SpurLens: Finding Spurious Correlations in Multimodal LLMs

Parsa Hosseini*, Sumit Nawathe*,
Mazda Moayeri, Sriram Balasubramanian, Soheil Feizi

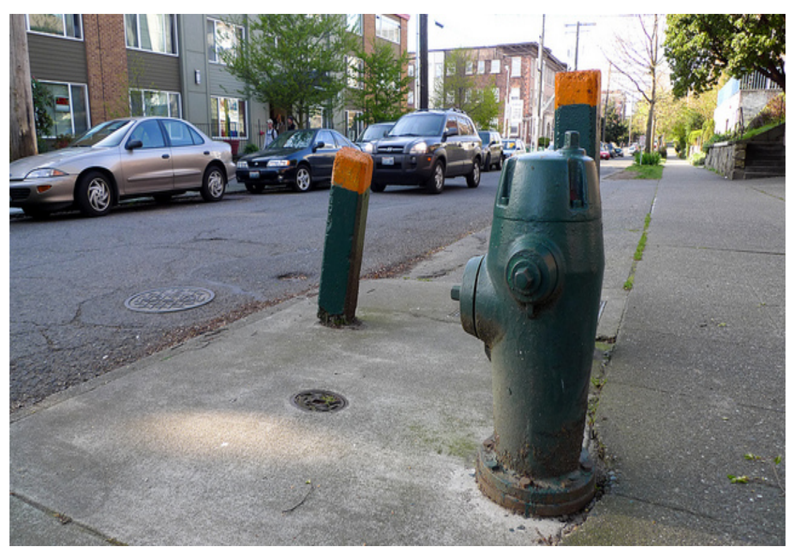
University of Maryland College Park

Problem Motivation

- Spurious bias is the tendency to rely on correlations between non-essential input attributes and target variables for prediction.
- This problem has been extensively studied in unimodal models for object detection/classification problems.
- We introduce *SpurLens*, a pipeline to automatically find and measure spurious correlations in MLLMs.
- We show that spurious biases persist in modern MLLMs for the object detection task, and identify spurious cues for COCO and HardImagenet classes.

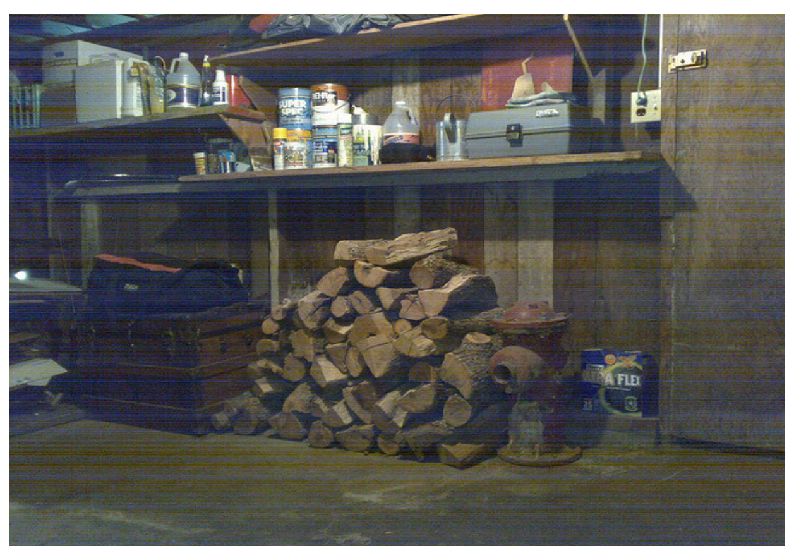
Is there a fire hydrant in the image?

With Spurious Cue



Yes, there is a fire hydrant in the image. It is located on the sidewalk, painted green and yellow, and appears to be old and rusty.

Without Spurious Cue



No, there is no fire hydrant in the image.

GPT-4o Failures Cases

The following are examples from COCO where GPT-4o incorrectly answers an object detection question because the image lacks a spurious feature (identified by SpurLens).

Object: Motorcycle
Spurious Cue Missing: Glove

Q: Is the person riding a motorcycle?

The person in the image is riding a scooter, not a motorcycle.

Object: Bus
Spurious Cue Missing: Curb

Q: Do you see a bus in the image?

I see a traffic jam with several yellow taxis ... but I can't specifically identify a bus.

Object: Vase
Spurious Cue Missing: Flower

Q: Is there a vase in the image? Explain.

In the image, there does not appear to be a vase. The image primarily features a plate of food with various meats ...

Object: Toaster
Spurious Cue Missing: Kitchen Counter

Q: Is the person wearing a backpack?

No, the person does not appear to be wearing a backpack.

Object: Bench
Spurious Cue Missing: Grass

Q: Do you see a red bench?

