Foundation Models Can Robustify Themselves, For Free

Anonymous Author(s)
Affiliation
Address
email

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 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 prediction is among the most exciting paradigms in machine learning. Zero-shot models obviate the need for data collection and training loops by simply asking for a prediction on any set of classes. Unfortunately, such models inherit biases or undesirable correlations from their large-scale training data [DLS+18, TE11]. In a now-canonical example [KSM+21], 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 [YNPM23, GKG+22, ZR22] 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 [AZS+, BCZ+16, DP19, LGPV20] 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 language models **contain actionable insights** that can be exploited to improve themselves or other 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.

Submitted to R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023. Do not distribute.

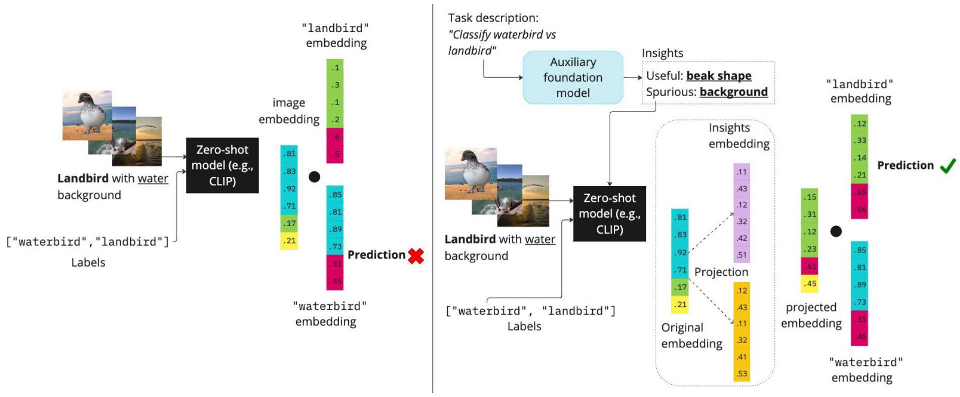


Figure 1: RoboShot pipeline (right) vs. vanilla zero-shot classification (left).

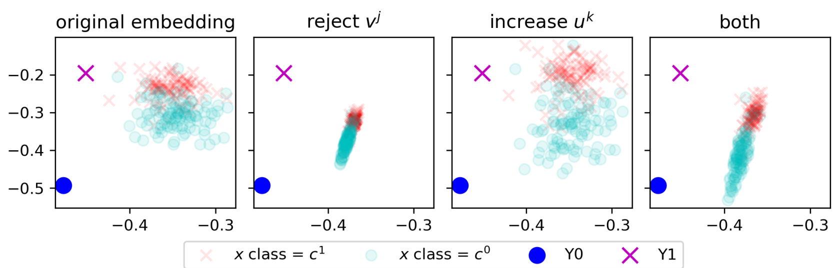


Figure 2: Visualization on CelebA (100 random samples from each class). L-R: (i) original embedding (ii) harmful concept removal (iii) helpful concept addition (iv) full RoboShot.

We propose RoboShot, a system that robustifies zero-shot models via language model-based insights *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 improved 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 are required as a function of the severity of harmful correlations. Empirically, RoboShot achieves 15.98% improvement across nine image and NLP datasets and has sufficient versatility to apply to a various 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. In summary, our contributions are:

1. A simple theoretical model describing zero-shot 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 RoboShot: Robustifying Zero-Shot Models

We are ready to provide our setup and describe the RoboShot algorithm. As mentioned before, we use embedding debiasing principles as building blocks. For our purpose, we utilize concepts obtained from language models and get their embeddings to build the beneficial and unwanted concept subspaces to work with. We call these embeddings the *insight representations*.

2.1 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 [DKA+22]. 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 γ_i are weights.

We work with the following theoretical model for zero-shot classification. For simplicity, we assume that there are two classes. It is straightforward to extend the analysis below to multi-class. 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. We assume that we have access to m answers $v^1, \ldots, v^m$ from a set of queries to the language model; we describe how these queries are used practically further on. These are given by $v^j = \sum_i \gamma_{i,j} z_i$ for $j \leq m$. We call these *insight representations*.

In the standard 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 the class that has the 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,

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

In other words, representations comprise of mixture of embeddings pertaining to harmful, helpful, and benign or neutral concepts—this holds for class and insight representations. In Appendix G.5, we empirically show that this assumption holds in real scenarios.

Example. We illustrate how harmful correlations produce errors on rare slices of data through a standard task setting, Waterbirds [KSM+21]. Here 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 from a data slice rarely seen in the training set, for instance, an image of landbird over water. Its embedding might be $x = 0.7 z_{\texttt{water}} + 0.3 z_{\texttt{landbird}}$. We may also have that $c^{\texttt{waterbird}} = 0.4 z_{\texttt{water}} + 0.6 z_{\texttt{waterbird}}$ and $c^{\texttt{landbird}} = 0.4 z_{\texttt{land}} + 0.6 z_{\texttt{landbird}}$. Then, $x^T c^{\texttt{waterbird}} = 0.1 > x^T c^{\texttt{landbird}} = -0.1$, which results in incorrect prediction: waterbird. Our goal is to *remove* harmful components (z_s's) and *boost* helpful ones (z_r's)—without labels or training. Our approach follows.

2.2 RoboShot: Robustifying zero-shot inference

We describe RoboShot in Algorithm 1. It uses representations of insights from LMs to shape input and class embeddings to remove harmful components and boost helpful ones. Figure 2 illustrates the intuition behind these procedures. Note how unhelpful directions are neutralized while perpendicular directions are boosted.

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 $\hat{y}$

Obtaining insight representations from LMs. The first question is how to obtain insight representations 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 F.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 $v = (g(s^1) - g(s^2)) / \left\| g(s^1) - g(s^2) \right\|$, 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 the harmful ones. We obtain insight representations using language models. For example, we ask "What are the true characteristics of waterbirds and landbirds?" and get e.g., {'`short beak`,' '`long beak`'}. The rest of the procedure is identical to that of harmful components. 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 harmful components (lines 2-5 of Algorithm 1) and boosts helpful ones (lines 6-9). To see the impact of doing so, we return to our earlier example. Suppose that we have a single harmful insight $v^{\text{harmful}} = 0.9z_{\texttt{water}} + 0.1z_{\texttt{landbird}}$ and a single helpful insight $v^{\text{helpful}} = 0.1z_{\texttt{water}} + 0.9z_{\texttt{landbird}}$. Note that even these insights can be imperfect — they have non-zero weights on other components.

