Mitigating Spurious Correlations in Multi-modal Models during Fine-tuning

Yu Yang [1] Besmira Nushi [2] Hamid Palangi [2] Baharan Mirzasoleiman [1]

Abstract

Spurious correlations that degrade model generalization or lead the model to be right for the wrong reasons are one of the main robustness concerns for real-world deployments. However, mitigating these correlations during pre-training for large-scale models can be costly and impractical, particularly for those without access to high-performance computing resources. This paper proposes a novel approach to address spurious correlations during fine-tuning for a given domain of interest. With a focus on multi-modal models (e.g., CLIP), the proposed method leverages different modalities in these models to detect and explicitly set apart spurious attributes from the affected class, achieved through a multi-modal contrastive loss function that expresses spurious relationships through language. Our experimental results and in-depth visualizations on CLIP show that such an intervention can effectively i) improve the model's accuracy when spurious attributes are not present, and ii) directs the model's activation maps towards the actual class rather than the spurious attribute when present. In particular, on the Waterbirds dataset, our algorithm achieved a worst-group accuracy 23% higher than ERM on CLIP with a ResNet-50 backbone, and 32% higher on CLIP with a ViT backbone, while maintaining the same average accuracy as ERM[1].

1. Introduction

Vision-Language models (e.g., CLIP, DALL-E, Stable Diffusion, Imagen) are becoming pervasive in real-world deployments and have transformed the way large-scale model architectures are trained and used in different applications. Their multi-modal nature has not only enabled a large variety of tasks (e.g. text-to-image generation, visual question answering, image captioning) but is also facilitating better learning techniques that take advantage of data in several modalities to jointly learn embeddings that can then be reused in downstream tasks (Radford et al., 2021; Kamath et al., 2021; Li et al., 2022; Zhou et al., 2020).

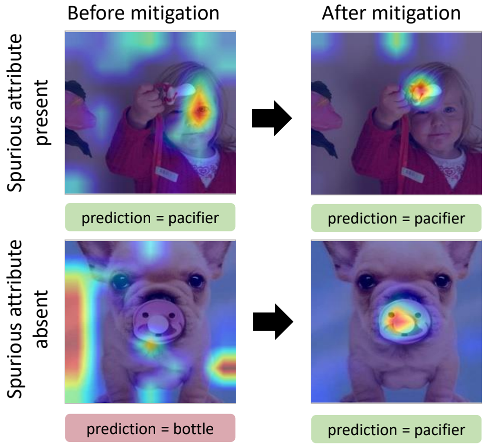

Figure 1. The *baby pacifier* class in ImageNet is spuriously correlated with the presence of *babies*, which leads the pre-trained model to be less accurate for cases when babies are absent in the image (bottom row) and also be right for the wrong reasons when babies are present (top row). Our approach mitigates both concerns by conveniently expressing and decorrelating the spurious relationships in the loss function via language.

While the multi-modal alignment increases the expectations about model reliability due to better grounding and larger availability of data in general, these models are still not immune to fundamental learning problems such as dealing with spurious correlations (Bommasani et al., 2021; Moayeri et al., 2022; Petryk et al., 2022; Agarwal et al., 2021). Therefore, when such models are used as a backbone to solve application-oriented tasks on a given domain, existing spurious correlations specific to that domain or the fine-tuning data that comes with it, may resurface in ways that are harmful to end users. At the same time, retraining large models from scratch to address such issues has become a less realistic avenue for two main reasons. First, stakehold-

[1] Department of Computer Science, University of California, Los Angeles, USA [2] Microsoft Research, Redmond, USA. Correspondence to: Yu Yang <yuyang@cs.ucla.edu>, Besmira Nushi <besmira.nushi@microsoft.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Code can be found at https://github.com/bigml-cs-ucla/clip-spurious-finetune

ers who need to adapt a model to a particular domain may not necessarily have access to large-scale computation. Second, the types of spurious correlations of interest are often domain-specific and not all of them can be anticipated ahead of time during pre-training of a general model. Furthermore, while previous work has studied spurious correlations in single-modal models trained with supervised learning, we note that spurious correlations learned in a joint multimodal embedding space with contrastive language image pretraining may not be the same due to differences in inputs and training objectives. For instance, we found that certain spurious correlations commonly studied in supervised learning of vision models, such as the correlation between gender and hair colors in the CelebA dataset (Liu et al., 2015), were not learned by multimodal models with contrastive language image pretraining. This suggests that spurious correlations in multimodal models may exhibit unique characteristics that require further investigation.

Building on the challenges of spurious correlations in vision-language models and the need for efficient mitigation methods, we introduce a contrastive learning approach that leverages the multi-modality of CLIP as a vision-language model to detect and mitigate spurious correlations through language in fine-tuning time. In the detection stage, our method extracts linguistic attributes from the image and tests whether their presence or absence affects model performance. If the accuracy of the model drops when a specific attribute is not present, it indicates that the attribute is either an *overemphasized but necessary attribute* (e.g., misclassifying taxi cabs that are not yellow) or a *spurious correlation* (e.g., misclassifying boats when there is no water in the background) (Singla et al., 2021). Assuming that a practitioner or domain expert in the loop can determine whether the attribute is healthy or spurious, in the next stage, our method mitigates the identified spurious correlation by extending the current contrastive language-vision learning techniques with a set of additional loss functions that explicitly i) decorrelate spurious attributes from the class names in language, and ii) push away both the vision representations across classes and language representations of templates substituted with different class labels. It is worth noting that our approach *only fine-tunes the projections to the joint embedding space*. Since the projection layers contain much fewer parameters than the full models, our method requires significantly less computational resources compared to extensive retraining from scratch without losing features learned in pretraining.

In contrast to previous work which requires human annotations about spurious or group attributes (Sagawa et al., 2019), our approach uses automatically detected language-based descriptions of spurious attributes that can then directly be expressed and used in optimization to set them apart from affected classes. While domain experts are still required in this method to judge whether a detected co-occurence is a spurious attribute or not, this still minimizes labeling human supervision per example. Fine-tuning experiments with two datasets, Waterbirds and Imagenet, show that the proposed approach offers a better trade off between the average accuracy and worst-group accuracy (i.e., examples when the spurious attribute is not present) and can better align model explanation maps to the class of interest.

It is worth noting that our work differs from existing studies that focus on spurious correlations learned by vision models (Sagawa et al., 2019; Nam et al., 2020; Creager et al., 2021; Liu et al., 2021; Nam et al., 2022; Izmailov et al., 2022). Instead, we investigate spurious correlations learned by multimodal models during pre-training with the contrastive language-image loss. Although larger models may be less accurate than specialized models on certain tasks, practitioners may still choose to use a pretrained model for reasons such as maintenance and data availability. In addition, having enough labeled data to train a specialized vision model may not always be possible. In such cases, the larger pretrained model trained on noisy image-caption pairs may have already encoded useful information about the concept, and our method is useful for scenarios where one needs to maintain this generality while mitigating found issues for a specific domain.