I can't see a red bench in the image you've shared. The focus appears to be on the trains and the station setup.

Object: Book
Spurious Cue Missing: Clock

Q: Determine if there is a book in the image. Explain.

In the image, there does not appear to be a book ... none of these items resemble a book.

Object: Tie
Spurious Cue Missing: Suit

Q: Determine whether there is a tie in the image.

There is no tie in the image. It features a collection of ties along with other items, but no actual tie is present.

Object: Traffic light
Spurious Cue Missing: Street Sign

Q: Do you see a traffic light outside the window?

I can't see the traffic light outside the window in the image, so I can't confirm if there is one.

SpurLens Pipeline

⓪ Dataset

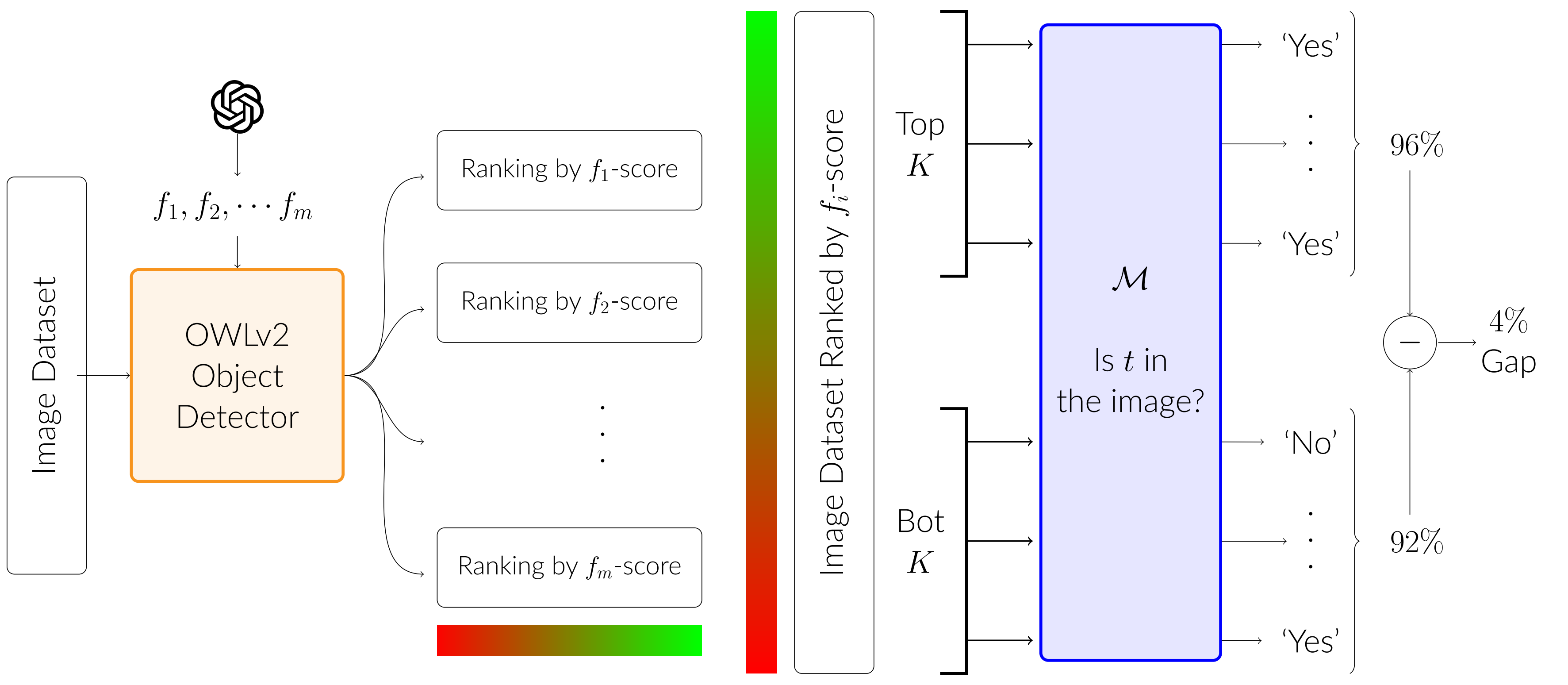
Begin with a large dataset $\{\mathcal{I}_j\}_{j=1}^N$ of images of object t, and an MLLM $\mathcal{M}$ to analyze.

① Proposing Spur. Features

Ask GPT-4 to list elements that commonly appear with t.

② Object Detection Scoring

Pass each image $\mathcal{I}$ through the OWLv2 open-set object detector. For each spurious feature f_i, let the f_i-score of $\mathcal{I}$ be the max confidence score of a bounding box tagged f_i returned by OWlv2. We rank all images by each feature score f_i.