From removing the harmful component (ignoring normalization for ease of calculation), we obtain $\hat{x} \leftarrow x - \langle x, v^{\text{harmful}} \rangle / \langle v^{\text{harmful}}, v^{\text{harmful}} \rangle v^{\text{harmful}} = -0.0244z_{\texttt{water}} + 0.2195z_{\texttt{landbird}}$. We already we have that $x^T c^{\texttt{waterbird}} = -0.1415 < x^T c^{\texttt{landbird}} = 0.1415$, thus the correct class is obtained. From a single insight we have neutralized a harmful correlation and corrected what had been an error. Adding in the helpful component further helps. Using vector addition equation in Algorithm 1 line 7, we obtain $-0.0006z_{\texttt{water}} + 0.4337z_{\texttt{landbird}}$. This further increases our margin. Note that it is not necessary 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 Section 3, 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 provide ablation experiments of each ROBOSHOT components (i.e., removing and boosting components) in Appendix B.2.

3 Theoretical Analysis

We provide an analysis that characterizes under what conditions ROBOSHOT can correct zero-shot errors. First, we consider the following error model on the weights of the representations. For all benign representations, we assume $\alpha_b, \beta_b, \gamma_b \sim \mathcal{N}(0, \sigma^2_{\text{benign}})$. The value of σ_{benign} is a function of the amount of data and the training procedure for the zero-shot model. Appendix G.5 empirically shows that in real scenarios, benign components can be canceled out.

Next, we assume that the insight embedding $v^s = \sum_{i=1}^k \gamma_{i,s} z_i$ (where $1 \leq s \leq S$) 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 $S+1 \leq r \leq S+R$) 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 targeted harmful concept coefficient after applying ROBOSHOT.

Theorem 3.1 *Under the noise model described above, the post-*ROBOSHOT *coefficient for harmful concept* s $(1 \leq s \leq S)$ *satisfies*

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

where k *is the number of concepts* ($k = S + R + B$).

The proof is included in Appendix E.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., more precise in specifying non-useful terms), ROBOSHOT

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

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**
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**
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
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**
CXR14	BiomedCLIP	55.3	28.9	26.4		-		56.2	**41.6**	**14.6**

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

4 Experimental Results

This section evaluates the following claims:

- **Improving multimodal models (Section 4.1)**: ROBOSHOT improves zero-shot classification robustness of various multimodal 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 4.2)**: ROBOSHOT improves zero-shot robustness using LM embeddings for text zero-shot classification, outperforming direct prompting to get predictions.
- **Extracting concepts from LM with varying capacities (Section 4.3)**: ROBOSHOT can extract insights from language models with varying capacities. Improvements persist with weaker LMs.

Metrics. 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.*
Baselines. We compare against the following sets of baselines:

1. **Multimodal baselines**: (i) vanilla zero-shot classification (**ZS**) and (ii) ZS with group information (**Group Prompt ZS**). We use a variety of models: CLIP (ViT-B-32 and ViT-L-14) [RKH+21], ALIGN [JYX+21], and AltCLIP [CLZ+22]. 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**: (i) zero-shot classification using language model embeddings, namely BERT [RG19] and Ada [NXP+22] (**ZS**), (ii) direct prompting to LMs, namely BART-MNLI [LLG+19, WNB18] and ChatGPT [ZSW+19] (**Direct prompting**). We also compare with calibration methods for zero-shot text classification [HWS+21], results in Appendix G.1.

4.1 Improving multimodal models

Setup. We experimented on 5 binary and multi-class datasets with spurious correlations and distribution shifts: **Waterbirds** [SKHL19], **CelebA** [LLWT15], **CXR14** [WPL+17], **PACS** [LYSH17], and **VLCS** [FXR13]. Appendix F.1 provides dataset details. For CXR14, we use BiomedCLIP [ZXU+23]– CLIP finetuned on biomedical data. We evaluate on two models: **CLIP** (ViT-B-32 and ViT-L-14). Additional results with CLIP variants (**ALIGN**, and **AltCLIP**) are given in Appendix B.1.

Table 2: ROBOSHOT text zero-shot classification. We use BERT embedding model Ada embedding model.

Dataset	Model	ZS			Direct prompting			**ROBOSHOT**		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
CivilComments	BERT	48.1	<u>33.3</u>	14.8	32.5	15.7	16.8	49.7	**42.3**	**7.4**
	Ada	56.2	<u>43.2</u>	13.0	85.6	19.2	66.4	56.6	**44.9**	**11.7**
HateXplain	BERT	60.4	0.0	60.4	61.2	<u>5.3</u>	55.9	57.3	**14.0**	**43.3**
	Ada	62.8	<u>14.3</u>	48.5	55.4	12.2	43.2	63.6	**21.1**	**42.5**
Amazon	BERT	81.1	<u>64.2</u>	16.8	74.9	36.0	38.9	81.0	**64.4**	**16.6**
	Ada	81.2	63.4	17.8	80.1	**73.5**	**6.6**	82.9	<u>63.8</u>	19.1
Gender Bias	BERT	84.8	83.7	1.1	86.1	78.4	7.6	85.1	**84.9**	**0.2**
	Ada	77.9	60.0	17.9	90.1	**86.6**	**3.5**	78.0	<u>60.1</u>	17.9

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

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

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.

4.2 Improving language models

Setup. We experimented on four text classification datasets: **CivilComments-WILDS** [BDS+19, KSM+21], **HateXplain** [MSY+21], **Amazon-WILDS** [NLM19, KSM+21] and **Gender Bias** classification dataset [DFW+20, MFB+17]. In text experiments, the distinctions between harmful and helpful insights are less clear than for images– so here we only use harmful vector rejection (line 3 in ROBOSHOT). Appendix F.1 and F.3 provides details on datasets and prompts.

Results. Table 2 shows that **ROBOSHOT also improves zero-shot text classification**, as shown by our consistent boost over the baselines across all datasets on BERT embedding model and BART-MNLI direct prompting. In the Gender Bias and Amazon experiments, RoboShot lifts weaker/older model performance to a level comparable to modern LLMs (ChatGPT).