Moreover, the multimodal nature of these models opens up new opportunities for detecting and mitigating failures without the need for additional annotation data, such as attributes or bounding boxes, to guide the model's attention. By leveraging the information encoded in the joint embedding space, our approach improves the model's attention in GradCAM explanations and quantitatively in AIoU scores, a new metric we proposed for evaluating the model's attention. This finding is particularly noteworthy as the need for metadata annotations and grounding has been a significant barrier for several applications, especially during cold starts.

In summary, our contributions are:

- A language-based approach that detects spurious correlations with practitioner supervision but no spurious attribute labeling.
- A loss function that extends current contrastive vision-language learning for mitigating spurious correlations in vision through language.
- A set of experiments that showcase how to use the proposed detection and mitigation approach in practice for the CLIP model as well as its effectiveness in datasets with known and unknown spurious correlations.

2. Related Work

Explaining and Debugging Trained Models. Several algorithms have been proposed to semantically explain and analyze trained neural networks, including distilling the decision modes into decision trees (Zhang et al., 2019; Singla et al., 2021), training classifiers in the latent space (Jain et al., 2022; Yang et al., 2022), and embedding inputs with joint vision-language representations to find the error slices with a mixture model (Eyuboglu et al., 2022). These methods usually accompany the semantic explanations with feature attention maps, e.g., GradCam (Selvaraju et al., 2017). The authors of (Shankar et al., 2020) conducted a comprehensive study on ImageNet (Russakovsky et al., 2015) by manually relabeling it and uncovered multiple instances of label noise and disagreement in the dataset. In this paper, we are only interested in discovering and mitigating *spurious correlations*, which are introduced next.

Enhancing Robustness to Spurious Correlations. We study spurious correlations in the context of deep learning, as they have been formally discussed in (Sagawa et al., 2019). Given a classification dataset $\mathcal{D}$ with labels $\mathcal{Y}$, if there exist spurious attributes $\mathcal{A}$ that are highly correlated with $\mathcal{Y}$, a deep neural network trained on this dataset is likely to learn $\mathcal{A}$ as features to distinguish $\mathcal{Y}$, even if the attribute is not conceptually part of the class concept. For example, in Figure 1, a pretrained CLIP-RN50 model (Radford et al., 2021) learned to use *baby* to identify *baby pacifier* because they often appear together in ImageNet (Russakovsky et al., 2015), instead of actually learning the baby pacifier itself.

To prevent deep learning models from learning such spurious correlations from biased data, recent work proposed training strategies robust to spurious attributes for either vision or language models (Sagawa et al., 2019; Nam et al., 2020; Creager et al., 2021; Liu et al., 2021; Nam et al., 2022; Izmailov et al., 2022). The spurious label of each training example (e.g., whether this example contains the spurious feature) is either provided (Sagawa et al., 2019; Izmailov et al., 2022) or inferred by training a reference model (Nam et al., 2020; Creager et al., 2021; Liu et al., 2021; Nam et al., 2022) until it learns the spurious correlations. Other approaches indirectly estimate and use the causal effect of hidden non-labeled spurious attributes in pre-training (Mao et al., 2022).

However, these studies all focus on training *unimodal* models with datasets that contain known spurious features, and spurious correlations learned by pretrained *multimodal* models have not been extensively studied. To the best of our knowledge, we are the first to propose a *fine-tuning* approach for mitigating spurious correlations in multimodal models. While (Zhang & Re, 2022) also studied CLIP's robustness to group shifts including spurious correlations, their method is designed for transfer learning rather than fine-tuning the learned embedding space.

Correcting Vision Models using Language. There is a line of recent work aiming to fix vision classifiers with language inputs. Petryk et al. (2022) uses attention maps from a pre-trained CLIP to supervise a CNN classifier's spatial attention. Zhang et al. (2023) probes a vision classifier trained on the joint vision-language embedding space of CLIP using language embeddings of attributes, identifies the attributes causing most failures, and generates a large set of natural language inputs with the influential attributes to rectify the model. However, this line of work aims to guide CNN classifiers rather than fixing CLIP models and does not prevent spurious feature usage.

3. Spurious-aware Contrastive Language Image Fine-tuning

Background. Contrastive Language-Image Pretraining (CLIP) learns from millions of image caption pairs, by maximizing the agreement between representations of every image and the representations of its corresponding caption. Specifically, the CLIP architecture consists of (i) an image encoder network, (ii) a text encoder network, and (iii) a contrastive objective that pulls the embeddings of every image and its corresponding caption together while pushing apart embeddings of the image from other captions in the same minibatch. Formally, for a minibatch of N image-captions pairs $\{I_j, T_j\}_{j=1}^N$, and their encoded embeddings $\{I_j^e, T_j^e\}_{j=1}^N$, the CLIP loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} = &-\frac{1}{2}\underset{(I_i,T_i)}{\mathbb{E}}\log\left[\frac{e^{\left\langle I_j^e, T_j^e\right\rangle/\tau}}{\sum_{k=1}^N e^{\left\langle I_j^e, T_k^e\right\rangle/\tau}}\right] \\ &-\frac{1}{2}\underset{(I_i,T_i)}{\mathbb{E}}\log\left[\frac{e^{\left\langle I_k^e, T_k^e\right\rangle/\tau}}{\sum_{j=1}^N e^{\left\langle I_j^e, T_k^e\right\rangle/\tau}}\right],\end{aligned} \tag{1}$$

where $\langle ., . \rangle$ represents the inner product, and τ is a trainable temperature parameter. For finetuning CLIP on a dataset of images and their labels, such as Waterbirds, the labels are replaced in the engineered prompt templates, such as "A photo of a {label}", "A photo of a {label}, a type of bird.", etc. Then, the loss is minimized on the images paired with templates built with image labels. We use all 80 templates described in (Radford et al., 2021).

For a given spurious attribute (e.g. water or land background in the Waterbirds dataset), we will use the following losses to eliminate the spurious correlation during fine-tuning. Please note that the contrastive losses below use the class information to pull together representations of examples from the same class label, and push away representations of examples from different class labels. The spurious losses use the spurious attribute detected in the spurious correlation detection stage (Section 4) to pull together representations of examples with the same spurious attribute (e.g. attribute present) and push away representations of examples with a different spurious attribute (e.g.

attribute absent).

