



Fighting Spurious Correlations Using Language Guided Abstractions

Knowledge distillation to tackle the presence of spurious correlations using text-based prompting.

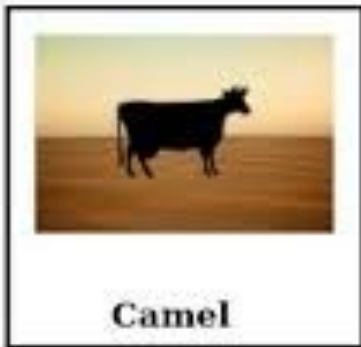
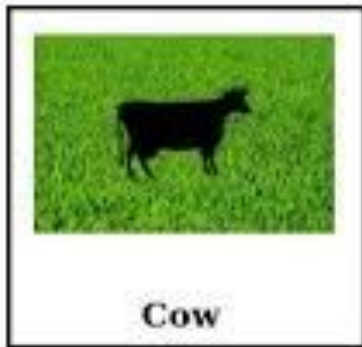
Shika Rao
Nikhil Kommineni
Musonda Sinkala
Manasvin Anand

Background

Introduction and Related Work

01

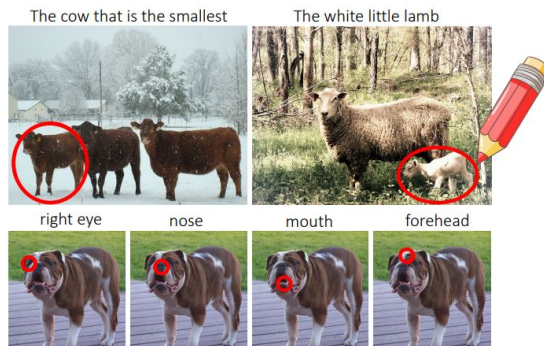
Introduction- Problem Setting



Neural Network Predictions

Spurious Correlations

Related Work



CLIP is better at visual classification when objects are highlighted with a red circle around it. [2]



Question: Choose the correct image for the caption. Caption: a big cat is next to a small dog. Options: (A) image 1(left) (B) image 2(right)

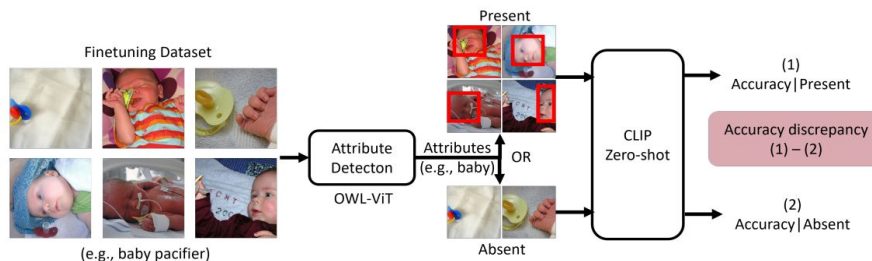
GPT-4V + SCAFFOLD:

Textual Prompt for SCAFFOLD: Two images are provided, each overlaid with a grid of dots arranged in a matrix with dimensions 6 by 6. Each dot on this grid is assigned a unique set of three-dimensional coordinates labeled as (t, x, y). ...

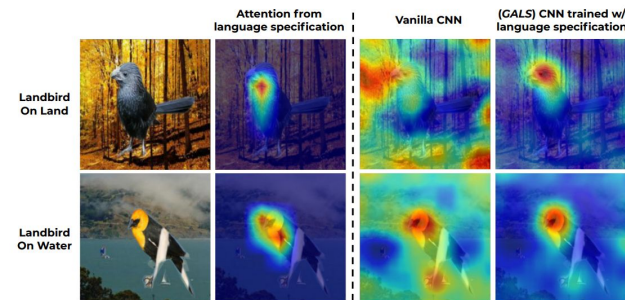
1. When you mention any key entities in the image, first output their nearest coordinates then identify them.

