.

4 The Multimodal Spurious Benchmark (MM-SpuBench)

4.1 Types of Spurious Correlations

We first define the types of spurious correlations in Table 1 to comprehensively cover the spurious correlations in real-world data. Note that there exist other research works [29] with similar definitions, such as shape bias and texture bias. In our work, we are interested in spurious correlations between attributes and the core object in the images rather than focusing on a single perspective. In the next section, we demonstrate the three steps for the construction of MM-SpuBench as shown in Fig. 3.

4.2 Construction of MM-SpuBench

Image pre-selection. We pre-select images with their class labels from various image classification datasets to ensure the diversity of our benchmark. ObjectNet [51] serves as our primary image source due to its numerous observable spurious biases. To supplement this dataset, we also collect data from other domain generalization datasets, including ImageNet-R (rendition)[52], ImageNet-Sketch[53], ImageNet-A [54], and ImageNet-C [55]. These datasets are derived from the superset ImageNet-Hard [56] for the ease of implementation. They add categories of spurious biases not present in ObjectNet, such as texture/noise and relative size. We choose existing datasets rather than using image generation techniques [30] to ensure our benchmark reflects realistic spurious biases found in the real world, avoiding additional biases that could render the benchmark results unrepresentative. The licenses of these datasets are provided in the Appendix.

To select image data without spurious correlations, we use the most commonly employed vision encoder in current open-source MLLMs, `CLIP-ViT-L/14@336px` [27], for zero-shot classification. We utilize the logit vectors from the classification output to find samples where CLIP's true class prediction is not in the top-k but is in the top-l, where k and l are hyperparameters to control whether the misclassification is due to spurious biases rather than potential annotation errors/no enough visual cues. For each image, we record the ground truth class and top misclassified classes. The pair of ground truth labels and misclassified labels can indicate the spurious correlations the vision encoder relies on during the training process, guiding the design of our benchmark. For image pre-selection, we deploy $k = 3, l = 20$ for ObjectNet and $k = 3, l = 40$ for ImageNet-Hard. With this selection strategy, we curate a dataset with a total of $10,773$ image samples. To retrieve a smaller VQA subset, we deploy $k = 5, l = 10$ for ObjectNet and $k = 3, l = 40$ for ImageNet-Hard with a total of $2,400$ image/labels samples.

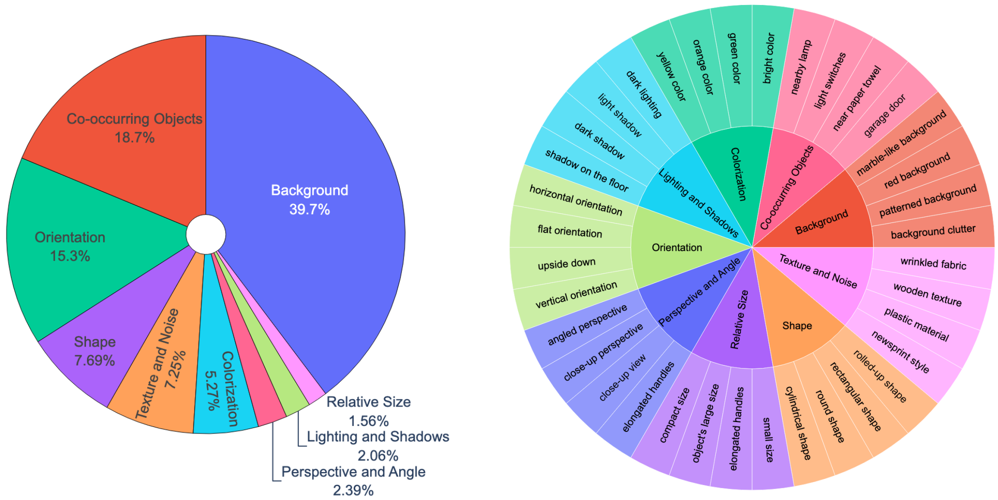


Figure 4: Overview of the MM-SpuBench. (a) Distribution of spurious correlation types. (b) Selected attributes within each spurious correlation type. Note that there might be shared attributes in different types since each image may contain at most two types of spurious correlations.