Here, we will use the following construct as a basis for the definition of all loss terms: a *cross-group representation similarity* term that pulls together representations from the same group and pushes away representations of different groups. The representations can be either in the vision or language space. We reuse this construct to extend CLIP contrastive learning to improve classification and also mitigate spurious correlations. Let $G_1 = \{(I_p, T_p)\}_{p=1}^{P}$ be the set of examples in one group of examples in the minibatch, and $G_2 = \{(I_q, T_q)\}_{q=1}^{Q}$ the set of examples in another group of the same minibatch, as defined by the relationship of a given example in the minibatch (I_i, T_i) to these groups. Depending on the loss term, the relationship between examples can be either due to examples belonging to the same class or having the same spurious attribute value. Then, the cross-group representation similarity defined across two modalities of representation embeddings A and B is:

$$\mathrm{CS} = -\mathop{\mathbb{E}}_{\substack{(I_i,T_i),\\ (I_p,T_p)\in G_1\\ (I_q,T_q)\in G_2}} \left[\log \frac{e^{\left\langle A_i^e, B_p^e \right\rangle/\tau}}{\sum_{p=1}^{P} e^{\left\langle A_i^e, B_p^e \right\rangle/\tau} + \sum_{q=1}^{Q} e^{\left\langle A_i^e, B_q^e \right\rangle/\tau}} \right]$$

Contrastive Image Loss The first term is a contrastive image loss which pulls together image representations of a class, and pushes away image representations of different classes in the vision model. Let $G_l = \{(I_p, T_p)\}_{p=1}^{P}$ be the set of examples in the minibatch with the **same label** as example (I_i, T_i), i.e., $T_i = T_p$, and $\hat{G}_l = \{(I_q, T_q)\}_{q=1}^{Q}$ be the set of examples with a **different label**. Then the contrastive image loss within the vision representation embeddings I is defined as:

$$\mathcal{L}_{vc} = \mathrm{CS}(G_l, \hat{G}_l, I, I) \tag{2}$$

Contrastive Language Loss The second term is a contrastive language loss which pulls together language representations of templates of a class in the language model, and pushes away language representations of different classes. Let $G_l = \{(I_p, T_p)\}_{p=1}^{P}$ be the set of examples in the minibatch with the **same label** as example (I_i, T_i), i.e., $T_i = T_p$, but with different templates. Let $\hat{G}_l = \{(I_q, T_q)\}_{q=1}^{Q}$ be the set of examples in the minibatch with a **different label**. Then the contrastive language loss within the language representation embeddings T is defined as:

$$\mathcal{L}_{lc} = \mathrm{CS}(G_l, \hat{G}_l, T, T) \tag{3}$$

Spurious Image Loss The third term is a spurious contrastive image loss which pulls together image representations of each group of examples in a class, and pushes away image representations of different groups of examples. For example, it pulls together images of waterbirds with water background, and pulls them away from images of waterbirds with land background and from landbird images with water or land background.

Assume $G_s = \{(I_p, T_p)\}_{p=1}^{P}$ is the set of images in the minibatch with the **same spurious attribute** as example (I_i, T_i), and $\{\hat{G}_s = (I_q, T_q)\}_{q=1}^{Q}$ is the set of examples with a **different spurious attribute** than example (I_i, T_i). A different spurious attribute here could also mean that the spurious attribute is absent. Then, the spurious image loss within the vision representation embeddings I is defined as:

$$\mathcal{L}_{vs} = \mathrm{CS}(G_s, \hat{G}_s, I, I) \tag{4}$$

Spurious Language Loss The last term is a spurious contrastive language loss which pulls together language representations of each group of examples in a class, and pushes away language representations of different groups of examples. Assume $G_s = \{(I_p, T_p)\}_{p=1}^{P}$ is the set of examples in the minibatch with the **same spurious attribute** with example (I_i, T_i), and $\hat{G}_s = \{(I_q, T_q)\}_{q=1}^{Q}$ is the set of examples with a **different spurious attribute** in the minibatch. Note that, a different spurious attribute here could also mean that a spurious attribute is absent. Then, the spurious language loss within the language representation embeddings T is defined as:

$$\mathcal{L}_{ls} = \mathrm{CS}(G_s, \hat{G}_s, T, T) \tag{5}$$

The final loss is the sum of all the terms above:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CLIP}} + \mathcal{L}_{vc} + \mathcal{L}_{lc} + \mathcal{L}_{vs} + \mathcal{L}_{ls}. \tag{6}$$

In practice, either $\mathcal{L}_{vs}$ or $\mathcal{L}_{ls}$ can be combined with $\mathcal{L}_{lc}$ to effectively eliminate the spurious correlation. If spurious attribute annotation labels are available, one can use $\mathcal{L}_{vs}$. If spurious attribute annotation labels are not available $\mathcal{L}_{ls}$ can provide a good separation between groups in different classes. In all experiments reported hereafter we show results for both, and the ablation study in Section A details the tradeoffs between these and other choices.

From an implementation perspective, all language-related losses could be implemented across examples or templates. Our implementation follows a template-based approach.

4. Spurious Correlation Detection

This section introduces our pipeline for detecting and evaluating spurious correlations learned by a pretrained model. While we apply this pipeline to CLIP models in this work, it can be generalized to other pretrained models as well. The approach closely follows previously discussed techniques (Singla et al., 2021; Nushi et al., 2018) but relies on automatically generated annotations for attributes.

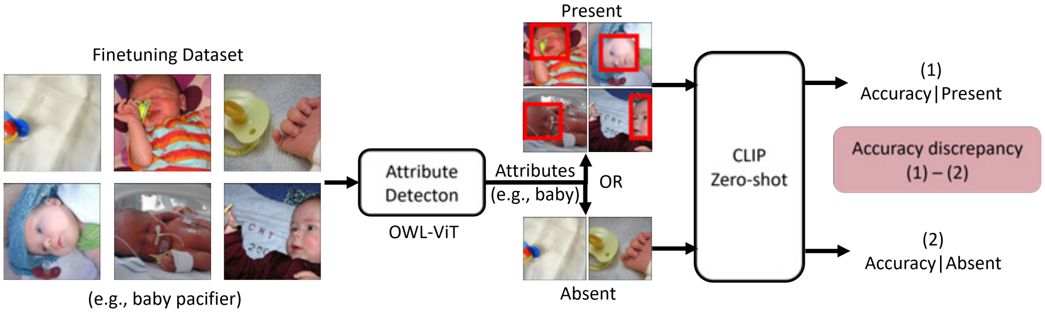

Figure 2. Spurious correlation detection based on attributes from an open-vocabulary detector and accuracy discrepancy scores of the model between examples when the spurious attribute is present or absent.

4.1. Methodology

For any given fine-tuning dataset, we are interested in knowing whether CLIP (or any other pretrained models) has learned any spurious correlations for the classes in the dataset. According to the definition of spurious attributes introduced in Section 3, models that have learned a certain spurious correlation usually show better performance (e.g., higher classification accuracy) on examples with that spurious attribute. For example, a model that majorly relies on the presence of an emergency vehicle to detect an accident, would have a lower accuracy in detecting accidents when there are no emergency vehicles around.

We use the pipeline depicted in Figure 2 to (1) find such spurious attributes for a class of interest if the spurious attributes are unknown, and (2) measure how much each spurious attribute negatively affects the model.

Spurious Detection. For the case where the spurious attribute is unknown, we first use an open-vocabulary detector, OWL-ViT (Minderer et al., 2022), to detect potential spurious attributes for examples in the fine-tuning data. We use the synsets of object names in Visual Genome (Krishna et al., 2016) as our list of attributes to detect after removing objects that are classes of the fine-tuning data.