2. You can use the coordinates to determine the spatial relationships of the objects. ...

Overlaying an image with coordinates helps a model associate relevant texts. [3]

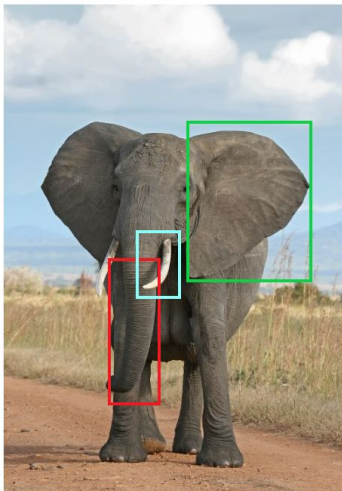


The authors label classes in image to detect spurious correlations. Apply contrastive loss on the detected classes. [4]



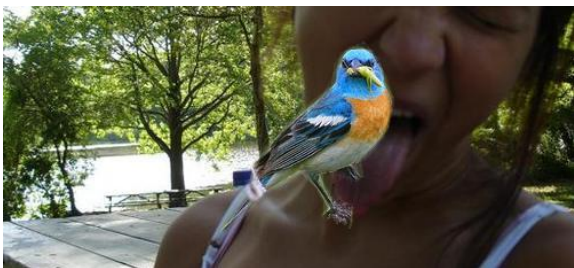
Distill activation maps from CLIP to smaller models. [1]

Introduction- Motivation



- Big Ears
- Trunk
- Tusks

1. We hypothesize that big MLLMs are able to perform better than smaller MLLMs on spurious correlation tasks because **bigger models have the underlying reasoning** information latent in the model.
2. Hence we aim to extract **latent reasoning** info from (bigger) teacher models to distill to student models to improve representation learning and alignment.
3. Humans learn distinguishing features with ease, allowing us to reason with less data.
4. Many-to-one mapping from image_features to text_features with pre-trained clip. We try to develop a one-to-one mapping between image and text.



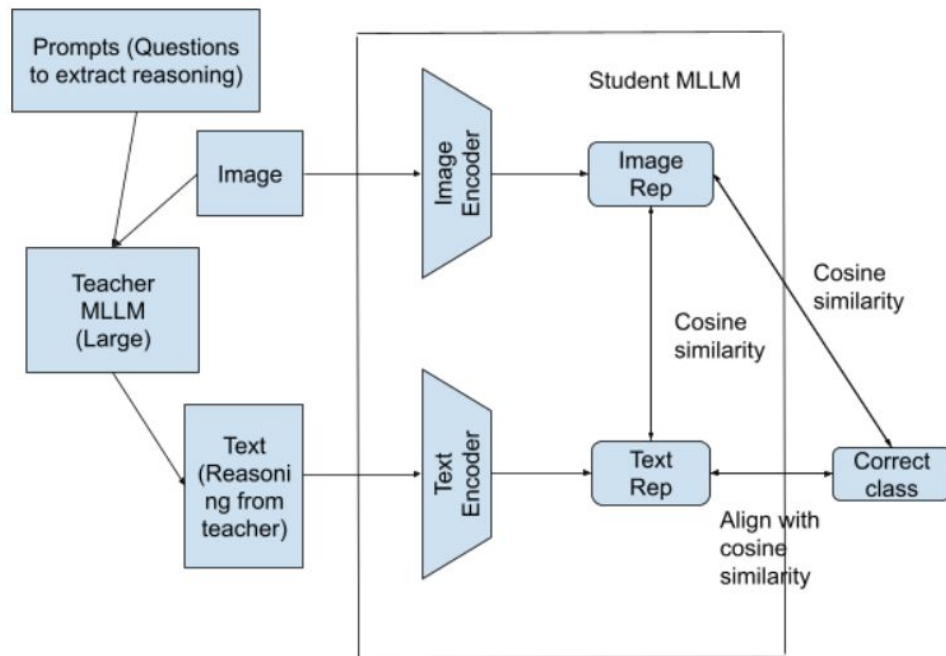
Focus on attributes of bird to align with class representation.

