Prompting is a Double-Edged Sword: Improving Worst-Group Robustness of Foundation Models

Amrith Setlur [*][1] Saurabh Garg [*][1] Virginia Smith [1] Sergey Levine [2]

Abstract

Machine learning models fail catastrophically under distribution shift, but a surprisingly effective way to empirically improve robustness to some types of shift (*e.g.*, Imagenet-A/C) is to use stronger open-vocabulary classifiers derived from foundation models. In this work, we first note that for shifts governed by spurious correlations (features spuriously correlated with the label on the training data, but not on test), the zero-shot and few-shot performance of foundation models is no better than ERM models, and remains unchanged when pretrained data/model size is scaled. Secondly, even in these situations, foundation models are quite accurate at predicting the value of the spurious feature. In a simplified setup, we theoretically analyze both these findings. Specifically, we show that during contrastive pretraining, the simplicity bias of foundation models tends to result in the learning of features that mostly rely on the spurious attribute, compared to more robust features. We leverage these observations to propose Prompting for Robustness (PfR) which first uses foundation models to zero-shot predict the spurious attribute on labeled examples, and then learns a classifier with balanced performance across different groups of labels and spurious attribute. Across 5 vision and language tasks, we show that PfR's performance nearly equals that of an oracle algorithm (group DRO) that leverages human labeled spurious attributes[1].

1. Introduction

Machine learning classifiers are often trained on datasets with hidden confounders that are spuriously correlated with the label. For example, waterbirds tend to occur in pictures with water backgrounds (Blodgett et al., 2016; Barocas & Selbst, 2016; Hovy & Søgaard, 2015; Tatman, 2017). Empirical risk minimization (ERM) latches onto these confounders and can consequently fail catastrophically on underrepresented (minority) groups where the confounder is uncorrelated with the label, *e.g.*, pictures of waterbirds on land (Lyu et al., 2021; Shah et al., 2020; Nagarajan et al., 2020). Numerous algorithmic interventions have been proposed to make ERM models more robust to such confounders (e.g., Ben-Tal et al., 2013; Arjovsky et al., 2019).

On the other hand, with the advent of foundation models trained on heterogeneous datasets, we are observing a paradigm shift in how we learn classifiers. Driven by their unprecedented zero-shot prediction capabilities, the common strategy of learning classifiers has been to simply prompt models with class names directly (Wei et al., 2020; Brown et al., 2020). In fact, zero-shot prompting sometimes yields classifiers that are more robust than ERM classifiers trained on downstream data (Hendrycks et al., 2020; Fang et al., 2022), *e.g.,* as seen in robustness gains observed on benchmarks like ImageNet with distribution shifts (Radford et al., 2021). However, as we show in our work, such gains do not proportionately transfer to other forms of distribution shift such as when confounders that are highly predictive of the label in training distribution are no longer correlated with the label on test (Yang et al., 2023; Tu et al., 2020; Hall et al., 2023). Thus, robustness to hidden confounders in the training data remains an open challenge.

In this work, we aim to improve the performance of foundation models on paritions of the distribution (groups) where the confounder is not correlated with the label (minority group). One way is to incorporate downstream labeled data. Unfortunately, unless we have access to deconfounded data (without the spurious correlation), simply fine-tuning naïvely would result in the same issues as standard ERM training, as we confirm experimentally. However, with open-vocabulary foundation models, we can provide for robustness by *telling* the model about the confounder directly (i.e., by describing it in a prompt). One natural way to use this knowledge is to incorporate the description into the classification prompt. However, we observe that even this doesn't improve zero-shot robustness (see Sec. 3.2).

[*]Equal contribution [1]Carnegie Mellon University [2]UC Berkeley. Correspondence to: Amrith Setlur <asetlur@cs.cmu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] The code and datasets are available here.

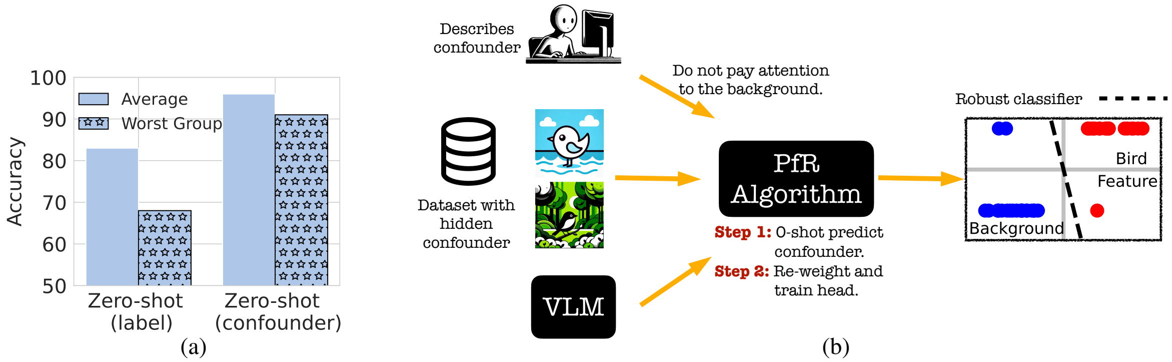


Figure 1: (a): *Foundation models are not robust to spurious correlations, but can predict them*; Averaged across four tasks with spurious correlations, we see that while zero-shot foundation models perform much worse on groups where the spurious correlation is absent, they are highly accurate at predicting the spurious attribute itself, across all groups. (b): *Prompting for Robustness (PfR)*: Leveraging findings in (a), we propose prompting for robustness (PfR), that learns robust classifiers from foundation models in two steps. In Step 1, armed with a text description of the spurious feature, PfR prompts foundation models to zero-shot predict the spurious attribute on a labeled dataset with spurious correlations, and in Step 2 it learns a robust classifier by minimizing worst group loss, across groups given by the combination of the predicted attribute and label.

We make an intriguing observation: while foundation models are not robust zero-shot classifiers of the true label, they perform remarkably well in predicting the *presence* of spurious attributes. Moreover, we observe that while scaling up the model size and pretraining data does not improve the performance of label prediction on minority groups, the worst group performance of spurious attribute prediction does. Motivated by these findings, we propose a simple technique that we call *Prompting for Robustness (PfR)*. PfR learns robust classifiers for downstream tasks with a few labeled examples and a language description of the confounding attribute. PfR first uses the language description to prompt for a zero-shot classifier that accurately predicts the spurious feature on each labeled example. The value of the label and the predicted confounder jointly define a set of disjoint groups in our data. Then, a robust predictor is learnt by minimizing worst group loss, similar to group DRO, as described by Sagawa et al. (2019), but without ground-truth knowledge of examples in the minority group. This simple method yields surprising performance gains of $\geqslant 40\%$ (averaged across datasets) relative to zero-shot performance of foundation mdoels and ERM on downstream data alone. We further illustrate the applicability of our findings by showcasing its efficacy in extracting group annotations for auditing zero-shot (or ERM) models to assess their robustness. Specifically, we prompt GPT-4V to annotate Chest-Xray 14 dataset (Wang et al., 2017b) for the presence of chest drains (the spurious attribute) and observe a significant robustness gap among ERM models.

Finally, in a simplified setup for multimodal contrastive pretraining, we show that when the spurious correlations in the downstream task are also present in the pretraining distribution over image, and text pairs, then contrastive pretraining learns: (i) image features that couple the spurious feature with other robust features, while placing a higher weight on the spurious one; and (ii) text features that are almost identical for the text descriptions of the label and the spurious attribute. As a consequence of this, we prove that even with infinite pretraining data, the zero-shot performance for the pretrained model would be provably worse than random on examples where label and spurious attributed are uncorrelated. On the other, when it comes to predicting the spurious attribute it has almost perfect accuracy on all examples — precisely the observations we make empirically as well.

In summary our key contributions are as follows. First, we study the performance of foundation models across five vision and language classification tasks with hidden confounders, and observe that while foundation models have poor zero-shot performance on minority examples (that does not improve with scale), they are accurate at predicting the value of the confounder. Second, we leverage this finding to propose a new and simple method: PfR which first zero-shot predicts the confounder when given a text description of it, and then learns a robust classifier across predicted groups. Theoretically, we tie the performance of PfR to the zero-shot accuracy of foundation models on tasks with spurious correlations. Thus, in a simplified setup we provide a theoretical analysis for the zero-shot performance of solutions learned by multimodal contrastive pretraining, and reconcile our theoretical insights with practical findings. Empirically, we show PfR's worst group performance nearly matches the oracle (group DRO) on all datasets.

2. Problem setup

We aim to study the robustness of zero/few-shot foundation models, to distribution shifts in classification tasks with spurious correlations. We ground this statement more formally by first defining the task distribution, the model of distribution shift, and what it means to be robust to it.

For a classification task, we use $\mathcal{X}$ to denote input text/image and $\mathcal{Y}$ for the set of labels. Additionally, we also define a set $\mathcal{C}$ of spurious attributes (or confounders). With $\mathcal{G} =: \{G_1, G_2, \ldots, G_k\}$, we define a set of disjoint subsets of $\mathcal{X} \times \mathcal{Y} \times \mathcal{C}$ where each G_i has distribution $P_i(x, y, c)$. Then, our task distribution is a mixture of distributions over $\mathcal{G}$, i.e, $\sum_i \alpha_i P_i(x, y, c)$ where α_i is the proportion of data from each group. In particular, each group G_i corresponds to a unique pair of label and confounder values (y_i, c_i), *i.e.,* $\mathbb{1}((x, y, c) \in G_i) = \mathbb{1}(y = y_i)\mathbb{1}(c = c_i)$. When the label y and spurious attribute c are heavily correlated, a classifier that *only* learns the spurious feature c can easily predict the label y. But, this creates a performance disparity across groups where correlations do not hold. For *e.g.,* in Waterbirds (Sagawa et al., 2019), the spurious attribute is the background of the bird, the labels are the category of the bird (landbird vs waterbird) and the groups are defined over the joint space of the bird category and its background.