4.3 Extracting concepts from LMs with varying capacities

Setup. We use **ChatGPT** [OWJ+22], **Flan-T5** [CHL+22], **GPT2** [RWC+19], and **LLaMA** [TLI+23], to obtain insights. **Results.** Table 3 shows that even though the LM strength/sizes correlate with the performance, ROBOSHOT with weaker LMs still outperforms zero-shot baselines. We hypothesize, based on Theorem 3.1 and D.1, that insights from smaller LMs are still precise in specifying the useful and non-useful terms and thus ROBOSHOT is able to use the insight embeddings.

5 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

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[AZS+] 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*.

[BCZ+16] 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.

[BDS+19] 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.

[BHB+22] 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.

[CCSE22] Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv:2210.12530*, 2022.

[CHL+22] 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.

[CJL+23] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[CLZ+22] 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.

[DFW+20] 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.

[DKA+22] 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*, 2022.

[DLS+18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018.

[DP19] 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.

[FCS+13] 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.

[FXR13] 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.

[GKG+22] 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.

[HWS+21] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*, 2021.

[JYX+21] 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.

[KCJ+21] 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.

[KIW22] 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.

[KNST23] 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.

[KSM+21] 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.

[LCLBC20] 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.

[LCT+22] 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.

[LGPV20] 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.

[LHC+21] 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.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[LLWT15] 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.

[LYSH17] 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.

[MFB+17] 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.

[MSY+21] 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.

[MV22] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[MVM+23] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. *arXiv preprint arXiv:2307.11661*, 2023.

[NLM19] 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.

[NXP+22] 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.

[OWJ+22] 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.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[RKH+21] 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.

[RWC+19] 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.

[SKHL19] 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.

[TE11] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011.

[TLI+23] 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.

[WLW21] 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.

[WNB18] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[WPL+17] 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.

[WZS22] 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.

[YNPM23] 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.

[ZR22] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *arXiv preprint arXiv:2207.07180*, 2022.

[ZSW+19] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[ZXU+23] 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.

Appendix

The appendix contains related work (Appendix C), additional theoretical (Appendix D), and experimental results (Appendix B.2 and G), details, proofs. The glossary contains a convenient reminder of our terminology (Appendix A) Appendix E provides the proofs of theorems that appeared in Section 3. In Appendix F, we give more details and analysis of the experiments and provide additional experiment results. Finally, Appendix G entails additional experiments combining RoboShot with other methods to highlight its versatility.

A Glossary

The glossary is given in Table 4.

Symbol	Definition
x	input vector
X	embedding matrix
X_{proj}	RoboShot projected embedding matrix
$y, \hat{y}$	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 4: Glossary of variables and symbols used in this paper.

B Extended Experimental Result

B.1 Full Main result

We provide full experimental results, with additional multi-modal models, **ALIGN** and **AltCLIP**

Table 5: Extended results. Best WG and Gap performance **bolded**, second best underlined.

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**

B.2 Ablation

Table 6: Ablation. Best WG and Gap performance **bolded**, second best underlined.

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

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. Due to space constraint, we only include CLIP-based models ablations. Results on all models can be found in Appendix G.

Results. The combination of both projections often achieves the best performance, as shown in Table 6. Figure 2 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.

C Related Work

We describe related work in zero-shot model robustness and 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 a heavily studied area [SKHL19, ABGLP19, KCJ+21, KIW22, LHC+21, LCT+22]. 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. [YNPM23] detects spurious attribute descriptions and fine-tunes using these descriptions. A specialized contrastive loss is used to fine-tune a pretrained architecture in [GKG+22] and to train an adapter on the frozen embeddings in [ZR22]. While promising, fine-tuning recreates traditional machine learning pipelines (e.g., labeling, training, etc.), which sacrifices some of the promise of zero-shot models. In contrast, our goal is to avoid any training and any use of labeled data. Concurrent work seeks to robustify CLIP zero-shot predictions against spurious features by debiasing the classifier (i.e., the labels embedding) against harmful concepts [CJL+23]—but does so via manual specification. In contrast, our work amplifies helpful concepts and automates the process of obtaining debiasing vectors.

Debiasing embeddings. A parallel line of work seeks to debias text embeddings [AZS+] [BCZ+16] [DP19] [LGPV20] and multimodal embeddings [WZS22, BHB+22, WLW21] by removing subspaces that contain unwanted concepts. We use a similar procedure as a building block. However, these methods either target specific fixed concepts (such as, for example, gender in fairness contexts) 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. Moreover, our goal is simply to add robustness at low or zero-cost; we do 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 [RKH+21, FCS+13, LCLBC20]. Most relevant are those that focus on robustness, such as [YNPM23] that uses text descriptions of spurious attributes in a fine-tuning loss to improve robustness. In contrast to these works, we focus on using textual concepts to improve zero-shot model robustness—without fine-tuning. The most similar to our work is [MV22, MVM+23], where GPT-3 generated class descriptors are first generated, then CLIP predictions scores are grounded by additive decomposition of scores from the prompts with the descriptors. Similarly, this method also does not require fine-tuning. However, this method focuses mainly on grounding through prompting with class descriptors, while ours focuses on removing harmful concepts and increasing helpful concepts in the embedding space.

Language models as priors. The basis of our work is the observation that language models contain information that can serve as a prior for other tasks. [KNST23] finds that LLMs can perform causal reasoning tasks, substantially outperforming existing methods. [CCSE22] 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. .

D Extended Theory Results

Theorem D.1 *With an additional assumption $\alpha_s \leq 0$ $(1 \leq s \leq S)$ under the described noise model, the post-*ROBOSHOT *coefficient for helpful concept r $(S+1 \leq r \leq S+R)$ satisfies*

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

Refer to Appendix E.3 for the proof. Theorem D.1 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.

E Theory details

E.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.

E.1.1 Effects on harmful coefficients

Now we prove the following theorem.

Theorem E.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 $(k = S+R+B)$.

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 $|\mathbb{E} A_s|$.

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

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

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

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

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

The upper bound for the first term can be obtained by

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

And, for the second term,

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

Combining two bounds, we get the proposed result.

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

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

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

where k is the number of concepts ($k = S + R + B$).