Spurious Evaluation. We define $\delta(\mathcal{D}, s)$ as the model accuracy discrepancy between examples in dataset $\mathcal{D}$ with the attribute s and those without it.

$$\delta(\mathcal{D}, s) = acc(\mathcal{D}|s = 1) - acc(\mathcal{D}|s = 0). \qquad (7)$$

Attributes detected in the fine-tuning dataset can then be ranked by their accuracy discrepancy scores. The higher the discrepancy, the more this attribute could harm the generalization performance of the pretrained model. Since model failure modes and in particular spurious correlations are often specific to the class (Nushi et al., 2018), for the ImageNet studies we compute and sort the discrepancy scores per class. While the drop in accuracy with the absence of the spurious attributes are good indicators of spurious correlations, such drops may also happen for healthy attributes that are part of the class definition (e.g., the yellow color for taxi cabs albeit not all taxis are yellow).

Thus, for practical usages of our approach, we imagine this step to involve some miminal human investigation from domain experts or ML practitioners to judge whether the attribute is healthy or a potential spurious correlation. Humans can make this call based on their domain knowledge or one of the vision interpretability techniques (e.g, Grad-CAM, Integrated Gradients etc.). Nevertheless, this kind of supervision is considerably more lightweight than annotating attributes or manually inspecting individual examples. Table 3 shows several examples of previously unknown spurious correlations we found for CLIP in ImageNet.

5. Experiments

5.1. Backbones

CLIP uses two main groups of visual backbones, ResNets (RN) and Visual Transformers (ViT), and reported model performance separately for models with these two types of backbones in (Radford et al., 2021). In particular, ResNet-50 (RN50) and ViT-L/14@336px [2] are used as the prototypes of these two groups of models. Therefore, we follow (Radford et al., 2021) and study CLIP models with RN50 and ViT-L/14@336px visual backbones in our experiments.

For all experiments, we freeze both the language and vision encoders and only fine-tune the projection layers. Keeping both encoders intact is not only more lightweight but also resulted in better overall and worst-group accuracy for all studied datasets in our preliminary experiments.

[2] ViT-L/14@336px refers to ViT-L/14 model fine-tuned on 336-by-336 pixel input images.

Table 1. Statistics of the Waterbirds training data.

	LAND	WATER
LANDBIRDS	3498	184
WATERBIRDS	56	1057

5.2. Datasets

Waterbirds. Waterbirds (Sagawa et al., 2019) is the most commonly used benchmark dataset for studying spurious correlations. It combines birds segmented from the CUB dataset (Wah et al., 2011) and the background in dataset (Zhou et al., 2017) in an imbalanced way such that the background can be used as a spurious attribute for bird classification. Table 1 shows the sample size of each class-background combination in the Waterbirds training set. As landbirds appear more with land background and waterbirds are more often on water background in the training set, models fine-tuned on this dataset often learn to rely on the background instead of the birds.

ImageNet-1K. Singla et al. (2021) found that some features are spuriously correlated with some categories in ImageNet-1K (Russakovsky et al., 2015). For example, 55% of training examples in the "Rhodesian ridgeback" class can be correctly classified by a robust ResNet-50 model but the accuracy drops significantly to 24% when the dogs are not wearing a collar. We use the spurious detection pipeline shown in Figure 2 to find top-5 attributes with the highest accuracy discrepancy on CLIP for each class and attribute, and then rank the top attributes from all classes. Based on our inspection of the top attributes, we find a number of previously unknown spurious attributes learned by CLIP-RN50 with ImageNet as shown in Table 3. Out of this list, in the mitigation experiments we choose to mitigate the first major spurious correlation that has a high accuracy discrepancy: *Baby pacifier* class where the spurious attribute is *baby face*. CLIP accuracy drops by 69.1% for classifying baby pacifiers when there is no baby in the image. Note that since the validation set for ImagenNet contains only 50 images per class, we run the spurious correlation detection and evaluation stages on the training data instead, while mitigation results are presented for the test data. Figures 3 and 6 show further evidence of the pre-trained model focusing on the spurious attributes rather than the class itself.

Another dataset we considered for evaluation is CelebA (Liu et al., 2015) for the task of hair color classification. Previous work (Mao et al., 2022; Sagawa et al., 2019) has shown that models trained on such data can have a lower accuracy for small groups defined by the gender attribute such as men with blond hair, since this group has a low representation in the training data. It turns out however that model accuracy does not degrade for this group using CLIP model, which is why we do not present results on CelebA in this paper.

5.3. Metrics

We use the following metrics to evaluate the **predictions** and **explanations** of each model. We argue that only by obtaining high performance in both aspects, an algorithm can be proven to address the spurious correlations and that the correct model predictions are "*right for the right reasons*".

1. **Average Accuracy.** Classification accuracy averaged over classes on the test set. For the Waterbirds dataset, the test data is enriched and balanced to improve the accuracy of the evaluation, but this can lead to a discrepancy between the distribution of the test data and the training data. Following previous works, we report the adjusted average accuracy suggested by (Sagawa et al., 2019), which weights the test accuracy of each group by their sizes in the training data.

2. **Worst-group Accuracy.** The lowest model accuracy across groups as defined by the spurious attribute and the class of interest.

3. **Adjusted Intersection-over-Union (AIoU).** Previous works have used binary attribute maps to compute an Intersection-over-Union (IoU) score with the ground-truth bounding box (Nguyen et al., 2021). While IoU is a standard metric for object localization, using the standard IoU to evaluate the quality of attribute maps can be less reliable because the score highly depends on the threshold used for binarizing the attribute maps. To circumvent threshold dependency, we adapt the formulation such that it instead uses a min operator (instead of the binary intersection) between a bounding box B_y and an explanation map M_y, where y is the ground truth class. Similarly, we use a max operator (instead of binary union) in the denominator between the bounding box and the map.

$$\mathrm{IoU}(M, B) = \frac{\sum_{j,k} \min(M_{jk}, B_{jk})}{\sum_{j,k} \max(M_{jk}, B_{jk})}, \quad 0 \leq j \leq h, 0 \leq k \leq w. \tag{8}$$

Equation 8 measures the alignment between an explanation map and the ground truth bounding box but it does not take into consideration that despite a good alignment with the bounding box for true class, the explanation maps of other classes may still span across the bounding box of the ground truth class. Therefore, we use a definition of IoU that adjusts its denominator to include the class whose explanation map most intersects with the ground truth bounding box.

$$\mathrm{AIoU} = \frac{\mathrm{IoU}(M_y, B_y)}{\mathrm{IoU}(M_y, B_y) + \max\limits_{y' \in [C \backslash y]} \mathrm{IoU}(M_{y'}, B_y)}. \tag{9}$$

Table 2. Accuracy of different groups of Waterbirds on pre-trained ResNet- and Transformer-based CLIP models.