Under distribution P, the average error of a label classifier f is $\mathrm{err}_{\mathrm{y}}^{\mathrm{avg}}(f) =: \mathbb{E}_P\left[\mathbb{1}(f(x) \neq y)\right]$ and spurious atribute classifier g is $\mathrm{err}_{\mathrm{sp}}^{\mathrm{avg}}(g) =: \mathbb{E}_P\left[\mathbb{1}(g(x) \neq c)\right]$. Similarly, their corresponding worst-case error counterparts, taken over groups is: $\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) =: \max_{G \in \mathcal{G}} \mathbb{E}_{P|G}\left[\mathbb{1}(f(x) \neq y)\right]$ and $\mathrm{err}_{\mathrm{sp}}^{\mathrm{wg}}(g) =: \max_{G \in \mathcal{G}} \mathbb{E}_{P|G}\left[\mathbb{1}(g(x) \neq c)\right]$. We define the *robustness gap* as the difference between the average case and worst case performance. Consequently, a classifier with low robustness gap for label prediction performs similarly on any distribution that only reweights group proportions α_i. Alternatively, robustness to such group shifts is achieved by having a low robustness gap.

In this work, our goal is to learn a label classifier with (i) high average accuracy, and (ii) low robustness gap. For this, we assume that we are given a text description t_c of the confounder c, along with a few *i.i.d.* labeled samples $\mathcal{D}$ from $P(x, y)$. Unless specified otherwise, we assume that group annotations are not given to us. Finally, we use FM to denote a foundation model, whose zero-shot prediction of the spurious attribute in input x is $\mathrm{FM}(x, t_c)$.

3. Zero-shot robustness of foundation models

In this section, we examine the zero-shot performance of open-vocabulary foundation models on common benchmarks for spurious correlations with known confounders.

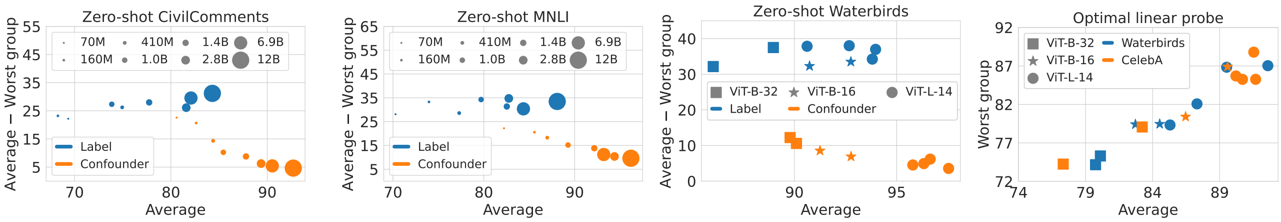
We find that the zero-shot performance of foundation models suffers from a large robustness gap, indicating a substantial difference between average-case and worst-group performance (also demonstrated by Yang et al. (2023); Tu et al. (2020); Lee et al. (2023)). As we increase the scale of pretraining datasets for foundation models, although the models might become better, the robustness gap stays the same or widens, indicating that scale alone does not provide robustness to confounders. Subsequently, we experiment with incorporating a natural language description of the spurious attribute when prompting the model to predict the label. Our findings indicate that while the inclusion of spurious attribute descriptions through naïve zero-shot prompting does not yield improvements, these models demonstrate high accuracy in predicting the presence of the spurious attribute itself. Building on these findings, we propose our method, Prompting for Robustness (PfR), in Section 5.

Finally, we test our observation of identifying spurious correlations using foundation models on a practical medical diagnosis task. In particular, we annotate Chest Xray-14 (Wang et al., 2017a) dataset for the presence of chest-drain which is a known spurious correlation when predicting pneumothorax (Oakden-Rayner et al., 2020). On the groups annotated by GPT-4V (Achiam et al., 2023), ERM models trained on MedCLIP features (Wang et al., 2022) show large difference between average case and worst-group performance. This exemplifies GPT-4V's ability to identify performance imbalances from descriptions of spurious attributes.

3.1. Setup

Datasets. We experiment with datasets in both language and vision modalities. For language, we experiment with: (i) MNLI (Williams et al., 2017), where the prediction task is relationship between two input sentences: contradiction, entailment, or none of the two. Here the spurious attribute is the presence of negation words, *e.g.*, 'no', and 'never'; (ii) CivilComments (Borkan et al., 2019; Koh et al., 2021), where the task is toxicity prediction and the spurious correlation lies with the underlying attribute annotating the comment, *e.g.*, male vs. female, Christian vs. Muslim, *etc*. For vision, we experiment with: (i) Waterbirds (Sagawa et al., 2019), where the prediction task is waterbird vs. landbird, and the spurious attribute is the background of the image (*i.e.*, land versus water background); (iv) CelebA (Sagawa et al., 2019), where the prediction task is gender and the spurious attribute is the color of hair. We also experiment with the CXR-drain dataset introduced in Sec. 3.3.

Experimental setup. For our zero-shot probing results on vision datasets, we experiment with CLIP model family (Radford et al., 2021; Gadre et al., 2023). For language, we use Llama-2 (Touvron et al., 2023) and Pythia model families (Biderman et al., 2023). We mainly pick these publicly available models for their reasonably good performance on standard benchmarks. Additionally, we can vary the model and pretraining dataset sizes for each family. For our ERM experiments, we train linear classifiers on the penultimate layer outputs, and for our zero-shot probes, we leverage standard prompts commonly used in the literature.



(a) (b) (c) (d)

Figure 2: *Robustness gap versus average performance as pretraining data and model sizes increase.* We observe that while the robustness gap for confounder prediction decreases the gap between average and worst case increases or remains the same for label prediction.

Precise details about prompts used for each dataset are in App. C. We report: (i) prediction accuracy of the label on the worst-group (across combinations of spurious attribute and label values), and (ii) average performance across groups. We also evaluate the performance of predicting the spurious attribute with zero-shot probes.

3.2. Observations

Large zero-shot performance gap between the average and worst group. Zero-shot results are in Table 2. When evaluating CLIP L/14 models on vision datasets, a notable drop of 32% is observed between average and worst group accuracy on Waterbirds dataset, and a drop of 3.5% is observed on CelebA. Turning to language datasets, the evaluation of the Llama-2 13b model indicates a significant 25% performance decline in CivilComments and a 7% drop in MNLI. Notably, the drops observed here are similar to the performance drops observed with models trained with ERM on their corresponding labeled data (Sagawa et al., 2019; Idrissi et al., 2022). The decline seen with ERM models is typically ascribed to the existence of hidden confounders in the training data (Sagawa et al., 2019), suggesting that pretraining datasets also frequently suffer from analogous spurious correlations. We formalize this intuition in Sec. 4.

Incorporating the group description naïvely does not help out of the box. We incorporate spurious attribute description in our zero-shot prompt to predict the label and the spurious attribute jointly. Results are shown in Table 1. However, the zero-shot performance for the worst-case group doesn't improve – there is less than a 1% change between the zero-shot and zero-shot with spurious attribute description rows in Table 1. We also evaluated other variants, where we explicitly instructed the model to ignore spurious attributes, but this did not substantively impact worst-group performance (details are in App. C).

Foundation models are surprisingly good at predicting the presence of hidden confounders. Results are in Table 1. Instead of incorporating spurious attribute description together with the label, we experiment with predicting the presence of a spurious attribute alone. On all standard spurious correlation benchmarks, we observe that the average performance of predicting the presence of the spurious attribute is around 95% with a similar worst-case group performance. This consistent performance is observed across different groups, emphasizing that, despite foundation models exhibiting significant robustness gaps in the joint prediction of spurious attributes and labels, the predictive accuracy for spurious attributes alone remains superior.

Scaling pretraining datasets and models does not improve zero-shot group robustness. The scaling trend results are presented in Fig. 2 (a)-(c), showcasing the performance plotted on average against the difference between average performance and worst-case performance. We analyze this difference in comparison to the average case for both zero-shot label and spurious attribute prediction. As we scale up the pretraining datasets and models, we observe that while the difference reduces for the cofounder prediction, the difference doesn't improve for the label prediction task. This highlights that the prediction performance on standard spurious correlation benchmarks don't improve with scaling and will require post-training interventions.

Scaling pretraining datasets and models does improve underlying representations. As expected we observe that the average and worst-case accuracy (trained with DRO on downstream labeled data) improves as we increase the scale of model size and pretraining data (Fig. 2 (d)).

3.3. CXR-Drain: Annotating confounders with GPT-4V

Previously, we used ground-truth annotations of spurious attributes to establish the high zero-shot accuracy of foundation models in predicting them. Now, we start with a language description of a real-world spurious attribute in a medical task and annotate examples by prompting GPT4-V with this description. By first predicting the spurious correlation in a zero-shot way on a task where annotations are not public (not even for validation), and then showing the performance imbalance of ERM models across annotated groups, we validate the ability of foundation models

Prompt	Predict	Waterbirds		CelebA		CivilComments		MNLI	
		WG	Avg	WG	Avg	WG	Avg	WG	Avg
Is this label L?	L	59.38	91.97	77.69	81.11	59.25	85.75	76.54	84.79
Is this label L? Ignore confounder C.	L	61.37	92.58	86.73	90.28	52.81	87.41	77.95	80.56
Is this label L and confounder C?	L,C	57.38	88.15	78.54	83.11	54.29	86.60	75.73	82.91
Is this confounder C?	C	90.55	96.33	95.01	99.15	86.73	92.70	92.37	96.19

Table 1: *Naively incorporating the confounder description into the label classification prompt does not improve robustness.* Results with leveraging natural language description of the group and label for zero-shot classification.