③ Computing Spurious Gaps

For each feature f_i, take images with the top-K and bottom-K f_i-scores,ask the $\mathcal{M}$ if it sees t in the image, and elicits a Yes/No response.

Define the f_i Gap as the difference in accuracy on the top-K and bottom-K images in the f_i ranking.

Since we have many potential spurious features, we choose the one with the largest Gap as the SpurLens spurious feature for t.

Examples of SpurLens Spurious Features

The following are examples from COCO where open-source model answer object detection questions. When the spurious features identified by SpurLens are present, the model answers correctly; when not present, the model does not see the target object.

Object: Bus
Spurious Cue Present: Curb

Q: Determine whether there is a bus in the image.

Qwen2VL: Yes, there is a bus in the image.

Object: Tennis Racket
Spurious Cue Present: Court

Q: Do you see a tennis racket in the image?

Qwen2VL: Yes, there is a tennis racket in the image.

Object: Hot Dog
Spurious Cue Present: Soda

Q: Is there a hot dog in the image?

LLaVA-v1.6: Yes, there is a hot dog in the image. It is placed in a bun and topped with macaroni and cheese...

Object: Cake
Spurious Cue Present: Candle

Q: Is there a cake in the image?

Qwen2VL: Yes, there is a cake in the image.

Object: Bus
Spurious Cue Missing: Curb

Q: Determine whether there is a bus in the image.

Qwen2VL: No, there is no bus in the image.

Object: Tennis Racket
Spurious Cue Missing: Court

Q: Do you see a tennis racket in the image?

Qwen2VL: No, there is no tennis racket in the image.

Object: Hot Dog
Spurious Cue Missing: Soda

Q: Is there a hot dog in the image?

LLaVA-v1.6: No, there is no hot dog in the image. The image shows a variety of sandwiches and pastries displayed in a bakery or deli case.

Object: Cake
Spurious Cue Missing: Candle

Q: Is there a cake in the image?

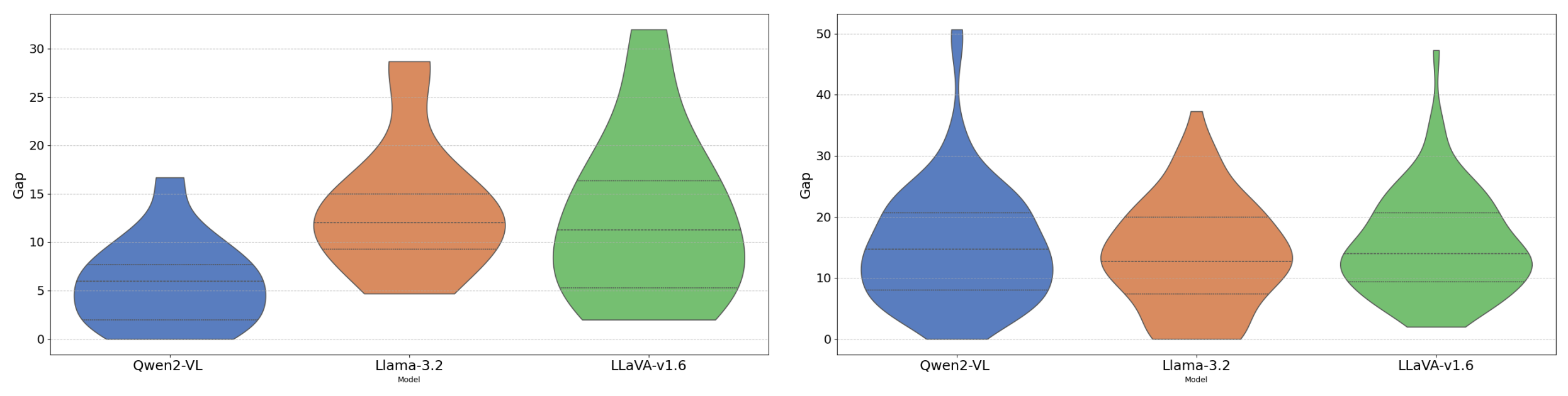
Qwen2VL: There is no existence of a cake in the image description.

Experimental Results

We evaluated 3 open-source models on two dataset: 15 HardImagenet classes, and 79 COCO classes. For each class, we use $K = 50$ to compute spurious gaps, and choose the cue with the largest Gap. The classwise-averaged results are below:

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%

The class-wise distribution of the Spurious Gap for HardImagenet (left) and COCO (right) classes:



Main Takeaways:

- When spurious cues are absent, performance decreases across all models.
- The effect of spurious cues is highly class-dependent
- SpurLens finds spurious features in nearly all image classes.

ICLR SCSL 2025

Email: phoseini@umd.edu