MASKING SPURIOUS CONCEPTS VIA SPARSE AUTOENCODERS FOR EFFICIENT BIAS MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks trained via Empirical Risk Minimisation (ERM) are prone to relying on *spurious correlations*: confounding features such as backgrounds that are statistically associated with a class label during training but are not causally relevant. When these correlations do not hold at test time, performance degrades on subgroups without spurious features. Existing mitigation methods either (a) require model retraining or large group label annotations, or (b) modify only the final classifier head, leaving spurious information encoded in the latent space and thus exhibiting a trade-off between worst-group and average accuracy. To circumvent these limitations, we propose SPOCK, a bias mitigation method that uses a Sparse Autoencoder to decompose latent representations into sparse concepts and learns a mask that suppresses spurious concepts while preserving those useful for the downstream task. By filtering concepts before reconstructing the representations, SPOCK reduces the model's reliance on spurious correlations while keeping the existing classifier fixed. In this way, SPOCK is compute- and label-efficient, requiring only a handful of group-annotated validation examples and avoiding end-to-end retraining. Our results across four synthetic and real-world datasets show that SPOCK substantially improves worst-group accuracy over ERM while maintaining high average accuracy, offering a more favourable trade-off than competing bias mitigation approaches.

1 INTRODUCTION

While Deep Neural Networks (DNNs) have achieved remarkable, often human-level performance across tasks (He et al., 2015; Wichmann & Geirhos, 2023; Jimenez et al., 2024; Liu et al., 2024), they are notorious for underperforming under distribution shifts: even small contextual changes, such as different backgrounds, can drastically alter their predictions (Taori et al., 2020; Ong Ly et al., 2024; Hermann et al., 2024; Yuan et al., 2024). This is because DNNs are typically trained via Empirical Risk Minimisation (ERM), which minimises the average loss over the training set and can encourage reliance on spurious correlations between confounding input features and class labels (Daube et al., 2021; Geirhos et al., 2020; Singla & Feizi, 2022; Zech et al., 2018). When these spurious correlations no longer hold, their performance degrades, leading to poor generalisation on *subgroups* without the spurious cues (Figure 1a) (Ming et al., 2022). For instance, such failures have been observed in healthcare tasks, where pneumonia models have been shown not to generalise across hospitals because they rely on hospital-specific cues (Zech et al., 2018), and skin-lesion models have been shown to perform worse on darker skin tones (Daneshjou et al., 2022).

A prominent approach to mitigate spurious correlations in DNNs is group-supervised learning (Sagawa et al., 2020; Levy et al., 2020; Zhou et al., 2021), which explicitly optimises for worst-group performance (performance on the subgroup with the lowest accuracy) rather than overall performance across the entire training dataset. To achieve this, these methods rely on per-example group annotations, typically constructed from human-specified spurious attributes, to train a more robust model from scratch. These requirements, however, are (1) costly to scale, as retraining may be infeasible and annotating the entire training set can be constraining, (2) prone to annotation errors, and (3) sometimes unrealistic when the spurious attribute is unknown (Sohoni et al., 2020; Creager et al., 2021). Moreover, collecting large sets of group labels, often correlating with protected attributes such as ethnicity, sex, or medical history, can be difficult and undesirable due to privacy and regulatory concerns (Coston et al., 2019; Lahoti et al., 2020).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

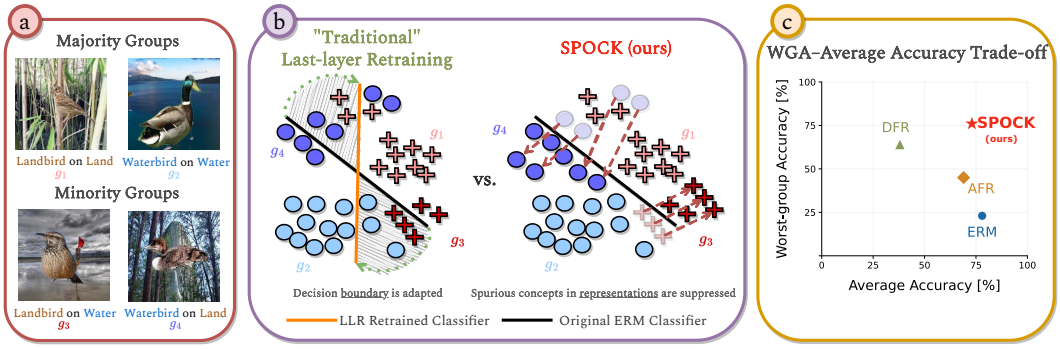


Figure 1: (a) In the Waterbirds dataset, most landbirds appear on land and most waterbirds on water, encouraging reliance on backgrounds rather than bird-specific features. The classifier may thus achieve high average accuracy while attaining low worst-group accuracy (WGA). (b) Last-layer retraining methods (e.g., DFR (Kirichenko et al., 2023) and AFR (Qiu et al., 2023)) improve performance on the minority groups by shifting the decision boundary, often trading off worst-group and average accuracy. In contrast, SPOCK suppresses spurious concepts, moving sample representations, while keeping the decision boundary fixed. (c) This enables SPOCK to improve worst-group performance while maintaining a favourable trade-off with average accuracy.

To address this, recent work (Qiu et al., 2023; Kirichenko et al., 2023; LaBonte et al., 2023; Kenfack et al., 2025) has explored *Last-Layer Retraining* (LLR) to reduce the labelling burden of bias mitigation and avoid end-to-end model retraining. These methods rely on the observation that, despite learning spurious correlations, DNNs still encode useful core concepts in their learned representations, even if they are often outweighed by spurious concepts in the last layer (Kirichenko et al., 2023). LLR methods address this by retraining the last layer using explicit or implicit balancing of a held-out set, without modifying the learned representations themselves (Hill et al., 2025). As a result, spurious concepts remain encoded in the representation, and robustness is achieved merely due to a shift in the classifier's decision boundary (Figure 1b, left). However, because the representation still reflects spurious correlations, improving minority-group performance requires shifting the decision boundary significantly away from the original boundary, introducing a trade-off with overall performance (Figure 1c). This motivates the need for a mechanism that operates directly in the latent space to suppress spurious concepts while preserving concepts useful for overall performance.

In this paper, we propose **S**parse s**P**uri**O**us **C**oncept mas**K**ing (**SPOCK**), a representation-level method designed to suppress spurious concepts encoded within a model's latent space, thereby preventing the model from relying on spurious correlations during prediction. As illustrated in Figure 2, SPOCK achieves this by (1) training a sparse autoencoder (SAE) (Huben et al., 2024) on frozen representations extracted from a pre-trained model, decomposing dense, entangled latents into distinct concepts, and then (2) learning a sparse mask using a *small* group-balanced held-out set, suppressing concepts associated with spurious behaviour while preserving task-relevant ones. The masked concepts are then decoded back into the latent space and passed to the original classification head. Our contributions are:

- We introduce SPOCK, an efficient bias mitigation method that reduces a model's reliance on spurious correlations by directly modifying its latent representations, without end-to-end retraining.
- We demonstrate that SPOCK substantially improves the worst-group performance of pre-trained models across vision and text benchmarks compared with standard Empirical Risk Minimisation, while preserving strong average accuracy relative to competing baselines.
- We further show that SPOCK remains effective with only a small group-annotated held-out set, thereby reducing the amount of group supervision required.

2 Problem Formulation

We consider the problem of learning a classification model given a training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(i)}, y^{(i)}) \mid \mathbf{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathcal{Y}\}_{i=1}^{N_{\text{train}}}$, where $\mathbf{x}^{(i)}$ and $y^{(i)} \in \mathcal{Y} = \{1, \cdots, L\}$ are the features

(e.g., pixels) and class label (e.g., "bird species") of the i-th sample, respectively. Specifically, we are interested in settings where there exists a set of *spurious* attributes $\mathcal{A} = \{a_1, \ldots, a_A\}$ (e.g., "beach", "forest") that are strongly correlated with some class labels in $\mathcal{Y}$ (e.g., "seagull", "crow"). Therefore, following Ye et al. (2024), we think of each sample in this setting as belonging to a *group* $g = (y, a) \in \mathcal{G} = \mathcal{Y} \times \mathcal{A}$, where all samples share the same task label and spurious attribute (e.g., "seagulls on a beach"). Here, we assume each group has its own distribution $\mathcal{P}_g(\mathbf{x}, y)$, and the overall data distribution is a weighted sum of group distributions: $\mathcal{P}(\mathbf{x}, y) = \sum_{g \in \mathcal{G}} \alpha_g \mathcal{P}_g(\mathbf{x}, y)$, where the weight α_g represents the proportion of group g in $\mathcal{D}_{\text{train}}$. In datasets with spurious correlations, these groups are often highly imbalanced (e.g., "seagulls" on a "beach" are more common), causing the attribute to strongly correlate with the label and become a predictive signal (Geirhos et al., 2020).