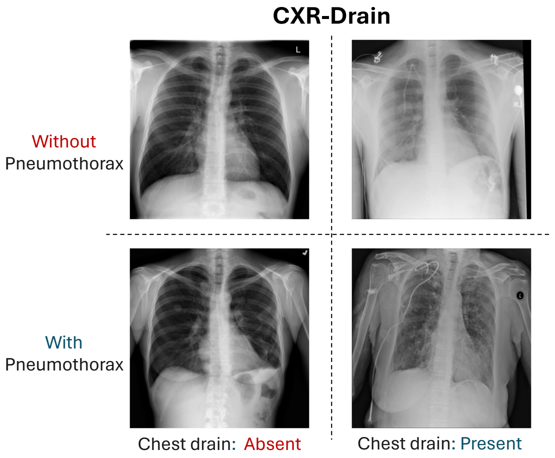


Figure 3: Samples from Chest-Drain dataset in each category. The presence of a drain is identified by prompting GPT-4V. Fig. 5 shows an annotated chest drain image.

to detect the presence of spurious features in practice. We annotate 2400 images from Chest Xray-14 dataset (Wang et al., 2017b) for the presence of chest drain with GPT-4V (details in App. C.3). On this dataset, the goal is to predict the whether the patient suffers from pneumothorax disease given their chest x-ray image and the presence of a chest tube in the chest cavity acts as a confounder. It is noteworthy that while previous studies have underscored the issue of spurious correlations in pneumothorax prediction (Oakden-Rayner et al., 2020), the spurious attributes pertinent to this task are not openly available. We refer to the subset of Chest Xray 14 with annotated spurious attributes as CXR-Drain.

While the annotations obtained with GPT-4V are expected to be noisy (different from ground truth annotations for the presence of chest drain), we observe that models trained with ERM show a significant performance gap on the constructed CXR-Drain dataset (Table 2). Next, we also note that CXR-drain differs from existing semi-synthetic spurious correlation benchmarks, e.g., the worst group is not the minority group which, and hence, re-weighting based methods (Idrissi et al., 2022; Kirichenko et al., 2022) that simply re-weight different groups may perform poorly when compared with DRO. Due to its unique properties, we believe that CXR-drain will also serve as a crucial benchmark for future research on spurious correlations, and we publicly release the dataset here.

4. Theoretical analysis of multimodal contrastive pretraining

In Sec. 3, we empirically identified that the worst group zero-shot performance for predicting the label of a task (with hidden confounders) never improves with scale. So, why does the worst-group performance for confounder prediction improve? In this section, we analyze both these trends theoretically when the label is correlated with the confounder in the pretraining data, similar to the task (Sec. 2). We analyse multimodal contrastive pretraining, mainly because: (i) we can derive and analyze the closed form solution for the population level contrastive pretraining objective; and (ii) it is commonly used in practice for training vision-language foundation models (*e.g.,* CLIP) that align features of image and text pairs (Radford et al., 2021; Wang et al., 2022).

Broadly speaking, we show that when spurious correlations in the downstream task are also present in pretraining, then contrastive learning learns an image encoder that almost fully *couples* (no linear separability) the spurious feature with other robust features predictive of the label. In this coupling, the component along the spurious feature is higher when the signal-to-noise ratio along the robust feature is relatively poor. Exacerbating this failure, the text encoder learns almost identical features (in ℓ_2) for the confounder and label. We show that even when training on population level pretraining data, the worst group accuracy of zero-shot label predictor is worse than random, while that of zero-shot confounder predictor is nearly perfect. Since the solution learnt on population data is itself "bad", under the following setup, our result highlights a more serious failure of the contrastive objective, than the one typically discussed for ERM (Nagarajan et al., 2020; Sagawa et al., 2020).

Setup. The downstream task T has joint distribution $P(x, y, c)$ over image x, label y, and confounder, where both y and c take values in $\{+1, -1\}$, see (1) for the data model. In this data model, the degree of spurious correlation between label y and confounder c increases when the random variable b is sampled from a Bernoulli with higher mean p. The input x is split into three components $[x_\mathrm{r}, x_\mathrm{c}, x_\mathrm{n}]$, where $x_\mathrm{r} \in \mathbb{R}$ is the robust feature determined solely by y, $x_\mathrm{c} \in \mathbb{R}$ equals the confounder c, and $x_\mathrm{n} \in \mathbb{R}^{d_\mathrm{n}}$ is high dimensional noise independent of y, c.

$$y \sim \mathrm{Unif}\{+1, -1\},\ b \sim \mathrm{Bern}(p),\ c = y(2b-1) \quad (1)$$
$$x_\mathrm{r} \sim \mathcal{N}(y, \sigma_\mathrm{r}^2),\ \ x_\mathrm{c} = c,\ \ x_\mathrm{n} \sim \mathcal{N}(\mathbf{0}_{d_\mathrm{n}}, \sigma_\mathrm{n}^2 \mathbf{I}_{d_\mathrm{n}}).$$

Contrastive pretraining. The pretraining distribution for multimodal learning is denoted by $Q(x,t)$ (with density $q(x,t)$), and is defined over $\mathcal{X} \times \mathcal{T}$, where $\mathcal{X}$ image set and $\mathcal{T}$ is the set of captios (text descriptions) for the images. Contrastive pretraining learns an image encoder $\phi : \mathcal{X} \mapsto \mathbb{R}^k$ and a text encoder $\omega : \mathcal{T} \mapsto \mathbb{R}^k$ by pushing together features of image and text pairs sampled from joint distribution $Q(x,t)$, and pulling apart representations of independently sampled images from $Q(x)$, and texts from $Q(t)$. The pretraining objective is in (2). We can view (2) as the multimodal version of the spectral contrastive loss (HaoChen et al., 2021), which is mathematically equivalent to more general contrastive and non-contrastive objectives (Johnson et al., 2022; Garrido et al., 2022).

$$-2\mathbb{E}_{(x,t)\sim Q}\phi(x)^\top \omega(t) + \mathbb{E}_{x\sim Q}\mathbb{E}_{t\sim Q}(\phi(x)^\top \omega(t))^2 \quad (2)$$

For simplicity, we consider a pretraining distribution $Q(x,t)$ that is most relevant for the downstream task T. Thus, the set of text descriptions $\mathcal{T}$ is: $\{t_{y,1}, t_{y,-1}, t_{c,1}, t_{c,-1}\}$, and the marginal $Q(t)$ is uniform. For the conditionals, given $a \in \{-1, 1\}$, the images are sampled from $Q(x \mid t_{y,a}) = P(x \mid y = a)$, and $Q(x \mid t_{c,a}) = P(x \mid c = a)$. Note that as p in (1) increases, not only does the downstream correlation between label and confounder ($\mathbb{E}_P[yc]$) increase, the overlap between $Q(x \mid t_{y,a})$ and $Q(x \mid t_{c,a})$ in the pretraining distribution also increases.

Zero-shot predictors. In practice, pretrained ϕ, ω are used as zero-shot classifiers by evaluating $\phi(x)^\top \omega(t)$, where t is the labels's text description. Adhering to this, we define zero-shot label classifier $f =: 2 \cdot \mathbb{1}(\phi(x)^\top(\omega(t_{y,1}) - \omega(t_{y,-1})) \geqslant 0) - 1$, and zero-shot confounder classifier $g =: 2 \cdot \mathbb{1}(\phi(x)^\top(\omega(t_{c,1}) - \omega(t_{c,-1})) \geqslant 0) - 1$.

4.1. Key insights and main result.

In Theorem 4.1 we provide an informal statement of our main result on the worst group zero-shot performance of label and confounder classifiers. We note that as the spurious correlation p increases, the worst group error worsens for the label predictor and improves for the confounder predictor.

Theorem 4.1. *(zero-shot robustness) Let the zero-shot label (f) and confounder classifier (g) be obtained by minimizing the loss in (2) on infinite pretraining data for linear functions ϕ, ω. Then, for $\sigma_\mathrm{r} = \Omega(1)$, label classifier is worse than random on the worst group, since* $\mathrm{err}_\mathrm{y}^\mathrm{wg}(f) = {}^1\!/\!_2\,\mathrm{erfc}(-c_1 p \sigma_\mathrm{r})$. *On the other hand, the confounder classifier suffers small error on all groups since* $\mathrm{err}_\mathrm{sp}^\mathrm{wg}(g) = {}^1\!/\!_2\,\mathrm{erfc}(c_2 p \sigma_\mathrm{r})$. *Here, $c_1, c_2 > 0$ are constants.*

Our analysis in Sec. 4.2 shows that the above result is a consequence of (i) image encoder relying more on non-robust x_c compared to robust x_r when σ_r is higher; (ii) text encoder failing to learn separate representations for the label and confounder descriptions.

Intuition. During multimodal contrastive pretraining feature alignment of the image and corresponding text features is achieved when images $x_i, x_j \sim Q(x \mid t)$ sampled from the text have well clustered representations, and the clusters of different text inputs are well separated. Our understanding relies on two key observations. First, when the pretraining distribution replicates the task distribution's spurious correlations (as $Q(x,t)$ does with $P(x,y,c)$), then the clusters learned for the label and confounder necessarily overlap since $Q(x \mid t_{y,a}) \approx Q(x \mid t_{c,a})$ (matches on all but the group where correlation is absent). Thus, given this distribution overlap the optimal text encoder's features for the label and the confounder would be very similar. Second, when the noise along the robust feature σ_r is high, the intra cluster variance along the non-robust feature x_c is relatively lower. This biases contrastive learning to place higher weight on the non-robust feature, in learning features that separate clusters corresponding to the different text inputs with large margins. Together, this would lead to poor robustness for the label predictor, and opposite for the spurious attribute predictor, as we note in Theorem 4.1.

