Zero-Shot Robustification of Zero-Shot Models With Auxiliary Foundation Models

Dyah Adila*† Changho Shin*† Linrong Cai† Frederic Sala†

†University of Wisconsin-Madison
{adila, cshin23, lcai54, fredsala}@wisc.edu

July 10, 2023

Abstract

Zero-shot inference is a powerful paradigm that enables the use of large pretrained models for downstream classification tasks without further training. However, these models are vulnerable to inherited biases that can impact their performance. The traditional solution is fine-tuning, but this undermines the key advantage of pretrained models, which is their ability to be used out-of-the-box. We propose ROBOSHOT, a method that improves the robustness of pretrained model embeddings in a fully zero-shot fashion. First, we use zero-shot language models (LMs) to obtain useful insights from task descriptions. These insights are embedded and used to remove harmful and boost useful components in embeddings—without any supervision. Theoretically, we provide a simple and tractable model for biases in zero-shot embeddings and give a result characterizing under what conditions our approach can boost performance. Empirically, we evaluate ROBOSHOT on nine image and NLP classification tasks and show an average improvement of 15.98% over several zero-shot baselines. Additionally, we demonstrate that ROBOSHOT is compatible with a variety of pretrained and language models.*

1 Introduction

Zero-shot models are among the most exciting paradigms in machine learning. These models obviate the need for data collection and model training loops by simply asking the model for a prediction on any set of classes. Unfortunately, such models inherit biases or undesirable correlations from their large-scale training data [12, 37]. In a now-canonical example [19], they often associate `waterbirds` with `water background`. This behavior leads to decreased performance, often exacerbated on rare data slices that break in-distribution correlations.

A growing body of literature [15, 42, 43] seeks to improve robustness in zero-shot models. While promising, these works require labeled data to train or fine-tune models, and so **do not tackle the zero-shot setting.** A parallel line of research seeking to debias word embeddings [1, 4, 10, 21] often sidesteps the need for labeled data. Unfortunately, these works often require domain expertise and painstaking manual specification in order to identify particular concepts that embeddings must be invariant to. As a result, out-of-the-box word embedding debiasing methods also cannot be applied to zero-shot robustification.

Can we robustify zero-shot models without (i) labeled data, (ii) training or fine-tuning, or (iii) manual identification? Surprisingly, despite this seemingly impoverished setting, it is often possible to do so. Our key observation is that zero-shot models **contain actionable insights** that can be exploited to improve themselves or other zero-shot models. These insights are noisy but cheaply available at scale—and can be easily translated into means of refinement for zero-shot representations. These refinements improve performance, particularly on underperforming slices—at nearly no cost.

We propose ROBOSHOT, a system that robustifies zero-shot models via auxiliary language models *without labels, training, or manual specification*. Using just the task description, ROBOSHOT obtains *positive and negative insights* from a language model (potentially the model to be robustified itself). It uses embeddings of these noisy insights to recover *harmful, beneficial*, and *benign* subspaces of zero-shot latent representation spaces. Representations are then modified to neutralize and emphasize their harmful and beneficial components, respectively.

Theoretically, we introduce a simple and tractable model to capture and quantify failures in zero-shot models. We provide a result that characterizes the *quantity and quality* of insights that must be obtained as a function of the

*These authors contributed equally to this work
*Code can be found in https://github.com/SprocketLab/roboshot

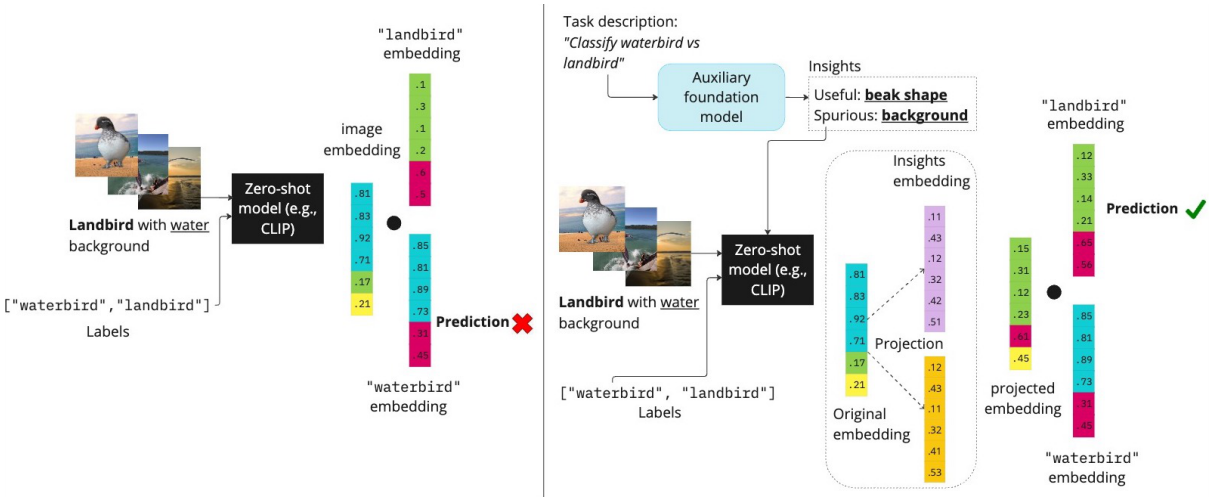


Figure 1: ROBOSHOT pipeline (right) vs. vanilla zero-shot classification (left).

severity of harmful correlations. Empirically, ROBOSHOT achieves 15.98% improvement across nine image and NLP datasets while offering sufficient versatility to apply to a diverse variety of base models. Most excitingly, in certain cases, it reaches comparable or greater improvements **even when compared to fine-tuned models** that rely on labeled data.

Our contributions include,

1. A simple theoretical model describing zero-shot model failures along with a theoretical analysis of our approach that characterizes the amount of information required for obtaining improvements as a function of the most harmful unwanted correlation,
2. ROBOSHOT, an algorithm that implements our core idea. It extracts insights from foundation models and uses them to improve zero-shot representations,
3. Extensive experimental evidence on zero-shot language and multimodal models, showing improved worst-group accuracy of 15.98% across nine image and NLP datasets.

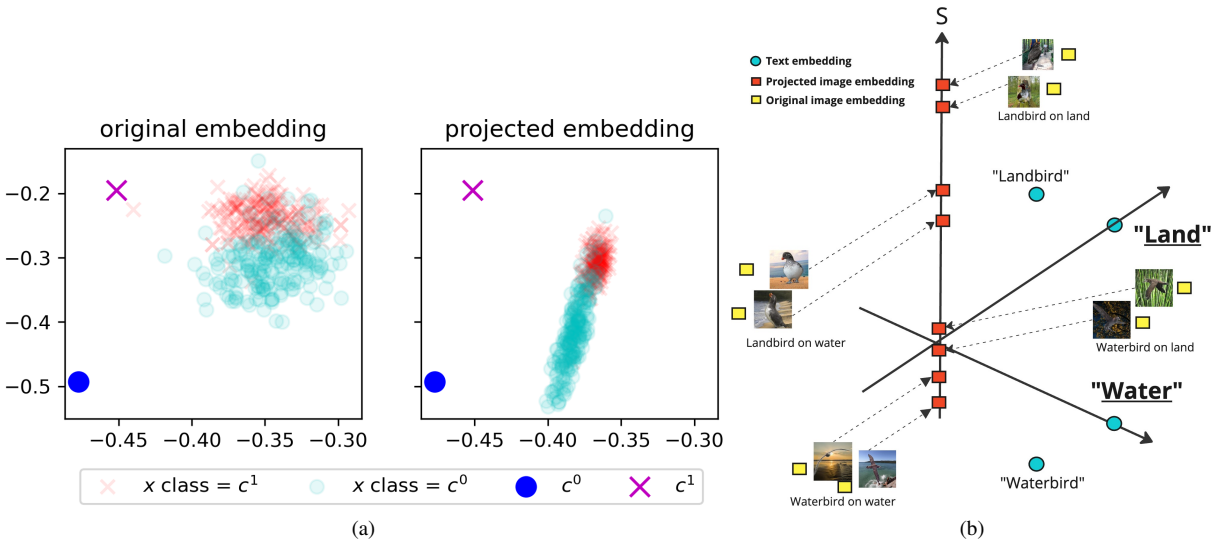
2 Related Work

We describe related work in zero-shot model robustness, debiasing embeddings, guiding multi-modal models using language, and using LMs as prior information.

Zero-Shot inference robustness. Improving model robustness to unwanted correlations is heavily studied [2, 18, 20, 23, 25, 36]. Some methods require training from scratch and are less practical when applied to large pretrained architectures. Existing approaches to improve robustness *post-pretraining* predominantly focus on fine-tuning. [42] detects spurious attribute descriptions and fine-tunes using these descriptions. Specialized contrastive loss is used to fine-tune a pretrained architecture in [15] and to train an adapter on the frozen embeddings in [43]. While promising, fine-tuning recreates traditional machine learning pipelines (e.g., labeling, training, etc.), which contradicts the promise of zero-shot models. In contrast, our goal is to avoid any training and any use of labeled data.

Debiasing embeddings. A parallel line of work seeks to de-bias text embeddings [1] [4] [10] [21] and multimodal embeddings [3, 39, 40] by removing subspaces that contain harmful or unwanted concepts. We use a similar procedure as a building block. However, these methods either target specific fixed concepts (such as gender) or rely on concept annotations, which limits their applicability across a wide range of tasks. In contrast, our method automates getting *both beneficial and unwanted concepts* solely from the task descriptions. An additional difference is that our goal is simply to add robustness at low or zero-cost; we not seek to produce fully-invariant representations as is often desired for word embeddings.