(RN50)	LAND	WATER
LANDBIRDS	93.44%	44.92%
WATERBIRDS	59.03%	91.59%

(VIT-L/14@336PX)	LAND	WATER
LANDBIRDS	99.29%	90.20%
WATERBIRDS	33.96%	55.61%

In our experiments, we used GradCAM (Selvaraju et al., 2017) for the explanation maps. While GradCAM explanations may not be perfectly aligned with the model's attention, their usage has shown practical benefits for model debugging (Yosinski et al., 2016; Simonyan et al., 2014; Mao et al., 2022).

5.4. Baselines

We compare our approach with pre-trained CLIP (Radford et al., 2021), fine-tuned CLIP with the training dataset in hand using the original contrastive vision and language loss as described in Equation 1, Empirical Risk Mimimization (ERM), and Group DRO (Sagawa et al., 2019). Group DRO is a distributionally robust optimization approach that minimizes worst-group loss and uses strong regularization. The methods requires attribute annotations to define groups being used during optimization. ERM instead is the standard empirical risk minimization technique for minimizing classification loss.

Reproducibility. Both CLIP models (CLIP-RN50 and CLIP-ViT) and prompt templates we use in our experiments are officially released by OpenAI[3](Radford et al., 2021). The Waterbirds dataset is from the WILDS library (Koh et al., 2021). We used the SGD optimizer for all the experiemnts, and tuned the learning rates and weight decays for ERM, GroupDRO and CLIP-based loss (CLIP-finetuning and our method) separately. Our method uses learning rate 1e-5 with weight decay 1e-4. The code will be publicly available upon publication.

5.5. Spurious Correlation Detection Results

Waterbirds. Table 2 shows model accuracy across the four groups as defined by class and spurious attribute definitions. The underlined groups show the worst-group accuracies for each model. For both models, there is a high accuracy discrepancy between groups from the same class. Figures 4 and 5 show examples of explanations from Pre-trained CLIP where explanations do not overlap with birds.

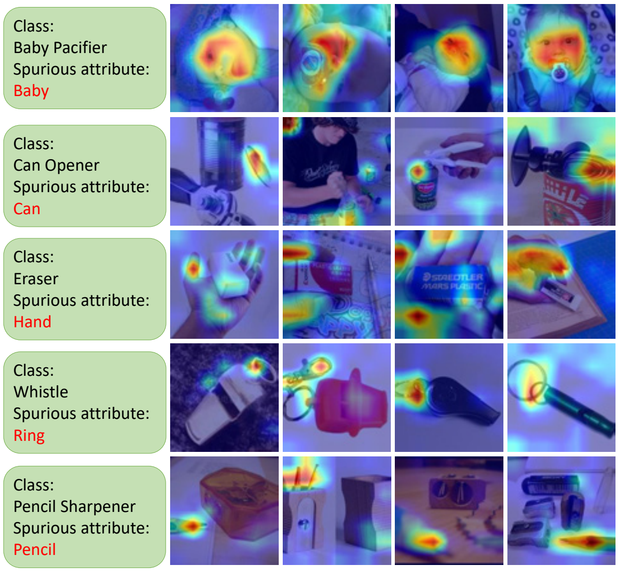

Figure 3. GradCAM explanations for cases when Pre-trained CLIP RN50 relies on the spurious classification described in Table 3.

Table 3. Spurious correlations found for CLIP RN50 on ImageNet.

Class	Spurious Attribute	Confused Class	Acc. Discrepancy
baby pacifier	baby	water bottle	62.1%
can opener	can	letter opener	45.2%
eraser	hand	pencil case	18.5%
whistle	ring	padlock	15.2%
pencil sharpener	pencil	pencil case	8.37%

Imagenet. Table 3 shows examples of prominent spurious correlations found for Pre-trained CLIP RN50. It is interesting to see how the found spurious attributes are concepts that are indeed highly related to the class but not necessarily part of the class definition. The natural co-occurence of these concepts leads the model to incorrectly rely rather on the attribute as shown in Figure 3.

5.6. Spurious Correlation Mitigation Results

Waterbirds. Table 4 summarizes our results on the Waterbirds dataset for both Resnet-50 and ViT-L/14@336px. Our method of mitigating spurious correlations through language has the best worst-group accuracy for ResNet-50 and second-best worst-group accuracy for ViT, maintaining a competitive average accuracy. What is of most interest from a mitigation perspective, is that the model ability to be right for the right reasons is indeed better for our method as indicated by the AIoU scores. These results are also qualitatively confirmed by visual explanation maps as shown in

[3] https://github.com/openai/CLIP

Table 4. Results of fine-tuning CLIP with Waterbirds. Average and worst-group performance is evaluated on the test set with models early stopped at the *highest worst-group accuracy* on the validation set. Worst groups: *Landbird on water* for RN50; *Waterbird on land* for ViT.

Model	ResNet-50				ViT-L/14@336px			
	Accuracy		AIoU		Accuracy		AIoU	
	Avg.	Worst-group	Avg.	Worst-group	Avg.	Worst-group	Avg.	Worst-group
Pre-trained CLIP	**90.8%**	44.9%	0.507	0.479	88.5%	34.0%	0.579	0.551
Fine-tuned CLIP	81.3%	**77.1%**	0.510	0.128	**97.2%**	89.7%	0.687	0.697
ERM	**93.5%**	54.4%	0.514	0.139	96.8%	58.1%	0.636	0.680
Group DRO	83.3%	73.7%	0.509	0.274	94.1%	**90.8%**	0.669	0.644
Ours($\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{ls}$)	84.7%	**77.5%**	**0.628**	**0.499**	**97.1%**	89.7%	**0.698**	**0.711**
Ours($\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{vs}$)	83.2%	**77.5%**	**0.654**	**0.587**	96.9%	**90.5%**	**0.716**	**0.709**

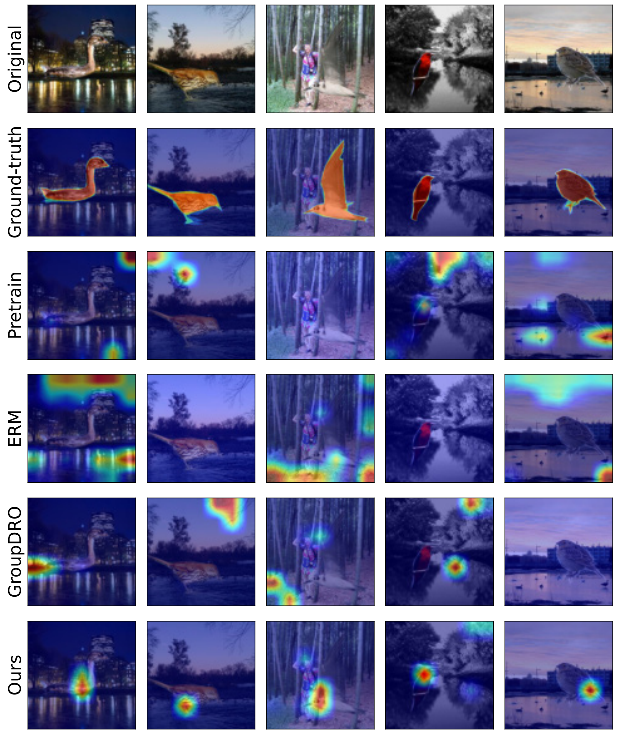

Figure 4. GradCAM explanations for different approaches based on CLIP RN50 for the Waterbirds dataset.