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

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

We improve the first term as follows.

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

E.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 $S+1 \leq q \leq k$. Essentially, 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 E.2 *Under the same noise model described above, the post-removal coefficient for helpful or benign concept q satisfies*

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

The proof technique is essentially identical to Theorem E.1.

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

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.

E.2 Helpful concept addition

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

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

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

. 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 $1 \le q \le S+R$, $q \ne t$ so that only $\gamma_{t,t}$ are constants.

E.2.1 Lower bound for the coefficient of helpful concept

Theorem E.3 *Under the described noise model, the post-addition coefficient for helpful concept r satisfies*

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

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

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

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,

$$
\begin{aligned}
\mathbb{E}A_r &= \mathbb{E}\alpha_r + \sum_{t=S+1}^{S+R}\sum_{i=1}^{k}\frac{\alpha_i\gamma_{i,t}\gamma_{r,t}}{T_t} \\
&= \alpha_r + \sum_{t=S+1}^{S+R}\sum_{i=1}^{k}\mathbb{E}\frac{\alpha_i\gamma_{i,t}\gamma_{r,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2} \\
&= \alpha_r + \mathbb{E}\frac{\alpha_r\gamma_{r,r}^2}{\sum_{l=1}^{k}\gamma_{l,r}^2} + \sum_{i=1,i\neq r}^{k}\mathbb{E}\frac{\alpha_i\gamma_{i,r}\gamma_{r,r}}{\sum_{l=1}^{k}\gamma_{l,r}^2} + \sum_{t=S+1,t\neq r}^{S+R}\sum_{i=1}^{k}\mathbb{E}\frac{\alpha_i\gamma_{i,t}\gamma_{r,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2} \\
&= \alpha_r + \mathbb{E}\frac{\alpha_r\gamma_{r,r}^2}{\sum_{l=1}^{k}\gamma_{l,r}^2} + \sum_{i=1,i\neq r}^{k}\gamma_{r,r}\mathbb{E}\frac{\alpha_i\gamma_{i,r}}{\sum_{l=1}^{k}\gamma_{l,r}^2} + \sum_{t=S+1,t\neq r}^{S+R}\sum_{i=1}^{k}\mathbb{E}\frac{\alpha_i\gamma_{i,t}\gamma_{r,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2} \\
&= \alpha_r + \mathbb{E}\frac{\alpha_r\gamma_{r,r}^2}{\sum_{l=1}^{k}\gamma_{l,r}^2} + \sum_{t=S+1,t\neq r}^{S+R}\sum_{i=1}^{k}\mathbb{E}\frac{\alpha_i\gamma_{i,t}\gamma_{r,t}}{\sum_{l=1}^{k}\gamma_{l,t}^2} \quad \text{by symmetry} \\
&= \alpha_r + \mathbb{E}\frac{\alpha_r\gamma_{r,r}^2}{\sum_{l=1}^{k}\gamma_{l,r}^2} \quad \text{by law of total expectation and symmetry} \\
&\geq \alpha_r + \alpha_r\gamma_{r,r}^2\mathbb{E}\frac{1}{\sum_{l=1}^{k}\gamma_{l,r}^2} \\
&\geq \alpha_r + \alpha_r\gamma_{r,r}^2\frac{1}{\mathbb{E}\sum_{l=1}^{k}\gamma_{l,r}^2} \quad \text{Jensen's inequality} \\
&= \alpha_r + \alpha_r\gamma_{r,r}^2\frac{1}{\gamma_{r,r}^2+(k-1)\sigma_{insight}^2}.
\end{aligned}
$$

Thus, we obtain the result.

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

E.2.2 Effects on harmful, benign coefficients

For notational convenience, let $I_{helpful}^c$ be the non-helpful concept index set such that $I_{helpful}^c = \{i \in \mathbb{N} | i \leq S \text{ or } S+R+1 \leq i \leq S+R+B\}$. 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 E.4 *Under the same noise model described above, the post-addition coefficient for helpful or benign concept q satisfies*

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

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

Theorem 3.1 Theorem D.1

E.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 (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 3.1, D.1 and provide proofs.

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

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

where k is the number of concepts $(k = S+R+B)$.

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

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

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

to keep the appearance of the result clear.

With an additional assumptions $\alpha_s \leq 0 \quad (1 \leq s \leq S)$ under the combined noise model, the post-coefficient for helpful concept $q(S+1 \leq q \leq S+R)$ satisfies

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

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

Here, $\mathbb{E}\sum_{s=1}^{S}\sum_{i=1,i\neq q}^{k}\dfrac{\alpha_i\gamma_{i,s}\gamma_{q,s}}{T_s} = 0$ by symmetry and law of total expectation, and $-\mathbb{E}\sum_{s=1}^{S}\dfrac{\alpha_s\gamma_{q,s}^2}{T_s} \geq 0$ since $\alpha_s \leq 0$ by assumption, which can be dropped for a lower bound.

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

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

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

F Experiments details

F.1 Datasets

Table 7 provides details of the datasets used in our experiments. For Gender Bias dataset [DFW+20, MFB+17], we test using the train set to get more data. For all other datasets, we use the default test set. For Amazon-WILDS [NLM19] 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.

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

F.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 [BDS+19, KSM+21] and HateXplain [MSY+21]. We provide prompt details to get $v^{harmful}$ for Amazon-WILDS [NLM19, KSM+21] and Gender Bias [DFW+20, MFB+17] datasets in Table 9. We also provide class prompts in Table 10.

F.3 Direct Prompting Prompt templates

Table 11 shows the prompts used for Direct Prompting ChatGPT baseline in Table 2. For BART-MNLI, we directly use the dataset labels as label input to the model.

Dataset	Prompt
CivilComments-WILDS	Is this a toxic comment? Answer with yes/no
HateXplain	Is this a toxic comment? Answer with yes/no
Amazon-WILDS	Is this a positive or negative product review? Answer with pos/neg
Gender Bias	Is this text about male/female? Answer with male/female

Table 11: Direct prompting prompts for ChatGPT

F.4 ROBOSHOT Experiment Details

All ROBOSHOT 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 ROBOSHOT experiment results reported in the main body of the paper.