Type identification and attribute extraction. We leverage images along with their corresponding ground truth and misclassified labels to identify the types of spurious biases and understand their underlying causes. To achieve this, we employ GPT-4 as a concept generator, utilizing the chain-of-thought strategy to extract detailed and useful concept-based information from both the ground truth and misclassified labels. For each image, we generate two types of attributes: core attributes and spurious attributes. Core attributes are generated based on the ground truth label. They describe the intrinsic properties of the core object within the image, such as shape, color, and specific distinguishable features inherent to the object. Spurious attributes are generated based on the misclassified labels. These attributes do not have direct correlations with the primary object but still influence the model's inference process, leading to spurious biases. To maintain a balanced and fair evaluation in our VQA benchmark, we limit the number of both core and spurious attributes to 5 per image, ensuring consistent evaluation and fair comparison across the dataset. Then we use the derived attributes together with the image to let the GPT-4V model to figure out the types of spurious biases (at most 2) in the image.

Visual Question Answering (VQA) generation. We build upon the identified core and spurious attributes to create VQA pairs that evaluate a model's robustness to multimodal spurious biases. Using the provided images and their core and spurious attributes, we design prompts that integrate spurious attributes into the question and use core attributes to generate one correct option referring to the main object. The GPT-4V model utilizes this information to produce multiple-choice questions that test whether a model can identify the true label based on core attributes while being misled by spurious ones. These questions avoid direct references to the core attributes or true label, instead describing the core object using spurious attributes and its spatial position. Each question may randomly incorporate the derived core and spurious attributes from the previous step, with only one correct answer and three misleading options. After generation, we filter out the VQAs that do not align with human knowledge. The overview of the MM-SpuBench is shown in Fig. 4. Panel (a) illustrates the distribution of spurious correlation types, while panel (b) displays the selected attributes within each type.

5 Experiments

5.1 Baselines

For close-sourced MLLMs, we selected Gemini 1.5 Pro [12], GPT-4V/GPT-4o [10] and the Claude 3 family models (Haiku, Sonnet, Opus)[11], which are the mainstream MLLMs in the AI community. The input for these models consists of a system prompt and a format prompt that describes the task and the question with four options, while the expected output includes the predicted option and an

MLLM	Method	MM-SpuBench									Average
		BG	TN	CO	RS	Col.	Ori.	LS	PA	Sha.	
Gemini 1.5 Pro [12]	zero-shot	60.12	55.35	63.46	50.28	53.25	62.86	60.38	48.15	54.79	58.06
	chain-of-thought	50.26	50.55	48.88	45.27	42.33	42.31	47.92	33.33	38.56	47.93
Claude 3 Haiku [11]	zero-shot	55.45	53.77	57.12	40.22	45.12	55.71	47.17	37.04	39.85	52.06
	chain-of-thought	58.12	59.59	59.81	51.40	43.09	58.57	52.83	41.98	32.18	54.76
Claude 3 Sonnet [11]	zero-shot	78.06	76.57	81.35	61.45	65.85	81.43	75.47	59.26	60.92	74.82
	chain-of-thought	76.91	75.43	77.84	59.22	60.98	72.86	77.36	53.09	51.72	72.08
Claude 3 Opus [11]	zero-shot	80.43	76.10	83.65	64.80	66.67	82.86	83.02	70.37	67.82	77.18
	chain-of-thought	85.54	83.94	87.16	67.33	66.67	79.66	82.00	69.35	69.14	81.68
GPT-4V [10]	zero-shot	83.58	82.39	85.33	67.60	72.65	81.43	84.91	70.37	73.36	80.90
	chain-of-thought	86.13	84.59	88.08	74.30	73.47	81.43	83.02	77.78	72.59	83.22
GPT-4o [10]	zero-shot	80.64	81.13	83.85	60.89	69.39	80.00	83.02	65.43	67.18	77.97
	chain-of-thought	80.53	76.50	83.65	62.36	69.39	85.71	79.25	69.14	63.95	77.05

Table 2: Benchmark results of different close-sourced MLLMs on MM-SpuBench. All numbers are accuracy in percentages. Higher accuracy is represented by lighter background color.