4.2. Optimal solutions for spectral contrastive loss.

In this subsection, we present Theorem 4.2 which states the solutions for the image and text encoders learned by minimizing the objective in (2), for linear ϕ, ω and $k = 2$. In Appendix A we prove results for more general families. We make two observations that are consistent with our intuition above. First, we see that when the noise along robust feature (σ_r) is large, then any increase in spurious correlation (p), increases the optimal image feature weights along spurious atttribute (x_c). Second, we see that the optimal solution for the text learns identical features for label and confounder. Thus, on any group that they disagree, the upweighted x_c feature contributes more to the prediction.

Theorem 4.2 (Optimal linear $\phi^\star, \omega^\star$)**.** *Let $p > 0.5$ and ϕ, ω be linear functions over $\mathcal{X}$, $\mathcal{T}$. Then, $\exists$ constants $c_1, c_2 > 0$ such that for the constraint set: $\int_\mathcal{X} \phi_i^2(x)\,\mathrm{d}Q(x) \leqslant c_1,\ \forall i$ and $\int_\mathcal{T} \omega^2(t)\,\mathrm{d}Q(t) \leqslant c_2,\ \forall i$, and ϕ, ω that are orthogonal in $L^2(Q)$, the optimal solutions for the objective in (2) are*

$$\phi_1 = \left[{}^{\cos(\theta)}\!/\!\sqrt{\sigma_\mathrm{r}^2+1},\ \ -\sin(\theta)\right]^\top,$$
$$\phi_2 = \left[{}^{\sin(\theta)}\!/\!\sqrt{\sigma_\mathrm{r}^2+1},\ \ \cos(\theta)\right]^\top,$$

where $\theta = O\left({}^1\!/\!_{p\sigma_\mathrm{r}^2}\right)$. Also, the text features match for

Algorithm 1 Prompting for Robustness (PfR)

```
Input: Foundation model $\mathrm{FM}$, text description of
counfounder $t_c$, labeled dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$.
Stage I: Predict confounder (spurious attribute)
• Prompt $\mathrm{FM}$ with $t_c$ to get zero-shot head $\mathrm{FM}(\cdot, t_c)$.
• For each datapoint predict confounder $\widehat{c}_i \leftarrow \mathrm{FM}(x_i, t_c)$.
• Partition dataset into set of disjoint groups $\widehat{\mathcal{G}}$ based on
  value of label and predicted confounder: $(y, \widehat{c})$.
Stage II: Optimize worst group loss with DRO
• Learn robust classifier $f$ by minimizing the worst loss
  over predicted groups in (3).
```

label and confounder, i.e., $\omega(t_{y,a}) = \omega(t_{c,a}) = [1, a]^\top$ *for* $a \in \{1, -1\}$.

5. Prompting for Robustness

Our results in Section 3 suggest that zero-shot classification with foundation models often attains high average group accuracy but low worst-group accuracy. However, we note that they are surprisingly accurate at predicting the presence of a confounder. We leverage this finding to propose a simple but effective method: Prompting for Robustness (PfR). PfR learns a robust classifier given a few labeled examples and a text description of the confounder. While standard techniques of using labeled data or foundation model alone fail, we show that PfR efficiently uses both to recover a classifier with worst group performance close to that of methods that have ground truth group information (i.e., Group DRO).

Prompting for Robustness (PfR). PfR (summarized in Algorithm 1) runs in two stages. In the first stage, PfR prompts an open vocabulary foundation model FM with the text description t_c of the confounding attribute and recovers a zero-shot prediction of the confounder c on any given input (for *e.g.,* in the case of CivilComments the confounder is described as “race, religion or gender”). Using this, each training example (x_i), which was previously annotated only for the label of interest (y_i), is additionally annotated with the value of the confounding attribute $(\widehat{c}_i)$ (for *e.g.,* “black/white and christian/muslim”). The training dataset is then split into disjoint groups $\widehat{\mathcal{G}}$ based on the paired value $(y_i, \widehat{c}_i)$ of the label and predicted confounder. In the second stage, PfR learns a robust classifier by minimizing the worst group loss over each predicted group, minimizing:

$$\min_{f} \max_{G \in \widehat{\mathcal{G}}} \; \mathbb{E}\left[\ell(f(x), y) \mid x \in G\right]. \tag{3}$$

The above objective can be optimized with an online algorithm that treats f and G as players in a minimax game, analogously to the group DRO algorithm described by Sagawa et al. (2020). Hence, we reuse their Algorithm 1 to optimize our objective in Equation (3). The key difference between our objective and standard Group DRO is that the latter minimizes worst group loss over ground truth groups obtained by using human annotations of the confounder attribute. Based on our findings from Section 3, we should expect that the confounder can be predicted accurately in zero shot, enabling PfR to possibly match the performance of Group DRO. This is indeed what we will see in experiments.

5.1. PfR is more robust than zero-shot and ERM

On the five datasets we introduced previously, we evaluate the performance of PfR and compare with both zero-shot and few-shot algorithms that have access to a few labels (but not the ground-truth group labels).

Setup and baselines. On the language tasks we use Llama2-7b and Llama2-13b models (Touvron et al., 2023) for zero-shot prediction (reporting max of the two), and on the vision tasks we use CLIP-ViT-L/16 (Radford et al., 2021). We compare PfR with standard ERM and four baseline methods: JTT (Liu et al., 2021), DebiAN (Li et al., 2022), EIIL (Creager et al., 2021), ReBias (Bahng et al., 2020) that were originally proposed to learn robust classifiers without relying on ground truth group annotations. Additionally, we also evaluate on two recent approaches (Yang et al., 2023; Zhang et al., 2022) that specifically aim to robustify training with contrastive learning objectives. For baselines excluding ERM, JTT and Zhang et al. (2022), we only evaluate on vision datasets, since they involve techniques that do not translate easily to language tasks. We also include Group DRO (Sagawa et al., 2019) as an oracle baseline that has access to true group labels. All few-shot methods including PfR are used to train a linear head over fixed features. In the language task we train a linear head on top of features learned by finetuning a RoBERTa encoder (Liu et al., 2019) on the MNLI/CivilComments dataset, and for vision tasks we train a linear head over CLIP’s image encoder.

Results. In Table 2, we compare average and worst group performance for different methods. First, we observe that averaged across datasets, PfR reduced worst group error by 47% compared to zero-shot, and 52% and 30% compared to ERM and JTT, respectively. On some datasets like Waterbirds, the worst group gains are as high as $> 75\%$. More importantly, PfR’s performance closely matches that of the oracle Group DRO algorithm across all datasets. Additionally, unlike overly pessimistic DRO objectives like CVaR-DRO (Hu et al., 2018), the average performance is not significantly compromised from trying to improve worst group accuracy. Thus, we see that PfR learns a classifier robust to spurious correlations without much human annotation overhead beyond a description of the confounder.

Method	Waterbirds		CelebA		CivilComments		MNLI		CXR-Drain	
	WG	Avg	WG	Avg	WG	Avg	WG	Avg	WG	Avg
Zero-shot	59.38	91.97	77.69	81.11	59.25	85.75	76.54	84.79	–	–
ERM	70.71	98.75	54.84	94.96	61.35	92.42	67.30	87.71	51.79	76.10
JTT	85.86	95.47	82.49	92.74	72.73	90.54	72.75	86.73	56.52	77.53
Zhang et al.	86.90	96.20	84.60	90.40	50.10	54.20	69.83	86.59	60.15	78.11
Yang et al.	90.13	95.80	**88.12**	91.64	–	–	–	–	59.37	74.58
ReBias	79.24	95.83	70.79	93.52	–	–	–	–	51.39	78.54
DebiAN	82.36	93.79	74.29	92.76	–	–	–	–	52.91	75.42
EIIL	81.18	96.84	79.53	91.75	–	–	–	–	56.62	79.45
PfR (ours)	**91.05**	94.32	88.05	91.97	**77.83**	88.70	**81.28**	84.60	**68.55**	76.73
Group DRO (oracle)	93.23	94.40	90.79	92.32	80.21	86.52	81.54	84.37	–	–

Table 2: *PfR improves worst group performance over ERM and zero-shot foundation models:* On five benchmarks from Section 3 we evaluate average and worst-group performance of PfR and compare it with baselines JTT, ERM, and zero-shot.

5.2. Comparing PfR with in-context learning

For language tasks, in-context learning (ICL) is a commonly used few-shot method to improve performance when zero-shot methods are poor (Brown et al., 2020). In ICL, some labeled training examples are fed along with a language description of the classification task to large language models (*e.g.,* GPT-3.5, Llama). Since PfR also uses labeled examples, we compare our method with ICL on CivilComments and MNLI (see Fig. 4). We observe that while ICL improves over zero-shot inference on average, the worst-group performance remains almost unchanged for CivilComments and worsens for MNLI. We can therefore see that ICL is not a viable alternative to PfR. One reason for why ICL can hurt worst group performance is prior works have shown ICL in language models to make predictions consistent with ERM models trained with gradient descent (Ahn et al., 2023; Akyürek et al., 2022; Von Oswald et al., 2023). Since such ERM models are known to latch onto spurious correlations in the training data (Shah et al., 2020; Nagarajan et al., 2020), we would expect ICL to improve average performance at the expense of worst group performance.