F.5 LFA Experiment Details

Dataset	Batch size	Learning rate
Waterbirds	$\{1.5e^{-8}, 2.5e^{-8}, 5e^{-8}, 2.5e^{-7}\}$	$\{16, 32, 64\}$
CelebA	$\{7.5e^{-9}, 1e^{-8}, 2.5e^{-8}\}$	$\{16, 32, 64\}$
PACS	$\{2.5e^{-9}, 5e^{-9}, 7.5e^{-9}, 1.5e^{-8}\}$	$\{16, 32, 64\}$
VLCS	$\{2.5e^{-9}, 5e^{-9}, 7.5e^{-9}, 1.5e^{-8}\}$	$\{16, 32, 64\}$

Table 12: LFA hyperparameter choices

Table 12 shows the choices of hyperparameters we tune over for LFA experiments. We use SGD optimizer with fixed default momentum form PyTorch. All training are run for a fixed maximum epoch of 300, and we choose model based on validation performance.

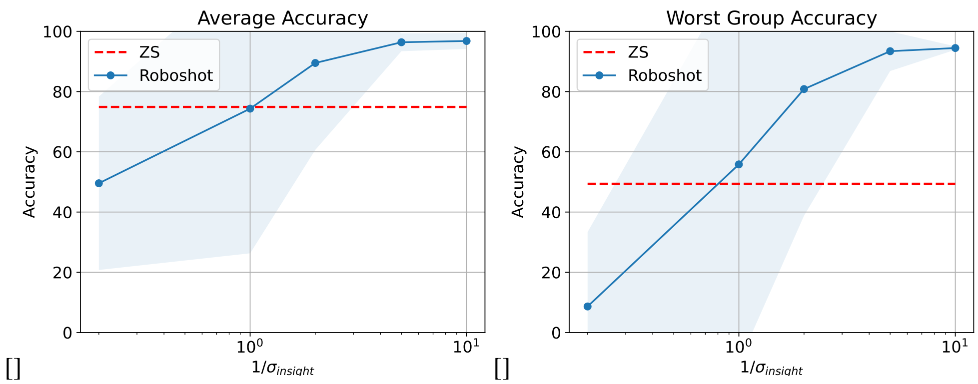


Figure 3: Synthetic experiment with varying σ_{noise}. As expected, the performance improves at a rate inversely proportional to σ_{noise}.

G Additional experiments

G.1 Combination with the calibration methods

Table 13: Additional baseline: text-classification calibration method [HWS+21]

Dataset	Model	Calibration			ROBOSHOT			Calibration + ROBOSHOT		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
CivilComments	BERT	51.0	37.3	13.7	49.7	**42.3**	**7.4**	53.4	36.9	16.5
	Ada	73.3	31.2	42.1	56.6	**44.9**	**11.7**	68.3	35.0	33.3
HateXplain	BERT	60.9	15.8	45.1	57.3	14.0	43.3	56.7	**22.8**	**33.9**
	Ada	61.9	31.6	30.3	63.6	21.1	42.5	59.6	**33.3**	**26.3**
Amazon	BERT	78.0	57.7	20.3	81.0	**64.4**	**16.6**	79.0	59.2	19.8
	Ada	71.2	50.5	20.7	82.9	63.8	**19.1**	83.2	**63.9**	19.3
Gender Bias	BERT	85.4	83.2	2.2	85.1	**84.9**	**0.2**	85.7	82.5	3.2
	Ada	84.2	77.8	6.4	78.0	60.1	17.9	84.2	**77.9**	**6.3**

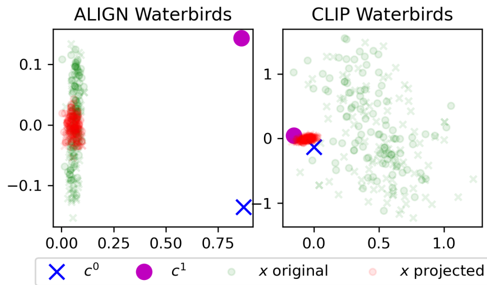
Table 13 shows that ROBOSHOT further benefits from the calibration methods. This further highlights the versatility of ROBOSHOT—we can combine it with such methods with no additional work. To showcase this, we show additional results from (1) applying the calibration method alone, (2) our method, (3) the combination.

This result show that the best performing method across the board is either ROBOSHOT or the combination. The underlying reason for this is that as the two methods are orthogonal, adding calibration can further improve the results.

G.2 Synthetic experiments

Setup. We validate our theoretical claims by performing a synthetic experiment where we vary the noise level in the insight vectors ($\sigma_{insight}$). Higher $\sigma_{insight}$ indicates more noise. We use the following basis vectors as concept vectors $z_{core} = (1, 0, 0), z_{spurious} = (0, 1, 0), z_{benign} = (0, 0, 1)$, and class embedding vectors $c_1 = z_{core} + z_{spurious} + z_{benign}$ and $c_0 = -z_{core} - z_{spurious} + z_{benign}$. Experiments are repeated 100 times.

- Synthetic data input distribution (s denotes spurious feature group)
  - $x|y = 1, s = 0 \sim \mathcal{N}([w_{core}, w_{spurious}, w_{benign}], \sigma_{input}I), n = 2500$
  - $x|y = 1, s = 1 \sim \mathcal{N}([w_{core}, -w_{spurious}, w_{benign}], \sigma_{input}I), n = 2500$
  - $x|y = 0, s = 0 \sim \mathcal{N}([-w_{core}, -w_{spurious}, w_{benign}], \sigma_{input}I), n = 2500$
  - $x|y = 0, s = 1 \sim \mathcal{N}([-w_{core}, w_{spurious}, w_{benign}], \sigma_{input}I), n = 2500$



[] []

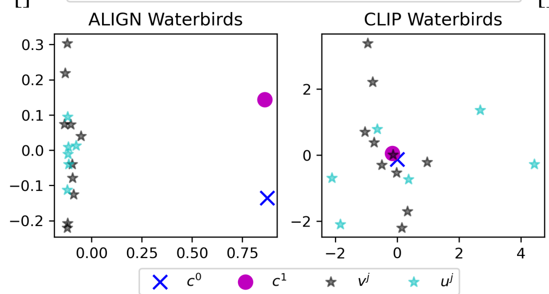


Figure 4: (a) Original (green) and projected (red) input embeddings x, and label embeddings c^0 and c^1. (b) label embeddings c^0 and c^1, harmful insight embeddings v^k (black star) and helpful insight embeddings u^j (blue star)

