SPURLENS: FINDING SPURIOUS CORRELATIONS IN MULTIMODAL LLMS

Parsa Hosseini*, Sumit Nawathe*, Mazda Moayeri, Sriram Balasubramanian & Soheil Feizi
Department of Computer Science
University of Maryland
{phoseini,snawathe, mmoayeri, sriramb, sfeizi}@umd.edu

ABSTRACT

While multimodal large language models (MLLMs) exhibit remarkable capabilities in visual and textual understanding, they remain highly susceptible to spurious correlations. We propose *SpurLens*, a novel pipeline leveraging LLMs and open-set object detectors to identify spurious cues and measure their effect on MLLMs in an object detection scenario. Furthermore, we tested different prompting strategies to mitigate this issue, but none proved effective. These findings highlight the urgent need for robust solutions to address spurious correlations in MLLMs.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) Wang et al. (2024); Liu et al. (2024); Meta (2024); OpenAI (2024a) have seen rapid advances in recent years. These models leverage the powerful capabilities of Large Language Models (LLMs) OpenAI (2024b); Touvron et al. (2023) to process diverse modalities, such as images and text. They have demonstrated significant proficiency in tasks such as image perception, visual question answering, and instruction following. However, despite these advancements, MLLMs still face critical visual shortcomings Tong et al. (2024a;b). Our study focuses on their visual shortcomings associated with spurious correlations.

Spurious bias is the tendency to rely on spurious correlations between non-essential input attributes and target variables for predictions, leading to poor generalization and unreliable predictions when such spurious cues are absent. While extensive work has focused on identifying and mitigating spurious correlations in unimodal models Sagawa et al. (2019); Kirichenko et al. (2022); Moayeri et al. (2023); Noohdani et al. (2024), addressing this issue in MLLMs is still an emerging research area.

We provide empirical evidence that MLLMs often exploit superficial cues, such as associating the presence of a fire hydrant with a street scene, rather than recognizing the fire hydrant as an independent object (Figure 1). This over-reliance on spurious cues can cause the models to hallucinate objects or fail in the absence of these cues, raising concerns about their robustness and generalization.

In this work, we highlight the persistent issue of spurious correlations in MLLMs and introduce SpurLens, a pipeline to systematically detect these failures. Our pipeline automatically identifies potential spurious cues, verifies their presence through object detection models, and ranks images based on these cues. Through this approach, we provide a structured framework to analyze and quantify spurious correlations in MLLMs.

2 RELATED WORKS

Spurious Correlation: Spurious correlation have been extensively studied in the context of deep neural network classifiers (e.g., ViT Alexey (2020)), with various approaches proposed to detect and mitigate the issue Sagawa et al. (2019); Kirichenko et al. (2022); Noohdani et al. (2024). However, these studies primarily focus on single-modality settings (image-only tasks). Some research Wang

*Equal contribution.

Figure 1: An Example of Spurious Bias in Llama-3.2 (Meta (2024)). SpurLens identifies ”storm drain” as a spurious cue for detecting a fire hydrant.

et al.; Varma et al. (2024); Kim et al. (2023) has explored spurious correlations in CLIP Radford et al. (2021), framing the problem in terms of zero-shot performance across vision and language modalities. Ye et al. (2024) introduces a visual question answering (VQA) benchmark designed to evaluate MLLMs' reliance on spurious correlations using open-source image datasets. Zheng et al. (2024) also proposes a framework to quantify the varying degrees of robustness of Vision-Language Models (used as few-shot image classifiers) against spurious bias.

Ranking Images by spuriosity: Similar to HardImageNet (Moayeri et al. (2022)), one approach to detecting spurious bias in models is to rank images within their classes based on spuriosity, the degree to which common spurious cues are present, using deep neural features from an interpretable network, combined with human supervision Moayeri et al. (2023; 2022). We followed the same idea, using object detectors *without* human supervision.

Failures of Multimodal Systems: Some studies have introduced frameworks to automatically identify critical shortcomings of MLLMs Tong et al. (2024a;b). In Tong et al. (2024b), the authors highlight MLLMs' struggles with basic visual understanding, attributing these issues to weaknesses in CLIP-based vision encoders. Conversely, Tong et al. (2024a) focuses on the language modality.

3 SPURLENS

The spuriosity rankings in the HardImagenet (Moayeri et al., 2022) dataset are constructed using the spurious neural features from the Salient Imagenet dataset (Singla & Feizi, 2022). This process requires human supervision to identify which features are spurious to each class. While this method generalizes well, it results in few potential spurious features for each class. To study spurious correlations in MLLMs for more objects and datasets, we develop a pipeline to produce interpretable spurious rankings of images, which we can use to compute object detection performance accuracy gaps due to those spurious features. Our method uses open-set object detectors to identify ChatGPT-suggested spurious objects; after running experiments with the MLLM based on the rankings from our pipeline, we obtain spuriosity gaps for specific spurious objects to a given target object.