MLLM	LLM Backbone	MM-SpuBench									Average
		BG	TN	CO	RS	Col.	Ori.	LS	PA	Sha.	
InstructBLIP [13]	Vicuna-7B [57]	22.54	23.43	21.15	26.82	18.29	21.43	33.96	20.99	23.75	22.59
MiniGPT4-v2 [14]	Llama-2-7B [58]	24.83	24.37	25.00	29.61	26.02	21.43	30.19	25.93	24.52	25.12
LLaVA-v1.5 [15]	Llama-2-7B [58]	34.17	32.86	32.69	37.43	35.37	28.57	45.28	24.69	32.57	33.71
LLaVA-v1.6 [15]	Mistral-7B [59]	32.39	32.70	31.73	34.64	35.37	28.57	43.40	25.93	32.18	32.59
Qwen-VL [60]	Qwen-7B [61]	28.98	31.76	28.08	25.70	29.27	41.43	28.30	27.16	20.69	28.82
CogVLM-v2 [62]	Vicuna-7B [57]	59.52	60.71	62.46	35.87	56.25	48.00	60.00	46.88	46.93	57.44
LLaVA-v1.5 [15]	Llama-2-13B [58]	52.71	50.63	54.62	35.75	46.75	42.86	60.38	37.04	42.53	50.06
LLaVA-v1.6 [15]	Vicuna-13B [57]	51.96	50.16	54.04	37.99	47.97	41.43	64.15	39.51	45.21	50.12
Intern-VL [63]	InternLM2-20B [64]	80.43	77.83	81.92	59.22	68.29	72.86	79.25	70.37	70.50	77.00
LLaVA-v1.6 [15]	Hermes-Yi-34B [65]	78.65	76.73	80.58	55.87	62.20	68.57	81.13	64.20	65.90	74.71
GPT-4V [10]	-	83.58	82.39	85.33	67.60	72.65	81.43	84.91	70.37	73.36	80.90
GPT-4o [10]	-	80.64	81.13	83.85	60.89	69.39	80.00	83.02	65.43	67.18	77.97

Table 3: Zero-shot results of different open-sourced MLLMs on MM-SpuBench. All numbers are accuracy in percentages. Higher accuracy is represented by lighter background color.

explanation to help us understand why some questions are not answered correctly. For open-sourced MLLMs, following previous works [44, 26], we select current state-of-the-art models that excel in general VQA tasks, including InstructBLIP [13], MiniGPT-4 [14], LLaVA [15], and Qwen-VL [60], with variants of LLM backbones. The input for these models is a system prompt that describes the task and the question with four options, with the expected output being only the option, since smaller models lack the reasoning ability to show what leads them to give a particular answer.

5.2 Implementation Details

To ensure a fair comparison, we shuffle the choices in each question to avoid option biases for each MLLM model during implementation. For the open-source models, all inference code is executed on four NVIDIA A100 GPUs. Each experiment is conducted with three different random seeds, and the reported value is the average of these runs. For all open-sourced models, we set the temperature to 0 to ensure reproducibility. Due to variations in the capabilities of each model, we design separate prompts to ensure the models can output the choices from our benchmark. To assess the performance on MM-SpuBench, we use accuracy as the metric to determine MLLMs' robustness to spurious biases as follows: Acc $= C/T$, where C denotes the number of image-text pairs correctly answered by the model, and T represents the total number of image-text pairs. To validate the usefulness of the concept information in our benchmark, we conduct experiments on two inference strategies in the open-sourced models, zero-shot and chain-of-thought. For zero-shot inference, we only input the system prompt with the question/choices in the language modality. For chain-of-thought inference, we first give the spurious bias type to the models, and let the MLLMs reason which attributes are core or spurious from the vision modality.

5.3 Main Results

Overall performance on MM-SpuBench. Based on the results in Table 2 and Table 3, we observe that MLLMs exhibit varying degrees of spurious bias. Generally, the close-sourced models outperform the open-source models. When examining the performance across different types of spurious bias, we found significant variations in the MLLMs' ability to address each type. They perform better in the `BG` and `Col.` types, while their performance is notably subpar in the `RS` and `PA` types. This suggests that while certain spurious correlations, such as backgrounds, are more easily perceived by these models, others, like relative size and perspective, present greater challenges. A potential explanation for the performance gap is that the models might heavily rely on specific visual cues that are more consistent in the training data while struggling with complex visual relationships.

Modality alignment plays a vital role. A deeper analysis of the open-source models reveals that models with larger sizes are more resilient to spurious biases. Models such as InstructBLIP [13], MiniGPT-4 [14], and LLaVA [15] employ similar modality alignment techniques: mapping the output from the vision encoder to the same space as the LLM input tokens. The poor results in their smaller models align with the proposition we derived in Sec. 3. Additionally, we note that better modality alignment techniques can significantly improve robustness to spurious biases. For example, InternVL [63] adopts an improved approach to scaling up the vision encoder and aligning it with the LLM. Consequently, even with 26B parameter size, it achieves competitive results compared to the LLaVA-1.6 model with 34B parameter size in our benchmark. This suggests that advanced modality alignment techniques, along with larger model sizes, contribute to better addressing spurious biases in multimodal learning.