- Insight vectors
  - $v_{helpful} = \gamma_{helpful} z_{core} + \gamma_s z_{spurious} + \gamma_b z_{benign}$, where $\gamma_s \sim \mathcal{N}(0, \sigma_{inisght})$, $\gamma_b \sim \mathcal{N}(0, \sigma_{benign})$
  - $v_{harmful} = \gamma_c z_{core} + \gamma_{harmful} z_{spurious} + \gamma_b z_{benign}$, where $\gamma_c \sim \mathcal{N}(0, \sigma_{inisght})$, $\gamma_b \sim \mathcal{N}(0, \sigma_{benign})$

For the experiment reported in Figure 3, we used $w_{core} = 1, w_{spurious} = 1, w_{benign} = 0.5, \gamma_{helpful} = 1, \gamma_{harmful} = 1, \sigma_{input} = 0.5, \sigma_{benign} = 0.01$

Results. In Figure 3, we observe that up to 10 - 20% of noise level to signal (harmful, helpful coefficients = 1), our algorithm works well, recovering worst group accuracy and improving average group accuracy. This result supports our claims in Theorems 3.1 and D.1.

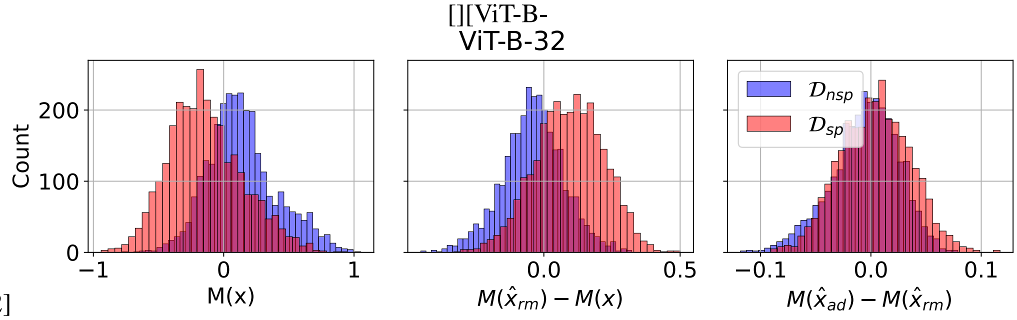
G.3 Embedding analysis

We provide insights into the case where our method does not improve the baseline (ALIGN model on Waterbirds) in Fig. 4. In Fig. **??**, 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. **??** (left) shows the 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 **??** (right)), we observe that the projected embeddings are significantly distant from the original ones. Unsurprisingly, Figure **??** (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.

G.4 Analysis on the robustness to spurious correlations.

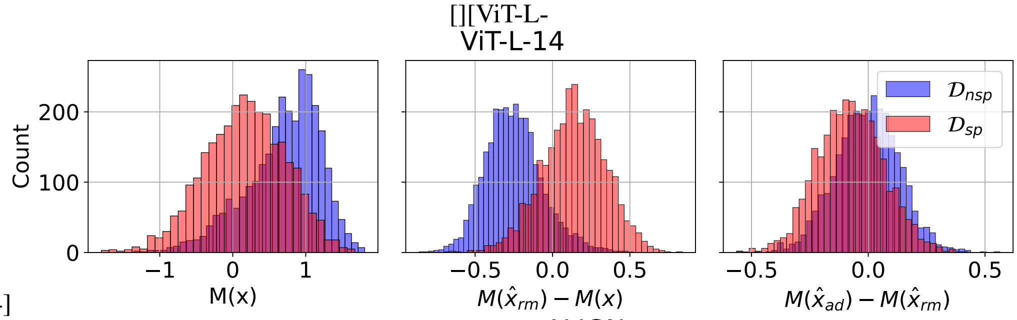
We provide in-depth result analysis to explain the performance changes in the average accuracy (AVG) and worst group accuracy (WG), especially with respect to spurious correlations. Concretely, consider the distribution of the margin $M : \mathcal{X} \to \mathbb{R}$ given by $M(x) := \langle c^+, x\rangle - \langle c^-, x\rangle$, where c^+, c^- are the correct/incorrect class embeddings. Accuracy can be expressed as $\mathbb{E}\mathbb{I}(M(x))$. The margin distributions and the margin changes by roboshot are illustrated in Figure 5 (Waterbirds), 6. We denotes data with spurious features as $\mathcal{D}_{sp}$ (i.e. waterbirds with land background, landbirds with water background), and data with non-spurious features as $\mathcal{D}_{nsp}$ (i.e. waterbirds with water background,

[][ViT-B-

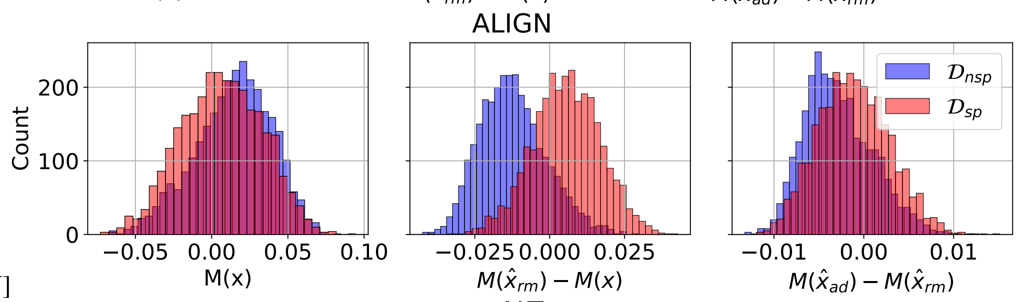


32]