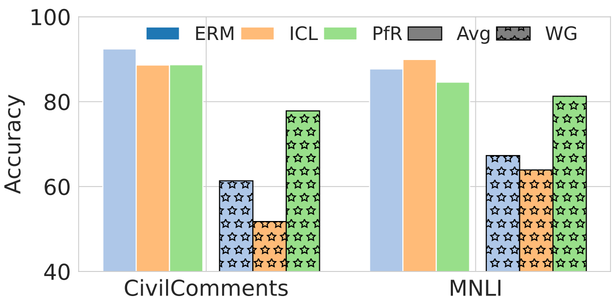


Figure 4: *In-context learning with 128 examples does not improve robustness gap, instead hurts it:* Average and worst-group performance of ICL, ERM and PfR on language tasks.

5.3. Theoretical analysis of PfR

PfR relies on foundation models to accurate predict the confounding attribute (Sec. 3), even when they cannot in zero shot disentangle this confounder from the class label. Given the description t_c, the confounder prediction error suffered by the zero-shot model in the first stage of PfR is $\mathrm{err}_c(\mathrm{FM}(\cdot, t_c))$. In Theorem 5.1 we provide worst-group generalization error guarantees for PfR (proof in Appendix B). Our shows that the worst group accuracy of PfR is upper bounded by two terms. The first term is the generalization error suffered by the oracle algorithm (Group DRO), and the second is the zero-shot error in predicting the confounder. Thus, as the the zero-shot accuracy of confounder prediction improves, it linearly affects worst-group error guarantees for PfR.

Theorem 5.1 (PfR's worst group error)**.** *For PfR output $\hat{f}$, w.h.p. $1-\delta$, worst group generalization error of $\hat{f}$ is $\lesssim \sqrt{\log \mathfrak{C}(\mathcal{F})K/\delta/n} + \mathrm{err}_c(\mathrm{FM}(t_c))$, where $\mathfrak{C}(\mathcal{F})$ is complexity of $\mathcal{F}$, K is number of groups and latter term is* FM*'s zero-shot performance on confounder prediction.*

6. Related Work

Zero-shot and few-shot robustness of foundation models. There has been a recent growth in the capabilities of pretrained *open vocabulary models* (Radford et al., 2021; Jia et al., 2021; Brown et al., 2020; Chowdhery et al., 2023; Rombach et al., 2022; Alayrac et al., 2022; Wei et al., 2021). In vision modality, models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) offer unprecedented zero-shot capabilities simply by assessing the relative compatibility of a given image with an arbitrary set of textual "prompts" Radford et al. (2021). For language modality, large language models have shown unprecedented capabilities on a wide range of tasks despite not being trained explicitly to do many of those tasks (Brown et al., 2020;

Chowdhery et al., 2023; Touvron et al., 2023; Wei et al., 2021; 2022). More recent GPT4-V (Bubeck et al., 2023) and Flamingo (Alayrac et al., 2022) models can take interleaved image-text input to generate text output. However, these models do suffer from robustness problems. For example, existing works have shown that during fine-tuning, the performance of models on distributions away from training data drops (Wortsman et al., 2022; Goyal et al., 2023; Zhang et al., 2022), including the scenarios where the downstream data contains spurious correlations (Yang et al., 2023; Tu et al., 2020; Hall et al., 2023; Lee et al., 2023). We evaluate zero-shot robustness models to spurious correlations and propose solutions to mitigate the observed robustness gap.

Robustness to spurious correlations. Several prior works use distribution robust optimization (DRO) to learn predictors robust to shifts in an uncertainty set (Ben-Tal et al., 2013; Blanchet & Murthy, 2019; Duchi et al., 2016; Duchi & Namkoong, 2021). For spurious correlation problems that result in more specific group shifts, DRO tends to be overly pessimistic (worse than ERM) (Hu et al., 2018). To address this, previous works assume knowledge of the spurious attribute, and either only minimize worst loss over known groups (Sagawa et al., 2019) or average loss over re-weighted ones (Idrissi et al., 2022; Kirichenko et al., 2022). Since it is restrictive to assume group knowledge, other works used relied on two observations: spurious attributes are easier to learn (than robust features) and ERM suffers from a simplicity bias (Shah et al., 2020; Sagawa et al., 2020). Using this, they either reconfigure DRO's uncertainty set (Setlur et al., 2023) (or make it random (Zhai et al., 2021)), while other works (Liu et al., 2021; Nam et al., 2020) exploit it to recover the hidden minority group with ERM losses. Finally, some other works on robustness to hidden confounders (Sohoni et al., 2021; Bao & Barzilay, 2022; Creager et al., 2021) either rely on dataset dependent heuristics, or the ability to query test samples (Lee et al., 2022). Different from the above, we assume a language description of the confounder (as opposed to groups). Armed with this, we use open vocabulary models to predict the presence of a confounder, and then learn robust predictors with DRO over predicted groups. Thus, while we leverage DRO formulation for robustness guarantees, we also avoid its pitfalls by relying on zero-shot foundation models.

Robustness of self-supervised learning: theoretical analysis. While several works theortically analyze (Tian et al., 2020; HaoChen et al., 2021; Mitrovic et al., 2020; Wang & Isola, 2020; Saunshi et al., 2022; HaoChen & Ma, 2022) models pretrained with contrastive learning, masked image and language modeling, they mainly do this for few-shot in-distribution generalization on downstream tasks. In contrast, there are fewer works that focus on out-of-distribution robustness (Shen et al., 2022; Kumar et al., 2022; HaoChen et al., 2022), and even fewer on robustness to spurious correlations (Garg et al., 2023), and all of them do this for unimodal few-shot settings. In contrast, we theoretically analyse zero-shot generalization for multimodal contrastive learning. (Zhang et al., 2023; Chen et al., 2023) are recent works that also theoretically analyze the multimodal setting, and the former only studies few-shot in-distribution generalization, similar to Lee et al. (2021). Closest to our analysis is Zhang et al. (2023), which analyzes zero-shot performance of CLIP, but unlike us they do not specifically model the pretraining distribution to also include spurious attributes from the downstream task, which we show impacts robustness to spurious correlations.

7. Conclusion and Limitations

In this work, we focus on the robustness of zero-shot models to tasks with spurious correlations. While foundation models have shown unprecedented zero-shot capabilities, we show that these models struggle when confounders lose correlation with labels. To address this, we propose Prompting for Robustness (PfR), leveraging language descriptions to prompt zero-shot classifiers and train robust models. Empirical results reveal significant performance gains in the worst accuracy groups. Overall, this work offers insights and a practical approach to enhance foundation model robustness against hidden confounders, contributing to bias mitigation and improved fairness in machine learning.

There are several directions for future work. Currently, we assume knowledge about what are potential contenders for "spurious attributes". Discovering spurious attributes in an automated manner is an interesting direction for future work. To improve the robustness of the classifier, we need some labeled downstream data for our post-training intervention. Near-perfect zero-shot accuracy in predicting groups, coupled with the presence of a robust linear classifier atop fixed features, hints that we should be able to improve post-training robustness in a zero-shot way This potential improvement represents an intriguing and valuable avenue for future inquiry.

Impact Statement

In this work we study the ability of foundation models to improve robustness to spurious features, and propose approach, PfR, that is highly effective in practice. A key insight in our work is the observation that foundation models are surprisingly good at predicting the presence of hidden confounders. While we explore using this for the beneficial purpose of improving classifier robustness, we note that it may be possible to exploit this same information to attack the model or degrade model performance. Understanding and mitigating such attacks is an important area of study, and we hope our approach provides a simple technique for identifying potential vulnerabilities.

Acknowledgements

The authors would like to thank Aviral Kumar, Aditi Raghunathan, Gaurav Ghosal, Christina Baek, Pratiksha Thaker, Steven Kolawole, Shengyuan Hu and Katie Kang for their thoughtful feedback, comments and discussion.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations, 2020.

Bao, Y. and Barzilay, R. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*, 2022.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California law review*, pp. 671–732, 2016.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, Z., Deng, Y., Li, Y., and Gu, Q. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

Deledalle, C.-A., Denis, L., Tabti, S., and Tupin, F. *Closed-form expressions of the eigen decomposition of 2 x 2 and 3 x 3 Hermitian matrices*. PhD thesis, Université de Lyon, 2017.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021. doi: 10.1214/20-AOS2004. URL https://doi.org/10.1214/20-AOS2004.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Garg, S., Setlur, A., Lipton, Z. C., Balakrishnan, S., Smith, V., and Raghunathan, A. Complementary benefits of contrastive learning and self-training under distribution shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and Lecun, Y. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19338–19347, 2023.

Hall, M., Gustafson, L., Adcock, A., Misra, I., and Ross, C. Vision-language models performing zero-shot tasks exhibit disparities between gender groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2778–2785, 2023.

HaoChen, J. Z. and Ma, T. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Hovy, D. and Søgaard, A. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pp. 483–488, 2015.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Johnson, D. D., Hanchi, A. E., and Maddison, C. J. Contrastive learning can find an optimal basis for approximately view-invariant functions. *arXiv preprint arXiv:2210.01883*, 2022.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

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.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022.

Lee, Y., Lam, M., Vasconcelos, H., Bernstein, M., and Finn, C. Interactive model correction with natural language. In *XAI in Action: Past, Present, and Future Applications*, 2023.

Li, Z., Hoogs, A., and Xu, C. Discover and mitigate unknown biases with debiasing alternate networks, 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, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34: 12978–12991, 2021.

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

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.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

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.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Setlur, A., Dennis, D., Eysenbach, B., Raghunathan, A., Finn, C., Smith, V., and Levine, S. Bitrate-constrained dro: Beyond worst case robustness to unknown group shifts. *arXiv preprint arXiv:2302.02931*, 2023.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33: 9573–9585, 2020.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 19847–19878. PMLR, 2022.