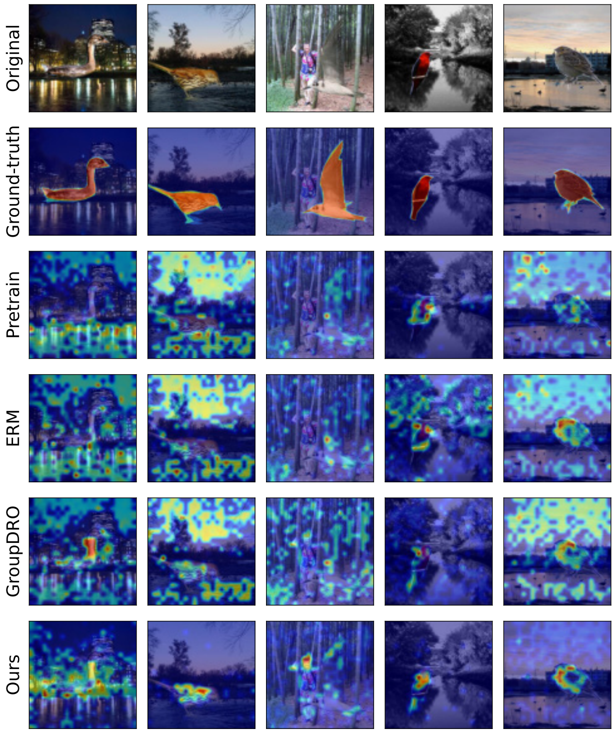

Figure 5. GradCAM explanations for different approaches based on CLIP ViT-L/14@336px for the Waterbirds dataset.

Figures 4 and 5, demonstrating that (i) the spurious correlation is present on the first place (pre-trained CLIP), (ii) it persists in the explanation maps of GroupDRO despite this method being competitive in both worst-group and average accuracy, and (iii) it is visibly alleviated though our approach whose explanations align with the available ground truth segmentations for the dataset. When comparing the two different variants of our method using the spurious language loss and image loss, we observe that the spurious image loss leads to better AIoU scores potentially because decorrelation is easier in the image representation, albeit for this method to work reliable attribute annotations are required. Using the spurious language loss is however still appealing with respect to both worst-group accuracy and AIoU. Note that implicitly, this method, and generally mitigating spurious correlations through language, relies on the capability of the model to map the spurious attribute from language to vision, which may not always be the case for the pre-trained vision-language models. The spurious attributes studied in this paper are based on the language concepts that are perhaps well-learned and mapped in a multi-modal way in CLIP (e.g., baby, water, land) but in other cases of less-frequent or domain-specific attributes, using the spurious image loss may be a more realistic avenue.

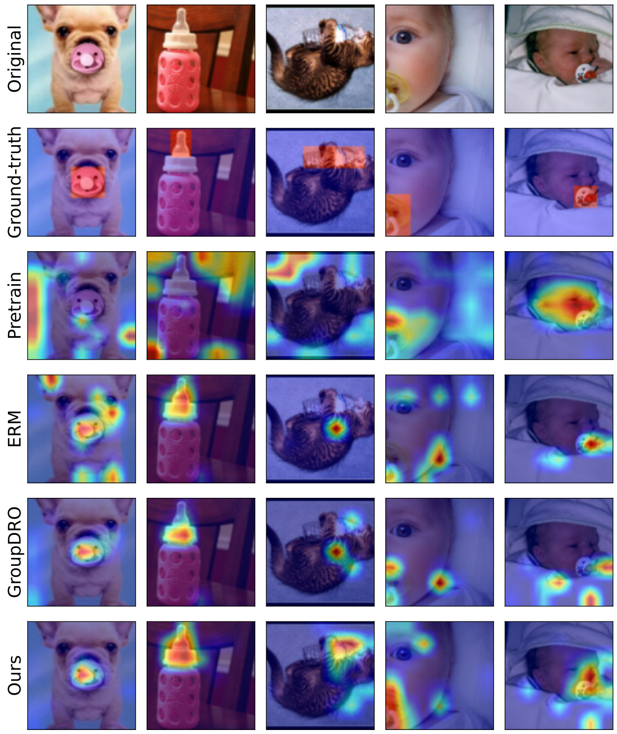

Figure 6. GradCAM explanations for different approaches based on CLIP RN50 for the ImageNet dataset.

When comparing these findings between the two different model backbones we observe that AIoU scores for the ViT model are higher for all methods than their corresponding ResNet versions, indicating that the larger transformer-based model is perhaps more prone to improve upon using such mitigation techniques or even standard fine-tuning.

ImageNet-1K. Here, we choose one of the most spurious correlations we found for the CLIP ResNet50: the *baby pacifier* class where the spurious attribute is *baby face*. The accuracy discrepancy between cases when there is a baby and no baby in the image is 69.2% in the validation set, with a worst-group accuracy of 30.8%. The most confusing class for baby pacifier is *water bottle*. For all methods, we then fine-tune the CLIP RN50 model with the training data from these two classes: *baby pacifier* and *water bottle* to understand if such an isolated mitigation could positively align the model. In Table 5 we see that in terms of both average accuracy and worst-group accuracy, the baseline ERM method performs just as well as our methods. However, since the test dataset in this case is rather small (only 50 images per class), it is useful to also look at the alignment of explanations. Figure 6 illustrates this visually, highlighting that GradCAM maps are not focused on the baby face for our approach, which is the case for other methods. The same result is confirmed by the higher AIoU scores.

Table 5. Results of fine-tuning CLIP-RN50 with a subset of ImageNet classes, “baby pacifier” and “water bottle”. Both average and worst-group performance are evaluated with models early stopped at the *highest worst-group accuracy* on the validation set.

Class 680	Accuracy		AIoU	
Baby Pacifier	Avg.	Worst	Avg.	Worst
Pre-trained	73.7%	30.8%	0.651	0.380
Fine-tuned	94.1%	91.7%	0.650	0.571
ERM	**94.9%**	**96.2%**	0.661	0.454
Group DRO	89.6%	93.1%	0.661	0.568
Ours($\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{ls}$)	**94.9%**	**96.2%**	**0.720**	**0.645**

Limitations. While the method proposed here shows promising results for mitigating spurious correlations, learning pipelines often face a combination of problems that go beyond spurious features and involve other out-of-distribution shifts. We illustrate these concerns through a running example from ImageNet in Appendix B and show that current decorrelation methods may not be sufficient when models deal with issues such as high concept variation, insufficient data, label noise, or visual commonalities between spurious and non-spurious features.