Using language to improve visual tasks A large body of work has shown the efficacy of using language to improve performance on vision tasks [14, 22, 34]. Most relevant are those that focus on robustness, like [32], where attention maps using multimodal models (like CLIP) are used as extra supervision to train a downstream



(a) (b)

Figure 2: (a) RoboShot debiases original input embedding (left). The projected embedding (right)'s variance in the unwanted direction is reduced, and in the relevant direction increases. (b) Embedding projection. We project embeddings to the space orthogonal to the embeddings of all unwanted insights (e.g., `water` and `land`)

image classifier. [42] uses text descriptions of spurious attributes in a fine-tuning loss to improve robustness against spurious correlations. In contrast to these works, we focus on using textual concepts to improve zero-shot model robustness—without fine-tuning.

Language model as prior The basis of our work comes from the observation that language models contain information that can serve as a prior for other learning tasks. [17] finds that LLMs can perform causal reasoning tasks, substantially outperforming existing methods. [7] explicitly prompts LLMs for task-specific priors, leading to substantial performance improvements in feature selection, reinforcement learning, and causal discovery. Our work shares the spirit of these approaches in using the insights embedded in language models to enhance zero-shot robustness.

3 RoboShot: Robustifying Zero-shot Models

We are ready to provide our setup and describe the algorithm.

3.1 Modeling and setup

Suppose that the zero-shot model's latent space contains an (unknown) *concept set*; similar notions have been studied frequently in the literature [9]. For simplicity, we assume that this concept set is given by the orthonormal vectors $\{z_1, \ldots, z_k\}$. The model's encoder produces, for a particular input a representation x that is a mixture of concepts $\sum_i \gamma_i z_i$, where $\gamma_i \geq 0$ are weights.

We shall work with the following theoretical model for zero-shot classification. It closely resembles models like CLIP. For simplicity, we assume that there are two classes. It is straightforward to extend the analysis below to multiple classes. We take $\sum_i \alpha_i z_i$ to be the embedding of a datapoint, while $c^0 = \sum_i \beta_{i,0} z_i$ is the embedding of the first class and $c^1 = \sum_i \beta_{i,1} z_i$ is that of the second. Finally, we assume that we have access to m answers $v^1, \ldots, v^m$ from the queries to the language model. These are given by $v^j = \sum_i \gamma_{i,j} z_i$ for $j \leq m$. We call these *insight representations*. Without our approach, the prediction is made by $\hat{y} = \mathbb{1}\{(\sum_i \alpha_i z_i)^T (\sum_i \beta_{i,0} z_i) < (\sum_i \alpha_i z_i)^T (\sum_i \beta_{i,1} z_i)\}$, so that we predict whichever class has higher inner product with the datapoint's embedding.

Next, we assume that each input representation x can be represented by partitioning the mixture components into three groups,

$$x = \sum_{s=1}^{S} \alpha_s^{\text{harmful}} z_s + \sum_{r=S+1}^{S+R} \alpha_r^{\text{helpful}} z_r + \sum_{b=S+R+1}^{S+R+B} \alpha_b^{\text{benign}} z_b. \tag{1}$$

The same holds for class and insight representations.

Algorithm 1: ROBOSHOT

Parameters: Input embedding x, class embeddings c^0, c^1, harmful insight representations $v^1, \ldots, v^S$, helpful insight representations $u^1, \ldots, u^R$
for $j \in \{1, 2, \ldots, S\}$ **do**
    Remove harmful insight v^j: set $x \leftarrow x - \langle x, v^j \rangle / \langle v^j, v^j \rangle v^j$
    Renormalize $x = x / ||x||$
end for
for $k \in \{1, 2, \ldots, R\}$ **do**
    Amplify helpful insight u_k: set $x \leftarrow x + \langle x, u^k \rangle / \langle u^k, u^k \rangle u^k$
end for
$\hat{y} = \mathbb{1}\{x^T c^0 < x^T c^1\}$
Returns: Robustified zero-shot prediction $\hat{y}$

Example We illustrate how harmful correlations produce errors on rare slices of data through a standard task setting, Waterbirds [19]. In this dataset, the goal is to classify `landbirds` versus `waterbirds`, and the background (`land` or `water`) is spurious. Suppose that we have these terms relate to concepts such that $z_{\texttt{water}} = -z_{\texttt{land}}$ and $z_{\texttt{waterbird}} = -z_{\texttt{landbird}}$.

Consider a datapoint coming from a rare slice infrequently encountered in the training set. This might be an image of a landbird over water. Its embedding might be $x = 0.7z_{\texttt{water}} + 0.3z_{\texttt{landbird}}$. We may also have that

$$c^{\texttt{waterbird}} = 0.4z_{\texttt{water}} + 0.6z_{\texttt{waterbird}} \text{ and } c^{\texttt{landbird}} = 0.4z_{\texttt{land}} + 0.6z_{\texttt{landbird}}.$$

Then, $x^T c^{\texttt{waterbird}} = 0.1 > x^T c^{\texttt{landbird}} = -0.1$, so that the prediction is waterbird, and thus incorrect. This is caused by the presence of harmful components in *both* the class embedding (caused by seeing too many images with water described as waterbirds) and the datapoint embedding (where the water background appears). Thus our goal is to *remove* harmful components (the z_s's) and *boost* helpful components (the z_r's). We explain our approach towards doing so next.

3.2 ROBOSHOT: Zeroshot robustification with LLM

We describe ROBOSHOT in Algorithm 1. It uses representations of insights from language models to shape input and class embeddings to remove harmful components and boost helpful ones. Figure 2 is helpful in understanding the intuition behind these procedures. The left part (a) illustrates the effect of ROBOSHOT on a true dataset. Note how unhelpful directions are neutralized while others are boosted. The illustration on the right (b) shows this effect on the waterbirds running example.

Obtaining insight representations from LMs The first question is how to obtain insight representations without training. To do so in a zero-shot way, we use *textual* descriptions of harmful and helpful concepts by querying language models using *only the task description*. For example, in the Waterbirds dataset, we use the prompt "What are the biased/spurious differences between waterbirds and landbirds?". We list the details of the prompts used in Appendix C.2. Let s^1, s^2 be the text insights obtained from the answer (e.g., {'`water background`,' '`land background`'}). We obtain a spurious insight representation by taking the difference of their embedding $v = \dfrac{g(s^1) - g(s^2)}{||g(s^1) - g(s^2)||}$, where g is the text encoder of our model.

In addition to attempting to discover harmful correlations, we seek to discover helpful components in order to boost their magnitudes past remaining harmful ones (or noise). The procedure is similar. We obtain insight representations using language models. For example, we ask "What are the true characteristics of waterbirds and landbirds?' and obtain e.g., {'`short beak`,' '`long beak`'}. The remainder of the procedure is identical to the case of harmful components. Note that since we are seeking to boost (rather than remove) components, it is also possible to fix a multiplicative constant (to be treated as a hyperparameter) for the boosting procedure. That is, we could take $x \leftarrow x + \nu \times \langle x, u^k \rangle / \langle u^k, u^k \rangle u^k$ for some $\nu > 0$. While this is possible if we have access to a labeled set that we can tune ν over, we *intentionally avoid doing so to ensure our procedure is truly zero-shot*.

Prompting a language model is typically inexpensive, which will enable obtaining multiple insight vectors $\tilde{v}^1, \ldots, \tilde{v}^m$. From these, we obtain an orthogonal basis $v^1, \ldots, v^m$ separately for harmful and helpful components using standard matrix decomposition methods. Thus we have access to recovered subspaces spanned by such components.

Removing and Boosting Components ROBOSHOT applies simple vector rejection to mitigate or remove harmful components, which is described in lines 2-5 of Algorithm 1. Similarly, it boosts helpful components as described in lines 6-9.

To see the impact of doing so, consider our earlier example. Suppose that $v^{\text{harmful}} = 0.9z_{\texttt{water}} + 0.1z_{\texttt{landbird}}$, and that this is our only harmful insight. Similarly, suppose that we obtain a single helpful insight given by $v^{\text{helpful}} = 0.1z_{\texttt{water}} + 0.9z_{\texttt{landbird}}$. Note that even these insights can be imperfect: they do not uniquely identify what are harmful or helpful concepts, as they have non-zero weights on other components.

We first obtain from removing the harmful component (ignoring normalization for ease of calculation) that

$$\hat{x} \leftarrow x - \frac{\left\langle x, v^{\text{harmful}} \right\rangle}{\left\langle v^{\text{harmful}}, v^{\text{harmful}} \right\rangle} v^{\text{harmful}} = -0.0244 z_{\texttt{water}} + 0.2195 z_{\texttt{landbird}}.$$

Then, we already we have that $x^T c^{\texttt{waterbird}} = -0.1415 < x^T c^{\texttt{landbird}} = 0.1415$, so that the correct class is obtained. In other words we have already, from having access to a single insight, neutralized a harmful correlation and corrected what had been an error. Adding in the helpful component further helps. We obtain

$$\hat{x} \leftarrow \hat{x} + \frac{\left\langle \hat{x}, v^{\text{helpful}} \right\rangle}{\left\langle v^{\text{helpful}}, v^{\text{helpful}} \right\rangle} v^{\text{helpful}} = -0.0006 z_{\texttt{water}} + 0.4337 z_{\texttt{landbird}}.$$

This further increases our margin. Note that it is not necessary to fully neutralize (i.e., to be fully invariant to) spurious or harmful components in our embeddings. The only goal is to ensure, as much as possible, that their magnitudes are reduced when compared to helpful components (and to benign components). In the following section, we provide a theoretical model for the magnitudes of such components and characterize the conditions under which it will be possible to correct zero-shot errors. We note that there is a variant of our approach that can also update class embeddings as well.