We want to learn the parameters $(\phi, \psi) \in \Theta$ of a DNN $f_{\phi,\psi} : \mathbb{R}^n \to \mathcal{Y}$, where $f_{\phi,\psi} = g_\phi \circ h_\psi$, that maps samples $\mathbf{x} \in \mathbb{R}^n$ to labels $y \in \mathcal{Y}$. Here, we think of $h_\psi : \mathbb{R}^n \to \mathbb{R}^d$ as a feature extractor that maps samples $\mathbf{x}$ to latent representations $\mathbf{z} \in \mathbb{R}^d$, and $g_\phi : \mathbb{R}^d \to \mathcal{Y}$ as a classification head mapping latent representations $\mathbf{z}$ to labels y. A standard approach is to minimise the expected loss over the underlying data distribution $\mathcal{P}$. Since $\mathcal{P}$ is unknown, this objective is typically approximated via *Empirical Risk Minimisation* (ERM) (Vapnik, 1991) over $\mathcal{D}_{\text{train}}$:

$$(\phi^*, \psi^*) = \underset{(\phi,\psi)\in\Theta}{\operatorname{argmin}} \mathbb{E}_{\mathbf{x},y\sim\mathcal{P}(\mathbf{x},y)}[\ell_{\text{CE}}(f_{\phi,\psi}(\mathbf{x}), y)] \approx \underset{(\phi,\psi)\in\Theta}{\operatorname{argmin}} \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \ell_{\text{CE}}(f_{\phi,\psi}(\mathbf{x}^{(i)}), y^{(i)}), \quad (1)$$

where ℓ_{CE} is the standard cross-entropy loss.

If the training set $\mathcal{D}_{\text{train}}$ contains spurious correlations, a model's performance may significantly drop on minority groups where these correlations do not hold. However, this failure mode goes unnoticed when we use average accuracy to evaluate the model. Therefore, we evaluate not only a model's average accuracy but also its *worst-group accuracy* (WGA):

$$\text{WGA}(f_{\phi,\psi}, \mathcal{P}) = \min_{g\in\mathcal{G}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_g(\mathbf{x},y)}[\mathbb{1}[f_{\phi,\psi}(\mathbf{x}) = y]], \quad (2)$$

where $\mathbb{1}$ is the indicator function and $\mathcal{P}_g(\mathbf{x}, y)$ represents the distribution of samples in group $g \in \mathcal{G}$.

3 Background and Related Work

Group-Supervised Methods. Group-supervised bias mitigation methods assume access to a fully group-annotated training set to improve the model's worst-group performance (Levy et al., 2020; Sagawa et al., 2020; Zhou et al., 2021). This encourages the model to perform well for underrepresented groups, rather than simply achieving good average performance. However, these gains often come at the expense of overall performance (Liu et al., 2021; Sagawa et al., 2020; Ye et al., 2026), creating a trade-off between WGA and average accuracy. Moreover, annotating an entire training set with group labels is costly, limiting these methods' practicality.

Group-Unsupervised Methods. Group-unsupervised bias mitigation methods (Liu et al., 2021; Nam et al., 2020; Sohoni et al., 2020; Han & Zou, 2024; Mulchandani & Kim, 2025; Arefin et al., 2024; Ye et al., 2025) do not require group annotations during training, typically leveraging a second model to identify spurious correlations (e.g., an ERM-trained model). For instance, some approaches infer pseudo-group training labels by exploiting the structure of ERM-learnt representations (e.g., via clustering) (Sohoni et al., 2020; Creager et al., 2021; Han & Zou, 2024) or by discovering concepts using a general-purpose external model (e.g., caption-to-image generators) (Arefin et al., 2024; Wu et al., 2023). Others instead mitigate biases through subnetwork pruning (Le et al., 2025) or by reweighting the training set (Liu et al., 2021; Qiu et al., 2023; Nam et al., 2020; Zhang et al., 2022; Espinosa Zarlenga et al., 2024). For instance, *Just Train Twice* (JTT) (Liu et al., 2021) upweights examples misclassified by an early-stopped ERM model, while *Learning from Failure* (LfF) (Nam et al., 2020) uses an intentionally biased model to identify "difficult" examples mid-training. All of these unsupervised methods, however, require retraining or modifying a model, making them computationally expensive. Moreover, they are highly sensitive to hyperparameter selection and, in practice, require validation group labels during model selection (Espinosa Zarlenga et al., 2024). This motivates the need for efficient methods that make better use of limited group annotations.

Last-Layer Retraining Methods. Recent work efficiently improves a pre-trained model's WGA via *Last-Layer Retraining* (LLR), where only the model's last layer is retrained (i.e., its

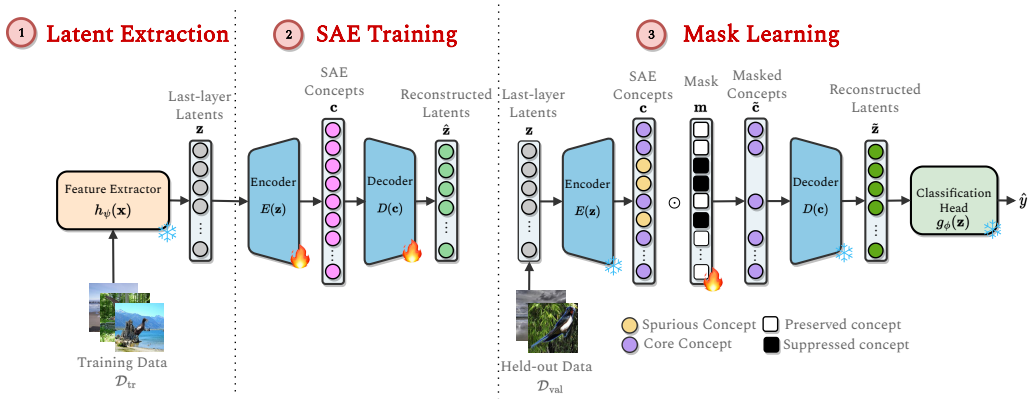


Figure 2: **Overview of SPOCK.** (1) The last-layer latent representations are extracted from a pre-trained model for all training samples. (2) These dense, entangled latents are then used to train an SAE that learns a sparse concept representation. (3) A mask is learned over the SAE concepts using a group-balanced held-out set to preserve useful information while suppressing spurious cues. During *inference*, the learned mask is applied to the sparse concepts to filter out spurious concepts, which are then decoded back into the latent space and passed through the classification head.

classification head). For example, *Deep Feature Reweighting* (DFR) (Kirichenko et al., 2023) builds on the observation that ERM-trained models still capture generalisable concepts in their learned representations, and reweights them by retraining the last layer using a small, group-balanced dataset. Follow-up methods (LaBonte et al., 2023; Kenfack et al., 2025), such as *Automated Feature Reweighting* (AFR) (Qiu et al., 2023), relax the need for a small group-balanced dataset. However, these methods still rely on implicit group balancing (Hill et al., 2025) and leave the underlying representations unchanged, instead shifting the classifier's decision boundary to better fit minority groups, thus introducing a trade-off between WGA and average accuracy (Kirichenko et al., 2023). This motivates our paper's goal of exploring whether directly suppressing spurious concepts in learned representations while preserving generalisable concepts can reduce this trade-off.

Sparse Autoencoders. Sparse Autoencoders (SAEs) (Huben et al., 2024) are an unsupervised approach for disentangling a model's latent representations in superposition (Park et al., 2024) into sparse, distinct concepts[1]. This helps address *polysemanticity* (Elhage et al., 2022; Bau et al., 2017; Fong & Vedaldi, 2018), whereby individual neurons encode multiple unrelated concepts, and allows for direct manipulation of individual concepts. Recent work has leveraged this property for post-hoc bias mitigation. For instance, Sparse Embedding Modulation (Guimard et al., 2026) and SAE Debias (Wu et al., 2025) use SAEs to suppress bias-sensitive concepts in vision-language models to improve the robustness of zero-shot classification and text-to-image generation. Inspired by this idea, we explore using SAEs to identify and suppress spurious concepts within a pre-trained model's representations to improve the trade-off between WGA and average accuracy observed in LLR methods.