6. Conclusion and Future Work

We proposed a language-based approach to mitigate spurious correlations of CLIP, as a contrastive learning vision and language model. Our focus on mitigations that can be initiated through language is motivated by the fact that spurious attribute annotations may not always be available. The contrastive loss function formulation guiding the spurious attribute decorrelation is applied at fine-tuning time and is effective even when the language and image encoders are excluded from the fine-tuning process. Besides the computational convenience, this is a promising finding speaking to the foundational nature of the larger representations. The work opens up several questions for future research, including the scalability of such methods when mitigating several spurious correlations at the same time. While this work focused on spurious correlations for classification tasks, studying the problem from a representational bias perspective and how spurious correlations may feed issues in representation fairness is an important relevant direction with several societal implications. Finally, we see opportunities in further leveraging model multi-modality and language-initiated mitigation actions to either generate teaching samples for mitigations or high-level instructions for the model to follow.

Acknowledgment. This research was partially supported by Cisco Systems and the National Science Foundation CAREER Award 2146492.

References

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., and Whang, S. E. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1550–1553. IEEE, 2019.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=FPCMqjI0jXN`.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=wKhUPzqVap6`.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7521–7531, 2022.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.

Moayeri, M., Pope, P., Balaji, Y., and Feizi, S. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19087–19097, 2022.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Nguyen, G., Kim, D., and Nguyen, A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=OKPS9YdZ8Va`.

Nushi, B., Kamar, E., and Horvitz, E. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pp. 126–135, 2018.

Petryk, S., Dunlap, L., Nasseri, K., Gonzalez, J., Darrell, T., and Rohrbach, A. On guiding visual attention with language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18092–18102, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on ImageNet. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8634–8644. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/shankar20c.html`.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Singla, S., Nushi, B., Shah, S., Kamar, E., and Horvitz, E. Understanding failures of deep networks via robust feature extraction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 2021.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yang, Y., Kim, S., and Joo, J. Explaining deep convolutional neural networks via latent visual-semantic filter attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8333–8343, 2022.

Yosinski, J., Clune, J., Nguyen, A. M., Fuchs, T. J., and Lipson, H. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*, 2016.

Zhang, M. and Re, C. Contrastive adapters for foundation model group robustness. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=uPdS_7pdA9p`.

Zhang, Q., Yang, Y., Ma, H., and Wu, Y. N. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6261–6270, 2019.

Zhang, Y., HaoChen, J. Z., Huang, S.-C., Wang, K.-C., Zou, J., and Yeung, S. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=D-zfUK7BR6c`.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049, 2020.

Table 6. Investigating the impact and necessity of different loss terms with CLIP-RN50 and Waterbirds. The last four columns show the AIoU scores on the four groups (e.g. WB-L is waterbird with a land background.)

	Accuracy		AIoU		AIoU			
$\mathcal{L}_{\text{CLIP}}$+	Avg.	Worst-group	Avg.	Worst-group	LB-L	LB-W	WB-L	WB-W
(1) $\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{vs}$+$\mathcal{L}_{ls}$	86.9%	**78.2%**	0.550	0.281	0.244	0.281	0.852	0.823
(2) $\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{vs}$	83.2%	**77.5%**	**0.654**	**0.587**	0.601	0.587	0.714	0.715
(3) $\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{ls}$	84.7%	**77.5%**	**0.628**	**0.499**	0.460	0.499	0.788	0.764
(4) $\mathcal{L}_{lc}$+$\mathcal{L}_{vs}$	83.8%	72.0%	0.525	0.310	0.329	0.310	0.762	0.697
(5) $\mathcal{L}_{lc}$+$\mathcal{L}_{ls}$	**88.0%**	75.1%	0.608	0.493	0.446	0.493	0.776	0.719
(6) $\mathcal{L}_{vc}$+$\mathcal{L}_{vs}$	86.8%	76.1%	0.609	0.461	0.441	0.461	0.768	0.763
Pre-trained CLIP	**90.8%**	44.9%	0.507	0.479	0.471	0.479	0.541	0.537
Fine-tuned CLIP	81.3%	77.1%	0.510	0.128	0.098	0.128	0.929	0.885
ERM	93.5%	54.4%	0.514	0.139	0.187	0.139	0.920	0.919
Group DRO	83.3%	73.7%	0.509	0.274	0.219	0.274	0.800	0.741

Table 7. Results of fine-tuning CLIP-RN50 with a subset of ImageNet classes, "can opener" and "letter opener". Both average and worst-group performance are evaluated with models early stopped at the *highest worst-group accuracy* on the validation set.

Class x	Accuracy		AIoU	
Can Opener	Avg.	Worst	Avg.	Worst
Pre-trained	86.4%	75.0%	0.568	**0.464**
Fine-tuned	76.8%	68.0%	0.561	0.339
ERM	80.1%	68.0%	0.436	0.295
Group DRO	78.4%	76.0%	0.426	0.205
Ours($\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{ls}$)	73.9%	68.0%	**0.585**	0.344
Ours($\mathcal{L}_{lc}$+$\mathcal{L}_{vc}$+$\mathcal{L}_{vs}$)	75.7%	72.0%	**0.612**	**0.356**

A. Ablation Study

To better understand the impact and necessity of each loss term we conducted ablation studies for different combinations of components. Table 6 summarizes the findings. First, we see that choosing between the spurious image loss (row 2) or the spurious language loss (row 3) leads to similar results in worst-group accuracy but AIoU scores are higher for the spurious image loss, perhaps because decorrelation directly in the image space, but of course this requires attribute annotations. Second, the image contrastive loss is necessary for any of the spurious losses to be effective. For example, when we compare rows 2 and 3 with their corresponding versions in rows 4 and 5 that do not have the contrastive image loss, we see both worst-group acuraccy and AIoU decreasing. While this dependency on the contrastive image loss calls for more investigation, it also shows evidence that for improving vision classification results, the process of separating spurious attributes needs to happen hand in hand with separating classes from each other.

B. Limitations

While our method has demonstrated promising results in mitigating spurious correlations in ImageNet, there are still limitations to consider. We have identified several circumstances where our method may not be effective, which we discuss in more detail below.

B.1. Spurious-aware Contrastive Language Image Fine-tuning

We will describe these limitations by using the example presented in Figure 7. In this example, the target class we are studying is "can opener" and the spurious feature being identified by our method is the "can" itself. The spurious correlation leads to a 45.2% discrepancy in accuracy and as shown in Figure 3, the model often focuses on the can rather than the opener.

High Concept Variation + Insufficient Data. One limitation is that the effectiveness of our method can depend on the level of variation within each class. For example, some classes may have many different variations, such as different breeds of dogs or types of flowers, while other classes may have less variation, such as types of paperclips or staplers. It is essential to note that increased variation within a class can pose a significant challenge in the learning process, even without spurious correlations. The presence of spurious correlations however can exacerbate these issues or even hide them when models are right for the wrong reasons. For example, the model can still be correct on an image with an unusual class concept variation when the spurious correlation is present (see the last two examples in Figure 8). Therefore, while correcting spurious correlations is vital, it alone may not address the inherent problem of object variation but the identification stage for spurious correlations itself may still be informative. Additionally, if there is insufficient data for a particular variation, our method may be unable to learn that variation effectively. For example, Figure 8) shows that even though our approach tends to put less emphasis on the "can" itself, it still cannot completely alleviate the problem or guarantee that the focus of the model will not shift to other parts of the image that are still spurious.