Suppose that, for a given MLLM $\mathcal{M}$ and target object t, we wish to determine what objects or image features are spurious to t. Further suppose that we have a large dataset $\{\mathcal{I}_j\}_{j=1}^N$ of images of target object t.

Proposing Spurious Features We use GPT-4 to generate list a list of objects or background elements that commonly appear in images of t. The number of features produced and their relation to t can be easily adjusted. We lemmatize each suggested object, remove duplicates, and remove any that share words with target object name t. We then use GPT-4 again to ensure that the proposed objects are truly spurious by asking the following Yes/No questions:

- "Can a {spurious feature} exist without a {target object}?" (Expected answer: "Yes".)

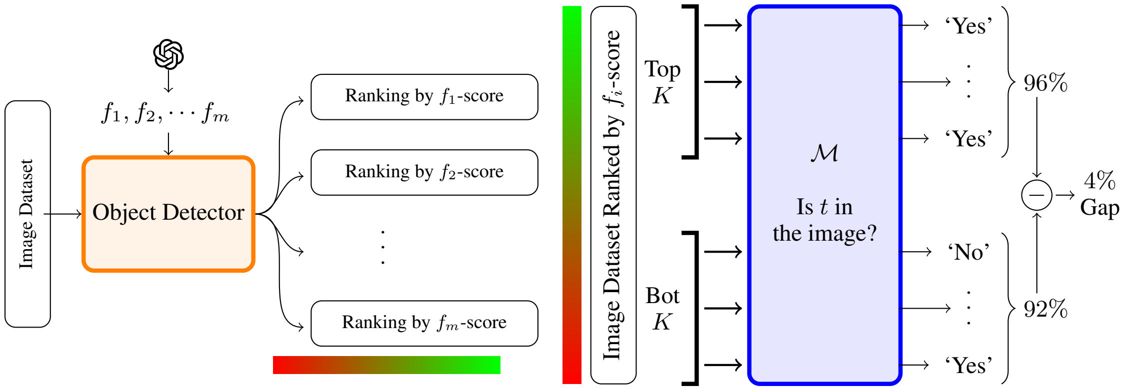

Figure 2: An overview of SpurLens. **Left**: proposing spurious objects, and ranking an image dataset by their presence. **Right**: computing spurious gaps for a given spurious feature and MLLM.

- "Is a {spurious feature} part of a {target object}?" (Expected answer: "No".)
- "Do all or almost all {spurious feature} have a {target object}?" (Expected answer: "No".)
- "Do all or almost all {target object} have a {spurious feature}?" (Expected answer: "No".)

The responses to these questions ensure that the propose features match the qualifications in the definition of a spurious feature. Works such as Leng et al. (2024); Zhou et al. (2023) identify spurious correlations through the frequent co-occurance of objects in MLLM-generated image captions. Our method avoids this computational cost, and the easily-modifiable prompt structure may suggest a more diverse pool of potential spurious objects. While our method may propose spurious objects not present in the dataset, this is generally not an issue with large datasets; furthermore, such cases are readily identifiable after following object detection step.

Identifying Spurious Objects To identify the presence of these spurious features f_i in the images $\mathcal{I}_j$, we use the OWLv2 open-set object detector Minderer et al. (2024). For each image, we query OWLv2 with all potential spurious features and obtain several triplets of consisting of a bounding box $b \in [0,1]^4$, label f_i, and confidence score $c \in [0,1]$. Let $\mathcal{O}(\mathcal{I}_i)$ denote the set of such triplets produced by OWLv2 for image $\mathcal{I}_i$. We define the f_i-score of $\mathcal{I}_j$ as

$$S(f_i, \mathcal{I}_j) = \max\left(\{0\} \cup \{c : (b, f_i, c) \in \mathcal{O}(\mathcal{I}_j)\}\right)$$

For each potential spurious feature f_i, we sort the images by f_i-score to obtain a ranking. (We randomize the order of 0-score images in each ranking before selection to avoid bias in ordering). Brief manual inspection can be performed at this stage to verify that the object detectors are reliable for the chosen spurious features by sampling images at the top and bottom of each ranking. In practice, the object detectors are fairly reliable for most potential spurious features.

Spuriosity Gaps For each ranking corresponding to feature f_i, let $\mathcal{U}^+_{t,f_i}, \mathcal{U}^-_{t,f_i} \subset \{\mathcal{I}_j\}_{j=1}^N$ be the images with the K-highest and K-lowest f_i-scores respectively. For each of these images, we apply the MLLM $\mathcal{M}$ paired with three prompts $p_k(t)$, $1 \leq k \leq 3$. Each prompt asks $\mathcal{M}$ if it sees the target object t in the image, and elicits a Yes/No response; we use three prompts to mitigate the bias due to word choice. We define the accuracy of $\mathcal{M}$ on image $\mathcal{I}$ depicting object t as

$$\text{Acc}(\mathcal{M}, \mathcal{I}, t) = \frac{1}{3}\sum_{k=1}^{3} \mathbf{1}\left(\mathcal{M}(\mathcal{I}, p_k(t)) = \text{“Yes”}\right)$$