Sohoni, N., Sanjabi, M., Ballas, N., Grover, A., Nie, S., Firooz, H., and Ré, C. Barack: Partially supervised group robustness with guarantees. *arXiv preprint arXiv:2201.00072*, 2021.

Tatman, R. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59, 2017.

Tian, Y., Yu, L., Chen, X., and Ganguli, S. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tu, L., Lalwani, G., Gella, S., and He, H. An empirical study on robustness to spurious correlations using pretrained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017a. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017b.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Yang, Y., Nushi, B., Palangi, H., and Mirzasoleiman, B. Mitigating spurious correlations in multi-modal models during fine-tuning. *arXiv preprint arXiv:2304.03916*, 2023.

Zhai, R., Dan, C., Suggala, A., Kolter, J. Z., and Ravikumar, P. Boosted cvar classification. *Advances in Neural Information Processing Systems*, 34:21860–21871, 2021.

Zhang, Q., Wang, Y., and Wang, Y. On the generalization of multi-modal contrastive learning. *arXiv preprint arXiv:2306.04272*, 2023.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

A. Analysis of multimodal contrastive pretraining

In Sec. 4 we analyzed the solutions of the multimodal spectral contrastive loss in (2), when optimized over the pretraining distribution in (1). Our setup mainly captures the spurious correlation between the label and confounder, that is present to the same extent in both the pretraining distribution, as well as the downstream task. In such cases, we show how the image encoder almost fully couples (no linear separability) the spurious feature with other robust features predictive of the label (Theorem 4.2). In this coupling, the component along the spurious feature is higher when the signal-to-noise ratio along the robust feature is relatively poor (σ_r is higher). Exacerbating this failure, the text encoder learns almost identical features (in ℓ_2) for the confounder and label.

Based on the above finding, we arrive at the zero-shot results in Theorem 4.1, characterizing the poor worst group accuracy of zero-shot label predictor (worse than random), and the near perfect zero-shot confounder prediction performance.

We shall firt present the proof of Theorem 4.2, followed by the zero-shot results i Theorem 4.1. But, before either of these, we will first prove a key result that presents a functional form of the solutions for multimodal spectral contrastive loss, when the image encoder ϕ and text encoder ω are constrained to be orthonormal in $L^2(Q(x))$ and $L^2(Q(t))$ respectively.

Theorem A.1 (Optimal $\phi^\star$ and $\omega^\star$ for objective in (2))**.** *When ϕ, ω are restricted to orthonormal functions in $L^2(Q)$ and $L^2(Q)$ respectively, then the objective in Equation (2) is equivalent to:*

$$\sup_{\phi,\omega} \sum_{i=1}^{k} \int_{\mathcal{T}} A\left(\phi_i(x)\sqrt{q(x)}\right) \cdot \omega_i(t)\sqrt{q(t)}\ \mathrm{d}t,$$
$$\textit{s.t.} \quad \int_{\mathcal{X}} \phi_i^2(x)\ dQ(x) = 1,\ \int_{\mathcal{T}} \omega_i^2(t)\ dQ(t) = 1\ \ \forall i, \tag{4}$$

where $A : L^2(Q) \mapsto L^2(Q)$ is the following linear operator and $\phi_i(x)$ and $\omega_i(t)$ are the i^{th} image and text features respectively. :

$$A(f) =: \int_{\mathcal{X}} \frac{q(x,t)}{\sqrt{q(x)q(t)}} \cdot f(x)\ \mathrm{d}x. \tag{5}$$

Furthermore, the optimal solutions for (2) are $\phi_i(x) = {}^{f_i(x)}/_{\sqrt{p(x)}}$ and $\omega_i(t) = {}^{g_i(t)}/_{\sqrt{p(t)}}$, where $\{f_i\}_{i=1}^k$ and $\{g_i\}_{i=1}^k$ are the top k eigen functions of self-adjoint operators AA^+ and $A^\dagger A$ respectively. Here, $A^\dagger$ is the adjoint of A and is defined as: $A^\dagger(g) =: \int_{\mathcal{T}} {}^{q(x,t)}/_{\sqrt{q(x)q(t)}} \cdot g(t)\ \mathrm{d}t$.

Proof. When ϕ, ω are orthonormal functions in $L^2(Q)$ and $L^2(Q)$, then:

$$\begin{aligned}
&\mathbb{E}_{x\sim Q(x)}\mathbb{E}_{t\sim Q(t)}(\phi(x)^\top\omega(t))^2 \\
&= \sum_{i=1}^{k} \mathbb{E}_{x\sim Q(x)}\mathbb{E}_{t\sim Q(t)}\phi_i(x)^2\omega_i(x)^2 + \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbb{1}(j \neq i)\mathbb{E}_{x\sim Q(x)}\mathbb{E}_{t\sim Q(t)}\phi_i(x)\phi_j(x)\omega_i(x)\omega_j(x) \\
&= k.
\end{aligned} \tag{6}$$

From the above result, we can redefine the objective in (2) as:

$$\sup_{\phi,\omega} \sum_{i=1}^{k} \int_{\mathcal{X}}\int_{\mathcal{T}} q(x,t)\ \omega_i(t)^\top\phi_i(x)\ \mathrm{d}x\mathrm{d}t,$$
$$\text{s.t.} \quad \int_{\mathcal{X}} \phi_i^2(x)\ dQ(x) = 1,\ \int_{\mathcal{T}} \omega_i^2(t)\ dQ(t) = 1\ \ \forall i. \tag{7}$$

The objective in (4) is obtained by substituting the definition of A into the above formulation.

Following Eckart & Young (1936), the solution to the above optimization problem is given by the eigenfunctions of the self-adjoint operators $AA^\dagger$ and $A^\dagger A$. Thus, the optimal solutions for $\phi_i^\star, \omega_i^\star$ are realized by ${}^{f_i}/_{\sqrt{q(x)}}$ and ${}^{g_i}/_{\sqrt{q(t)}}$ where $\{f_i\}_{i=1}^k$ and $\{g_i\}_{i=1}^k$ are the top k eigen functions of self-adjoint operators AA^+ and $A^\dagger A$ respectively. □

Leveraging the result in Theorem A.1, we can now analyze the impact the of the spurious correlations in pretraining data in a special case, when ϕ, ω when are linear functions. Note, given the one hot encoding of the text in $\mathcal{T}$ the linearity assumption in no way restricts the class of text encoders. We now present our proof for the result in Theorem 4.2.

Theorem A.2 (Optimal linear $\phi^\star, \omega^\star$)**.** *Without loss of generality, let $p > 0.5$. Let, ϕ, ω be linear functions over $\mathcal{X}$, $\mathcal{T}$, i.e., $\phi = \mathbf{A}^\top x$, $\omega = \mathbf{B}^\top t$, with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{k \times d}$. Then, there exists constants c_1, c_2 such that for the constraint set: $\int_{\mathcal{X}} \phi_i^2(x)\ \mathrm{d}Q(x) \leqslant c_1,\ \forall i$ and $\int_{\mathcal{T}} \omega^2(t)\ \mathrm{d}Q(t) \leqslant c_2,\ \forall i$, and ϕ, ω that are orthogonal in $L^2(Q)$, the optimal solutions $\mathbf{A}^\star, \mathbf{B}^\star$ for the objective in (2) are the top k columns of the matrices:*

$$\mathbf{A}^\star = \begin{bmatrix} {}^{\cos(\theta)}\!/\!\sqrt{\sigma_{\mathrm{r}}^2+1} & {}^{\sin(\theta)}\!/\!\sqrt{\sigma_{\mathrm{r}}^2+1} & \mathbf{0}_{d_{\mathrm{n}}}^\top \\ -\sin(\theta) & \cos(\theta) & \mathbf{0}_{d_{\mathrm{n}}}^\top \\ \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{U}_{d_{\mathrm{n}}} \end{bmatrix}, \quad \mathbf{B}^\star = {}^1\!/\!{}_2 \cdot \begin{bmatrix} +1 & +1 & +1 & -1 \\ +1 & +1 & -1 & +1 \\ +1 & -1 & -1 & -1 \\ +1 & -1 & +1 & +1 \end{bmatrix}, \tag{8}$$

where $\tan(2\theta) = \frac{4({}^1\!/\!\sigma_r^2+1)}{2p-1+{}^1\!/\!{}_{2p-1}}$.

Proof. We start with the equivalence established between the constrained spectral contrastive loss (in (2)) and the objectives in (4) and (7). Using the result in Theorem A.1, and plugging in the definitions of the linear operator A, we can redefine the objective in (7). Before that, we first note that for linear ϕ, ω, the orthonormality constraint translates to:

$$\mathbb{E}[\phi(x)\phi(x)^\top] = \mathbf{I}_k \quad \text{and} \quad \mathbb{E}[\omega(t)\omega(t)^\top] = \mathbf{I}_k,$$

Now, we are ready to redefine the multimodal contrastive objective for linear ω, ϕ that are constrained to be orthonormal in $L^2(Q)$. Since, the solutions are given by the eigenvectors of $AA^\dagger$ and $A^\dagger A$ matrices, we can write down the optimization over ϕ and ω as two separate optimization objectives. For simplicity, we start with the case where $k = 1$, and then show how we can obtain the result for higher values of k.

We will start with the objective for ϕ:

$$\max_{\phi : \phi^\top \Sigma_x \phi = 1} \phi^\top \widetilde{\Sigma}_x \phi$$
$$\Sigma_x = \mathbb{E}[xx^\top] \quad \widetilde{\Sigma}_x = \mathbb{E}_t[\mathbb{E}[x|t]\mathbb{E}[x|t]^\top]. \tag{9}$$

Here, we encode text as a one-hot vector: Thus, the set of text descriptions $\mathcal{T}$ is: { "y is $+1$", "c is $+1$", "c is -1" and "y is -1" }, which we input as one hot encodings $[1,0,0,0]^\top$, $[0,1,0,0]^\top$, $\left[0,0,1,0^\top\right]$ and $[0,0,0,1]^\top$ respectively to the text encoder ω.

