

# Bias Leaves a Gradient Trail: Label-Free Bias Identification via Gradient Probes on Concept Decompositions

Thomas Vitry<sup>1,2,\*</sup>[0009-0008-3329-2026], Kieran Edgeworth<sup>1</sup>[0009-0008-0310-7088],  
Stefan Wermter<sup>1</sup>[0000-0003-1343-4775], and Jae Hee Lee<sup>1,\*</sup>[0000-0001-9840-780X]

<sup>1</sup> University of Hamburg, Hamburg, Germany

<sup>2</sup> Ecole Normale Supérieure de Rennes, Rennes, France

{thomas.vitry,kieran.edgeworth}@studium.uni-hamburg.de

{stefan.wermter,jae.hee.lee}@uni-hamburg.de

**Abstract.** Vision classifiers can exploit spurious correlations, achieving high in-distribution accuracy yet failing under distribution shift. Existing approaches to bias mitigation and analysis often depend on curated datasets, spurious-attribute or group labels, or retraining, which may be infeasible once a model is deployed or the relevant bias is unknown. We present a bias-label-free, post-hoc method for identifying spurious concepts in frozen vision models, relying only on standard class labels from a held-out audit dataset. For each target class, we collect patches from inputs predicted as that class and apply non-negative matrix factorization to intermediate activations to obtain a bank of interpretable concept vectors. Candidate concepts are then ranked with a bias estimator derived from their interaction with backpropagated gradients on misclassified examples: bias concepts tend to get activated when correcting false negatives and suppressed when correcting false positives. On Colored MNIST and Waterbirds the method recovers concepts aligned with the known spurious cue, and on CelebA it surfaces decision-relevant directions that only partially coincide with the annotated gender attribute; suppressing the top-ranked concepts at inference time improves worst-group accuracy by up to 17.9 percentage points on Waterbirds and 10.4 on CelebA without any retraining or parameter updates. Our method identifies decision-relevant spurious directions that need not coincide with annotated ones, providing both an interpretable auditing tool and an actionable debiasing handle for frozen vision models. Code is available at <https://github.com/vitryt/label-free-bias-identification>.

**Keywords:** Bias identification · Bias mitigation · Concept-based XAI

## 1 Introduction

Modern vision models can achieve high in-distribution accuracy while relying on spurious correlations: attributes that predict the label in the training data

---

\* Corresponding authors

but are not causally related to the task. This shortcut learning behavior [12] can lead to brittle generalization and systematic failures when the correlation shifts at deployment. For example, in fine-grained recognition the model may over-rely on context: on Waterbirds [39], background type is spuriously correlated with the bird label during training. Such shortcuts may remain hidden when the available held-out set (e.g., the validation set) shares the same correlation, yet cause systematic errors when the correlation breaks at deployment.

To counter shortcut learning, methods often mitigate spurious correlations during data collection or training [29]. When the relevant bias concept is known, it can be targeted for mitigation via retraining [5,9,10,22] or post-hoc interventions on inputs or representations [2]. These pipelines typically assume that the spurious attribute has already been identified and therefore do not solve the discovery problem. In practice, however, we only have access to a deployed frozen model and a held-out set, but not to the original training pipeline, and rarely to bias or group annotations. In this setting, analyzing a deployed model is difficult because the spurious attribute is often unknown, and collecting new annotations can be costly. This motivates methods that can propose plausible spurious attributes post-hoc, while returning a concept vector in the latent space that is directly compatible with downstream interventions.

We study whether unsupervised, representation-level concept discovery can detect spurious attributes in frozen vision models when bias annotations are unavailable. Our approach decomposes intermediate activations into non-negative concept vectors and uses a held-out set (without group or bias labels) to probe the model via gradients (Sec. 3). We first test whether the spurious attribute learned by a biased model reliably appears as a distinct concept direction in this basis (Sec. 4.1). Building on this, we propose a bias identification criterion based on how concept activations change under a gradient descent step: bias concepts tend to be *activated* when correcting false negatives and *suppressed* when correcting false positives, while concepts encoding intrinsic class evidence change much less asymmetrically. We validate the scoring criterion against ground-truth bias (Secs. 4.2 and 4.3) and then test whether the identified concepts are actionable, by suppressing them at inference time without any parameter updates (Sec. 4.4). We evaluate these on Colored MNIST, Waterbirds and CelebA, three standard benchmarks for spurious correlation learning, and find that concept suppression substantially improves worst-group accuracy on Waterbirds and CelebA. The resulting approach provides both an interpretable auditing tool and a concrete debiasing handle for frozen vision models when bias annotations and retraining are unavailable.