Concept information helps mitigate spurious biases. In Table 2, we employ a simple chain-of-thought technique for the close-sourced models, providing the spurious bias type to the model and allowing it to reason. We avoid using the core/spurious attributes directly, as the VQAs are constructed based on these attributes, which could lead to information leakage. The results show that the most advanced models (GPT-4V and Claude 3 Opus) demonstrate larger performance improvement. This suggests that integrating concept information and strong reasoning capabilities can effectively mitigate spurious biases, enhancing the models' overall robustness and accuracy. Future work may explore the reasoning strategies together with the core/spurious attributes to learn better multimodal representation and mitigate the multimodal spurious biases.

6 Conclusion

In this work, we investigate the prevalence and impact of spurious biases in multimodal large language models (MLLMs). Our findings reveal that current MLLMs, particularly those relying solely on existing Vision Foundation Models (VFMs) for visual understanding, often fail to achieve effective alignment between visual and language components in multimodal tasks, indicating that these models are not yet fully equipped for robust vision-language integration. To address this gap, we introduce MM-SpuBench, a comprehensive benchmark designed to evaluate the robustness of MLLMs to spurious biases. This benchmark systematically assesses how well these models distinguish between core and spurious features, providing a detailed framework for understanding and quantifying spurious biases. Our results indicate that both open-source and close-sourced MLLMs continue to rely on spurious correlations to varying degrees, underscoring the need for improved multimodal alignment techniques and more robust architectures. We hope that MM-SpuBench will drive further research in this field, leading to the development of more robust and reliable multimodal models.

References

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

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[5] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[6] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

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

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

[11] Anthropic. Claude 3 family. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: 2024-05-27.

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

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

[16] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

[19] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[21] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2019.

[22] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, pages 6781–6792. PMLR, 2021.

[23] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *ICLR*, 2022.

[24] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.

[25] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Learning robust classifiers with self-guided spurious correlation mitigation. In *The 33rd International Joint Conference on Artificial Intelligence*, 2024.

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

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

[28] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

[29] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are vision language models texture or shape biased and can we steer them? *arXiv preprint arXiv:2403.09193*, 2024.

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

[31] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019.

[32] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

[33] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 126–135, 2018.

[34] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

[35] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

[36] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[37] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022. Expert Certification.

[38] Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *ICML*, pages 66–88. PMLR, 2022.

[39] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023.

[40] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

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

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

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

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

[45] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

[46] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

[47] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR, 2024.

[48] Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. *arXiv preprint arXiv:2402.14899*, 2024.

[49] Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multi-modal contrastive learning to distribution shift. In *The Twelfth International Conference on Learning Representations*, 2024.

[50] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: a closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39584–39622, 2023.

[51] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019.

[52] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[53] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[54] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021.

[55] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[56] Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. Zoom is what you need: An empirical study of the power of zoom and spatial biases in image classification. *arXiv preprint arXiv:2304.05538*, 2023.

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*, 2023.

[59] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv*, October 2023.

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

[61] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen Technical Report. *arXiv*, September 2023.

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

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

[64] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. InternLM2 Technical Report. *arXiv*, March 2024.

[65] 01 AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open Foundation Models by 01.AI. *arXiv*, March 2024.

Supplementary Materials

Table of Contents

A Broader Impacts 15

B Limitations 16

C More Information on the Data 16
- C.1 Public Availability 16
- C.2 Data Sources and Licenses 16

D Derivation of Proposition 3.1 17

E More Experimental Details 18
- E.1 Zero-shot Classification 18
- E.2 Prompt Engineering 18

F Data Instances in MM-SpuBench 20

A Broader Impacts

Social Impacts. We summarize the following aspects of social impacts of our work.

1. ***Enhanced Model Robustness:*** By identifying and addressing spurious biases in multimodal large language models (MLLMs), our work can lead to the development of more robust AI models. These models will perform more reliably across diverse real-world scenarios, benefiting applications in healthcare, autonomous driving, and education where robustness is critical.
2. ***Transparency and Trustworthiness:*** We provide a well-designed framework to evaluate and understand spurious biases in MLLMs, which can increase transparency in AI systems. The transparency is crucial for gaining public trust and for ensuring that AI systems are held accountable for their decisions, especially for MLLMs.