4 Analysis

Next, we provide an analysis that characterizes under what conditions ROBOSHOT is capable of correcting zero-shot errors. First, we consider the following error model on the weights of the various representations. For all benign representations, we assume that $\alpha_b, \beta_b, \gamma_b \sim \mathcal{N}(0, \sigma^2_{\text{benign}})$. That is, the magnitudes of benign components are drawn from a Gaussian distribution. The value of σ_{benign} is a function of the amount of data and the training procedure for the zero-shot model.

Next, we assume that the embedding insight $v^s = \sum_{i=1}^{k} \gamma_{i,s} z_i$ (where $1 \le s \le S$) satisfies the property that for $i \neq s$, $\gamma_{i,s} \sim \mathcal{N}(0, \sigma^2_{\text{insight}})$, while $\gamma_{s,s}$ is a constant. In other words, the vectors $v^1, \ldots, v^S$ spanning the harmful component subspace are well-aligned with genuinely harmful concepts, but are also affected by noise. Similarly, we assume that helpful insights $v^r = \sum_{i=1}^{k} \gamma_{i,r} z_i$ (where $S+1 \le r \le S+R$) satisfy the same property. We seek to understand the interplay between this noise, benign noise, and the coefficients of the other vectors (i.e., helpful components). Let the result of ROBOSHOT with insight representations $v^1, \ldots, v^{S+R}$ be

$$\hat{x} = x - \sum_{s=1}^{S} \frac{x^T v^s}{||v^s||^2} v^s + \sum_{r=S+1}^{S+R} \frac{x^T v^r}{||v^r||^2} v^r = \sum_{i=1}^{S+R+B} A_i z_i.$$

We first provide a bound on A_s, the coefficient of a targeted harmful concept after applying ROBOSHOT algorithm.

Theorem 4.1. *Under the noise model described above, the post-*ROBOSHOT *coefficient for harmful concept* s $(1 \le s \le S)$ *satisfies*

$$|\mathbb{E}[A_s]| \le \left| \frac{(k-1)\alpha_s \sigma^2_{insight}}{\gamma^2_{s,s}} \right| + \left| \sum_{t=1, t\neq s}^{S+R} \frac{\alpha_s \sigma^2_{insight}}{\gamma^2_{t,t}} \right|,$$

where k *is the number of concepts.*

The proof is included in Appendix B.3. The theorem illustrates how and when the rejection component of ROBOSHOT works—it scales down harmful coefficients at a rate inversely proportional to the harmful coefficients of the insight embeddings. As we would hope, when insight embeddings have larger coefficients for harmful vectors (i.e., are more precise in specifying terms that are not useful), ROBOSHOT yields better outcomes. In addition, we observe that the harmful coefficients decrease when the insight embeddings have less noise. In fact, we have that $\lim_{\sigma_{insight} \to 0} A_s = 0$ — the case of perfectly identifying harmful, helpful concepts.

Next, we provide a bound on A_r, the post-ROBOSHOT coefficient of a targeted helpful concept.

Theorem 4.2. *With additional assumptions* $\alpha_s \leq 0$ $(1 \leq s \leq S), \alpha_r \geq 0$ $(S+1 \leq r \leq S+R), \gamma_{t,t}^2 \geq \sigma_{insight}^2$ *under the described noise model, the post-*ROBOSHOT *coefficient for helpful concept* r $(S+1 \leq r \leq S+R)$ *satisfies*

$$\mathbb{E}\left[A_r\right] \geq \left(1 + \frac{\gamma_{r,r}^2}{\gamma_{r,r}^2 + (k-1)\sigma_{insight}^2}\right)\alpha_r$$

Refer to Appendix B.3 for the proof. Theorem 4.2 implies the helpful coefficients are scaled up at a rate inversely proportional to the noise rate $\sigma_{insight}$. When concepts are perfectly identified, i.e. $\sigma_{insight} = 0$, the coefficient α_r is doubled, yielding more emphasis on the concept z_r as desired.

5 Experimental Results

This section evaluates the following claims about ROBOSHOT:

- **Improving multi-modal models (Section 5.1)**: ROBOSHOT improves zero-shot classification robustness of various multi-modal models, even outperforming prompting techniques that include spurious insight descriptions (which we do not have access to) in the label prompts.
- **Improving language models (Section 5.2)**: ROBOSHOT improves zero-shot robustness when using language model embeddings for text zero-shot classification.
- **Extracting concepts from LM with varying capacities (Section 5.3)**: ROBOSHOT can extract insights from language models with varying capacities. Improvements persist with weaker LMs.
- **Ablations (Section 5.4)** ROBOSHOT benefits from both removing harmful and boosting helpful representations (line 3 and line 7 in ROBOSHOT Algorithm 1).
- **Isolating concepts by averaging relevant concepts (Section 5.5)** We conduct proof of concept experiments to test the viability of our concept modeling in equation 1.

Metrics and how to interpret the results. We use three metrics: average accuracy % (AVG), worst-group accuracy % (WG), and the gap between the two (Gap). While a model that relies on harmful correlations may achieve high AVG when such correlations are present in the majority of the test data, it may fail in settings where the correlation is absent. **A robust model should have high AVG and WG, with a small gap between them**.

Baselines We compare against the following sets of baselines:

1. **Multimodal baselines**: We compare against: (i) vanilla zero-shot classification (**ZS**) and (ii) zero-shot classification with group information (**Group Prompt ZS**). We do so across a variety of models: CLIP (ViT-B-32 and ViT-L-14) [34], ALIGN [16], and AltCLIP [6]. Group Prompt ZS assumes access to spurious or harmful insight annotations and includes them in the label prompt. For instance, the label prompts for waterbirds dataset become [`waterbird with water background`, `waterbird with land background`, `landbird with water background`, `landbird with land background`]. We only report Group Prompt ZS results on datasets where spurious insight annotations are available.
2. **Language model baselines**: We compare against zero-shot classification using multiple language model embeddings, including BERT [35] and Ada [29] (**ZS**).

5.1 Improving multi-modal models

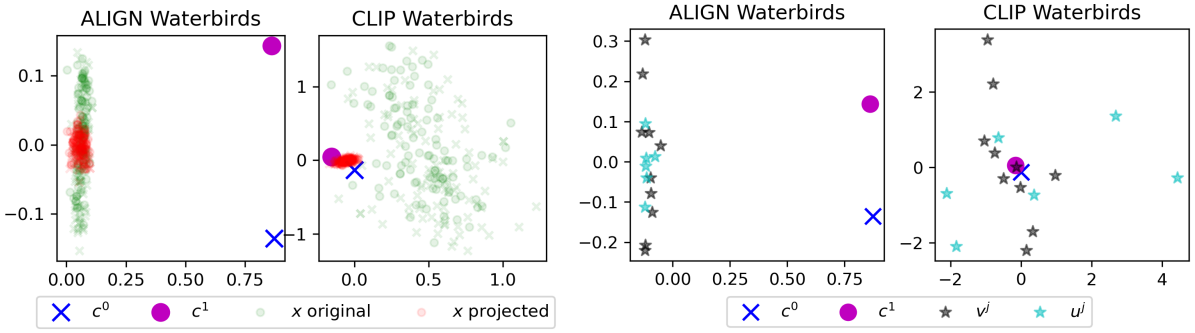
Setup. We experimented on five binary and multi-class datasets with spurious correlations and distribution shifts, coming from a variety of domains: **Waterbirds** [36], **CelebA** [26], **CXR14** [41], **PACS** [24], and **VLCS** [13]. We use the default test splits of all datasets. Dataset details are provided in Appendix C.1. For CXR14, we use BiomedCLIP [44], which is a variant of CLIP finetuned on biomedical images and articles. All experiments are conducted using frozen pretrained models.

Results. Table 1 shows that **ROBOSHOT significantly improves the worst group performance (WG)** and maintains (and sometimes also improves) the overall average (AVG) without any auxiliary information (in contrast to Group Prompt, which requires access to spurious insight annotation).

Improved robustness nearly across-the-board suggests that both the insights extracted from LMs and the representation modifications are useful. We also provide insights insights into the case where our method does not improve the baseline (ALIGN model on Waterbirds) in Fig. 3. In Fig. 3a, we visualize the original and projected input embeddings (x in green and red points, respectively), and the label embeddings (c^0 and c^1). Fig. 3a (left) shows the

Dataset	Model	ZS			GroupPrompt ZS			**RoboShot**		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
Waterbirds	CLIP (ViT-B-32)	80.7	27.9	52.8	81.6	43.5	38.1	82.0	**54.4**	**28.6**
	CLIP (ViT-L-14)	88.7	27.3	61.4	70.7	10.4	60.3	79.9	**45.2**	**34.7**
	ALIGN	72.0	**50.3**	21.7	72.5	5.8	66.7	50.9	41.0	**9.9**
	AltCLIP	90.1	35.8	54.3	82.4	29.4	53.0	78.5	**54.8**	**23.7**
CelebA	CLIP (ViT-B-32)	80.1	72.7	7.4	80.4	74.9	5.5	84.8	**80.5**	**4.3**
	CLIP (ViT-L-14)	80.6	74.3	6.3	77.9	68.9	9.0	85.5	**82.6**	**2.9**
	ALIGN	81.8	77.2	4.6	78.3	67.4	10.9	86.3	**83.4**	**2.9**
	AltCLIP	82.3	**79.7**	**2.6**	82.3	79.0	3.3	86.0	77.2	8.8
PACS	CLIP (ViT-B-32)	96.7	82.1	14.6	97.9	82.7	15.2	97.0	**86.3**	**10.7**
	CLIP (ViT-L-14)	98.1	79.8	18.3	98.2	**86.6**	**11.6**	98.1	83.9	14.2
	ALIGN	95.8	**77.1**	**18.7**	96.5	65.0	31.5	95.0	73.8	21.2
	AltCLIP	98.5	82.6	15.9	98.6	85.4	13.2	98.7	**89.5**	**9.2**
VLCS	CLIP (ViT-B-32)	75.6	20.5	55.1		-		76.5	**33.0**	**43.5**
	CLIP (ViT-L-14)	72.6	4.20	68.4		-		71.1	**12.6**	**58.5**
	ALIGN	78.8	33.0	45.8		-		77.6	**39.8**	**37.8**
	AltCLIP	78.3	24.7	**53.6**		-		78.9	**25.0**	53.9
CXR14	BiomedCLIP	55.3	28.9	26.4		-		56.2	**41.6**	**14.6**

Table 1: Main results. Best WG and Gap performance **bolded**, second best underlined.