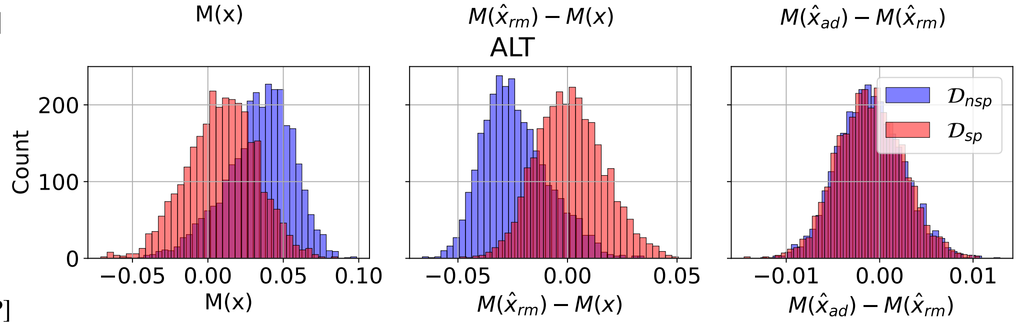
[][ViT-L-



14]



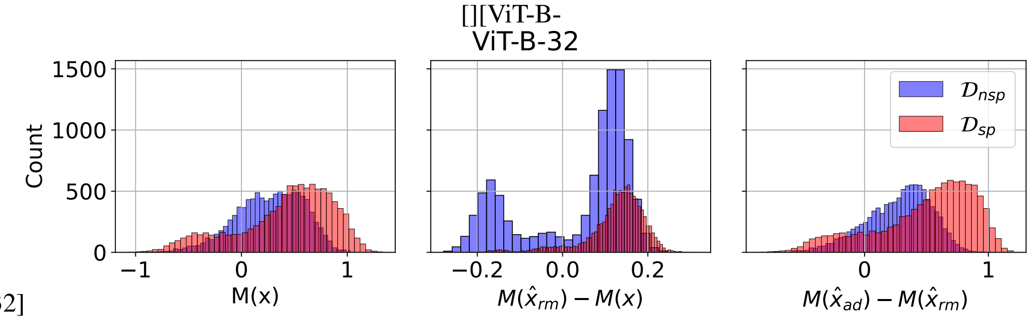
[][ALIGN]



[][AltCLIP]

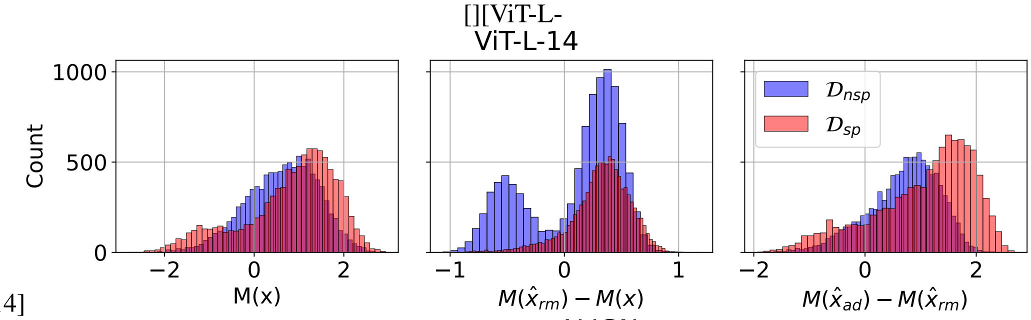
Figure 5: Margin analysis in Waterbirds dataset. Typically, inputs with spurious features $\mathcal{D}_{sp}$ tend to be closer to the decision boundary, inducing more errors. As expected, we can observe that harmful insight removal procedure increases the margin of $\mathcal{D}_{sp}$, but decreases the margin of inputs with non-spurious features $\mathcal{D}_{nsp}$. This can explain the potential tradeoff between the accuracy of $\mathcal{D}_{sp}$ and $\mathcal{D}_{nsp}$. If the gain in $\mathcal{D}_{sp}$ outweights the loss in $\mathcal{D}_{nsp}$, the average accuracy increases as in most cases. However, if the gain in $\mathcal{D}_{sp}$ is less the loss in $\mathcal{D}_{nsp}$, the average accuracy decreases as in ALIGN. In either case, the model performance in $\mathcal{D}_{sp}$ is improved by this procedure. In addition step, we expect that margin improves in both of D_{sp}, D_{nsp} on average as in ViT-B-32. However, in most cases, the margin changes are not that crucial, implying extracting helpful insights is not easy in Waterbirds dataset.

[][ViT-B-

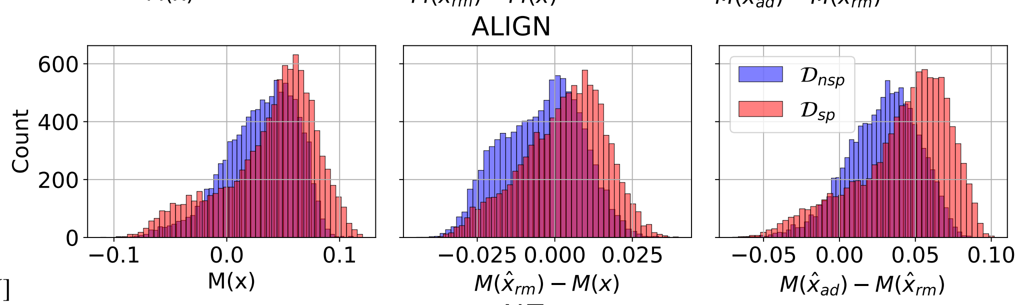


32]

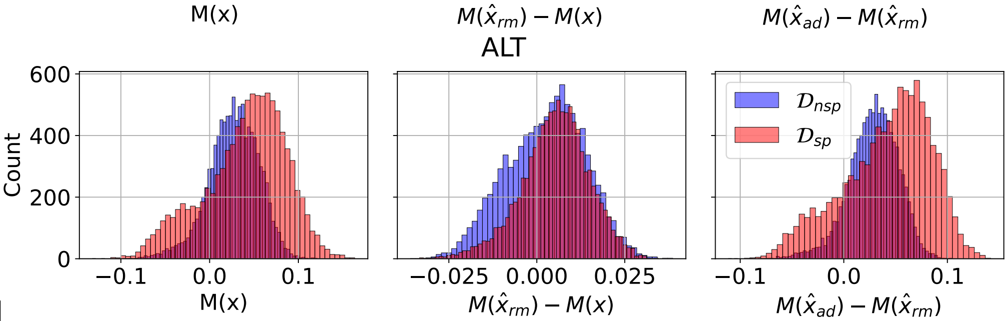
[][ViT-L-



14]



[][ALIGN]



[][AltCLIP]