4 Sparse Spurious Concept Masking (SPOCK)

We introduce Sparse sPuriOus Concept masKing (SPOCK), which builds on the same premise as LLR methods: DNNs capture both spurious and task-relevant concepts in their latent representations. Instead of shifting the classifier's decision boundary, as in LLR methods, SPOCK uses an SAE to efficiently modify the latent representations while keeping the original classification head fixed (Figure 1b). This leads to a classifier that operates predominantly on generalisable core concepts, reducing the trade-off between WGA and average performance. We summarise our method in Figure 2 and describe it in further detail below.

[1]The mechanistic interpretability literature typically refers to these units as "features". However, we use the term "concepts" to avoid confusion with a latent representation's dimensions (also typically called "features").

4.1 TRAINING

Given a potentially biased pre-trained DNN $f_{\phi,\psi} = g_\phi \circ h_\psi$ and its training set $\mathcal{D}_{\text{train}}$, SPOCK assumes access to a *small* group-annotated held-out dataset $\mathcal{D}_{\text{val}}$ to mitigate the effect of spurious correlations in the model. This is done via a three-step process:

(Step 1) Latent Representation Extraction. Using the DNN's feature extractor h_ψ, we first collect the latent representations $\mathbf{z}^{(i)} \in \mathbb{R}^d$ of all training samples $\mathbf{x}^{(i)} \in \mathcal{D}_{\text{train}}$ (Figure 2, Step 1):

$$\mathcal{Z}_{\text{train}} = \left\{ \mathbf{z}^{(i)} \mid \mathbf{z}^{(i)} = h_\psi(\mathbf{x}^{(i)}),\ \forall \mathbf{x}^{(i)} \in \mathcal{D}_{\text{train}} \right\}.$$

(Step 2) Sparse Autoencoder Training. Next, we want to distinguish spurious from task-relevant, *core* concepts in the latents $\mathbf{z}$. However, individual latent dimensions are *polysemantic* and tend to encode multiple concepts (see Section 3). We therefore train an SAE $(D \circ E) : \mathbb{R}^d \to \mathbb{R}^d$ to disentangle these representations into distinct concepts (Figure 2, Step 2). The encoder $E : \mathbb{R}^d \to \mathbb{R}^M$ projects each dense latent representation $\mathbf{z}$ to a high-dimensional sparse concept vector $\mathbf{c}$:

$$\mathbf{c} = E(\mathbf{z}), \quad \mathbf{c} \in \mathbb{R}^M, \quad M \gg d.$$

These are subsequently mapped back to the original latent space by the decoder $D : \mathbb{R}^M \to \mathbb{R}^d$:

$$\hat{\mathbf{z}} = D(\mathbf{c}), \quad \hat{\mathbf{z}} \in \mathbb{R}^d.$$

We train the SAE by minimising the combined objective following Huben et al. (2024):

$$\mathcal{L}_{\text{SAE}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(\|\mathbf{z}^{(i)} - \hat{\mathbf{z}}^{(i)}\|_2^2 + \lambda_s \|\mathbf{c}^{(i)}\|_1 \right). \tag{3}$$

Here, the first term is the reconstruction loss, which ensures that information from the original latent representation is preserved, while the second term is an ℓ_1 penalty that encourages sparse concept activations. To avoid overfitting, we split $\mathcal{Z}_{\text{train}}$ into 80:20 training and validation sets, $\mathcal{Z}_{\text{tr}}$ and $\mathcal{Z}_{\text{val}}$, respectively, monitoring reconstruction quality on $\mathcal{Z}_{\text{val}}$ to select λ_s. We further analyse the importance of the SAE to SPOCK's performance in Appendix A.

(Step 3) Mask Learning. Without modifying the SAE concepts, the reconstructed representation $\hat{\mathbf{z}}$ retains both core and spurious concepts in $\mathbf{z}$, implying $\hat{f}_{\phi,\psi} := g_\phi \circ D \circ E \circ h_\psi$ will likely achieve similar average accuracy and WGA as $f_{\phi,\psi}$. To debias this model, one would ideally not include concepts that capture spurious information when reconstructing the latent features. SPOCK does this by learning a sparse continuous mask $\mathbf{m} \in [0,1]^M$ over SAE concepts (Figure 2, Step 3):

$$\tilde{\mathbf{c}} = \mathbf{m} \odot \mathbf{c}, \quad \tilde{\mathbf{c}} \in \mathbb{R}^M$$

where $\odot$ denotes the Hadamard product. The decoder can then be used to map these masked activations back to the original latent space:

$$\tilde{\mathbf{z}} = D(\tilde{\mathbf{c}}), \quad \tilde{\mathbf{z}} \in \mathbb{R}^d.$$

Following previous LLR approaches (Kirichenko et al., 2023), we construct $\mathcal{D}_{\text{balanced}}$ from a *small* group-annotated held-out set $\mathcal{D}_{\text{val}}$ by subsampling all groups to match the size of the smallest group. This prevents the mask from preserving spurious concepts associated with overrepresented groups. Throughout this step, the feature extractor h_ψ, SAE encoder E and decoder D, and classification head g_ϕ remain frozen.

To learn the concept mask $\mathbf{m}$, we adapt the differentiable ℓ_0-regularised stochastic gating formulation of Louizos et al. (2018), yielding a sparse gate per SAE concept:

$$\mathcal{L}_{\text{mask}} = \frac{1}{N_{\text{balanced}}} \sum_{i=1}^{N_{\text{balanced}}} \ell_{\text{CE}}\big(g_\phi(\tilde{\mathbf{z}}^{(i)}), y^{(i)}\big) + \lambda_m \|\mathbf{m}\|_0, \tag{4}$$

where the first term encourages downstream task accuracy, while the second term encourages sparsity via an ℓ_0 penalty. We tune λ_m as described in Appendix E.

Like competing methods that use group-annotated validation data for hyperparameter selection (Espinosa Zarlenga et al., 2024; Kirichenko et al., 2023), SPOCK requires held-out group annotations without assuming greater privileged access to group information than these methods. However, it requires group labels only for a *small* held-out set, and does *not* retrain the base model. Crucially, the learned mask suppresses spurious concepts while preserving task-relevant ones, improving WGA while largely preserving the original model's average accuracy.

4.2 INFERENCE

The SAE $(D \circ E)$, along with the mask $\mathbf{m}$, can be thought of as a modular component that can be seamlessly integrated between the feature extractor and the classification head of the pre-trained DNN. At test time, the parameters of this module and the base network remain frozen, while latent representations are encoded, masked, decoded, and then passed to the original classification head:

$$\hat{y} = g_\phi\Big(D\big(\mathbf{m} \odot E(h_\psi(\mathbf{x}))\big)\Big).$$

This makes SPOCK an efficient post-hoc bias mitigation method, unlike most group-unsupervised approaches that require computationally expensive retraining.

5 EXPERIMENTS

In this section, we evaluate SPOCK by exploring the following research questions:

- RQ1 **(Worst-Group Performance)** Does SPOCK improve WGA over the baseline ERM model on both synthetic and real-world datasets? Is it competitive w.r.t. comparable methods?
- RQ2 **(Trade-off)** Does SPOCK achieve a more favourable trade-off between WGA and average accuracy than competing methods?
- RQ3 **(Group Label Efficiency)** How does SPOCK perform compared to state-of-the-art LLR baselines when only a limited amount of group-annotated validation data is available?
- RQ4 **(Spurious Suppression)** Do SPOCK's suppressed concepts capture spurious information?

5.1 SETUP

Datasets. We evaluate SPOCK on four standard spurious-correlation benchmarks: Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015), BAR (Nam et al., 2020), and MultiNLI (Williams et al., 2018). Waterbirds classifies birds as *waterbirds* or *landbirds*, with the label spuriously correlated with background (*land* or *water*), yielding four groups (e.g., *waterbird* on *water*). CelebA classifies hair colour as *blond* or *non-blond*, with the label spuriously correlated with perceived gender (*male* or *female*), also yielding four groups. BAR contains six action classes spuriously correlated with scenes, such as *climbing* with *rocky* backgrounds, while test-set background scenes differ. Without explicit group annotations, we report worst-class accuracy as a proxy for worst-group accuracy, following Espinosa Zarlenga et al. (2024). MultiNLI classifies sentence-pair relationships as *contradiction*, *entailment*, or *neutral*. Negation words (e.g., "*no*", "*never*") are spuriously correlated with contradiction, yielding six groups. Full dataset details are provided in Appendix B.