(a) (b)

Figure 3: (a) Original (green) and projected (red) input embeddings x, and label embeddings c^0 and c^1. (b) label embeddings c^0 and c^1, harmful insight embeddings v^k (black star) and helpful insight embeddings u^j (blue star) embeddings from the ALIGN model. We observe that the projected embeddings (red) still lie within the original embedding space, even with reduced variance. In contrast, when examining the CLIP model embeddings (Figure 3a (right)), we observe that the projected embeddings are significantly distant from the original ones. Unsurprisingly, Figure 3b (left) reveals that v^j and u^k (harmful and helpful insight embeddings in black and blue stars, respectively) are not distinguishable in the text embedding space of ALIGN, collapsing the input embeddings after RoboShot is applied.

5.2 Improving language models

Setup. We experimented on four text classification datasets: **CivilComments-WILDS** [5, 19], **HateXplain** [27], **Amazon-WILDS** [19, 30] and **Gender Bias** classification dataset [11, 28]. We use the default test splits of all datasets. In text experiments, the distinctions between harmful and helpful insights are less clear than for images. For this reason, we only use harmful vector rejection (line 3 in RoboShot) in text experiments. CivilComments and HateXplain are toxic classification datasets with unwanted correlation between toxicity labels and mentions of demographics (e.g., male, female, mentions of religions). The datasets are annotated with demographic mentions of each text, and we directly use them to construct v^j. For Amazon and Gender Bias datasets, we query LMs with task descriptions. All experiments are conducted using frozen pretrained models.

Results. Table 2 shows that RoboShot also improves zero-shot text classification in text datasets, as shown by our consistent boost over the baselines across all datasets.

Dataset	Model	ZS			RoboShot		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
CivilComments	BERT	48.1	33.3	14.8	49.7	**42.3**	**7.4**
	Ada	56.2	43.2	13.0	56.6	**44.9**	**11.7**
HateXplain	BERT	60.4	0.0	60.4	57.3	**14.0**	**43.3**
	Ada	62.8	14.3	48.5	63.6	**21.1**	**42.5**
Amazon	BERT	81.1	64.2	16.8	81.0	**64.4**	**16.6**
	Ada	81.2	63.4	**17.8**	82.9	**63.8**	19.1
Gender Bias	BERT	84.8	83.7	1.1	85.1	**84.9**	**0.2**
	Ada	77.9	60.0	17.9	78.0	**60.1**	17.9

Table 2: RoboShot text zero-shot classification. Best WG in **bold**.

Dataset	ZS		Ours (ChatGPT)		Ours (Flan-T5)		Ours (GPT2)		Ours (LLaMA)	
	AVG	WG	AVG	WG	AVG	WG	AVG	WG	AVG	WG
Waterbirds	80.7	27.9	82.0	**54.4**	72.1	32.4	88.0	<u>39.9</u>	84.8	36.5
CelebA	80.1	72.7	84.8	<u>80.5</u>	77.5	68.2	80.3	74.1	84.2	**82.0**
PACS	96.7	<u>82.1</u>	97.0	**86.3**	96.2	80.3	97.2	74.0	94.8	71.9
VLCS	75.6	20.5	76.5	**33.0**	69.6	20.5	75.5	<u>26.1</u>	72.0	18.2

Table 3: RoboShot with LMs of varying capacity. Best WG **bolded**, second best <u>underlined</u>

5.3 Extracting concepts from LMs with varying capacities

Setup. We use LMs with different capacities: **ChatGPT** [31], **Flan-T5** [8], **GPT2** [33], and **LLaMA** [38], to get harmful and helpful features insights (v^j and u^k).

Results. Table 3 shows that RoboShot can get insights on v^j and u^k from LMs of various capacities and improves zero-shot performance. Even though the the LM capacity correlates with the zero-shot performance, RoboShot with weaker LMs still outperforms zero-shot (ZS) baseline.

5.4 Ablations

Setup. We run RoboShot with only harmful component mitigation (reject v^j: RoboShot line 3), only boosting helpful vectors (amplify u^k: RoboShot line 7), and both.

Results. The combination of both projections often achieves the best performance, as shown in Table 4. Figure 4 provides insights into the impact of each projection. Rejecting v^j reduces variance in one direction, while increasing u^k amplifies variance in the orthogonal direction. When both projections are applied, they create a balanced mixture. We note that when doing both projections does not improve the baseline, using only u^k or v^j still outperforms the baseline. For instance, the ALIGN model in the Waterbirds dataset achieves the best performance with only u^k projection. This suggests that in certain cases, harmful and helpful concepts are intertwined in the embedding space, and using just one projection can be beneficial. We leave further investigation to future work.

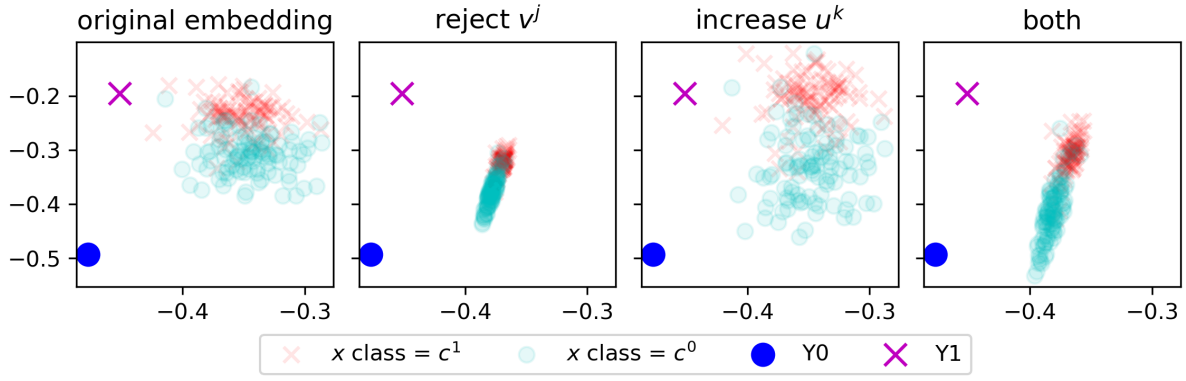


Figure 4: The effect of v^j (reject), u^j (increase), and both projections

Dataset	Model	ZS			Ours (v^j only)			Ours (u^k only)			Ours (both)		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
Waterbirds	CLIP (ViT-B-32)	80.7	27.9	52.8	82.0	50.4	31.6	82.6	30.2	52.4	83.0	**54.4**	**28.6**
	CLIP (ViT-L-14)	88.7	27.3	61.4	82.7	35.8	46.9	88.3	29.8	58.5	79.9	**45.2**	**34**.7
	ALIGN	72.0	50.3	21.7	56.4	41.6	14.8	62.8	**56.4**	**6.4**	50.9	41.0	9.9
	AltCLIP	90.1	35.8	54.3	81.4	**59.0**	**22.4**	89.1	35.2	53.9	78.5	54.8	23.7
CelebA	CLIP (ViT-B-32)	80.1	72.7	7.4	85.2	**81.5**	**3.7**	79.6	71.3	8.3	84.8	80.5	4.3
	CLIP (ViT-L-14)	80.6	74.3	6.3	85.9	**82.8**	3.1	80.0	73.1	6.9	85.5	82.6	**2.9**
	ALIGN	81.8	77.2	4.6	83.9	78.0	5.7	83.9	81.4	**2.5**	86.3	**83.4**	2.9
	AltCLIP	82.3	**79.7**	**2.6**	86.1	75.6	10.5	81.9	79.0	2.9	86.0	77.2	8.8
PACS	CLIP (ViT-B-32)	96.7	82.1	14.6	97.0	83.7	13.3	96.6	84.2	12.4	97.0	**86.3**	**10.7**
	CLIP (ViT-L-14)	98.1	79.8	18.3	98.0	79.8	18.2	98.1	83.8	14.3	98.1	**83.9**	**14.2**
	ALIGN	95.8	77.1	18.7	95.8	**78.0**	**17.8**	95.1	71.1	24.0	95.0	73.8	21.2
	AltCLIP	98.5	82.6	15.9	98.4	83.0	15.4	98.6	88.8	9.8	98.7	**89.5**	**9.2**
VLCS	CLIP (ViT-B-32)	75.6	20.5	55.1	75.6	22.7	52.9	76.4	29.5	46.9	76.5	**33.0**	**43.5**
	CLIP (ViT-L-14)	72.6	4.2	68.4	70.9	6.8	64.1	73.4	8.9	64.5	71.1	**12.6**	**58.5**
	ALIGN	78.8	33.0	45.8	78.2	30.7	47.5	78.0	**43.2**	**34.8**	77.6	39.8	37.8
	AltCLIP	78.3	24.7	**53.6**	77.5	24.4	53.1	79.0	20.5	58.5	78.9	**25.0**	53.9
CXR14	BiomedCLIP	55.3	28.9	26.4	55.7	**41.8**	**13.9**	54.8	21.8	33.0	56.2	41.6	14.6

Table 4: Main results. Best WG and Gap performance **bolded**, second best underlined.