## 2 Preliminaries

We consider a multi-class classification task with inputs  $x \in \mathcal{X} \subseteq \mathbb{R}^d$  and labels  $y \in \mathcal{Y} = \{1, \dots, C\}$ , and a hypothetical ideal classifier  $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ . Following [29], an *attribute* is a function  $b : \mathcal{X} \rightarrow \{0, 1\}$ ; it is *intrinsic* to class  $y$  if it defines membership (e.g., digit shape), and a *bias attribute* for  $y$  if it is spuriously

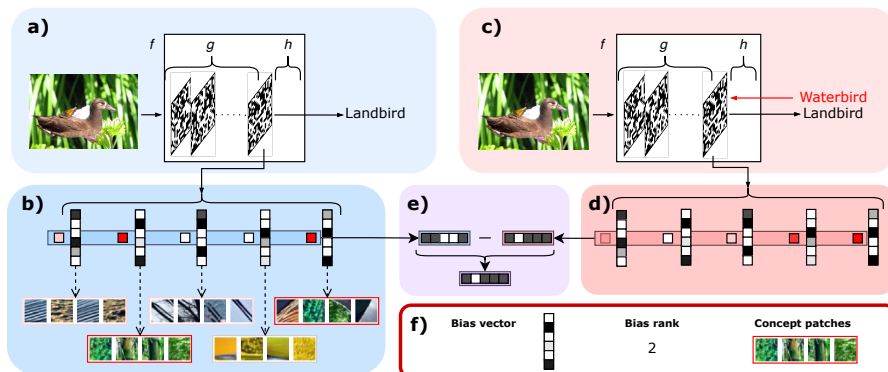


Fig. 1: Diagram of the bias identification method. (a) We collect false negatives and false positives from a held-out set (the *bias-audit* set) and pass them through the model. Here, studying class  $y=0$  (landbird), the sample is a false positive: a waterbird misclassified as a landbird. (b) Using non-negative concept decomposition trained beforehand on the bias-audit set, we decompose activations of the sample into its concept coefficients. (c) Using the ground-truth label from the bias-audit set, we backpropagate the gradient to the activations. (d) We decompose the updated activations (after the gradient step) using the same concept bank. (e) We compare which concepts are activated before and after the gradient step. Concepts that are removed (for false positives) or added (for false negatives) by the gradient step are candidate bias concepts. (f) The identified bias concepts come with their encoding in the activation space and their most activating patches, enabling interpretation. Note that bias concepts are identified by aggregating over multiple samples, not from a single example.

correlated with  $y$  without being intrinsic. We are given a frozen classifier  $f$  with class scores  $f(x) = (f_c(x))_{c=1}^C$  and a held-out *bias-audit* set  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  with  $y_i = f^*(x_i)$  but no attribute or group annotations.  $\mathcal{D}$  need not be from the training set and may be unbiased or distribution-shifted; it may come from a validation split, relabeled deployment data, or targeted stress tests. Our goal is to detect, for each class  $y$ , concept directions in the representation space that plausibly encode bias attributes used by the model and that are empirically implicated in its decisions.

*Non-Negative Concept Decomposition.* Following CRAFT [11], we write  $f = h \circ g$  where  $g : \mathcal{X} \rightarrow \mathcal{A}$  maps an input to an intermediate representation  $a = g(x) \in \mathcal{A} \subseteq \mathbb{R}^p$  and  $h : \mathcal{A} \rightarrow \mathbb{R}^C$  maps the representation to class scores. We choose a representation layer after a ReLU so that  $a \in \mathbb{R}_{\geq 0}^p$ , enabling non-negative matrix factorization (NMF). For each class  $y$ , let  $\hat{y}(x) = \arg \max_{c \in \mathcal{Y}} f_c(x)$  be the predicted label and collect inputs predicted as  $y$ :

$$\mathcal{X}_y = \{x_i : \hat{y}(x_i) = y\}.$$

A patch dataset is then created by applying a crop-and-resize operator  $\pi_s$  (patch size  $s$ ) to these images, yielding patches  $\mathcal{P}_y = \{\pi_s(x) : x \in \mathcal{X}_y\}$ . Let  $A_y = g(\mathcal{P}_y) \in \mathbb{R}_{\geq 0}^{n_y \times p}$  be the resulting activations. We compute a class-conditional concept bank