Baselines. We include Group DRO (Sagawa et al., 2020) as an approximate upper bound for WGA, as it uses training-set group annotations to optimise worst-group performance, and ERM as an approximate upper bound for average accuracy and a lower bound for WGA. For LLR baselines, we include DFR (Kirichenko et al., 2023), which retrains the last layer using the same held-out set configuration as SPOCK and AFR (Qiu et al., 2023), which uses group information only for model selection. Finally, we include JTT (Liu et al., 2021) and LfF (Nam et al., 2020) as unsupervised baselines to assess whether their higher computational costs yield performance gains.

Base Models and Model Selection. For the vision benchmarks, we use a ResNet-18 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009), and for MultiNLI, pre-trained BERT (Devlin et al., 2019). For SPOCK, we use a single-hidden-layer SAE and an optimisable mask with one parameter per concept. For fair comparison across all baselines, we perform a hyperparameter sweep following their respective works (Kirichenko et al., 2023; Qiu et al., 2023; Sagawa et al., 2020; Nam et al., 2020; Liu et al., 2021): we select ERM hyperparameters using mean validation accuracy and all other methods using validation WGA (i.e., we assume validation group labels for *all* methods). We evaluate each hyperparameter configuration over five seeds for LLR methods and SPOCK, and three seeds for all other baselines due to their higher computational costs. We report the mean and standard deviation across multiple runs of the selected configuration. Full hyperparameter search spaces and training details are in Appendix C.

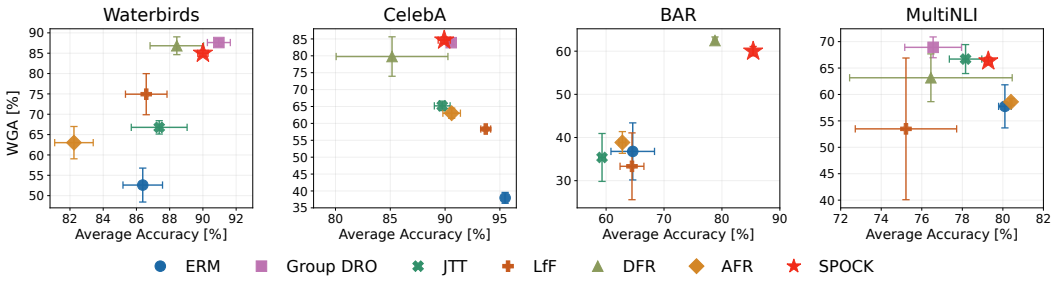


Figure 3: **Trade-off between worst-group and average accuracy.** Points nearer the top-right indicate a better trade-off. SPOCK consistently achieves strong WGA while retaining high average accuracy across all four datasets. Group DRO requires group labels for the *entire* training set (unavailable for BAR). Therefore, it serves as a group-supervised reference rather than a comparable baseline. Full numerical results are provided in Table 2 in Appendix D.

5.2 Worst-Group Performance (RQ1)

SPOCK improves worst-group accuracy across all datasets. Figure 3 shows that SPOCK (red) substantially improves WGA over ERM (blue) across all four benchmarks. In terms of WGA, SPOCK outperforms all baselines on CelebA. On MultiNLI, it is on par with JTT (teal); however, JTT requires up to 2.6× the ERM runtime, whereas SPOCK remains close to it (Figure 4). On Waterbirds and BAR, SPOCK falls slightly short of DFR (olive), but DFR has wider error bars across runs on most datasets, suggesting greater sensitivity to the composition of the group-balanced dataset. While both methods downsample to construct this dataset, DFR retrains the classifier head on it, making its decision boundary more susceptible to sample variance. This is undesirable when collecting additional representative data is costly or infeasible. SPOCK instead learns a mask to filter out spurious concepts and therefore remains more stable across seeds. SPOCK also outperforms AFR (brown), an LLR method that does not require an explicit group-balanced held-out set.

These results support our hypothesis that suppressing spurious concepts in last-layer representations improves worst-group performance without classifier retraining. Therefore, SPOCK provides an efficient and generalisable framework for mitigating spurious correlations across both vision and language domains.

5.3 Average Performance vs Worst Group Performance Trade-off (RQ2)

SPOCK improves robustness while maintaining competitive average accuracy against baselines. Figure 3 shows that SPOCK (red) occupies a "sweet spot" achieving strong WGA and high average accuracy across all four datasets. Compared with LLR baselines, SPOCK consistently achieves (1) higher average accuracy than DFR (olive) while maintaining competitive WGA, and (2) substantially higher WGA than AFR (brown) while retaining similar or higher average accuracy across most datasets.

Figure 3 further shows that JTT (teal) and LfF (rust) provide a less favourable balance between WGA and average accuracy across all datasets. More importantly, these unsupervised approaches also incur significantly higher end-to-end training costs than other methods, requiring roughly 1.5–3× the ERM runtime (Figure 4). This suggests that their additional computational cost does not necessarily translate into better performance. Moreover, LfF's performance varies substantially across runs on BAR and MultiNLI, as bias-aligned examples dominate the training sets, leaving LfF's debiased model to rely on a high-variance set of bias-conflicting examples.

Finally, we observe that, compared with Group DRO (purple), the upper-bound reference for WGA, SPOCK remains competitive, with overlapping error bars on CelebA and MultiNLI. Against ERM (blue), the reference for average accuracy, SPOCK remains competitive on Waterbirds and MultiNLI, with only a 5% drop on CelebA. Surprisingly, ERM does not achieve the highest average accuracy on Waterbirds or BAR, reflecting the distribution shift between training and evaluation sets. Overall, these results suggest that SPOCK offers the best trade-off between WGA and average accuracy across the evaluated benchmarks.

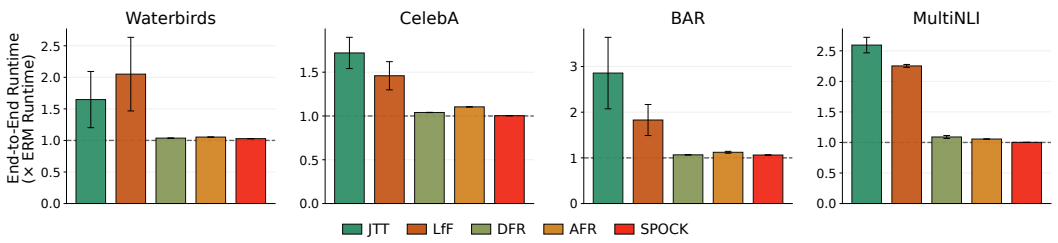


Figure 4: **End-to-end runtime relative to ERM.** We report the runtime of methods that do not require group supervision during training as a multiple of the ERM runtime. SPOCK and LLR-based methods achieve runtimes close to that of ERM, while JTT and LfF incur higher computational costs.

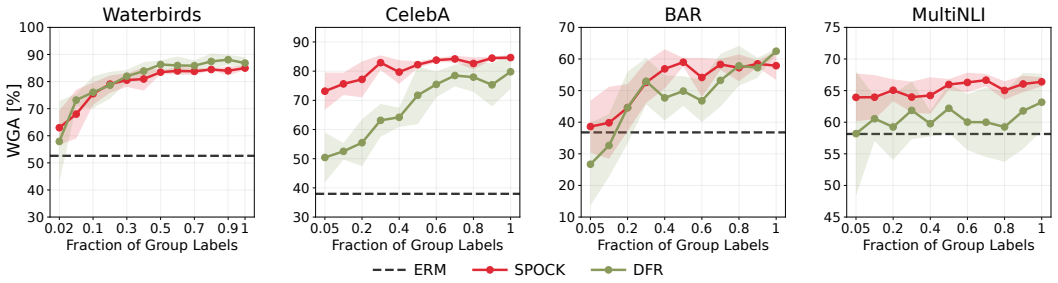


Figure 5: **Group label efficiency.** SPOCK maintains strong WGA with limited group-annotated validation data, outperforming DFR on CelebA and MultiNLI while remaining competitive on Waterbirds and BAR. This demonstrates its effectiveness when group annotations are costly or scarce.