5.5 Isolating concepts by averaging relevant concepts

Concept	Original	Average
Green	0.237	**0.241**
Red	0.236	**0.240**
Blue	0.213	**0.229**
Yellow	0.237	**0.246**
Square	0.214	**0.220**

ZS			RoboShot Original			RoboShot Average		
AVG	WG	Gap	AVG	WG	Gap	AVG	WG	Gap
86.6	29.6	57.0	87.1	31.5	55.6	78.8	**55.1**	**23.7**

Table 5: Left: Cosine similarity between concept images and original embedding vs. averaged embedding. Right: RoboShot on Waterbirds with original vs. averaged embedding

We conduct experiments to test the viability of our concept modeling (equation 1 in section 3.1). Specifically, we aim to find out if CLIP input representation x contains harmful, helpful, and benign components (z_s, z_r, and z_b respectively in equation 1).

Can we partition CLIP input representation into harmful, helpful, and benign concepts? For a particular concept (e.g., "land"), we hypothesize that the true concept component is mixed with other concept components due to the signal in training data. For instance, land often co-occurs with sky, cattle, and other objects. Thus, the CLIP representation of "land" is entangled with these other concepts. To potentially isolate the helpful concept, we ask LM for an exhaustive list of concepts related to "land" and average the embedding of all related concepts. The intuition here is that a clean "land" component exists in each individual embedding, and the remaining is likely to be random, which can be averaged out and leave us with the true concept.

To verify this intuition, we compare the original and averaged embeddings of concepts listed in Table 5 (left). For each concept, we get 100 Google image search results and filter out noisy images (e.g., images with large text and artifacts) by eyeballing. We then report the average cosine similarity between the images and original embedding vs. the embedding from our averaging procedure. Averaged embedding has higher cosine similarity across the board than original CLIP embedding. To some extent, this indicates that the averaging procedure isolates the true concept.

Does RoboShot gain improvement with isolated concept? Table 5 (right) compares RoboShot with removing harmful insights using original CLIP embedding vs. averaged embedding. We use Waterbirds dataset because the harmful insights are known in prior. To isolate the effect of our averaging procedure, we use "landbird" and "waterbird" as labels without additional prompts (e.g., "a picture of [label]"), and we only use "land" and

"water" as the harmful insights to remove, which causes slight difference with the results reported in Table 1. Confirming our intuition, using the averaged embedding results in better WG performance and smaller Gap.

6 Conclusion

We introduced ROBOSHOT, a fine-tuning-free system that robustifies zero-shot pretrained models in a truly zero-shot way. Theoretically, we characterized the quantities required to obtain improvements over vanilla zero-shot classification. Empirically, we found that ROBOSHOT improves both multi-modal and language model zero-shot performance, has sufficient versatility to apply to various base models, and can use insights from less powerful language models.

References

[1] Prince Osei Aboagye, Yan Zheng, Jack Shunn, Chin-Chia Michael Yeh, Junpeng Wang, Zhongfang Zhuang, Huiyuan Chen, Liang Wang, Wei Zhang, and Jeff Phillips. Interpretable debiasing of vectorized language representations with iterative orthogonalization. In *The Eleventh International Conference on Learning Representations*.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[5] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[6] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

[7] Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv:2210.12530*, 2022.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[9] Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. Discovering latent concepts learned in bert. In *International Conference on Learning Representations*.

[10] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR, 2019.

[11] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://www.aclweb.org/anthology/2020.emnlp-main.23.

[12] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018.

[13] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[15] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. *arXiv preprint arXiv:2212.00638*, 2022.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[17] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

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

[19] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[20] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[21] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138, 2020.

[22] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. Using sentences as semantic representations in large scale zero-shot learning. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 641–645. Springer, 2020.

[23] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*, 2022.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[25] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/liu21f.html`.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[27] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[28] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2014. URL `https://aclanthology.org/D17-2014`.

[29] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

[30] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language*

processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pages 188–197, 2019.

[31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[32] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18092–18102, 2022.

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[35] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

[37] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

[40] Junyang Wang, Yi Zhang, and Jitao Sang. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*, 2022.

[41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.

[42] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. *arXiv preprint arXiv:2304.03916*, 2023.

[43] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *arXiv preprint arXiv:2207.07180*, 2022.

[44] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing, 2023. URL `https://arxiv.org/abs/2303.00915`.

A Glossary

The glossary is given in Table 6.

Symbol	Definition
x	input vector
y, $\hat{y}$	class label, prediction
c^i	embedding of class i
$z_1, \ldots, z_k$	The concept vectors consisting of orthonormal vectors
v^i, u^j	insight representations
α_j	The coefficient of input x with respect to the concept z_j (before ROBOSHOT)
A_j	The coefficient of transformed input $\hat{x}$ with respect to the concept z_j (after ROBOSHOT)
$\beta_{i,j}$	The coefficient of j-th class embedding with respect to the concept z_i
$\gamma_{i,j}$	The coefficient of j-th insight vector with respect to the concept z_i
S	the number of harmful concepts
R	the number of helpful concepts
B	the number of benign concepts
g	text encoder to get embeddings
s^i	text string for insight vectors
$\sigma_{\text{benign}}, \sigma_{\text{insight}}$	noise rates in the coefficients of benign/insight concepts

Table 6: Glossary of variables and symbols used in this paper.

B Theory details

B.1 Harmful concept removal

As the simplest form of ROBOSHOT, we consider the case of ROBOSHOT the harmful concept removal only, without boosting helpful concepts. Recall our noise model:

$$x = \sum_{s=1}^{S} \alpha_s z_s + \sum_{r=S+1}^{S+R} \alpha_r z_r + \sum_{b=S+R+1}^{S+R+B} \alpha_b z_b$$

$$v^t = \sum_{s=1}^{S} \gamma_{s,t} z_s + \sum_{r=S+1}^{S+R} \gamma_{r,t} z_r + \sum_{b=S+R+1}^{S+R+B} \gamma_{b,t} z_b \qquad (1 \leq t \leq S)$$

. Again, we assume that benign coefficients are drawn from a zero-centered Gaussian distribution, i.e. $\alpha_b, \gamma_{b,t} \sim \mathcal{N}(0, \sigma_{benign})$ and also helpful coefficients and non-target harmful coefficients are assumed to be drawn from a Gaussian distribution, i.e. $\gamma_{q,t} \sim \mathcal{N}(0, \sigma_{insight})$, where $1 \leq q \leq R$, $q \neq t$ so that only $\gamma_{t,t}$ is a constant.

B.1.1 Effects on harmful coefficients

Now we prove the following Theorem.

Theorem B.1. *Under the noise model described above, the post-removal coefficient A_s for harmful concept z_s satisfies*

$$|\mathbb{E}[A_s]| \leq \left| \frac{(k-1)\alpha_s \sigma_{insight}^2}{\gamma_{s,s}^2} \right| + \left| \sum_{t \neq s}^{S} \frac{\alpha_s \sigma_{insight}^2}{\gamma_{t,t}^2} \right|,$$

where k is the number of concepts.

Proof. Let $\hat{x}$ be the output of harmful concept removal procedure such that

$$\begin{aligned}\hat{x} &= x - \sum_{s=1}^{S} \frac{x^T v^s}{||v^s||^2} v^s \\ &= \sum_{i=1}^{k} \alpha_i z_i - \sum_{s=1}^{S} \frac{\sum_i^k \alpha_i \gamma_{i,s}}{\sum_{l=1}^{k} \gamma_{l,s}^2} (\sum_{j=1}^{k} \gamma_{j,s} z_j)\end{aligned}$$

As the first step, we sort out the coefficients of features. For notational convenience, let $T_s = \sum_{l=1}^{k} \gamma_{l,s}^2$. Then,

$$\begin{aligned}\hat{x} &= \sum_{i=1}^{k} \alpha_i z_i - \sum_{s=1}^{S} \frac{\sum_{i=1}^k \alpha_i \gamma_{i,s}}{T_s} (\sum_{j=1}^{k} \gamma_{j,s} z_j) \\ &= \sum_{i=1}^{k} \alpha_i z_i - \sum_{s=1}^{S} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\alpha_i \gamma_{i,s} \gamma_{j,s}}{T_s} z_j \\ &= \sum_{j=1}^{k} \alpha_j z_j - \sum_{j=1}^{k} \sum_{s=1}^{S} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,s} \gamma_{j,s}}{T_s} z_j \\ &= \sum_{j=1}^{k} \left(\alpha_j - \sum_{s=1}^{S} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,s} \gamma_{j,s}}{T_s} \right) z_j\end{aligned}$$

Thus we can get the expression for the coefficient of the target feature z_s $(1 \leq s \leq S)$,

$$A_s = \alpha_s - \sum_{t=1}^{S} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{s,t}}{T_t}$$

Next, we get the bound of the absolute expectation $|\mathbb{E}[A_s]|$.