$$(U_y, W_y) = \arg \min_{U_y \geq 0, W_y \geq 0} \frac{1}{2} \|A_y - U_y W_y^\top\|_F^2, \quad (1)$$

via NMF, where  $W_y \in \mathbb{R}_{\geq 0}^{p \times r}$  contains  $r$  concept vectors  $w_{y,1}, \dots, w_{y,r}$  and  $U_y \in \mathbb{R}_{\geq 0}^{n_y \times r}$  the corresponding non-negative concept coefficients. As in [11], this typically yields semantically meaningful, relatively disentangled concepts; the decomposition is approximate and its residual is captured by Eq. (1). In practice, we solve this NMF objective by alternating between two non-negative least-squares (NNLS) subproblems, fixing  $W_y$  while optimizing  $U_y$  and vice versa. This monotonically decreases the objective and converges to a stationary point under standard assumptions.

Given  $W_y$ , we obtain concept coefficients for any representation vector  $a \in \mathbb{R}^p$  (in particular,  $a = g(x) \in \mathbb{R}_{\geq 0}^p$ ) by solving an NNLS problem,

$$u_y(a) = \arg \min_{u \geq 0} \|a - u W_y^\top\|_2^2, \quad (2)$$

and write  $u_{y,k}(a)$  for its  $k$ -th component.

### 3 Gradient-Based Bias Identification

We hypothesize that spurious attributes learned by a biased model appear as distinct concept directions in the non-negative concept decomposition, and that a gradient-based criterion can identify them among the extracted concepts.

Our intuition is as follows. A partially biased model uses both intrinsic and spurious attributes. Intrinsic concepts for class  $y$  should be active on true positives, while bias concepts need not be. On a false negative of class  $y$ , missing evidence is therefore more likely due to missing spurious support than missing intrinsic evidence; on a false positive, intrinsic evidence is typically absent and the error is more likely driven by a spurious attribute. Since NMF enforces non-negativity, evidence can only be added via positive concept activations. Because activity alone is noisy, we use gradient information to isolate the concepts the model itself would add or remove to correct its prediction: by propagating the loss gradient to activations (instead of updating weights), we measure which concepts change under one probe step. A bias concept should be systematically absent then added for false negatives, and present then removed for false positives.

To identify bias concepts without access to  $b(x)$ , we therefore measure how concept coefficients change under a single gradient probe step that improves the model’s prediction on a given example. For a labeled example  $(x_i, y_i)$  with representation  $a = g(x_i)$ , let  $L(h(a), y_i)$  be the cross-entropy loss and compute

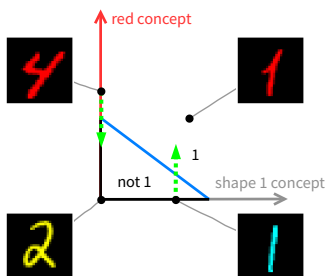


Fig. 2: Intuition of the gradient-concept interaction on class  $y$  ( $= 1$ ). The vertical axis is the bias concept **red** and the horizontal axis is the relevant concept **shape 1**. The blue line is the model’s decision boundary. For false negatives, i.e., image (**cyan**, 1), the gradient step increases bias attribute; for false positives, i.e., image (**red**, 4), the same probe removes the bias attribute. Intrinsic attribute, i.e., **shape 1**, is expected to show a weaker asymmetric pattern.

the gradient w.r.t. the representation,  $\nabla_a L(h(a), y_i)$ . We define a perturbed representation

$$a' = a - d \nabla_a L(h(a), y_i), \quad (3)$$

with gradient step size  $d > 0$ . After making  $a'$  non-negative, we compute  $u_{y,k}(a')$  via the same non-negative least squares problem. We treat a concept as *active* if its coefficient is positive and write

$$I_{y,k}(a) = \mathbb{1}[u_{y,k}(a) > 0],$$

where  $\mathbb{1}[\cdot]$  is the indicator function.

For a class  $y$ , we define false negatives and false positives with respect to  $y$ :

$$\text{FN}_y = \{(x_i, y_i) \mid y_i = y, \hat{y}(x_i) \neq y\}, \quad \text{FP}_y = \{(x_i, y_i) \mid \hat{y}(x_i) = y, y_i \neq y\}.$$

More explicitly, we define the false-negative and false-positive estimators for concept  $k$  as:

$$E_{\text{FN}}^{y,k} = \sum_{(x_i, y_i) \in \text{FN}_y} \frac{I_{y,k}(a'_i) - I_{y,k}(a_i)}{|\text{FN}_y|}, \quad E_{\text{FP}}^{y,k} = \sum_{(x_i, y_i) \in \text{FP}_y} \frac{I_{y,k}(a_i) - I_{y,k}(a'_i)}{|\text{FP}_y|},$$

where  $a_i = g(x_i)$  and  $a'_i$  is obtained from Eq. (3). The two terms measure how consistently concept  $k$  is *added* on false negatives and *removed* on false positives under the gradient step.