Technical Impacts. We summarize the technical impacts as follows:

1. ***Improving Multimodal Learning:*** Our work pushes the boundaries of multimodal learning by defining an overlooked robustness issue, multimodal spurious bias, and providing a comprehensive benchmark for evaluating and improving the alignment between visual and language modalities. This can lead to advancements in fields like visual question answering (VQA) and multimodal reasoning.
2. ***Benchmark using Concept Information***: MM-SpuBench sets a new standard for evaluating spurious biases in multimodal models using the core/spurious attribute information. This could inspire future research directions and benchmarking practices, ensuring that new models are evaluated in a diverse perspective for the robustness to spurious correlations.
3. ***Inspiration on Better Model Design:*** Insights gained from our benchmark can inform the design of future MLLMs, leading to architectures that are inherently more robust to spurious biases. This could result in better performance in real-world applications where robustness is essential.

Potential Negative Impacts.

1. ***Over-reliance on Benchmarks:*** There's a risk that focusing on specific benchmarks might lead researchers to optimize models solely for benchmark performance rather than general robustness. This could result in models that perform well on MM-SpuBench but still exhibit other robustness issues in untested scenarios.

B Limitations

1. ***Dynamic Nature of Spurious Biases:*** Spurious biases in AI models can evolve over time, especially as models are exposed to new data and the biases are based on human perception. MM-SpuBench provides a snapshot based on current understandings of spurious biases, but it may need updates to remain relevant with the development of new models and data.
2. ***Granularity of Categorization:*** While we have comprehensively categorized spurious biases into 9 distinct types, this categorization is coarse considering the wide range of such biases. There is a possibility that within each of these categories, there exist more subtle, fine-grained spurious biases that are not explicitly accounted for. This limitation means that our benchmark might not fully capture the complexity and nuances of all spurious correlations that can occur in multimodal data. Future work could involve developing more granular classifications and corresponding evaluation metrics to provide a deeper understanding of these biases.

C More Information on the Data

C.1 Public Availability

We have made the MM-SpuBench dataset publicly available at `https://huggingface.co/datasets/mmbench/MM-SpuBench`.

C.2 Data Sources and Licenses

ObjectNet ObjectNet is a vision dataset with 50,000 images, specifically designed to test object recognition systems under varied conditions. It includes 313 object classes and controls for rotation, background, and viewpoint. This dataset reveals significant performance drops, showing real-world challenges and difficulties in transfer learning. ObjectNet is free for both research and commercial use, with the following restrictions:

1. ObjectNet cannot be used to tune the parameters of any model.
2. Individual images from ObjectNet must include their 1-pixel red border when posted online.

The license details can be found at `https://objectnet.dev/download.html`.

ImageNet ImageNet is a comprehensive visual database used for visual object recognition research, containing millions of labeled images across thousands of categories. It serves as a key benchmark for evaluating computer vision algorithms and advancing deep learning research. The license details for ImageNet are available at `https://www.image-net.org/download.php`.

ImageNet-R(endition) ImageNet-R is a subset of ImageNet-1K classes with art, cartoons, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game renditions of ImageNet classes. It contains renditions of 200 ImageNet classes, with a total of 30,000 images. This dataset is available under the MIT License at `https://github.com/hendrycks/imagenet-r`.

ImageNet-A ImageNet-A contains real-world, unmodified examples that cause significant performance degradation in machine learning models. The dataset is available under the MIT License at `https://github.com/hendrycks/natural-adv-examples`.

ImageNet-C The ImageNet-C dataset consists of 15 types of corruptions applied to ImageNet validation images, categorized into noise, blur, weather, and digital, each with five severity levels, resulting in 75 distinct corruptions. This dataset is available under the Apache License 2.0 at `https://github.com/hendrycks/robustness`.

ImageNet-Sketch ImageNet-Sketch includes 50,000 images, with 50 sketches for each of the 1,000 ImageNet classes. These images are gathered using Google Image searches with the query "sketch of CLASS" in black and white. The dataset is under the MIT License at `https://github.com/HaohanWang/ImageNet-Sketch`.

ImageNet-ReaL ImageNet-ReaL offers "Re-Assessed" (ReaL) labels with multi-label and more accurate annotations from the "Are we done with ImageNet" paper. The dataset is available under the Apache License 2.0 at `https://github.com/google-research/reassessed-imagenet`.