$$\begin{aligned}|\mathbb{E}[A_s]| &= \left| \mathbb{E}\left[\alpha_s - \sum_{t=1}^{S} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{s,t}}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right| \\ &\leq \left| \mathbb{E}\left[\alpha_s - \sum_{t=1}^{S} \frac{\alpha_s \gamma_{s,t}^2}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right| + \left| \sum_{t=1}^{S} \mathbb{E}\left[\frac{\sum_{i=1, i \neq s}^{S} \alpha_i \gamma_{i,t} \gamma_{s,t}}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right|\end{aligned}$$

Here, the second term on RHS is 0 by independence, i.e.

$$\begin{aligned}\left| \mathbb{E}\left[\frac{\sum_{i=1, i \neq s}^{S} \alpha_i \gamma_{i,t} \gamma_{s,t}}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right| &\leq \left| \mathbb{E}\left[\frac{\sum_{i=1, i \neq s}^{k} \alpha_i \gamma_{i,t} \gamma_{s,t}}{\gamma_{t,t}^2} \right] \right| \\ &= \left| \sum_{i=1, i \neq s}^{k} \frac{\alpha_i}{\gamma_{t,t}^2} \mathbb{E}[\gamma_{i,t} \gamma_{s,t}] \right| = 0\end{aligned}$$

since $\mathbb{E}[\gamma_{s,t} \gamma_{j,t}] = 0$ by independence. Now we split the first term and get the bounds separately.

$$\begin{aligned}|\mathbb{E}[A_s]| &\leq \left| \mathbb{E}\left[\alpha_s - \sum_{t=1}^{S} \frac{\alpha_s \gamma_{s,t}^2}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right| \\ &\leq \left| \mathbb{E}\left[\alpha_s - \frac{\alpha_s \gamma_{s,s}^2}{\sum_{l=1}^{s} \gamma_{l,s}^2} \right] \right| + \left| \sum_{t=1, t \neq s}^{S} \mathbb{E}\left[\frac{\alpha_s \gamma_{s,t}^2}{\sum_{l=1}^{k} \gamma_{l,t}^2} \right] \right|\end{aligned}$$

The upper bound for the first term can be obtained by

$$\begin{aligned}\left|\mathbb{E}\left[\alpha_s - \frac{\alpha_s\gamma_{s,s}^2}{\sum_{l=1}^{s}\gamma_{l,s}^2}\right]\right| &= \left|\mathbb{E}\left[-\frac{\sum_{i\neq s}^{k}\alpha_s\gamma_{i,s}^2}{\sum_{l=1}^{s}\gamma_{l,s}^2}\right]\right| \\ &\leq \left|\mathbb{E}\left[\frac{\sum_{i\neq s}^{k}\alpha_s\gamma_{i,s}^2}{\gamma_{s,s}^2}\right]\right| \\ &\leq \left|\frac{\alpha_s}{\gamma_{s,s}^2}\sum_{i\neq s}^{k}\mathbb{E}\left[\gamma_{i,s}^2\right]\right| \\ &\leq \left|\frac{(k-1)\alpha_s\sigma_{insight}^2}{\gamma_{s,s}^2}\right|\end{aligned}$$

. And, for the second term,

$$\begin{aligned}\left|\sum_{t=1,t\neq s}^{S}\mathbb{E}\left[\frac{\alpha_s\gamma_{s,t}^2}{\sum_{i=1}^{k}\gamma_{i,t}^2}\right]\right| &\leq \left|\sum_{t=1,t\neq s}^{S}\mathbb{E}\left[\frac{\alpha_s\gamma_{s,t}^2}{\gamma_{t,t}^2}\right]\right| \\ &= \left|\sum_{t=1,t\neq s}^{S}\frac{\alpha_s}{\gamma_{t,t}^2}\mathbb{E}\left[\gamma_{s,t}^2\right]\right|\end{aligned}$$

Combining two bounds, we get the proposed result.

$$|\mathbb{E}\left[A_s\right]| \leq \left|\frac{(k-1)\alpha_s\sigma_{insight}^2}{\gamma_{s,s}^2}\right| + \left|\sum_{t\neq s}^{S}\frac{\alpha_s\sigma_{insight}^2}{\gamma_{t,t}^2}\right|,$$

□

While the constant $(k-1)$ can look daunting since it actually increases as the number of concepts increases, a bound less affected by $\sigma_{insight}^2$ exists as well, scaling down the target coefficient α_s.

Corollary B.1.1. *Under the noise model of Theorem B.1, the post-removal coefficient for harmful concept s satisfies*

$$|\mathbb{E}\left[A_s\right]| \leq \left|\alpha_s\frac{(k-1)\sigma_{insight}^2}{\gamma_{s,s}^2 + (k-1)\sigma_{insight}^2}\right| + \left|\sum_{t\neq s}^{S}\frac{\alpha_s\sigma_{insight}^2}{\gamma_{t,t}^2}\right|,$$

where k is the number of concepts.

Proof. With the identical steps to the proof of Theorem B.1, we can obtain

$$\begin{aligned}|\mathbb{E}\left[A_s\right]| &\leq \left|\mathbb{E}\left[\alpha_s - \sum_{t=1}^{S}\frac{\alpha_s\gamma_{s,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\ &\leq \left|\mathbb{E}\left[\alpha_s - \frac{\alpha_s\gamma_{s,s}^2}{\sum_{l=1}^{k}\gamma_{l,s}^2}\right]\right| + \left|\sum_{t=1,t\neq s}^{S}\mathbb{E}\left[\frac{\alpha_s\gamma_{s,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\ &\leq \left|\mathbb{E}\left[\alpha_s - \frac{\alpha_s\gamma_{s,s}^2}{\sum_{l=1}^{k}\gamma_{l,s}^2}\right]\right| + \left|\sum_{t=1,t\neq s}^{S}\frac{\alpha_s}{\gamma_{t,t}^2}\mathbb{E}\left[\gamma_{s,t}^2\right]\right|\end{aligned}$$

We improve the first term as follows.

$$
\begin{aligned}
\left|\mathbb{E}\left[\alpha_s - \frac{\alpha_s\gamma_{s,s}^2}{\sum_{l=1}^{s}\gamma_{l,s}^2}\right]\right| &= \left|\alpha_s - \alpha_s\mathbb{E}\left[\frac{\gamma_{s,s}^2}{\sum_{l=1}^{s}\gamma_{l,s}^2}\right]\right| \\
&\leq \left|\alpha_s - \alpha_s\frac{\gamma_{s,s}^2}{\mathbb{E}\left[\sum_{l=1}^{s}\gamma_{l,s}^2\right]}\right| \quad \because \text{Jensen's inequality} \\
&= \left|\alpha_s\left(1 - \frac{\gamma_{s,s}^2}{\mathbb{E}\left[\sum_{l=1}^{s}\gamma_{l,s}^2\right]}\right)\right| \\
&= \left|\alpha_s\left(1 - \frac{\gamma_{s,s}^2}{\gamma_{s,s}^2 + (k-1)\sigma_{insight}^2}\right)\right| \\
&= \left|\alpha_s\left(\frac{(k-1)\sigma_{insight}^2}{\gamma_{s,s}^2 + (k-1)\sigma_{insight}^2}\right)\right|
\end{aligned}
$$

□

B.1.2 Effects on helpful, benign coefficients

Based on the coefficient expression

$$
A_q = \alpha_q - \sum_{t=1}^{S}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,t}\gamma_{q,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}
$$

, we analyze the bound of $|\mathbb{E}[A_q]|$ for $S+1 \leq q \leq k$. Basically, the following theorem implies helpful, benign coefficients are less affected than harmful coefficients as long as the harmful coefficients of insight embeddings are significant and the noise is small.

Theorem B.2. *Under the same noise model described above, the post-removal coefficient for helpful or benign concept q satisfies*

$$
|\mathbb{E}[A_q] - \alpha_q| \leq \left|\sum_{t=1}^{S}\frac{\alpha_q\sigma_{insight}^2}{\gamma_{t,t}^2}\right|.
$$

Proof. The proof technique is essentially identical to Theorem B.1.

$$
\begin{aligned}
|\mathbb{E}[A_q] - \alpha_q| &= \left|\alpha_q - \mathbb{E}\left[\alpha_q - \sum_{t=1}^{S}\frac{\alpha_q\gamma_{q,t}^2 + \sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\
&\leq \left|\mathbb{E}\left[\sum_{t=1}^{S}\frac{\alpha_q\gamma_{q,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| + \left|\mathbb{E}\left[\frac{\sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\
&= \left|\mathbb{E}\left[\sum_{t=1}^{S}\frac{\alpha_q\gamma_{q,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \quad \because \left|\mathbb{E}\left[\frac{\sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| = 0 \\
&\leq \left|\sum_{t=1}^{S}\frac{\alpha_q}{\gamma_{t,t}^2}\mathbb{E}\left[\gamma_{q,t}^2\right]\right| \\
&= \left|\sum_{t=1}^{S}\frac{\alpha_q\sigma_{insight}^2}{\gamma_{t,t}^2}\right|
\end{aligned}
$$

□

This bound implies the differences of helpful or benign features by harmful concept removal are proportional to the noise of insight embeddings $\sigma_{insight}^2$, and inversely proportional to the coefficients of harmful coefficients of insight embeddings.