As illustrated in Fig. 2, we hypothesize that (i) bias concepts are more likely to increase under the gradient probe update in Eq. (3) on  $\text{FN}_y$  (i.e., the model prefers the spurious cue to predict  $y$ ) and decrease on  $\text{FP}_y$  (i.e., reducing features that spuriously support  $y$ ), and (ii) that this asymmetric behavior does *not* hold strongly for intrinsic concepts, i.e., false negatives possess intrinsic attributes of class  $y$  (otherwise they would not have ground-truth label  $y$ ), and false positives lack them. More precisely, for false negatives we expect the probability of a bias concept  $k$  being activated to increase after the gradient step:

$$\mathbb{P}(I_{y,k}(a') = 1) - \mathbb{P}(I_{y,k}(a) = 1) > 0,$$

whereas for false positives we expect the *opposite*. To rank concepts, we define the bias score of concept  $k$  for class  $y$  as

$$S_{y,k} = \frac{1}{2} \left( E_{\text{FN}}^{y,k} + E_{\text{FP}}^{y,k} \right). \quad (4)$$

For each example in  $\text{FN}_y \cup \text{FP}_y$ , scoring requires one backward pass to compute  $\nabla_a L$  and two non-negative least squares solves to obtain  $u_y(a)$  and  $u_y(a')$ . The non-negative concept decomposition is computed once per class. Figure 1 gives a visual overview of the pipeline, and Alg. 1 in Appendix A summarizes the full procedure, from class-conditional concept extraction to bias scoring and ranking.

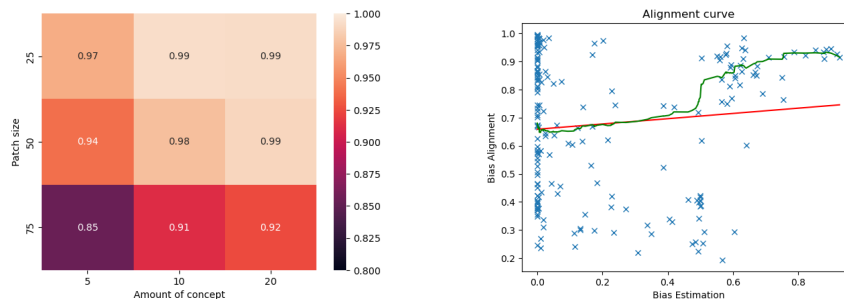
## 4 Experiments

We evaluate our method on standard benchmarks for spurious correlation learning and the hypotheses from Sec. 3.

**Datasets.** We use three standard spurious-correlation benchmarks with a 0.95 train-time correlation. **CMNIST** is a colored variant of MNIST where color acts as the spurious attribute (implementation adapted from [5]); we also construct an *unbiased* variant and a **shifted-bias** variant as distribution-shifted audit sets. **Waterbirds** [39] classifies waterbirds vs. landbirds [41] with water/land background [46] as the spurious attribute. **CelebA** [31] is a blond-hair vs. not-blond binary task with gender as the spurious attribute [42]; it is particularly challenging due to a severe class imbalance in the test split (only 180 blond men vs. 9767 not-blond women), making overall accuracy an unreliable metric. Split construction follows standard practice (details in Appendix B).

**Audit bias.** In all experiments, we vary the bias structure of the bias-audit set  $\mathcal{D}$ , a categorical we refer to as the *audit bias* throughout. We consider three values: **biased** (same spurious correlation as the training distribution), **unbiased** (correlation broken), and **shifted bias** (a different spurious assignment than training). CMNIST evaluates all three; Waterbirds and CelebA are only available in biased form, so we restrict them to **biased**. For all audit variants, the underlying images are taken from the held-out validation split of the corresponding dataset.