5.4 Group Label Efficiency (RQ3)

SPOCK remains effective with only limited group-annotated validation data. Figure 5 shows that SPOCK achieves better WGA than ERM with *as few as 2 examples per group* on Waterbirds and *only 1 example per group* on CelebA. With only 10% of the validation set, SPOCK surpasses the WGA of both LfF and JTT (Table 2 in Appendix D) on the vision benchmarks, despite these baselines being tuned using the complete validation set.

Since DFR and SPOCK both use a group-balanced held-out set, we compare their WGA as we vary available group annotations. For BAR, fractions correspond to the approximately group-balanced set described in Appendix B. Our results in Figure 5 show that, across data sizes, SPOCK consistently outperforms DFR on CelebA and MultiNLI while remaining competitive on Waterbirds and BAR. Moreover, DFR is more sensitive on CelebA and MultiNLI than on Waterbirds, even with access to the complete validation set. We note that, because these real-world datasets have validation sets that follow the training distribution, balancing requires substantial downsampling, leaving far fewer examples per group, and consequently a smaller balanced set. This supports our earlier observation (Section 5.2) that DFR is sensitive to the quality of the group-balanced held-out set and suggests that SPOCK is better suited to settings where group annotations are limited or costly.

5.5 Spurious Concept Suppression (RQ4)

SPOCK suppresses concepts that encode spurious information. Lastly, we examine whether concepts suppressed by SPOCK capture more spurious information than those preserved. For this, we train linear probes to predict the spurious attribute from two sets of equal size k: the k most suppressed and k most preserved concepts, according to the learned mask $\mathbf{m}$. We do this for all tasks but BAR, for which group annotations are not available. Intuitively, if $\mathbf{m}$ suppresses spurious concepts, the suppressed set should predict the spurious attribute better than the preserved set. As shown in Figure 6, the bottom-k (most suppressed) concepts consistently yield higher accuracy than the top-k (most preserved) concepts on Waterbirds and CelebA, providing evidence that SPOCK selectively suppresses concepts capturing spurious information. Qualitatively, Figure 7 shows that

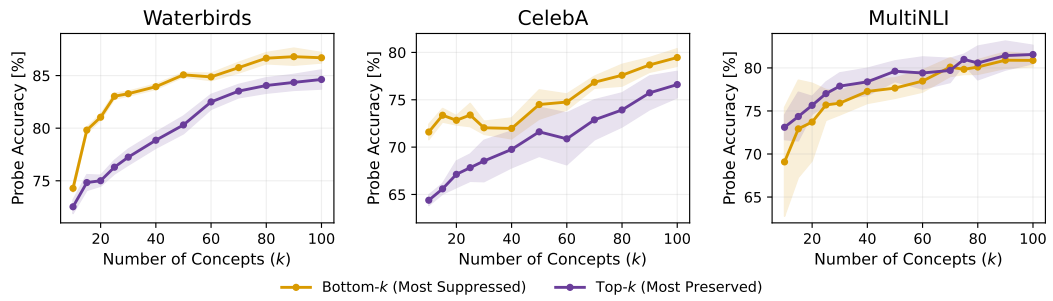


Figure 6: **Spurious information across suppressed and preserved SAE concepts.** For each k, linear probes predict the spurious attribute from the k most suppressed or most preserved concepts. Suppressed concepts yield higher accuracy on Waterbirds and CelebA, suggesting selective suppression of spurious information. We do not observe a significant distinction on MultiNLI.

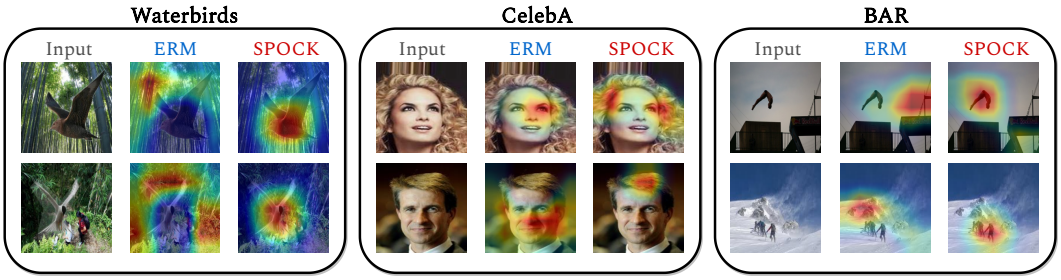


Figure 7: **Saliency visualisations.** SPOCK attends more strongly to task-relevant concepts than to spurious concepts compared to ERM. On Waterbirds, attention shifts from background to bird; on CelebA, from facial features to hair; and on BAR, from background to action.

SPOCK shifts the ERM model's attention, as determined by Grad-CAM (Selvaraju et al., 2017), from spurious concepts to task-relevant concepts.

Finally, in contrast to visual tasks, the separation in probe accuracy in Figure 6 is less pronounced on MultiNLI, with overlapping error bars between the bottom-k and top-k concepts. This suggests that negation-related language information may be more broadly distributed across concepts rather than concentrated in those suppressed by SPOCK. Therefore, we cannot draw the same conclusion from these results and leave a better understanding of this effect to future work.

6 Discussion and Conclusion

Limitations. We point out two important limitations of our work: first, SAEs are not guaranteed to discover semantically disentangled concepts (Kantamneni et al., 2025; Bricken et al., 2024; Leask et al., 2025), so the mask may act on features only partially corresponding to distinct spurious concepts. Second, although SPOCK does not require a large group-labelled training set, it requires group labels to construct a balanced set for mask optimisation. Future work could reduce this dependence through implicit balancing based on per-sample loss (Qiu et al., 2023; Liu et al., 2021; Nam et al., 2020), clustering (Sohoni et al., 2020; Le et al., 2025; Espinosa Zarlenga et al., 2024), or multimodal foundation models (Tan et al., 2024; He et al., 2024) to assist with group annotation.

Conclusion. This work introduced SPOCK, an efficient bias mitigation method that uses SAE-discovered concepts to modify pre-trained latent representations, suppressing spurious concepts while preserving task-relevant ones and keeping the classifier fixed. Across several benchmarks, SPOCK substantially improves worst-group accuracy while maintaining competitive average accuracy, even with limited group-annotated validation data. Overall, these results suggest that directly modifying learned representations can reduce reliance on spurious correlations and improve the trade-off between worst-group and average accuracy, all without expensive end-to-end retraining.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and improving the clarity and grammar of this manuscript. Additionally, we used generative AI tools to generate boilerplate code, bash scripts for hyperparameter sweeps, and to assist with processing and presenting experimental results. We also used generative AI tools to analyse existing literature and verify our interpretation of prior methodologies. We reviewed all AI-assisted work and verified the generated code for correctness. All core research ideas, experimental design, and interpretation and analysis of the results were conducted by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

Our work proposes a method for improving the fairness of deep neural networks by mitigating their reliance on spurious correlations. We hope that SPOCK may contribute to more robust debiasing practices in real-world applications, particularly in high-stakes domains such as medicine, where model predictions can directly affect people's lives. Nevertheless, we recognise potential negative societal implications. Since SPOCK learns which concepts to suppress using a held-out set designed to represent all subgroups equally, its effectiveness depends on the quality and representativeness of this set. In principle, a biased or poorly constructed held-out set could instead encourage the suppression of useful concepts and reinforce or exacerbate existing biases. This risk, however, is not unique to SPOCK and more broadly applies to all bias mitigation methods.

REPRODUCIBILITY STATEMENT

We have made a comprehensive effort to ensure the reproducibility of our results. The architecture and optimisation procedure for SPOCK are described in Section 4, while dataset construction and preprocessing are described in Appendix B. Experimental settings, model selection, and implementation details are provided in Appendix C. Moreover, we provide the source code required to reproduce our experiments as supplementary material, including preprocessing scripts, training code, and execution instructions. A cleaned and fully documented version of the codebase will be released publicly as an MIT-licensed repository upon publication.

REFERENCES

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *International Conference on Machine Learning*, 2024.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, Thomas Henighan, and Adam Jermyn. Using dictionary learning features as classifiers. *Transformer-Circuits. pub*, 2024.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200, 2021.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances*, 8(31): eabq6147, 2022.