ImageNet-Hard ImageNet-Hard is a new benchmark featuring challenging images curated from various ImageNet validation datasets. It challenges state-of-the-art vision models as simply zooming in often fails to improve classification accuracy. The dataset is available under the MIT License at `https://github.com/taesiri/ZoomIsAllYouNeed`.

D Derivation of Proposition 3.1

In the vision encoder and the LLM from the MLLM, we represent the training data probability with one spurious attribute $\mathbf{z}$, the core object c, and the core features $x_{\text{core}}, y_{\text{core}}$ as follows.

$$\textbf{In Vision Encoder: } p(\mathbf{z}|c, x_{\text{core}}) \gg p(\mathbf{z}|x_{\text{core}}) \tag{4}$$

$$\textbf{In LLM: } p(\mathbf{z}|c, y_{\text{core}}) \approx p(\mathbf{z}|y_{\text{core}}) \tag{5}$$

The conditional probability on the spurious attribute $\mathbf{z}$, given the core features and object, is:

$$\begin{aligned}
p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}, c) &= \frac{p(x_{\text{core}}, y_{\text{core}}|\mathbf{z}, c)p(\mathbf{z}|c)}{p(x_{\text{core}}, y_{\text{core}}|c)} && (6)\\
&= \frac{p(x_{\text{core}}|\mathbf{z}, c)p(y_{\text{core}}|\mathbf{z}, c)p(\mathbf{z}|c)}{p(x_{\text{core}}|c)p(y_{\text{core}}|c)} && (7)\\
&= \frac{p(\mathbf{z}|x_{\text{core}}, c)p(\mathbf{z}|y_{\text{core}}, c)p(\mathbf{z}|c)}{p(\mathbf{z}|c)p(\mathbf{z}|c)} && (8)\\
&= \frac{p(\mathbf{z}|x_{\text{core}}, c)p(\mathbf{z}|y_{\text{core}}, c)}{p(\mathbf{z})} && (9)
\end{aligned}$$

Without considering the core object c, the conditional probability on the spurious attribute $\mathbf{z}$ is:

$$\begin{aligned}
p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}) &= \frac{p(x_{\text{core}}, y_{\text{core}}|\mathbf{z})p(\mathbf{z})}{p(x_{\text{core}}, y_{\text{core}})} && (10)\\
&= \frac{p(x_{\text{core}}|\mathbf{z})p(y_{\text{core}}|\mathbf{z})p(\mathbf{z})}{p(x_{\text{core}}, y_{\text{core}})} && (11)\\
&= \frac{p(\mathbf{z}|x_{\text{core}})p(\mathbf{z}|y_{\text{core}})p(\mathbf{z})p(x_{\text{core}})p(y_{\text{core}})}{p(\mathbf{z})p(\mathbf{z})p(x_{\text{core}}, y_{\text{core}})} && (12)\\
&= \frac{p(\mathbf{z}|x_{\text{core}})p(\mathbf{z}|y_{\text{core}})}{p(\mathbf{z})} \cdot \frac{p(x_{\text{core}})p(y_{\text{core}})}{p(x_{\text{core}}, y_{\text{core}})} && (13)\\
&\approx \frac{p(\mathbf{z}|x_{\text{core}})p(\mathbf{z}|y_{\text{core}})}{p(\mathbf{z})} && (14)
\end{aligned}$$

By (2) and (3), we can get inequality (1) in the multimodal case.

$$\begin{aligned}
p(\mathbf{z}|x_{\text{core}}, c)p(\mathbf{z}|y_{\text{core}}, c) &\gg p(\mathbf{z}|x_{\text{core}})p(\mathbf{z}|y_{\text{core}}) && (15)\\
\frac{p(\mathbf{z}|x_{\text{core}}, c)p(\mathbf{z}|y_{\text{core}}, c)}{p(\mathbf{z})} &\gg \frac{p(\mathbf{z}|x_{\text{core}})p(\mathbf{z}|y_{\text{core}})}{p(\mathbf{z})} && (16)\\
p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}, c) &\gg p(\mathbf{z}|x_{\text{core}}, y_{\text{core}}) && (17)
\end{aligned}$$