**Models and Training.** For CMNIST, we use a three-layer MLP with ReLU activations, trained for 100 epochs with stochastic gradient descent (SGD) and batch size 128. For the gradient descent, we pick a learning rate of 0.01 halved every 25 epochs. Waterbirds experiments are run on a ResNet-18 with pretrained weights [17]. The model is trained with SGD (learning rate  $10^{-1}$  divided by 10 every 30 epochs) for 100 epochs with batch size 1024. Finally, we use a ResNet-50 with pretrained weights [17] trained again with SGD (learning rate  $10^{-4}$ ) for 20 epochs with batch size 512 for CelebA. For non-negative concept decomposition, we extract activations from the final ReLU layer before the classifier head to ensure non-negativity. Experiments are repeated 10 times with different random seeds; we report means and standard errors.



(a) Bias-alignment score (fraction of background pixels in most-activating patches) of the most background-dominated concept as a function of patch size and number of concepts (both classes, 5 runs per class).

(b) Waterbirds bias score against bias-alignment score. Blue crosses represent concepts, red curves are linear regressions, and green curves show the average bias-alignment score of concepts above the current bias score threshold.

Fig. 3: Waterbirds concept bias analysis.

#### 4.1 Concept-Bias Alignment

We first investigate whether the spurious attributes learned by the model appear as distinct concept directions in the non-negative concept decomposition. We hypothesize that if the model relies on a spurious attribute to predict class  $y$ , then at least one concept in  $W_y$  aligns with that attribute. To test this, we use datasets with known ground-truth spurious attributes. On CMNIST, we measure cosine similarity between each concept vector and an estimated color-bias direction following [10], treating concepts with similarity  $\geq 0.55$  as bias-aligned. On Waterbirds, we compute the average fraction of *background* pixels in each concept’s most activating patches using foreground masks; concepts with  $\geq 85\%$  background pixels are considered bias-aligned. We sweep the patch size  $s$  and number of concepts  $r$ .

**Results.** On Waterbirds (Fig. 3a), we reliably find a background-dominated concept with very low variance (below 2% standard error), confirming that a bias-aligned concept is consistently recovered. The best results are obtained with patch size around 50 and  $r \geq 10$ : patches at this scale capture enough background structure to form a coherent concept while remaining small relative to the full image, and sufficiently many concepts are needed for a dedicated bias direction to emerge. CMNIST experiments confirm these findings across multiple audit-set distributions and show that unbiased models do not produce spurious concept directions (see Appendix C). These experiments confirm that non-negative concept decomposition reliably captures the bias attribute in biased models. Balancing interpretability and decomposition quality, we set  $(r = 8, s = 6)$  for CMNIST and  $(r = 10, s = 50)$  for Waterbirds and CelebA; gradient step size is set to  $d = 2 \times 10^4$  (sensitivity analysis in Appendix C).

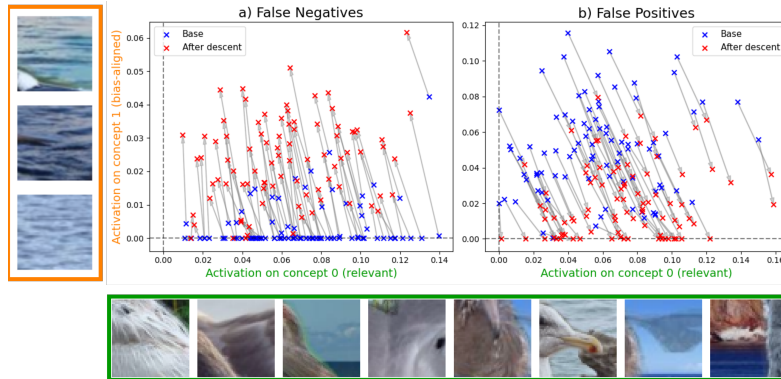


Fig. 4: Empirical validation of the gradient-concept interaction on a ResNet-18 trained on Waterbirds, to be compared with Fig. 2. Each panel scatters concept activations for the waterbird class, with the relevant concept (concept 0, bird features; green frame) on the horizontal axis and the bias-aligned concept (concept 1, sea backgrounds; orange frame) on the vertical axis. Blue crosses: base activations  $u_y(a)$ ; red crosses: activations  $u_y(a')$  after one gradient probe step (Eq. (3)); grey segments connect the two. (a) On false negatives, the bias concept is systematically added. (b) On false positives, it is systematically suppressed. The relevant concept shows a markedly weaker and less asymmetric change.

## 4.2 Bias Score Validation

Having confirmed that bias directions emerge in the NMF basis, we now test whether our scoring criterion ranks them automatically. We apply the gradient-concept interaction score from Eq. (4) to rank concepts as candidate bias directions on misclassified and distribution-shifted examples, and evaluate if the high-ranked concepts correspond to the ground-truth bias concepts from Sec. 4.1. Figure 4 first illustrates this interaction directly on Waterbirds: for a bird-focused relevant concept and a sea-focused bias concept, the gradient probe step consistently increases the bias-concept activation on false negatives and decreases it on false positives, empirically confirming the intuition of Fig. 2 on real data.