Christoph Daube, Tian Xu, Jiayu Zhan, Andrew Webb, Robin AA Ince, Oliver GB Garrod, and Philippe G Schyns. Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, 2(10), 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Mateo Espinosa Zarlenga, Swami Sankaranarayanan, Jerone TA Andrews, Zohreh Shams, Mateja Jamnik, and Alice Xiang. Efficient bias mitigation without privileged information. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8730–8738, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Quentin Guimard, Federico Bartsch, Simone Caldarella, Rahaf Aljundi, Elisa Ricci, and Massimiliano Mancini. SEM: Sparse embedding modulation for post-hoc debiasing of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8101–8110, 2026.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *International Conference on Machine Learning*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, 2024.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, pp. 43832–43868, 2024.

John Collins Hill, Tyler LaBonte, Xinchen Zhang, and Vidya Muthukumar. On the unreasonable effectiveness of last-layer retraining. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, pp. 7827–7845, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *International Conference on Machine Learning*, 2025.

Patrik Kenfack, Ulrich Aïvodji, and Samira Ebrahimi Kahou. Gradtune: Last-layer fine-tuning for group robustness without group annotation. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36: 11552–11579, 2023.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 2020.

Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Out of spuriousity: Improving robustness to spurious correlations without group annotations. *Transactions on Machine Learning Research*, 2025.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, pp. 53617–53642, 2025.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

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 *International Conference on Machine Learning*, pp. 6781–6792, 2021.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024.

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*, pp. 3730–3738, 2015.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018.

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10051–10059, 2022.

Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. In *International Conference on Learning Representations*, 2025.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, et al. Shortcut learning in medical AI hinders generalization: method for estimating ai model generalization without external data. *NPJ digital medicine*, 7(1):124, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pp. 39643–39666, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pp. 28448–28467, 2023.

Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

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. In *International Conference on Learning Representations*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sahil Singla and Soheil Feizi. Salient ImageNet: How to discover spurious features in Deep Learning? In *International Conference on Learning Representations*, 2022.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 930–957, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4, 1991.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual review of vision science*, 9(1):501–524, 2023.

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)*, pp. 1112–1122, 2018.

Chao Wu, Zhenyi Wang, Kangxian Xie, Naresh Kumar Devulapally, Vishnu Suresh Lokhande, and Mingchen Gao. Model-agnostic gender bias control for text-to-image generation via sparse autoencoder. *arXiv preprint arXiv:2507.20973*, 2025.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, 2023.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. Spurious correlations in machine learning: A survey. *CoRR*, abs/2402.12715, 2024.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Improving group robustness on spurious correlation via evidential alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3610–3621, 2025.

Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Christopher W. Porter, Megan Coffee, Zhe Zeng, Sheng Li, Ziran Wang, Ting-Hao Kenneth Huang, James Matthew Rehg, Henry Kautz, and Aidong Zhang. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12188–12200, 2024.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-N-Contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pp. 26484–26516, 2022.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pp. 12857–12867, 2021.

A Disentangling Concepts Using Sparse Autoencoders

As discussed in Section 3, individual dimensions of a model's latent representation are *polysemantic*, encoding multiple concepts simultaneously (Elhage et al., 2022; Park et al., 2024). We therefore use an SAE to disentangle the dense latent representation into distinct concepts. To assess whether this step is crucial for SPOCK, we compare our standard approach, which learns a mask over SAE concepts, against directly learning a sparse mask over the original latent dimensions. Figure 8 shows that SPOCK benefits from SAE-based masking on the vision benchmarks. This suggests that masking on SAE space enables more selective suppression of spurious information, whereas masking polysemantic latent dimensions may also remove task-relevant information.

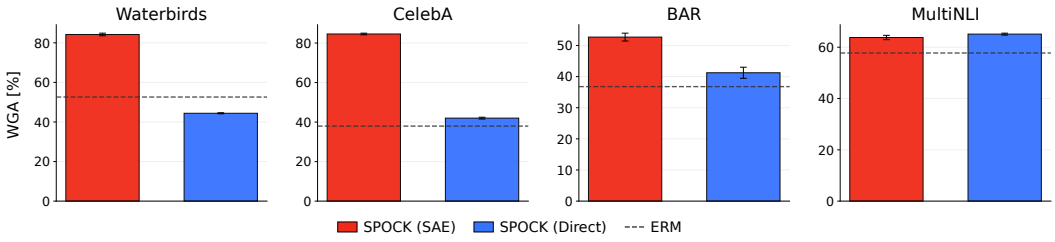


Figure 8: **Masking SAE Concepts vs. Direct Latents.** Masking SAE concepts consistently yields higher WGA than directly masking latents on Waterbirds, CelebA and BAR. Performance on MultiNLI is comparable to direct latent masking.

B Dataset Details

In this section, we provide additional information on the datasets used for our evaluation in Section 5. We summarise their composition in Table 1 and describe the corresponding prediction tasks below.

Waterbirds. This is a binary image classification dataset introduced by Sagawa et al. (2020), where the goal is to classify a bird as either a *waterbird* or a *landbird*. This dataset is constructed by cropping birds from the CUB dataset (Wah et al., 2011), and placing them onto land or water backgrounds from the Places dataset (Zhou et al., 2017). This construction enables explicit control over the strength of spurious correlations between the background and the class label.

We use the official train-validation-test split to ensure a fair comparison. In the training set, 95% of birds are placed on class-aligned backgrounds. As a result, the background acts as a spurious attribute, with the majority groups being (*landbird*, *land*) and (*waterbird*, *water*), leaving (*landbird*, *water*) and (*waterbird*, *land*) as minority groups. On the other hand, the validation and test sets are group-balanced within each label.

CelebA. This is a real-world dataset of celebrity faces with over 40 binary attributes, such as *Mustache*, *Blond_Hair*, *Male*, *Young*, etc., for each image. We use the official train-validation-test split; however, we only use 20% of the training and validation sets so that we can perform a hyperparameter grid search for all methods, given our limited computational resources. When subsampling, we ensured that the underlying group distribution remains identical to that of the complete set. The test set is used as is.

In line with previous works (Sagawa et al., 2020; Qiu et al., 2023; Nam et al., 2020; Liu et al., 2021; Kirichenko et al., 2023), we consider a binary classification task where the goal is to classify hair colour as *blond* or *non-blond*. In the training set, hair colour is strongly correlated with gender (Sagawa et al., 2020; Liu et al., 2015): approximately 94% of individuals with *blond* hair are female. Therefore, gender acts as a spurious attribute, with the majority groups being (*blond*, *female*) and (*non-blond*, *male*), leaving (*blond*, *male*) and (*non-blond*, *female*) as minority groups. Furthermore, the validation and test sets follow the same distribution as the training set.

BAR. Biased Action Recognition (BAR) is a real-world action-recognition dataset introduced by Nam et al. (2020). It contains six action classes: *climbing*, *diving*, *fishing*, *vaulting*, *racing*, and *throwing*. The task is to predict the action depicted in an image. In the training set, each action class is spuriously correlated with a background scene; for example, *climbing* images commonly contain

Table 1: **Dataset statistics.** Group compositions, empirical training distributions, and sample counts across the train, validation, and test splits for Waterbirds, CelebA, BAR, and MultiNLI. Note that the probabilities may not sum exactly to 1 due to rounding. For BAR, spurious-attribute annotations are unavailable, so corresponding statistics are omitted.

Dataset	Group g		Training distribution $\hat{p}$			Data quantity		
	Class y	Spurious a	$\hat{p}(y)$	$\hat{p}(g)$	$\hat{p}(y\|a)$	Train	Val	Test
Waterbirds	landbird	land	.768	.730	.984	3498	467	2225
	landbird	water		.038	.148	184	466	2225
	waterbird	land	.232	.012	.016	56	133	642
	waterbird	water		.220	.852	1057	133	642
CelebA	non-blond	female	.851	.440	.758	14326	1707	9767
	non-blond	male		.411	.980	13375	1655	7535
	blond	female	.149	.141	.242	4576	575	2480
	blond	male		.009	.020	277	36	180
BAR	climbing	–	.168	–	–	306	40	85
	diving	–	.275	–	–	500	40	139
	fishing	–	.079	–	–	143	40	22
	vaulting	–	.142	–	–	259	40	111
	racing	–	.174	–	–	316	40	112
	throwing	–	.163	–	–	297	40	65
MultiNLI	contradiction	no negation	.333	.278	.300	21879	541	2692
	contradiction	negation		.055	.759	4301	102	548
	entailment	no negation	.333	.326	.351	25587	676	3390
	entailment	negation		.008	.105	593	20	73
	neutral	no negation	.333	.324	.349	25408	606	3043
	neutral	negation		.010	.136	772	19	86

rock walls, *vaulting* images often contain sky backgrounds, and *diving* images typically contain underwater scenes.