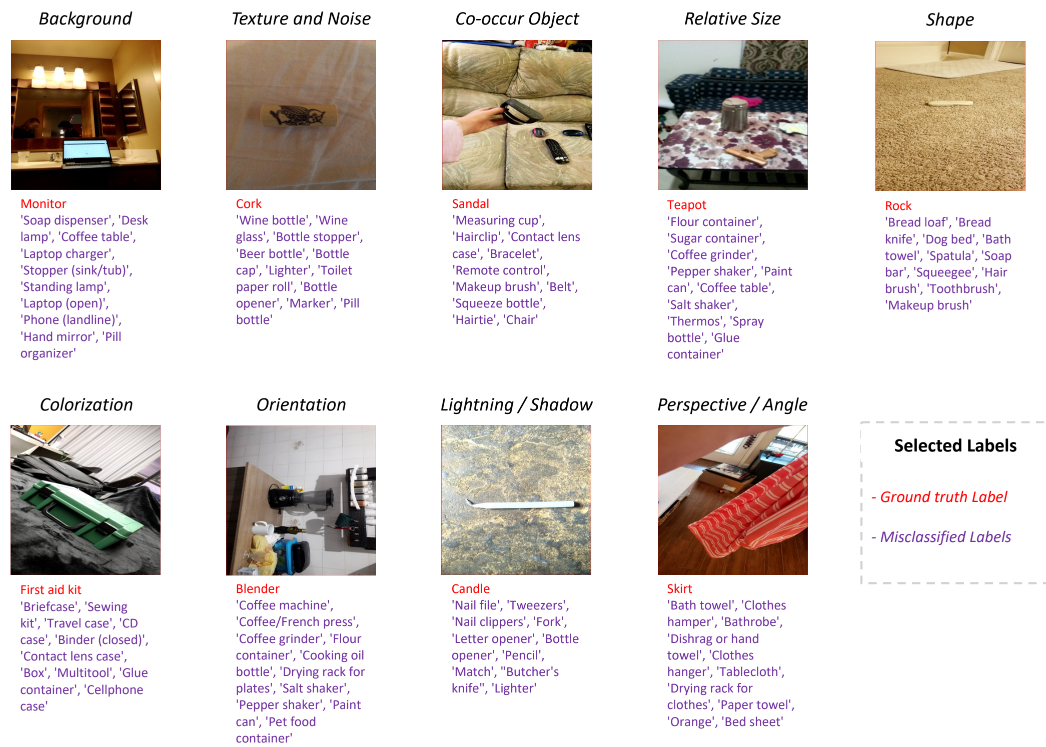


Figure 5: Examples of ground truth and misclassified labels, illustrating spurious correlations.

E More Experimental Details

E.1 Zero-shot Classification

In Figure 5, we see that misclassified labels often include spurious information unrelated to the core object but present in the image or contextually linked to the object (e.g., Cork vs. Wine Bottle). This validates our zero-shot classification approach for extracting potential spurious attributes. Zero-shot classification lets models predict without prior exposure to specific examples. By examining misclassified labels, we can identify the spurious attributes based on the main object (ground truth label). For instance, 'Cork' is often misclassified as 'Wine bottle' or 'Wine glass,' showing the model's reliance on the contextual cues rather than intrinsic features. Using `CLIP-ViT-L/14@336px`, we identified spurious correlations hurt model performance. For example, 'Monitor' was confused with 'Soap dispenser' and 'Desk lamp' due to background information, while 'Sandal' was misclassified as 'Measuring cup' or 'Hairclip' due to shape and orientation.

E.2 Prompt Engineering

To ensure the effective generation and evaluation of questions for analyzing spurious correlations in images, we design four prompts for the MLLMs. In Table 4, we created a system message prompt to guide the assistant in identifying spurious correlations, and deriving core and spurious attributes. We then formulate multiple-choice questions that test a model's ability to distinguish these attributes. This ensures challenging and accurately reflective questions of spurious biases. For zero-shot evaluation on open-sourced models in Table 5, we only ask the model to select the best answer. For zero-shot evaluation on close-sourced models in Table 6, we designed a straightforward prompt instructing the assistant to answer questions based on the provided image and four answer options, with a focus on selecting the best answer and providing a brief explanation. Additionally, we used the chain-of-thought prompt to enhance the assistant's reasoning capability by considering the type of spurious correlation provided in the benchmark and thinking step-by-step before choosing the best answer in Table 7.