B.2 Helpful concept addition

With a similar fashion to the harmful concept removal, we consider the following noise model for the helpful concept addition.

$$x = \sum_{s=1}^{S} \alpha_s z_s + \sum_{r=S+1}^{S+R} \alpha_r z_r + \sum_{b=S+R+1}^{S+R+B} \alpha_b z_b$$

$$v^t = \sum_{s=1}^{S} \gamma_{s,t} z_s + \sum_{r=S+1}^{S+R} \gamma_{r,t} z_r + \sum_{b=S+R+1}^{S+R+B} \gamma_{b,t} z_b \qquad (S+1 \leq t \leq S+R)$$

. Again, we assume that benign coefficients are drawn from a zero-centered Gaussian distribution, i.e. $\alpha_b, \gamma_{b,t} \sim \mathcal{N}(0, \sigma_{benign})$ and also harmful coefficients and non-target helpful coefficients are assumed to be drawn from another Gaussian distribution, i.e. $\gamma_{q,t} \sim \mathcal{N}(0, \sigma_{insight})$, where $1 \leq q \leq S+R$, $q \neq s$ so that only $\gamma_{t,t}$ are constants.

B.2.1 Lower bound for the coefficient of helpful concept

To show the lower bound for the coefficient of helpful concepts, we need additional mild assumptions. For $S+1 \leq r \leq S+R$

1. $\alpha_r \geq 0$
2. $\gamma_{r,r}^2 \geq \sigma_{insight}^2$

The first assumption can be interpreted that the input embedding is already somewhat aligned with the label embeddings' concepts — since typically pretrained models provide embeddings aligned with class text, it can be justified. The second assumption is also a natural assumption: the signal is stronger than noise. Now we state Theorem and show the proof of the theorem.

Theorem B.3. *Assuming $\alpha_r \geq 0, \gamma_{r,r}^2 \geq \sigma_{insight}^2$ for $S+1 \leq r \leq S+R$ under the described noise model, the post-addition coefficient for helpful concept r satisfies*

$$\mathbb{E}\left[A_r\right] \geq \left(1 + \frac{\gamma_{r,r}^2}{\gamma_{r,r}^2 + (k-1)\sigma_{insight}^2}\right) \alpha_r$$

Proof. Let $\hat{x}$ be the output of helpful concept addition procedure such that

$$\begin{aligned}
\hat{x} &= x + \sum_{t=S+1}^{S+R} \frac{x^T v^s}{||v^s||^2} v^s \\
&= \sum_{i=1}^{k} \alpha_i z_i + \sum_{t=S+1}^{S+R} \frac{\sum_{i=1}^{k} \alpha_i \gamma_{i,s}}{\sum_{l=1}^{k} \gamma_{l,s}^2} (\sum_{j=1}^{k} \gamma_{j,s} z_j)
\end{aligned}$$

As the first step, we sort out the coefficients of concepts. For notational convenience, let $T_t = \sum_{l=1}^{k} \gamma_{l,t}^2$. Then,

$$\begin{aligned}
\hat{x} &= \sum_{i=1}^{k} \alpha_i z_i + \sum_{t=S+1}^{S+R} \frac{\sum_{i=1}^{k} \alpha_i \gamma_{i,t}}{T_t} (\sum_{j=1}^{k} \gamma_{j,t} z_j) \\
&= \sum_{i=1}^{k} \alpha_i z_i + \sum_{t=S+1}^{S+R} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{j,t}}{T_t} z_j \\
&= \sum_{j=1}^{k} \alpha_j z_j + \sum_{j=1}^{k} \sum_{t=S+1}^{S+R} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{j,t}}{T_t} z_j \\
&= \sum_{j=1}^{k} \left(\alpha_j + \sum_{t=S+1}^{S+R} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{j,t}}{T_t} \right) z_j
\end{aligned}$$

Thus we can get the expression for the coefficient of the target concept $z_r \quad (S+1 \le r \le S+R)$,

$$A_r = \alpha_r + \sum_{t=S+1}^{S+R} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{r,t}}{T_t}$$

Then,

$$\begin{aligned}
\mathbb{E}[A_r] &= \mathbb{E}\left[\alpha_r + \sum_{t=S+1}^{S+R} \sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,t} \gamma_{r,t}}{T_t}\right] \\
&= \alpha_r + \mathbb{E}\left[\frac{\alpha_r \gamma_{r,r}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] + \mathbb{E}\left[\sum_{t=S+1, t\neq r}^{S+R} \frac{\alpha_t \gamma_{r,t}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] + \mathbb{E}\left[\sum_{t=S+1}^{S+R} \frac{\sum_{i=1, i\neq r}^{k} \alpha_i \gamma_{i,t} \gamma_{r,t}}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] \\
&\ge \alpha_r + \mathbb{E}\left[\frac{\alpha_r \gamma_{r,r}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] + \mathbb{E}\left[\sum_{t=S+1}^{S+R} \frac{\alpha_i \gamma_{i,t} \gamma_{r,t}}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] \\
&\ge \alpha_r + \mathbb{E}\left[\frac{\alpha_r \gamma_{r,r}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] + \mathbb{E}\left[\sum_{t=S+1}^{S+R} \frac{\alpha_i \gamma_{i,t} \gamma_{r,t}}{k \gamma_{r,r}^2}\right]
\end{aligned}$$

Here, the third term can be dropped since $\gamma_{i,t}$ and $\gamma_{r,t}$ are independent. Thus,

$$\begin{aligned}
\mathbb{E}[A_r] &\ge \alpha_r + \mathbb{E}\left[\frac{\alpha_r \gamma_{r,r}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] \\
&\ge \alpha_r + \alpha_r \gamma_{r,r}^2 \mathbb{E}\left[\frac{1}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right] \\
&\ge \alpha_r + \alpha_r \gamma_{r,r}^2 \frac{1}{\mathbb{E}\left[\sum_{l=1}^{k} \gamma_{l,r}^2\right]} \\
&= \alpha_r + \alpha_r \gamma_{r,r}^2 \frac{1}{\gamma_{r,r}^2 + (k-1)\sigma_{insight}^2}
\end{aligned}$$

Finally, we obtain the result.

$$\mathbb{E}[A_r] \ge \left(1 + \frac{\gamma_{r,r}^2}{\gamma_{r,r}^2 + (k-1)\sigma_{insight}^2}\right) \alpha_r$$

□

Note that the nonnegative condition can be dropped by keeping $\mathbb{E}\left[\frac{\alpha_t \gamma_{r,t}^2}{\sum_{l=1}^{k} \gamma_{l,r}^2}\right]$ where $\alpha_t < 0$ terms, which linearly loosens the lower bound.

B.2.2 Effects on harmful, benign coefficients

For the notational convenience, let $I_{helpful}^c$ be the non-helpful concept index set such that $I_{helpful}^c = \{i \in \mathbb{N} | i \le S \text{ or } S+R+1 \le i \le S+R+B\}$. For $q \in I_R^c$, we obtain the bound of effects on harmful, benign coefficients with a similar fashion to the harmful concept removal case.

Theorem B.4. *Under the same noise model described above, the post-addition coefficient for helpful or benign concept q satisfies*

$$|\mathbb{E}[A_q] - \alpha_q| \le \left| \sum_{t=S+1}^{S+R} \frac{\alpha_q \sigma_{insight}^2}{\gamma_{t,t}^2} \right|.$$

Proof.

$$
\begin{aligned}
|\mathbb{E}\left[A_q\right] - \alpha_q| &= \left|\alpha_q - \mathbb{E}\left[\alpha_q + \sum_{t=1}^{S} \frac{\alpha_q\gamma_{q,t}^2 + \sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\
&\leq \left|\mathbb{E}\left[\sum_{t=S+1}^{S+R}\frac{\alpha_q\gamma_{q,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| + \left|\mathbb{E}\left[\frac{\sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \\
&= \left|\mathbb{E}\left[\sum_{t=S+1}^{S+R}\frac{\alpha_q\gamma_{q,t}^2}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| \quad \because \left|\mathbb{E}\left[\frac{\sum_{j=1,j\neq q}\alpha_q\gamma_{q,t}\gamma_{j,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2}\right]\right| = 0 \\
&\leq \left|\sum_{t=S+1}^{S+R}\frac{\alpha_q}{\gamma_{t,t}^2}\mathbb{E}\left[\gamma_{q,t}^2\right]\right| \\
&= \left|\sum_{t=S+1}^{S+R}\frac{\alpha_q\sigma_{insight}^2}{\gamma_{t,t}^2}\right|
\end{aligned}
$$

□

B.3 Combined main results

Now, we are ready to provide the combine main result, i.e. the coefficient bounds with harmful concept removal and helpful concept addition. The noise model can be described as follows.

$$
x = \sum_{s=1}^{S}\alpha_s z_s + \sum_{r=S+1}^{S+R}\alpha_r z_r + \sum_{b=S+R+1}^{S+R+B}\alpha_b z_b
$$

$$
v^t = \sum_{s=1}^{S}\gamma_{s,t} z_s + \sum_{r=S+1}^{S+R}\gamma_{r,t} z_r + \sum_{b=S+R+1}^{S+R+B}\gamma_{b,t} z_b \qquad (1 \leq t \leq\ S+R)
$$

$$
\alpha_b, \gamma_{b,t} \sim \mathcal{N}(0, \sigma_{benign})
$$

$$
\gamma_{q,t} \sim \mathcal{N}(0, \sigma_{insight})
$$

, where $1 \leq q \leq S+R$, $q \neq s$ so that only $\gamma_{t,t}$ is a constant. We can obtain the expression for each coefficient as before.

$$
\hat{x} = \sum_{j=1}\left(a_j - \sum_{s=1}^{S}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,s}\gamma_{j,s}}{T_s} + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,r}\gamma_{j,r}}{T_r}\right) z_j
$$

$$
A_q = a_q - \sum_{s=1}^{S}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,s}\gamma_{q,s}}{T_s} + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,r}\gamma_{q,r}}{T_r}
$$

, where A_q is the coefficient of $z_q (1 \leq q \leq k)$ after ROBOSHOT(ignoring normalization) and $T_t = \sum_{l=1}^{k}\gamma_{l,t}^2$. Using the results from the previous subsections, we provide an upper bound on harmful coefficients, a lower bound on helpful coefficients, and an upper bound on the change in the benign coefficients. We restate Theorem 4.1, 4.2 and provide proofs.