We use the official train–evaluation split. To approximate a group-balanced validation set despite the absence of explicit group annotations, we sample 20 examples per class from each of the official training and evaluation sets, thereby including equal numbers of bias-aligned and bias-conflicting examples. The remaining training examples form our training split, while the remaining evaluation examples form the test split. We report worst-class accuracy as a proxy for WGA, following Espinosa Zarlenga et al. (2024).

MultiNLI. This is a text classification dataset introduced by Williams et al. (2018), where the goal is to classify the relationship between a pair of sentences as *contradiction*, *entailment*, or *neutral*. It contains samples from ten different genres, such as fiction, government, and travel. In the training set, the presence of negation words such as "*no*" and "*never*" is spuriously correlated with the *contradiction* label: approximately 76% of examples with negation are labelled *contradiction*.

We use the official training, matched development, and mismatched development splits as our train–validation–test splits; however, we only use 20% of the training and matched development sets so that we can perform a hyperparameter grid search for all methods given our limited computational resources. The mismatched development set is used as is. When subsampling, we ensured that the underlying group distribution remains identical to that of the complete set.

C Experimental Details

In this section, we provide additional details on the experimental setup described in Section 5, along with the specific procedures for hyperparameter selection.

C.1 Implementation and Training Details

Base Models. For all vision tasks, we use a ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). We resize all input samples to $3 \times 224 \times 224$ and apply random flip and crop data augmentation during training. Additionally, all images are converted to their floating-point representation and normalised using ImageNet's (Deng et al., 2009) mean and standard deviation. We train these models using an SGD optimiser with a learning rate selected from $\{10^{-4}, 10^{-3}, 10^{-2}\}$, weight decay selected from $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 1\}$, momentum of 0.9, and a batch size of 64, for up to 200 epochs. We reduce the learning rate by a factor of ten when the training loss plateaus for 10 epochs. Finally, we use early stopping to stop training if a model's selection metric does not increase by more than 10^{-3} over the last 5 epochs.

For MultiNLI, we use a pretrained BERT-base-uncased (Devlin et al., 2019) with the default tokeniser and model settings from the original implementation. We train the model using the AdamW optimiser with a learning rate of 2×10^{-5}, zero weight decay, and a batch size of 128 for 5 epochs. For LfF, we reduce the batch size to 64 due to computational constraints, as it requires training two models simultaneously.

SPOCK. The SAE is implemented as a single-hidden-layer autoencoder, where the encoder consists of a linear layer followed by a ReLU activation, and the decoder consists of a single linear layer. We use an expansion factor of 8. We train the SAE for up to 200 epochs using the Adam optimiser (Kingma & Ba, 2014), with a batch size of 256, a learning rate of 10^{-3} and an ℓ_1 penalty of 0.01 on the vision benchmarks; for MultiNLI, we use a learning rate of 10^{-4} and an ℓ_1 penalty of 0.003. The decoder weights are normalised after each optimisation step. We also use early stopping based on validation loss, stopping training if there is no improvement for 20 consecutive epochs.

The mask is implemented as an optimisable parameter with the same dimensionality as the SAE hidden representation, such that each entry corresponds to a single SAE concept and controls whether that concept is preserved or suppressed. We initialise the mask parameters from a Gaussian distribution $\mathcal{N}(0, 0.01^2)$, giving all concepts approximately equal initial probability of being retained. We encourage sparsity using an ℓ_0 penalty on the number of active concepts. Since this objective is discrete and non-differentiable, we use the hard-concrete relaxation of Louizos et al. (2018) to obtain a smooth, differentiable approximation. We optimise the mask for 500 epochs using the Adam optimiser (Kingma & Ba, 2014), with a learning rate of 0.1.

C.2 Model Selection

To ensure a fair evaluation, we perform a hyperparameter grid search over all combinations of candidate hyperparameter pairs for each method. However, to reduce computational costs, we fix the learning rates and weight decay for all methods to the ERM ones, except for Group DRO, whose performance is strongly dependent on these hyperparameters (Sagawa et al., 2020). We outline the specific hyperparameter search spaces for each method below:

ERM. We trained models across 3 independent random seeds, considering learning rates $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ and weight decay parameters $\lambda_{\ell_2} \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 1\}$, resulting in a total of 54 runs. Models were selected based on the highest average validation accuracy. The selected hyperparameters were $(\eta, \lambda_{\ell_2}) = (10^{-2}, 10^{-4})$ for Waterbirds, $(\eta, \lambda_{\ell_2}) = (10^{-3}, 10^{-4})$ for CelebA and $(\eta, \lambda_{\ell_2}) = (10^{-2}, 10^{-2})$ for BAR. For MultiNLI, we use the fixed BERT training configuration of $(\eta, \lambda_{\ell_2}) = (2 \times 10^{-5}, 0)$.

Group DRO. We paired learning rates $\eta \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ with weight decays $\lambda_{\ell_2} \in \{0, 10^{-4}, 10^{-2}, 1\}$ across 3 independent random seeds. Furthermore, we fixed the exponential step-size decay factor to $\gamma = 0.01$, as it is known to ensure stable training (Sagawa et al., 2020). We selected models based on the highest validation WGA, resulting in the following parameters: $(\eta, \lambda_{\ell_2}) = (10^{-4}, 1)$ for Waterbirds, $(\eta, \lambda_{\ell_2}) = (10^{-4}, 10^{-2})$ for CelebA and $(\eta, \lambda_{\ell_2}) = (10^{-2}, 10^{-2})$ for BAR. For MultiNLI, we use the fixed BERT training configuration of $(\eta, \lambda_{\ell_2}) = (2 \times 10^{-5}, 0)$.

JTT. JTT uses two parameters: T, the number of epochs used to train the early-stopped model, and λ_{up}, the weight by which each sample misclassified by the auxiliary model is upweighted during retraining (Liu et al., 2021). We performed a grid search over $T \in \{1, 2, 3, 5, 7, 9, 10\}$ and $\lambda_{\text{up}} \in \{10, 25, 50, 100\}$ across 3 independent random seeds and selected the model with the highest validation WGA. We additionally tune MultiNLI over smaller values of $\lambda_{\text{up}} \in \{3, 4, 5\}$, as Liu et al. (2021) report that these values perform better. The selected hyperparameters were: (T, λ_{up}) = (7, 50) for Waterbirds, (T, λ_{up}) = (1, 10) for CelebA, (T, λ_{up}) = (2, 100) for BAR and (T, λ_{up}) = (1, 4) for MultiNLI.

LfF. LfF involves training a biased model using generalised cross-entropy loss (Nam et al., 2020), which dynamically reweights samples for the debiased model. This loss depends on the hyperparameter $q \in (0, 1]$, which controls the model's ability to learn complex features. We searched over $q \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ across 3 independent random seeds and selected the model based on the highest validation WGA. The selected hyperparameters were $q = 0.75$ for Waterbirds, $q = 0.5$ for CelebA, $q = 0.25$ for BAR and MultiNLI.

DFR & SPOCK. We followed the same model selection procedure for DFR and SPOCK. DFR uses a single hyperparameter, C, which corresponds to the inverse ℓ_1 regularisation strength. Meanwhile, for SPOCK, we only tune the sparsity penalty λ_m of the mask. For each seed, we first randomly split the group-balanced set into two equal halves, where the first half was used for training and the second half was used to tune the hyperparameters based on the highest WGA. We then used the chosen hyperparameter to reoptimise over the complete group-balanced set, which was subsequently evaluated on the test set. We considered the hyperparameter space for C as suggested by Kirichenko et al. (2023) and $\lambda_m \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1\}$, and reported the average performance across 5 independent random seeds.

AFR. AFR has two hyperparameters: a temperature parameter, γ, which controls the sample upweighting strength based on base model predictions, and an ℓ_2 regularisation parameter (Qiu et al., 2023). We used an 80:20 split of the training data, with 80% used to train the base model via ERM and the remaining 20% used to retrain the last layer. We used the validation set to perform a grid search over the hyperparameter space suggested by Qiu et al. (2023) across 5 independent random seeds and selected the optimal configuration based on the highest validation WGA.