Methodology

Approach and Results

01

Methodology



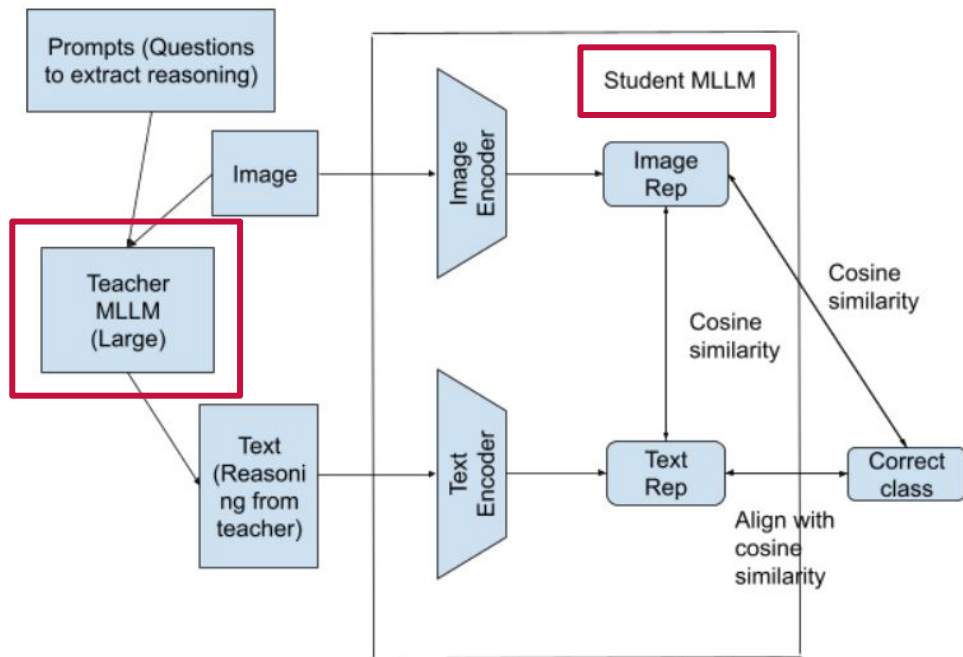
Loss Functions

Supervised Contrastive Loss

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

$$\mathcal{L}_{in}^{sup} = \sum_{i \in I} \mathcal{L}_{in,i}^{sup} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \right\}$$

Methodology



Models

Teacher Model

1

Pre-trained LLaVA: v1.5-7b

Student Model

2

Fine-tuned Clip: ViT-L/14 trained on 2000 samples

Experiment Setup

Dataset

The dataset contains 11,788 images featuring various land and waterbirds set against diverse backgrounds

All experiments were conducted using a sample of 2,000 images - with the next steps involving further reducing the sample size



LLaVA Prompts

Object

Identify the bird in the image and describe distinguishing features of the bird in the image. Ignore the background, be concise.



The bird in the image is a brown and white bird with a black head. It is flying over a swimming pool, possibly landing on the edge of the pool.

What is the common name of the bird in the image? Give keywords to identify and describe physical traits of the waterbird or landbird class it belongs to. Ignore the background. Be concise.



Seagull, pointed beak, streamlined body, webbed feet.

Background

Describe the background in the image, ignore the bird. Be concise.



The background in the image is a body of water, which appears to be a lake or a pond.

Give keywords describing the background. Ignore the bird. Be concise.



Mountains, water, city, trees.

Results

	Zero shot ViT-L/14 Clip	Binary Clip ViT-B/32	Binary Clip ViT-L/14	Model 1	Model 2
Overall Accuracy	83.69%	80.29%	89.35%	90.61%	96.01%
Waterbird Accuracy	48.44%	65.11%	76.79%	82.32%	87.62%
Landbird Accuracy	93.73%	84.61%	92.92%	92.97%	98.51%
Waterbird on water background	63.55%	89.56%	92.52%	93.46%	90.50%
Waterbird on land background	33.02%	40.65%	61.06%	68.07%	84.74%
Landbird on water background	88.69%	69.98%	85.99%	75.48%	97.83%
Landbird on land background	98.94%	99.25%	99.87%	97.52%	99.20%