We define the accuracy of $\mathcal{M}$ on spurious (by feature f_i) and non-spurious (by feature f_i) as

$$\text{Acc}_s = \frac{1}{K}\sum_{I \in \mathcal{U}^+_{t,f_i}} \text{Acc}(\mathcal{M}, \mathcal{I}, t) \qquad \text{Acc}_c = \frac{1}{K}\sum_{I \in \mathcal{U}^-_{t,f_i}} \text{Acc}(\mathcal{M}, \mathcal{I}, t)$$

Finally, we define the spurious gap $\text{Gap} = \text{Acc}_s - \text{Acc}_c$. That is, the Gap is the difference in object detection accuracy between images with f_i and images without f_i, as measured by the top-K and bottom-K images in the f_i-score ranking. A positive gap is evidence that f_j is truly spurious for t. After computing the Gap for all potential spurious features, we choose the one with the largest Gap.

Dataset	**HardImageNet**			**COCO**		
Model	Acc_s	Acc_c	Gap	Acc_s	Acc_c	Gap
Qwen2-VL	98.1%	92.3%	5.8%	95.3%	80.2%	15.1%
Llama-3.2	92.5%	80.2%	12.3%	84.6%	70.4%	14.3%
LLaVA-v1.6	90.7%	83.5%	7.2%	95.4%	80.0%	15.4%

Table 1: Accuracy across different datasets and models. The Gap indicates that, in the absence of spurious cues, all models struggle to detect the main object. We set $K = 50$ for all experiments, and the results are averaged class-wise.

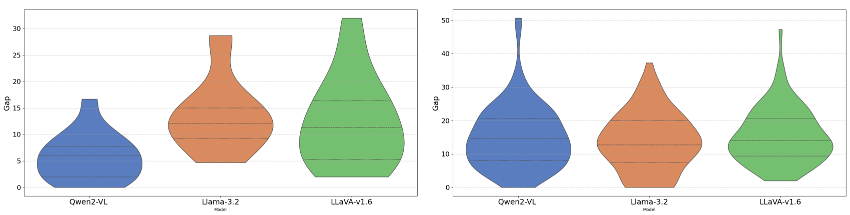

Figure 3: Comparison of Gap distributions over different classes from HardImageNet (**Left**) and COCO (**Right**) across models. The results show that spurious bias is very class dependent.

4 Results

Models For our experiments, we accessed three open-source MLLMs, Qwen2-VL-7B-Instruct Wang et al. (2024), Llama-3.2-11B-Vision-Instruct Meta (2024), and LLaVA-v1.6-mistral-7B Liu et al. (2023), all accessed through HuggingFace.

Evaluation Settings We utilized two open-source image datasets: HardImageNet Moayeri et al. (2022) and COCO Lin et al. (2014). We applied our pipeline to each dataset to generate spuriosity rankings for each class. Subsequently, we calculated the Accuracy (see the previous section) separately for the top 50 images (high spurious, Acc_s) and the bottom 50 images (low spurious, Acc_c) and then computed the **spuriosity gap**, the difference between the two. In all the experiments in this section, we used three different prompts to ask the model whether it detected the object. We averaged the results across different prompts and classes to compute the aggregated accuracy.

Results The results of applying SpurLens to HardImagenet and COCO are presented in Table 1, which shows the performance on spurious images, non-spurious images, and the performance Gap, averaged class-wise. We see that when spurious cues are absent, performance decreases across all models. It is surprising that the average performance of these large reputable models are quite low for non-spurious images. The distribution of Gaps across classes for each model and dataset are visualized in Figure 3. We see that the effect of spurious cues is highly class-dependent, but is significantly present in both datasets.

5 Conclusion

We have presented a scalable and easily adjustable method to identify and evaluate spurious correlations in MLLMs. We apply our system, SpurLens, on two large image datasets and found significant evidence that modern MLLMs are still reliant on spurious correlations.

References

Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2585–2595, 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Meta. Llama-3.2-11b-vision-instruct, 2024. URL `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`. Accessed: 2025-01-14.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. URL `https://arxiv.org/abs/2306.09683`.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues, June 2022.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriosity rankings: sorting data to measure and mitigate biases. *Advances in Neural Information Processing Systems*, 36:41572–41600, 2023.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27662–27671, 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024a. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed: 2025-01-14.

OpenAI. Gpt-4 technical report, 2024b. URL `https://arxiv.org/abs/2303.08774`.

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*, pp. 8748–8763. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning?, 2022. URL `https://arxiv.org/abs/2110.04301`.

Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

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

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. *arXiv preprint arXiv:2411.04097*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*, 2024.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Benchmarking spurious bias in few-shot image classifiers. In *European Conference on Computer Vision*, pp. 346–364. Springer, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.