The same discussion applies to label noise, which can also be compounded by increased class variation or label subjectivity (Shankar et al., 2020).

Feature Definition Challenges. Our method relies on being able to identify and isolate spurious features from non-spurious ones. However, in some cases, the spurious feature may be difficult to define or may have parts that are similar to non-spurious features. For example, in the case of a can and a can opener, the metal parts of each object may look similar and contain similar features (Figure 9). This can make it more challenging to separate the spurious correlations from the legitimate ones.

Due to these challenges in combination, in Table 7 we see that almost all methods, including ours, are not able to improve the accuracy for the "can opener" class. It is important to note that these limitations are not unique to our method, but rather are common issues that arise with spurious correlation mitigation methods and their interactions with general learning objectives. Nevertheless, the approach proposed in this work can be beneficial in many cases, in particular when significant performance drops are mainly due to spurious correlations and when spurious features do not share a visual commonality with non-spurious ones. Further research is needed to address these limitations and improve the effectiveness of spurious correlation mitigation methods in combination with other forms of out-of-distribution shifts on concept variation.

B.2. Spurious Correlation Detection

One limitation of our spurious correlation detection method is that it relies on the object detector to identify relevant attributes. However, certain factors such as lighting and contrast may not be captured by the object detector, which can limit the effectiveness of our approach. To mitigate this limitation, our method can be used in conjunction with previous work that utilizes system/content metadata or discovered visual features for general failure analysis (Nushi et al., 2018; Singla et al., 2021; Chung et al., 2019; Jain et al., 2022; Eyuboglu et al., 2022). For instance, one can enrich the test data with additional meta-data, such as contrast, blur, lighting, and camera angle, and apply our method as well as previous approaches to understand if the model's performance drops for some of these conditions.

Note that our method leverages language attributes to describe spurious correlations but does not exclude the use of other metadata for the same purpose. While language attributes can capture a broad range of failure modes, incorporating all relevant information is crucial for comprehensive debugging.

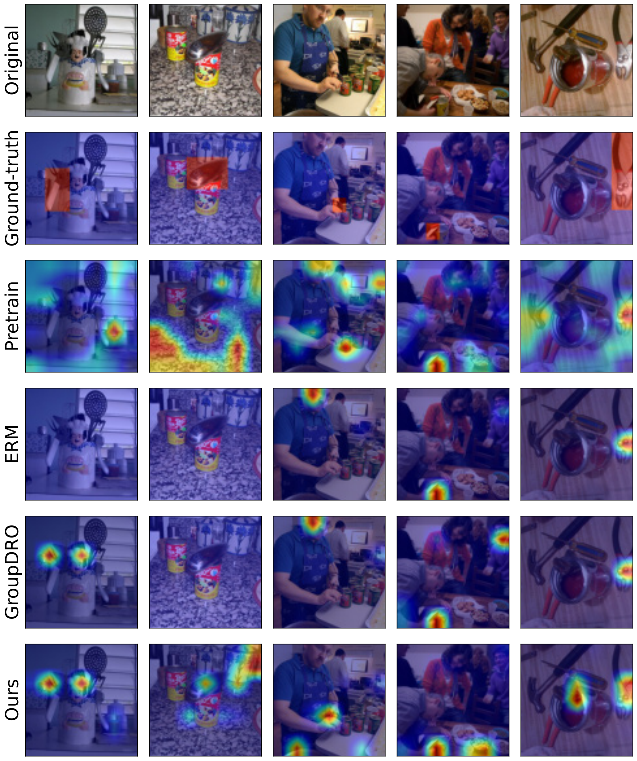

Figure 7. GradCAM explanations for different approaches based on CLIP RN50 for the ImageNet dataset, "can opener" class.

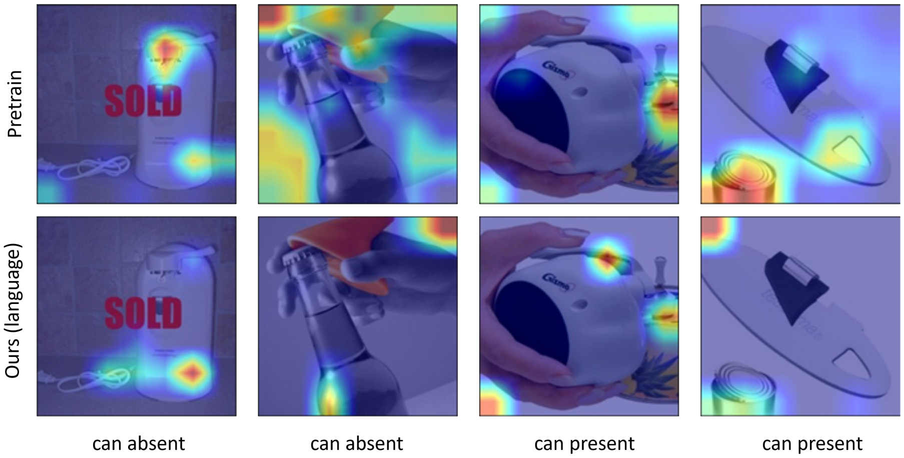

Figure 8. The "can opener" class in ImageNet has several concept variations, posing additional challenges in the learning process. The presence of spurious correlations can in addition exacerbate these issues or even hide them when models are right for the wrong reasons (e.g., the last two examples in this figure). While our method reduces the focus on "cans", it is still not able to completely alleviate the problem for examples with unusual concept variation.

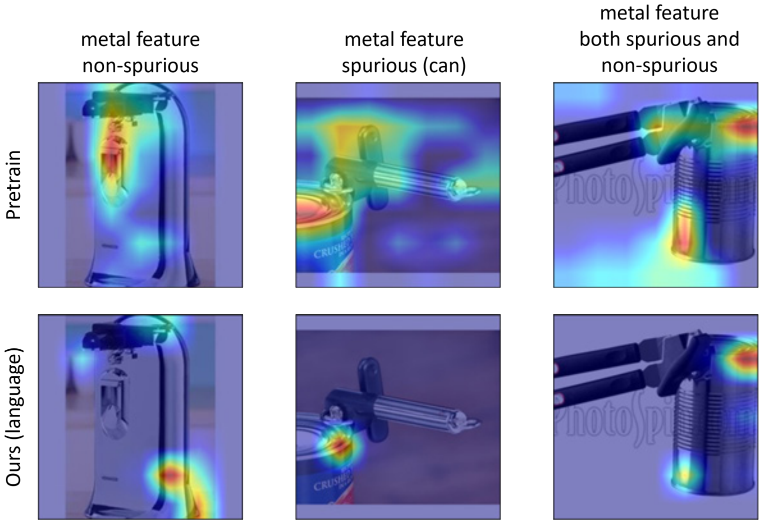

Figure 9. Metal features of can openers share visual commonalities with cans, which makes the problem of mitigating spurious correlations more difficult for such cases.