The objective for ω is defined symetrically:

$$\max_{\phi : \omega^\top \Sigma_t \omega = 1} \omega^\top \widetilde{\Sigma}_t \omega,$$

$$\Sigma_t = \mathbb{E}[tt^\top] \qquad \widetilde{\Sigma}_t = \mathbb{E}_x[\mathbb{E}[t|x]\mathbb{E}[t|x]^\top].$$

Since both the above objectives are similar but involve different matrices, we show our working for one, and plug in values from the distribution for the other.

First we note that changing the constraint from $\phi^\top \Sigma_x \phi = 1$ to $\phi^\top \Sigma_x \phi \leqslant 1$, does not change the optimal solution, since these are eigen vectors and Σ is full rank in both cases. Second, we use the identity:

$$\phi^\top \Sigma_x \phi \leqslant 2 \cdot \phi^\top \mathrm{diag}(\Sigma_x)\phi.$$

Thus, we replace the constraint on ϕ, with the right right hand side of the above expression. Thus, when $\phi^\top \mathrm{diag}(\Sigma_x)\phi \leqslant 1$, we satisfy the constraint in Theorem A.2 with $c_1 = 2$. Thus, we are optimizing over a constraint set of orthogonal functions in $L^2(Q)$, where $\forall i, \int_{\mathcal{X}} \phi_i^2(x)\ \mathrm{d}Q(x) \leqslant 2$.

Recall that in our setup both $\widetilde{\Sigma}_x$ and Σ_x are positive definite and invertible matrices. To solve the above problem, let's consider a re-parameterization: $\phi' = \mathrm{diag}(\Sigma_x)^{1/2}\phi$, thus $\phi^\top \mathrm{diag}(\Sigma_x)\phi = 1$, is equivalent to the constraint $\|\phi'\|_2^2 = 1$. Based on this re-parameterization we are now solving:

$$\underset{\|\phi'\|_2^2 \leqslant 1}{\arg\max} \quad \phi'^{\top} \operatorname{diag}(\Sigma_x)^{-1/2} \cdot \widetilde{\Sigma}_x \cdot \operatorname{diag}(\Sigma_x)^{-1/2} \phi'. \tag{10}$$

which is nothing but the top eigenvector for $\operatorname{diag}(\Sigma_x)^{-1/2} \cdot \widetilde{\Sigma}_x \cdot \operatorname{diag}(\Sigma_x)^{-1/2}$.

Now, to extend the above argument from $k = 1$ to $k > 1$, we need to care of one additional form of constraint in the form of feature diversity, or orthogonality: $\phi_i^{\top} \Sigma_x \phi_j = 0$ when $i \neq j$. For this, we can simply repeat the steps above and arrive at the following reformulated optimization problem:

$$\underset{\substack{\|\phi_i'\|_2^2 \leqslant 1, \ \forall i \\ \phi_i'^{\top} \phi_j' = 0, \ \forall i \neq j}}{\arg\max} \quad \left[\phi_1', \phi_2', \ldots, \phi_k'\right]^{\top} \operatorname{diag}(\Sigma_x)^{-1/2} \cdot \widetilde{\Sigma}_x \cdot \operatorname{diag}(\Sigma_x)^{-1/2} \left[\phi_1', \phi_2', \ldots, \phi_k'\right], \tag{11}$$

where $\phi_i' = \operatorname{diag}(\Sigma)^{1/2} \phi_i$. The solution for the above is nothing but the top k eigenvectors for the matrix $\operatorname{diag}(\Sigma_x)^{-1/2} \cdot \widetilde{\Sigma}_x \cdot \operatorname{diag}(\Sigma_x)^{-1/2}$.

Let SVD_k is the top k singular vectors of an SVD decomposition. Now, from our problem description we state values of the four matrices above. For the image encoder, the solution is given by:

$$\mathbf{A}^{\star} = \operatorname{diag}(\Sigma_x)^{-1/2} \cdot \mathrm{SVD}_k \left(\operatorname{diag}(\Sigma_x)^{-1/2} \cdot \widetilde{\Sigma}_x \cdot \operatorname{diag}(\Sigma_x)^{-1/2}\right) \tag{12}$$

where $\Sigma, \widetilde{\Sigma}$ are defined as follows:

$$\Sigma_x =: \begin{bmatrix} 1 + \sigma_{\mathrm{r}}^2 & 2p - 1 & \mathbf{0}_{d_n} \\ 2p - 1 & 1 & \mathbf{0}_{d_n} \\ \mathbf{0}_{d_n}^{\top} & \mathbf{0}_{d_n}^{\top} & I_k \end{bmatrix} \tag{13}$$

$$\widetilde{\Sigma}_x =: \begin{bmatrix} (1 + (2p-1)^2)/2 & 2p - 1 & \mathbf{0}_{d_n} \\ 2p - 1 & (1 + (2p-1)^2)/2 & \mathbf{0}_{d_n} \\ \mathbf{0}_{d_n}^{\top} & \mathbf{0}_{d_n}^{\top} & I_k \end{bmatrix}. \tag{14}$$

Similarly, the optimal text encoder is given by:

$$\mathbf{B}^{\star} = \operatorname{diag}(\Sigma_t)^{-1/2} \cdot \mathrm{SVD}_k \left(\operatorname{diag}(\Sigma_t)^{-1/2} \cdot \widetilde{\Sigma}_t \cdot \operatorname{diag}(\Sigma_t)^{-1/2}\right) \tag{15}$$

Here, $\Sigma_t = \mathbf{I}_4$ and $\widetilde{\Sigma}_t$ is:

$$\widetilde{\Sigma}_t =: \begin{bmatrix} 1 & p & 1-p & 0 \\ p & 1 & 0 & 1-p \\ 1-p & 0 & 1 & p \\ 0 & 1-p & p & 1 \end{bmatrix} \tag{16}$$

Plugging the values of $\Sigma_x, \widetilde{\Sigma}_x, \Sigma_t, \widetilde{\Sigma}_t$ into the equations for $\mathbf{A}^{\star}$ and $\mathbf{B}^{\star}$, and using Lemma A.3, we get the final result:

$$\mathbf{A}^{\star} = \begin{bmatrix} \cos(\theta)/\sqrt{\sigma_{\mathrm{r}}^2+1} & \sin(\theta)/\sqrt{\sigma_{\mathrm{r}}^2+1} & \mathbf{0}_{d_{\mathrm{n}}}^{\top} \\ -\sin(\theta) & \cos(\theta) & \mathbf{0}_{d_{\mathrm{n}}}^{\top} \\ \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{U}_{d_{\mathrm{n}}} \end{bmatrix}, \quad \mathbf{B}^{\star} = 1/2 \cdot \begin{bmatrix} +1 & +1 & +1 & -1 \\ +1 & +1 & -1 & +1 \\ +1 & -1 & -1 & -1 \\ +1 & -1 & +1 & +1 \end{bmatrix},$$

where $\tan(2\theta) = \frac{4(1/\sigma_r^2+1)}{2p-1+1/2p-1}$. □

Lemma A.3 (closed-form eigenvalues and eigenvectors of 2×2 real symmetric matrices (Deledalle et al., 2017)). *For a 2×2 real symmetric matrix $\begin{bmatrix} a, & b \\ c, & d \end{bmatrix}$ the eigenvalues λ_1, λ_2 are given by the following expressions:*

$$\lambda_1 = \frac{(a+b+\delta)}{2}, \quad \lambda_2 = \frac{(a+b-\delta)}{2},$$

where $\delta = \sqrt{4c^2 + (a-b)^2}$. Further, the eigenvectors are given by

$$U = \begin{bmatrix} \cos(\theta), -\sin(\theta) \\ \sin(\theta), \quad \cos(\theta) \end{bmatrix},$$

where $\tan(\theta) = \frac{b-a+\delta}{2c}$.

In summary, we defined functional forms for optimal orthogonal ϕ, ω in Theorem A.1. Next, we presented closed form solutions for optimal linear and “nearly” orthonormal ϕ, ω in Theorem A.2. Now, we can easily characterize the zero-shot performance of these learned feature extractors. Following presents the proof of our result in Theorem 4.1.

Theorem A.4. *(zero-shot robustness; restated) Let the zero-shot label (f) and confounder classifier (g) be obtained by minimizing the loss in (2) on infinite pretraining data. Then, for $\sigma_{\mathrm{r}} = \Omega(1)$, label classifier is worse than random on the worst group, since $\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) = {}^1\!/\!_2\, \mathrm{erfc}(-c_1 \sigma_{\mathrm{r}} p)$. On the other hand, the confounder classifier suffers small error on all groups since $\mathrm{err}_{\mathrm{sp}}^{\mathrm{wg}}(g) = {}^1\!/\!_2\, \mathrm{erfc}(c_2 \sigma_{\mathrm{r}} p)$. Here, $c_1, c_2 > 0$ are constants .*

Proof. First, we state the formal version of the theorem statement. Let f be zero-shot label predictor, and g be the zero-shot confounder predictor extracted from ϕ, ω in Theorem A.2. Then, the worst group error for f is:

$$\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) = {}^1\!/\!_2 \cdot \mathrm{erfc}\left({}^{\rho}\!/\!\sqrt{2}\right),$$

and for g is:

$$\mathrm{err}_{\mathrm{sp}}^{\mathrm{wg}}(g) = {}^1\!/\!_2 \cdot \mathrm{erf}\left({}^{\rho}\!/\!\sqrt{2}\right),$$

where $\rho = -{}^1\!/\!_{\sigma_{\mathrm{r}}} - \cot(\theta)\sqrt{{}^1\!/\!_{\sigma_{\mathrm{r}}^2} + 1}$. Here, θ is the value defined in Theorem A.2.