Model 1:
ViT-B/32 Clip

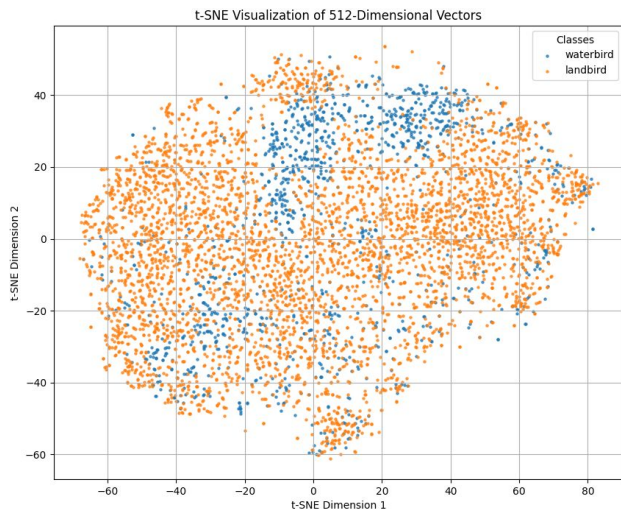
Model 2:
ViT-L/14 Clip

	Zero shot ViT-L/14 Clip	Binary Clip ViT-L/14	Model 2	Fine-tuned Clip *[4]
Overall Accuracy	83.69%	89.35%	96.01%	97.2%
Worst Group Accuracy	33.02%	61.06%	84.74%	89.7%

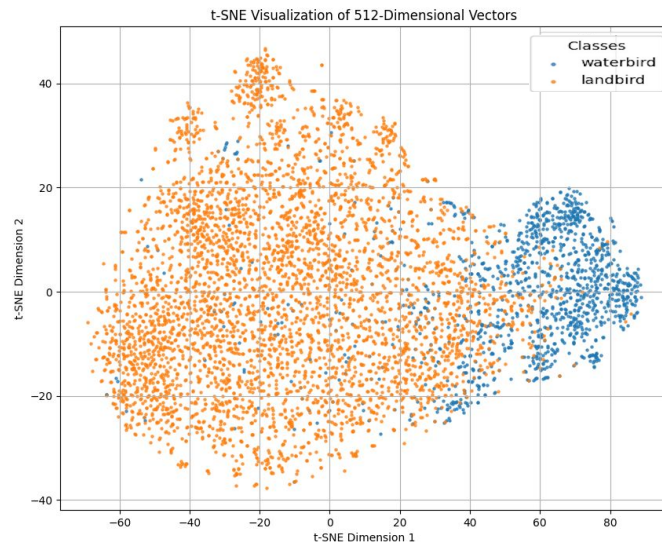
Worst Group for fine-tuned clip = Waterbird on land background