Figure 6: Margin analysis in CelebA dataset. Again, inputs with spurious features "blond" tend to induce errors ("men"-"blond", "girl"-"non-blond"). As expected, we can observe that harmful insight removal procedure increases the margin of $\mathcal{D}_{sp}$, but decreases the margin of inputs with non-spurious features $\mathcal{D}_{nsp}$, which may lead to the potential tradeoff. However, in CelebA dataset, the helpful insight addition step turns out to be helpful, increasing the margins of both distributions much. It can be interpreted as helpful insights can be captured easily in images.

landbirds with land background). In the first column, $M(x)$ denotes the margin distribution of zeroshot prediction. In the second column, $M(\hat{x}_{rm}) - M(x)$ represents the margin changes by the roboshot harmful concept removal procedure. In the third column, $M(\hat{x}_{ad}) - M(\hat{x}_{rm})$ represents the margin changes by the roboshot helpful concept addition. Typically, inputs with spurious features $\mathcal{D}_{sp}$ tend to be closer to the decision boundary, inducing more errors. As expected, we can observe that harmful insight removal procedure increases the margin of $\mathcal{D}_{sp}$, but decreases the margin of inputs with non-spurious features $\mathcal{D}_{nsp}$. This can explain the potential tradeoff between the accuracy of $\mathcal{D}_{sp}$ and $\mathcal{D}_{nsp}$. If the gain in $\mathcal{D}_{sp}$ outweights the loss in $\mathcal{D}_{nsp}$, the average accuracy increases as in most cases. However, if the gain in $\mathcal{D}_{sp}$ is less the loss in $\mathcal{D}_{nsp}$, the average accuracy decreases as in ALIGN. In either case, the model performance in $\mathcal{D}_{sp}$ is improved by this procedure. In addition step, we expect that margins improve in both of D_{sp}, D_{nsp} on average. Helpful insight addition procedure turns out be quite effective in CelebA dataset, where visual features can be described more easily by language models.

G.5 Isolating concepts by averaging relevant concepts

Table 14: Left: Cosine similarity between concept images and original embedding vs. averaged embedding. Right: ROBOSHOT on Waterbirds with original vs. averaged embedding

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**

We conduct experiments to test the viability of our concept modeling. Specifically, we want 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) and whether it is reasonable to assume benign components as noise.

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 14 (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. And thus, *benign components in embeddings can be canceled out*.

Does ROBOSHOT gain improvement with isolated concept? Table 14 (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*.

G.6 Roboshot without decomposition

To see the effectiveness of QR decomposition of insight vectors, we conduct additional ablation experiment of decomposition method. In Table 15, w/o QR (v^j only), w/o QR (u^k only), and w/o QR (both) represents roboshot rejection only, addition only, both without QR decomposition step. Contrary to our expectation, in binary classification (Waterbirds, CelebA), Roboshot method works well without QR decomposition. This can be interpreted as insights from LLM provide almost orthogonal vectors. However, in multiclass classification, where rejection, addition vectors are generated by combinatorially paring insights for each class, Roboshot method get worse. Especially, addition step collapse. While rejection step wears off the subspace that the insight vectors span and there couldn't be more difference, addition steps can push multiple times to the similar directions. From this ablation experiment, the benefits of obtaining subspace via decomposition can be explained by two ways. First, in removal step, it provides a clean way to remove the subspace that spurious features span. Secondly, int addition step, it prevents overemphasis on some helpful insight directions.

Table 15: Ablation of QR decomposition

Dataset	Model	Roboshot w/ QR			w/o QR (v^j only)			w/o QR (u^k only)			w/o QR (both)		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
Waterbirds	CLIP (ViT-B-32)	83.0	54.4	28.6	79.5	58.3	21.2	83.0	31.2	51.8	79.6	**62.5**	**17.1**
	CLIP (ViT-L-14)	79.9	45.2	34.7	79.3	36.3	43.0	88.8	31.6	57.2	75.0	**45.8**	**29.2**
	ALIGN	50.9	41.0	9.9	53.3	36.6	16.7	62.0	**50.9**	11.1	38.2	36.5	**1.7**
	AltCLIP	78.5	54.8	23.7	70.8	**56.1**	14.7	89.0	35.0	54.0	64.3	52.8	**11.5**
CelebA	CLIP (ViT-B-32)	84.8	80.5	4.3	85.3	81.6	3.7	80.5	73.2	7.3	86.5	**83.5**	**3.0**
	CLIP (ViT-L-14)	85.5	**82.6**	**2.9**	86.1	81.7	4.4	79.7	72.5	7.2	85.8	80.0	5.8
	ALIGN	86.3	83.4	2.9	84.4	78.9	5.5	83.9	81.5	2.4	86.8	**84.5**	**2.3**
	AltCLIP	86.0	77.2	8.8	86.5	75.6	9.9	80.4	75.6	**4.8**	86.0	**77.8**	8.2
PACS	CLIP (ViT-B-32)	97.0	**86.3**	**10.7**	97.0	82.9	14.1	85.5	37.8	47.7	83.8	33.0	50.8
	CLIP (ViT-L-14)	98.1	**83.9**	**14.2**	98.0	79.8	18.2	84.9	13.4	71.5	85.8	11.8	74.0
	ALIGN	95.0	73.8	21.2	95.7	**75.9**	**19.8**	56.9	0.2	56.7	58.0	0.2	57.8
	AltCLIP	98.7	**89.5**	**9.2**	98.4	83.1	15.3	67.8	4.0	63.8	65.0	2.8	62.2
VLCS	CLIP (ViT-B-32)	75.6	**33.0**	43.5	75.5	20.5	55.0	21.4	0.0	**21.4**	30.7	0.0	30.7
	CLIP (ViT-L-14)	71.1	**12.6**	58.5	71.1	6.9	64.2	22.3	0.0	22.3	22.1	1.3	**20.8**
	ALIGN	77.6	**39.8**	37.8	78.1	33.0	45.1	36.2	0.0	36.2	32.7	0.1	**32.6**
	AltCLIP	78.9	25.0	53.9	77.5	**25.1**	52.4	31.4	0.0	31.4	30.6	2.0	**28.6**
CXR14	BiomedCLIP	56.2	**41.6**	**14.6**	55.9	36.6	19.3	55.2	23.9	31.3	56.1	37.2	18.9