Using our expressions for the zero-shot predictor in Sec. 4, we use the result from Theorem A.2 to define:

$$f([x_{\mathrm{r}}, x_{\mathrm{c}}]) = g([x_{\mathrm{r}}, x_{\mathrm{c}}]) = 2\mathbb{1}(-\frac{2x_{\mathrm{r}} \sin\theta}{\sqrt{1+\sigma^2}} + 2x_{\mathrm{c}} \cos\theta) - 1$$

Now, based on the signs along x_{r} and x_{c}, we conclude that the worst group for f is $y = 1, c = -1$.

$$\begin{aligned}
&\Pr(f([x_{\mathrm{r}}, x_{\mathrm{c}}]) \leqslant 1 \mid (y, c) = (1, -1)) \\
&= \Pr(\frac{-2\sin\theta}{\sqrt{1+\sigma_{\mathrm{r}}^2}} \leqslant -2\cos\theta) \\
&= \Pr(\frac{x_{\mathrm{r}} - 1}{\sigma_{\mathrm{r}}} \geqslant -\frac{1}{\sigma_{\mathrm{r}}} + \sqrt{\frac{1}{\sigma_{\mathrm{r}}^2} + 1 \cot\theta\cos\theta}) \\
&= \frac{1}{2}\mathrm{erfc}(\frac{-1}{\sigma_{\mathrm{r}}} - \cot\theta(\sqrt{1 + 1/\sigma_{\mathrm{r}}^2})) \\
&= \frac{1}{2}\mathrm{erfc}(\rho/\sqrt{2}).
\end{aligned}$$

On the other hand the worst group for the confounder is $(y, c) = (1, 1)$, but even here, the error is negligible.

$$\begin{aligned}
&\Pr(f([x_\mathrm{r}, x_\mathrm{c}]) \leqslant 1 \mid (y, c) = (1, 1)) \\
&= \frac{1}{2}\mathrm{erfc}(\frac{-1}{\sigma_\mathrm{r}} + \cot\theta(\sqrt{1 + 1/\sigma_\mathrm{r}^2})) \\
&= \frac{1}{2}\mathrm{erf}(\rho/\sqrt{2}).
\end{aligned}$$

This completes our proof of zero-shot performance guarantees. □

B. Worst group guarantees for PfR

Theorem B.1 (PfR's worst group error)**.** *For PfR output $\widehat{f}$, w.h.p. $1 - \delta$, worst group generalization error of $\widehat{f}$ is $\lesssim \sqrt{\log \mathfrak{C}(\mathcal{F})K/\delta/n} + \mathrm{err}_c(\mathrm{FM}(t_c))$, where $\mathfrak{C}(\mathcal{F})$ is complexity of $\mathcal{F}$, K is number of groups and latter term is FM's zero-shot performance on confounder prediction.*

Proof. Recall the objective for PfR which minimizes worst group loss over predicted groups $\widehat{G}_1, \ldots, \widehat{G}_K$. Let,

$$f^\star \;:=\; \inf_{f \in \mathcal{F}} \sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(h(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in \widehat{G}_k\right] \tag{17}$$

Lemma B.2 (worst-case risk generalization (Group DRO))**.** *With probability $\geqslant 1 - \delta$ over dataset $\mathcal{D} \sim P^n$, the worst group risk for $f^\star$ can be upper bounded by the following, where opt is the minimum on the training objective,*

$$\sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(h(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in \widehat{G}_k\right] \;\lesssim\; \mathrm{opt} + \sqrt{\frac{\log\left(\frac{\mathfrak{C}K}{\delta}\right)}{n}}\,,$$

where $\mathfrak{C}$ is the complexity of class $\mathcal{F}$ (e.g., the covering number (Wainwright, 2019)).

Proof. We first apply the generalization bound for a single group, which is given by $\sqrt{\frac{\log\left(\frac{\mathfrak{C}}{\delta}\right)}{n}}$ (Wainwright, 2019), followed by a union bound over the K groups. □

We can break down down the worst group loss for the learned function $\widehat{f}$ on the true groups $G_1, \ldots, G_K$ in the following way, where we assume loss ℓ is M bounded:

$$\sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right] \leqslant \sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k \cap \widehat{G}_k\right] \tag{18}$$

$$+ M\mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in G_k\right] \tag{19}$$

$$+ M\mathbb{E}_{P_T}\left[\mathbb{1}(x \in G_k) \mid x \in \widehat{G}_k\right] \tag{20}$$

Since $\max_{1,2}(a_1 + b_1, a_2 + b_2) \leqslant \max_{1,2}(a_1, a_2) + \leqslant \max_{1,2}(b_1, b_2)$ for some scalars a_1, a_2, b_1, b_2, we can upper bound $\sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right]$ as:

$$\begin{aligned}
\sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right] &\leqslant \sup_{k \in [K]} \mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in \widehat{G_k}\right] + \mathbb{E}\left[\mathbb{1}(\mathrm{FM}(x, t_c) \neq c)\right] \\
&= \sup_{k \in [K]} \mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in \widehat{G_k}\right] + \mathrm{err}_{\mathrm{sp}}^{\mathrm{avg}}(\mathrm{FM}(x, t_c)).
\end{aligned}$$

for positive losses. Above, we replaced the group mixmatch error with the error of the zero-shot classifier $\mathrm{FM}(x, t_c)$. Further, in our case $M = 1$.

The above result when used in a simple triangle inequality with the result in Lemma B.2 completes the proof of Theorem B.2.

□

C. Details on prompts used

First we provide examples on the prompts used to zero-shot predict labels on one language and one vision benchmark, followed by those used to predict the values of the spurious attribute. Finally, we provide the prompt used for annotation of chest drain for our CXR-Drain dataset.

C.1. Zero-shot prediction of labels

CivilComments:

Predict label

> You are an AI bot that is an expert at detecting toxic comments. Evaluate the following USER COMMENT on a news website, and classify the comment as either ’TOXIC’ or ’SAFE’, ignoring references to religion, race, and gender:
>
> *USER COMMENT:* {}
>
> In one word, the above comment is TOXIC (ONLY as ’YES’ or ’NO’):

Predict label, ignore confounder

> You are an AI bot that is an expert at detecting toxic comments. You are not biased by references to religion, race, and gender. Evaluate the following USER COMMENT on a news website, and classify the comment as either ’TOXIC’ or ’SAFE’, ignoring references to religion, race, and gender:
>
> *USER COMMENT:* {}
>
> In one word, the above comment is TOXIC (ONLY as ’YES’ or ’NO’):

Waterbirds:

Predict label

> This is the image of a waterbird.
>
> This is the image of a landbird.

Predict label and confounder

> This is the image of a waterbird in front of forest.
>
> This is the image of a waterbird in front of water.
>
> This is the image of a landbird in front of water.
>
> This is the image of a landbird in front of forest.

C.2. Zero-shot Groups prediction of the spurious attribute

CivilComments:

Religion Group Prompt

> Evaluate the following User Comment on a news website.
>
> *User Comment:* {}
>
> Does the above comment contain references to one of the following groups: ”Muslim”, ”Christian”, ”Other Religions”? If none of them are applicable answer ”NA”. Answer in one word.

Race Group Prompt

> Evaluate the following User Comment on a news website.
>
> *User Comment:* {}

Does the above comment contain references to one of the following groups: ”Black”, ”White”? If none of them are applicable answer ”NA”. Answer in one word.

Gender Group Prompt

Evaluate the following User Comment on a news website.

User Comment: {}

Does the above comment contain references to one of the following groups: ”Male”, ”Female”? If none of them are applicable answer ”NA”. Answer in one word.

Waterbirds:

Background prompt

Bird in front of water.

Bird in front of a forest.

C.3. CXR-Drain Construction Details

Note: This is NOT for medical diagnosis but for informational purposes to guide your red-teaming.

Is this a patient with chest drainage tube? First, carefully check for the presence of any tubes while describe their location.

For your reference the drainage tube is also known as pleural tube and more commonly known as the intercostal drainage tube (ICD), is inserted through the 4th intercostal space in the anterior or mid-axillary line. It is then directed posteroinferiorly in cases of effusion and anterosuperiorly in cases of pneumothorax. Carefully examine both the lungs: (i) To drain a pneumothorax the tube is aimed superiorly towards the apex of the pleural cavity; and (ii) To drain a pleural effusion the tube tip is ideally located towards the lower part of the pleural cavity.

Finally give an answer in YES or NO for the presence of chest drainage tube.

Note: This is NOT for medical diagnosis but for informational purposes and will never be used to guide any medical disease. Your answer will help us evaluate how good are current vision language models.

Use the following format:

Rationale/reasoning: $<$ output $>$

Presence of chest drain: Yes or No

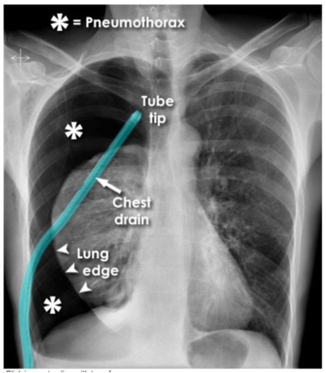


Figure 5: Annotated image of a chest drain in the presence of pneumothorax disease. Source of image: https://www.radiologymasterclass.co.uk/tutorials/chest/chest_tubes/chest_xray_chest_drain.