System Message
You are a helpful assistant that analyze images.
I will give you image... true label: ... misclassified labels: ...
Spurious correlations are brittle associations learned by the models between non-essential spurious attributes of inputs and the corresponding core learning attributes in the training dataset.
Based on the provided information 1. Figure out what kind of spurious correlations is performing in the given image. 2. Based on the true label and the image, generate what are the core attributes of this true object label. Based on the misclassified labels and the image, generate what are the spurious attributes that are causing the misclassification. 3. Generate a multiple choice question based on the analysis to test the capability of a model whether it can identify the true label based on the spurious attributes. Among the choices, there should be only one correct answer related to the core attributes. Make the other choices as misleading as possible so that the model may fail on it. 4. Do not provide the true label or the core attributes of the main object in the question. Only use its visible spurious attributes or its spatial position in the image to refer to the object.
The max words for each attribute is {max_words_per_attribute}.
The max number of core attributes is {num_core_attributes}.
The max number of spurious attributes is {num_spurious_attributes}.
For the generated multiple choice questions, the number of correct options is 1, and the number of wrong options is {num_wrong_options}.
You should only respond in the format as described below:

Response Format
Explanation: The explanation of the attributes.
Core Attributes: The core attributes of the main object, must be visible in the image.
Spurious Attributes: The spurious attributes in the image.
Spurious Correlation Type: Should be from the 9 possible categories: Background; Texture and Noise; Co-occurring Objects; Relative Size; Colorization; Orientation; Lighting and Shadows; Perspective and Angle; Shape. Two at most.
Questions: The question to ask about the image.
Choices: The choices for the question, indexed by a single letter.
Answer: The index of the correct answer, as a single letter.

Table 4: System message and response format for the QA generation with GPT-4V.

System Message
You are a helpful assistant that can answer question for an image. I will provide you 4 options.
Response Format
Choice: A single character from A, B, C, D.

Table 5: System message and response format for the zero-shot evaluation on open-sourced models.

System Message
You are a helpful assistant that can answer question for an image. I will provide you 4 options.
Response Format
Explanation: Explanation text in one sentence.
Choice: A single character from A, B, C, D.

Table 6: System message and response format for the zero-shot evaluation on close-sourced models.

System Message
You are a helpful assistant that can answer question based on the image. I will provide you 4 options and {the type of spurious correlations} that exist in the image. Think step by step and then choose the best answer.
Response Format
Explanation: Explanation text in one sentence.
Choice: A single character from A, B, C, D.

Table 7: System message and response format for the chain-of-thought evaluation on close-sourced models.

F Data Instances in MM-SpuBench

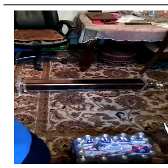
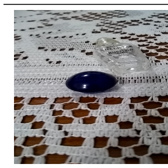
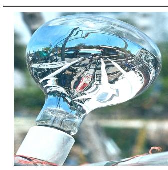
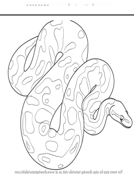
Image	Question	Choices	Answer	Type
	Which feature best indicates the identity of the object held in hand on the countertop?	A. The countertop B. The lighting C. The metallic parts D. The multiple tools	D	Texture and Noise Shape
	Which feature best indicates the identity of the object that is lying horizontally on the bathroom sink?	A. The object's control panel B. The bathroom mirror C. The toothbrushes D. The soap dispenser	A	Background Orientation
	Which feature best indicates the identity of the object that is lying horizontally on the floor?	A. Office chair B. Water bottles C. Table D. Light bulb socket	D	Orientation Co-occurring Objects
	Which feature best indicates the identity of the small cylindrical object next to the small bottle?	A. Its small size B. Its cylindrical shape C. The lace tablecloth D. The nearby bottle	B	Co-occurring Objects Relative Size
	Which feature best indicates the identity of the object being held?	A. The glitter bottle B. The object's circular shape C. The shadow D. The surface texture	B	Background Lighting and Shadows
	Which feature best indicates the identity of the object that is upside down with two sticks on top?	A. Circular shapes B. The shark-like form C. The cinnamon sticks D. The curved handle	D	Orientation Shape
	Which feature best indicates the identity of the object with a distorted reflection?	A. The bulb shape B. The background buildings C. The outdoor setting D. The lightening	A	Background Perspective and Angle
	Which feature best indicates the identity of the animal with black and white patterns?	A. Small size B. Dog-like head shape C. Black and white pattern D. Elongated body	B	Colorization Orientation
	Which feature best indicates the identity of the coiled object in the drawing?	A. The patterns on the object B. The serpentine body C. The outline drawing D. The coiled shape	B	Shape Texture and Noise

Table 8: Data instances in MM-SpuBench. Images are cropped and resized to fit in the table. Red denotes the spurious attributes and green denotes the core attributes.