C.3 Software and Hardware Used

Software. We initialised our codebase from the MIT-licensed official Group DRO repository (Sagawa et al., 2020), which provided a useful starting point for experiment tracking, logging, and evaluation. We subsequently adapted the official implementations of the remaining baselines into this framework (Qiu et al., 2023; Kirichenko et al., 2023; Nam et al., 2020; Liu et al., 2021). Our

implementation of SPOCK and all underlying models is built on top of PyTorch 2.9.1 (Paszke et al., 2019).

All figures except Figures 1, 2, and 7 were generated using Matplotlib 3.10.8; the remaining figures were created with draw.io, an Apache 2.0-licensed open-source drawing tool. All code, configurations, and scripts required to reproduce our results will be released in an MIT-licensed public repository upon publication.

Resources. We ran all experiments on a shared compute machine equipped with eight NVIDIA L4 GPUs with 24 GB of memory each and 32 Intel Xeon Silver 4514Y CPU cores.

D Additional Results

Table 2 reports the numerical results corresponding to the WGA–average accuracy trade-off shown in Figure 3.

Table 2: **Performance overview.** Worst-group and average test accuracy of SPOCK and the baselines across the benchmark datasets. The Group Annot. column indicates whether group labels are available during training and validation. **Bold** indicates the highest performance among bias mitigation methods that do not use training group annotations, while underlined results are not statistically significantly different ($p \geq 0.05$) from the best-performing method in the corresponding setting, evaluated using Welch's t-test (Ruxton, 2006).

	Group Annot.	Waterbirds		CelebA		BAR		MultiNLI	
	Train/Val	Worst (%)	Average (%)	Worst (%)	Average (%)	Worst (%)	Average (%)	Worst (%)	Average (%)
ERM (baseline)	✗ / ✗	52.60 ± 4.17	86.38 ± 1.19	37.96 ± 1.61	95.48 ± 0.03	36.78 ± 6.61	64.60 ± 3.76	57.75 ± 4.08	80.10 ± 0.31
Group DRO (Sagawa et al., 2020)	✓ / ✓	87.62 ± 0.95	90.96 ± 0.69	83.89 ± 0.56	90.56 ± 0.16	N/A	N/A	68.92 ± 1.97	76.56 ± 1.40
JTT (Liu et al., 2021)	✗ / ✓	66.78 ± 1.64	87.37 ± 1.68	65.19 ± 1.40	89.74 ± 0.72	35.38 ± 5.55	59.30 ± 0.58	**66.70 ± 2.74**	78.16 ± 0.80
LfF (Nam et al., 2020)	✗ / ✓	74.93 ± 5.05	86.59 ± 1.25	58.33 ± 0.56	**93.71 ± 0.45**	33.33 ± 7.74	64.47 ± 2.06	53.49 ± 13.41	75.22 ± 2.50
DFR (Kirichenko et al., 2023)	✗ / ✓	**86.83 ± 2.19**	88.43 ± 1.61	79.79 ± 5.84	85.16 ± 5.10	**62.46 ± 0.84**	78.80 ± 0.36	63.16 ± 4.52	76.45 ± 4.01
AFR (Qiu et al., 2023)	✗ / ✓	62.67 ± 3.80	82.24 ± 1.16	79.43 ± 4.06	89.05 ± 0.78	38.85 ± 2.52	62.81 ± 0.60	58.60 ± 0.64	**80.40 ± 0.05**
SPOCK	✗ / ✓	84.98 ± 0.72	**90.00 ± 0.31**	**84.65 ± 0.92**	89.92 ± 0.55	60.00 ± 1.09	**85.39 ± 0.00**	66.34 ± 0.72	79.28 ± 0.19

For Waterbirds, following prior work (Sagawa et al., 2020; Liu et al., 2021; Nam et al., 2020; Kirichenko et al., 2023; Qiu et al., 2023), we additionally report the mean weighted accuracies across methods in Table 3. We calculate mean accuracy by weighting the test-group accuracies according to the group distribution in the training set, since the test-set group distribution differs from the training distribution.

Table 3: **Waterbirds performance across methods.** Following prior work, we report mean accuracy by weighting test-group accuracies according to the training-group distribution, alongside worst-group accuracy. **Bold** indicates the highest performance among bias mitigation methods that do not use training group annotations, while underlined results are not statistically significantly different ($p \geq 0.05$) from the best-performing method in the corresponding setting, evaluated using Welch's t-test (Ruxton, 2006).

	ERM	Group DRO	JTT	LfF	DFR	AFR	SPOCK
WGA (%)	52.60 ± 4.17	87.62 ± 0.95	66.78 ± 1.64	74.93 ± 5.05	**86.83 ± 2.19**	62.67 ± 3.80	84.98 ± 0.72
Mean Accuracy (%)	96.69 ± 0.15	89.54 ± 0.71	94.12 ± 0.65	94.51 ± 0.81	88.18 ± 1.82	**95.81 ± 0.30**	94.23 ± 0.93

E Ablation Studies

Mask Sparsity. To tune the mask sparsity penalty λ_m, we split the balanced set $\mathcal{D}_{\text{balanced}}$ into disjoint subsets $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{tune}}$, using the former for mask learning and the latter for selecting λ_m. Figure 9 shows performance across a range of λ_m. Waterbirds and CelebA remain largely stable across different sparsity penalties, while BAR shows some variation in WGA across settings. MultiNLI, on the other hand, is more sensitive to λ_m, with larger changes in WGA while average accuracy remains relatively stable.

SAE Expansion Factor. Throughout our evaluation, we use an SAE expansion factor of $8\times$, following prior work (Bricken et al., 2023), where expanding a 512-dimensional latent representation

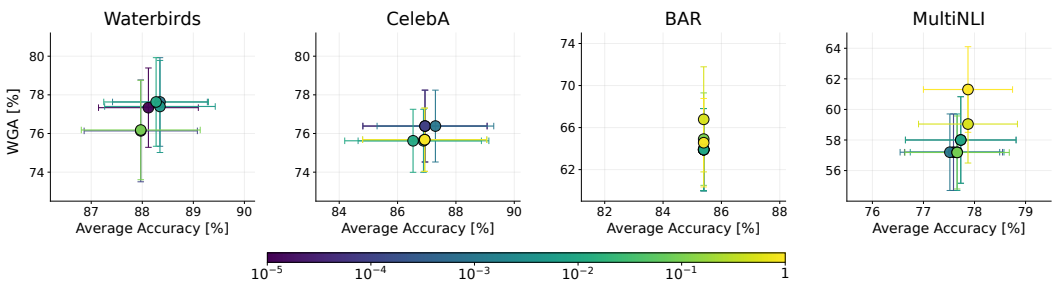


Figure 9: **WGA–accuracy trade-off across sparsity levels.** Performance is relatively stable across the vision benchmarks, while the text benchmark MultiNLI is more sensitive to λ_m.

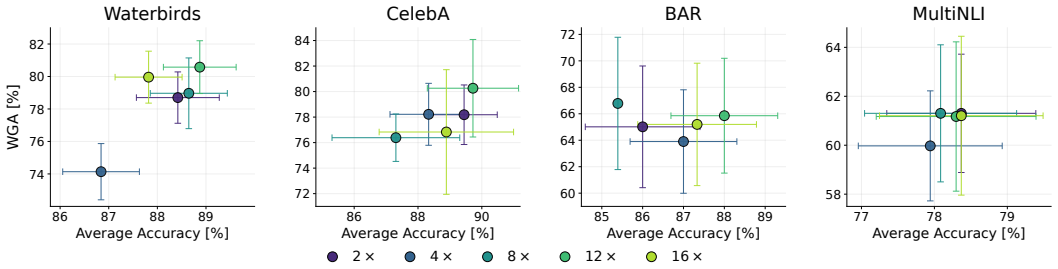


Figure 10: **WGA–accuracy trade-off across expansion factors.** Performance remains broadly competitive across expansion factors, with $8\times$ and $12\times$ showing the least compromise between WGA and average accuracy across benchmarks.

to 4,096 dimensions was shown to facilitate the isolation of meaningful concepts. To assess whether SPOCK depends strongly on this choice, we additionally evaluate a range of expansion factors on $\mathcal{D}_{\text{tune}}$. Figure 10 shows that $12\times$ generally provides a favourable trade-off between WGA and average accuracy across benchmarks, although $8\times$ remains competitive. Given that larger expansion factors increase the SAE representation size and that selecting them would require an additional, computationally expensive hyperparameter tuning step for relatively modest gains, we retain $8\times$ as a practical default.