Theorem 4.1. *Under the combined noise model described above, the post-*ROBOSHOT *coefficient for harmful concept q $(1 \leq q \leq S)$ satisfies*

$$
|\mathbb{E}\left[A_q\right]| \leq \left|\frac{(k-1)\alpha_q\sigma_{insight}^2}{\gamma_{q,q}^2}\right| + \left|\sum_{t=1,t\neq q}^{S+R}\frac{\alpha_q\sigma_{insight}^2}{\gamma_{t,t}^2}\right|,
$$

where k is the number of concepts.

Proof.

$$
\begin{aligned}
|\mathbb{E}[A_q]| &= \left| \mathbb{E}\left[a_q - \sum_{s=1}^{S}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] \right| \\
&\leq \left| \frac{(k-1)\alpha_q \sigma_{insight}^2}{\gamma_{q,q}^2} \right| + \left| \sum_{s=1, s\neq q}^{S} \frac{\alpha_q \sigma_{insight}^2}{\gamma_{s,s}^2} \right| + \left| \sum_{t=S+1}^{S+R} \frac{\alpha_q \sigma_{insight}^2}{\gamma_{t,t}^2} \right| \\
&= \left| \frac{(k-1)\alpha_q \sigma_{insight}^2}{\gamma_{q,q}^2} \right| + \left| \sum_{t=1, t\neq q}^{S+R} \frac{\alpha_q \sigma_{insight}^2}{\gamma_{t,t}^2} \right| \quad \because \text{two terms have the same sign by } a_q
\end{aligned}
$$

□

Next, we state the lower bound for the helpful features. Still, we assume the signs of helpful, harmful concepts in input embeddings for the clarity of theorem.

$$\alpha_s \leq 0 \quad (1 \leq s \leq S)$$

$$\alpha_r \geq 0 \quad (S+1 \leq r \leq S+R)$$

Also, we assume $\gamma_{t,t}^2 \geq \sigma_{insight}^2 \quad (1 \leq t \leq S+R)$

Theorem 4.2. *With additional assumptions* $\alpha_s \leq 0 \quad (1 \leq s \leq S), \alpha_r \geq 0 \quad (S+1 \leq r \leq S+R), \gamma_{t,t}^2 \geq \sigma_{insight}^2$ *under the combined noise model, the post-*ROBOSHOT *coefficient for helpful concept* $q (S+1 \leq q \leq S+R)$ *satisfies*

$$\mathbb{E}[A_q] \geq \left(1 + \frac{\gamma_{q,q}^2}{\gamma_{q,q}^2 + (k-1)\sigma_{insight}^2} \right) \alpha_q$$

Proof.

$$
\begin{aligned}
\mathbb{E}[A_q] &= \mathbb{E}\left[a_q - \sum_{s=1}^{S}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] \\
&= \mathbb{E}\left[a_q + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] - \mathbb{E}\left[\sum_{s=1}^{S}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} \right] \\
&= \mathbb{E}\left[a_q + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] - \mathbb{E}\left[\sum_{s=1}^{S} \frac{\alpha_s \gamma_{q,s}^2}{T_s} \right] - \mathbb{E}\left[\sum_{s=1}^{S}\sum_{i=1, i\neq q}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} \right]
\end{aligned}
$$

Here, $\mathbb{E}\left[\sum_{s=1}^{S}\sum_{i=1, i\neq q}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} \right]$ is zero by independence, and $-\mathbb{E}\left[\sum_{s=1}^{S} \frac{\alpha_s \gamma_{q,s}^2}{T_s} \right] \geq 0$ since $\alpha_s \leq 0$ by assumption, which can be dropped for a lower bound.

$$
\begin{aligned}
\mathbb{E}[A_q] &= \mathbb{E}\left[a_q + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] - \mathbb{E}\left[\sum_{s=1}^{S} \frac{\alpha_s \gamma_{q,s}^2}{T_s} \right] - \mathbb{E}\left[\sum_{s=1}^{S}\sum_{i=1, i\neq q}^{k} \frac{\alpha_i \gamma_{i,s}\gamma_{q,s}}{T_s} \right] \\
&\geq \mathbb{E}\left[a_q + \sum_{r=S+1}^{S+R}\sum_{i=1}^{k} \frac{\alpha_i \gamma_{i,r}\gamma_{q,r}}{T_r} \right] \\
&\geq \left(1 + \frac{\gamma_{q,q}^2}{\gamma_{q,q}^2 + (k-1)\sigma_{insight}^2} \right) \alpha_q
\end{aligned}
$$

□

Now, we state the upper bound on the changes in benign concepts. The proof is straightforward from the previous ones in harmful concept removal and helpful concept addition.

Corollary B.4.1. *Under the same combined noise model, the post-*ROBOSHOT *coefficient for benign concept q satisfies*

$$|\mathbb{E}[A_q] - \alpha_q| \leq \left| \sum_{t=1}^{S+R} \frac{\alpha_q \sigma_{insight}^2}{\gamma_{t,t}^2} \right|.$$

C Experiments details

C.1 Datasets

Table 7 provides details of the datasets used in our experiments. For Gender Bias dataset [11, 28], we test using the train set to get more data. For all other datasets, we use the default test set. For Amazon-WILDS [30] dataset, we convert the original 5-class rating classification into binary, by removing all samples with rating 3, and convert rating 1 and 2 into *bad* label, and 4 and 5 into *good* label.

C.2 Prompt templates

We provide details on prompts used to get the $v^{harmful}$ and $v^{helpful}$ on image datasets in Table 8. As mentioned in the main body, for NLP datasets we only used $v^{harmful}$. Additionally, we use the demographic mentions annotations to construct $v^{harmful}$ in CivilComments-WILDS [5, 19] and HateXplain [27]. We provide prompt details to get $v^{harmful}$ for Amazon-WILDS [19, 30] and Gender Bias [11, 28] datasets in Table 9. We also provide class prompts in Table 10.

C.3 Model and hyperparameters

All experiments are carried out using frozen weights and embeddings from huggingface (ALIGN, AltCLIP) and open-clip (CLIP ViT-B-32 and ViT-L-14, BiomedCLIP), and no training is involved. There is no randomness in the experiment results reported in the main body of the paper.

Dataset	Groups	N_{all}	N_{wg}	n_{class}	classes
Waterbirds	{ landbird in land, landbird in water, waterbird on land, waterbird on water }	5794	642	2	{landbird, waterbird }
CelebA	{ male & not blond, female & not blond, male & blond , female & blond }	19962	180	2	{not blond, blond}
PACS	{ art, cartoons, photos, sketches,}	9991	80	7	{dogs, elphant, giraffe, guitar, house, person }
VLCS	{ Caltech101, LabelMe, SUN09, VOC2007 }	10725	20	5	{bird, car, chair, dog, person}
CXR14	{ no-pneumothorax, pneumothorax }	2661	20	2	{no-pneumothorax, pneumothorax}
CivilComments-WILDS	{male, female, LGBTQ, christian, muslim, other religions, black, white }	133782	520	2	{non-toxic, toxic }
HateXplain	{hindu, islam, minority, refugee, indian, caucasian, hispanic, women, disability, homosexual, arab, christian, jewish, men, african, nonreligious, asian, indigenous, heterosexual, buddhism, bisexual, asexual}	1921	6	2	{normal, offensive}
Amazon-WILDS	{beauty, garden, books, luxury beauty, kindle store, movies and TV, pet supplies, industrial and scientific, office products, CDs and vinyl, electronics, cell phones, magazine, clothing, groceries, music, instruments, tools, sports, automotive, toys, arts crafts, kitchen, video games, pantry, software, gift cards }	90078	25	2	{good,bad}
Gender Bias	{male, female }	22750	3594	2	{female, male}

Table 7: Dataset details

Dataset	Model	$v^{harmful}$ prompt	$v^{helpful}$ prompt
All	ChatGPT	"List the biased/spurious differences between [classes]."	"List the true visual differences between [classes]."
	Flan-T5 & GPT2	{"[class] typically", "[class] usually"}	{"a characteristic of [class]: ", "[class] are", ""a [class] is", "Charactericstics of [class]" "Stereotype of [class]" "Typical characteristic of [class]"}
	LLaMA	"List the biased/spurious characteristics of [class]"	"List the visual characteristics of [class]"

Table 8: Image dataset prompt details

Dataset	Model	$v^{harmful}$ prompt
Amazon-WILDS	ChatGPT	"what are the biased differences between good and bad amazon reviews?"
Gender bias	ChatGPT	"what are the biased differences between comments about female and comments about male?"

Table 9: NLP dataset prompt details

Dataset	Class prompt
Waterbirds	["a landbird", "a waterbird"]
CelebA	["person with dark hair", "person with blond hair"]
PACS	"an image of [class]"
VLCS	"this object is [class]"
CXR14	["non-pneumothorax", "pneumothorax"]
CivilComments-WILDS	["non-toxic", "toxic"]
HateXplain	["normal", "offensive"]
Amazon-WILDS	["negative", "positive"]
Gender Bias	["female", "male"]

Table 10: Class prompt details