Experiments

Pre-trained CLIP

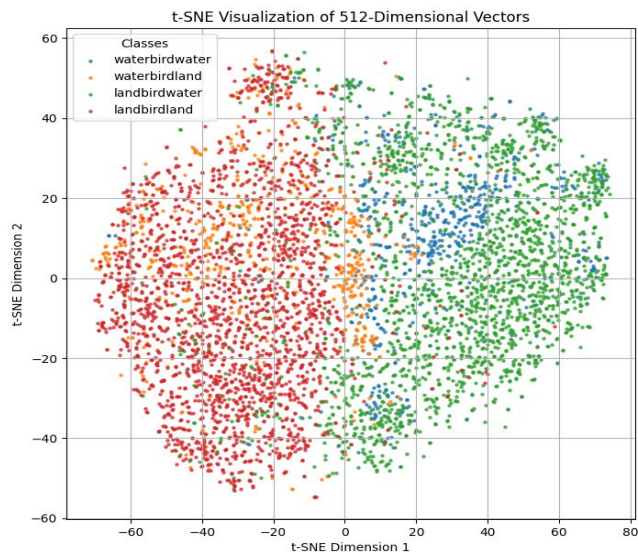


Fine-tuned CLIP

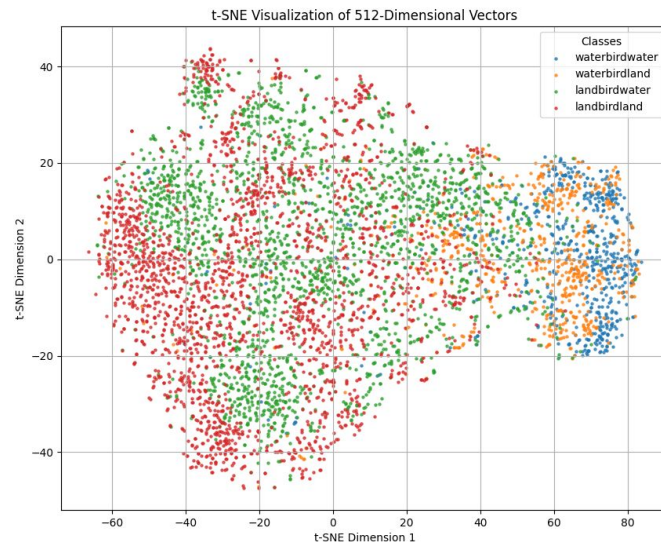


Experiments

Pre-trained CLIP



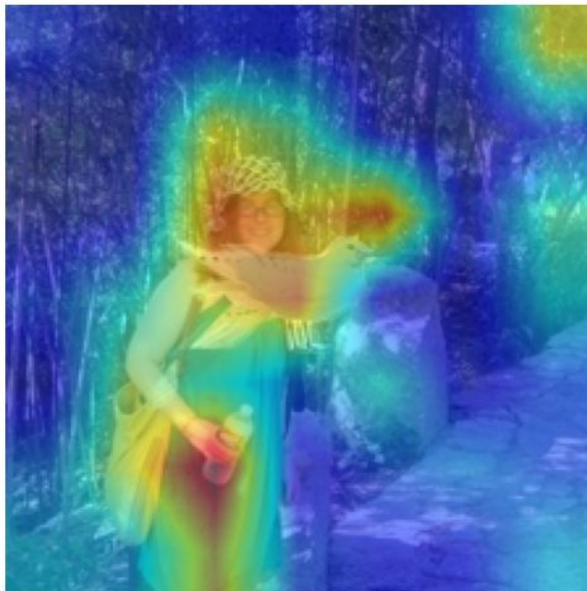
Fine-tuned CLIP



Experiments

Pre-trained CLIP

a photo of a landbird.
0.626



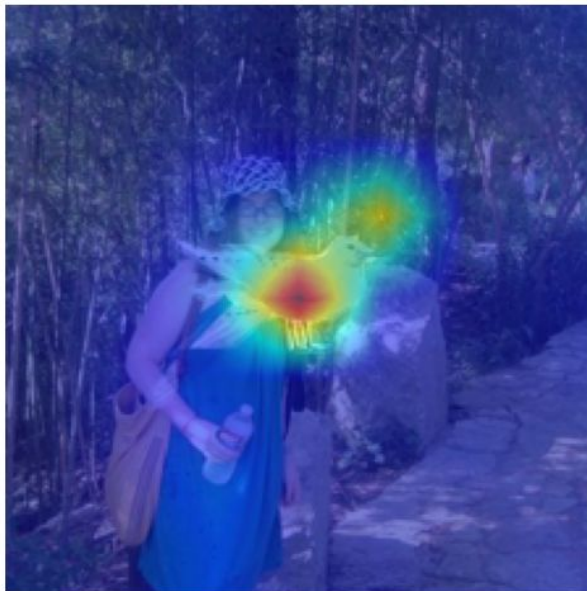
a photo of a waterbird.
0.374



Experiments

Fine-tuned CLIP

a photo of a landbird.
0.709



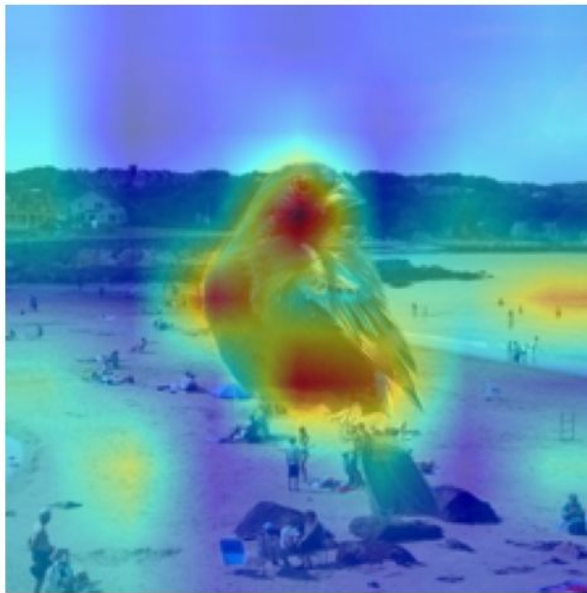
a photo of a waterbird.
0.291



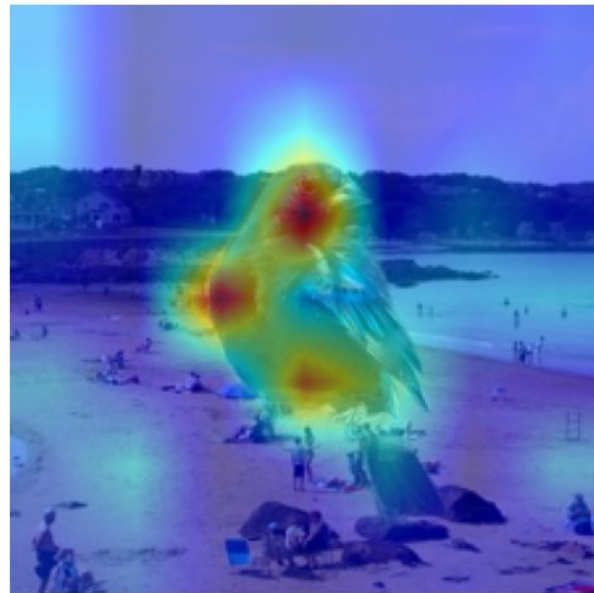
Experiments

Pre-trained CLIP

a photo of a waterbird.
0.514



a photo of a landbird.
0.486



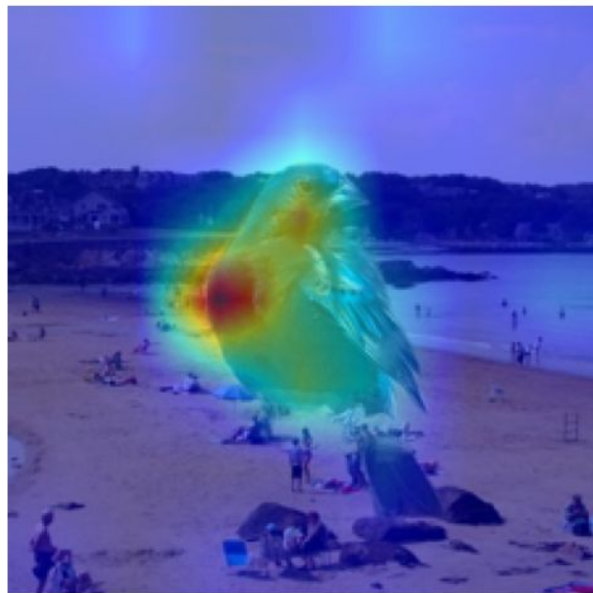
Experiments

Fine-tuned CLIP

a photo of a landbird.
1.000



a photo of a waterbird.
0.000



Next Steps

01

Intermediate Conclusions and Next Steps

- 1 Use Other datasets in experimentation - e.g., WiLDS.
- 2 Further convert the problem to a small data problem by constraining the sample size.
- 3 Experiment with alternative multimodal models (not solely LLaVA) and alternative Vision Language Models (not solely Clip).
- 4 How finetuning impacts generalizability of CLIP: Stable diffusion Text Encoder (TE) replace with our TE.
- 5 Understand alignment.

References

01

[1] Petryk, S., Dunlap, L., Nasser, K., Gonzalez, J., Darrell, T., & Rohrbach, A. (2022). On guiding visual attention with language specification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18092-18102).

[2] Shtedritski, A., Rupprecht, C., & Vedaldi, A. (2023). What does clip know about a red circle? visual prompt engineering for vlms. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11987-11997).

[3] Lei, X., Yang, Z., Chen, X., Li, P., & Liu, Y. (2024). Scaffolding coordinates to promote vision-language coordination in large multi-modal models. arXiv preprint arXiv:2402.12058.

[4] Yang, Y., Nushi, B., Palangi, H., & Mirzasoleiman, B. (2023). *Mitigating Spurious Correlations in Multi-modal Models during Fine-tuning*. arXiv preprint arXiv:2305.00000