Figure 3b further demonstrates that the relation between bias score and bias alignment is strong: concepts with bias scores above 0.55 are reliably background-dominated (mean alignment above 0.8), supporting the use of this threshold for bias identification. On CMNIST, the same pattern holds for biased models, while unbiased models show no strong correlation and no concept exceeds a bias score of 0.7 (see Appendix C). Across both datasets, concepts with bias scores above 0.55 are consistently aligned with the ground-truth bias. We therefore use a bias score threshold of  $\tau = 0.55$  for the remainder of this study; this threshold can be increased to reduce the false-positive rate at the cost of sensitivity.

Table 1: Correlation between identified bias concepts and ground-truth bias labels. #bias: number of identified bias concepts;  $f_{\text{bias}} / f_{\text{other}}$ : proportion of statistically significant pairs ( $\chi^2$  test) for bias concepts vs. other concept–bias pairs;  $\bar{\Phi}_{\text{bias}} / \bar{\Phi}_{\text{other}}$ : mean absolute MCC of significant pairs;  $p$ : one-sided Mann–Whitney  $U$ -test. CMNIST results are reported in Appendix C.

Train dataset	Audit bias	#bias	$f_{\text{bias}}$	$f_{\text{other}}$	$\bar{\Phi}_{\text{bias}}$	$\bar{\Phi}_{\text{other}}$	$p$
Waterbirds	biased	4.3±0.6	0.977	0.790	0.389±0.175	0.154±0.123	0.001
CelebA	biased	3.2±0.9	0.969	0.768	0.163±0.076	0.124±0.083	0.010

### 4.3 Correlation with Ground-Truth Bias

To quantitatively validate that the identified concepts correspond to the ground-truth bias, we measure the correlation between concept activations and ground-truth bias labels on the test sets. For each concept in the merged concept bank and each bias attribute, we compute the absolute Matthews Correlation Coefficient (MCC) between the binarized concept activation  $I_{y,k}$  and the binary bias label.<sup>3</sup> The absolute MCC values of identified bias concepts with their corresponding bias attribute are then compared against those of all other concept–bias pairs. We report the proportion of statistically significant pairs under a  $\chi^2$  independence test and conduct a one-sided Mann–Whitney  $U$ -test to assess whether identified bias concepts exhibit significantly higher correlation than other pairs. Results are reported in Table 1.

On Waterbirds, activations of identified bias concepts are highly correlated with the background attribute ( $f_{\text{bias}} = .977$ ,  $p < 0.001$ ), consistently exceeding other concept–bias pairs. On CelebA, identified concepts ( $f_{\text{bias}} = .969$ ) are statistically associated with bias but MCC values remain low, and no concept exceeds an MCC of 0.4. The Mann–Whitney test still shows a significant advantage over other concepts ( $p = 0.010$ ), indicating that the identified concepts are relatively more bias-aligned than the remaining ones. The low absolute values suggest that the single-scale NMF decomposition struggles to capture a high-level attribute such as gender, possibly compounded by the severe class imbalance (blond:not blond  $\approx 1:6.5$ ). CMNIST results (Appendix C) confirm strong correlation for biased models and no correlation for unbiased models.

### 4.4 Inference-Time Bias Mitigation

To test whether identified bias directions are actionable, we apply a simple inference-time concept suppression inspired by projection-based methods such as P-CIArC [2]. Adapting such projection-style interventions to NMF-based

<sup>3</sup> We use the absolute value because the sign of a concept direction under NMF is determined by the factorization and is not semantically meaningful: a concept that anti-correlates with the binary bias label encodes the same bias information as one that correlates with it.

Table 2: Effect of suppressing the identified bias concepts at inference time. Symbol “–” under *Audit bias* denotes the base frozen model without mitigation. CMNIST results are reported in [Appendix C](#).

Train dataset	Audit bias	Accuracy	Worst-class acc	Worst-group acc
Waterbirds	–	82.1±0.3	69.1±0.8	45.9±1.5
	biased	<b>86.4±1.7</b>	<b>77.6±2.5</b>	<b>63.8±4.0</b>
	ablation	84.8±1.4	66.8±5.2	43.7±8.5
CelebA	–	<b>95.3±0.1</b>	81.3±1.3	43.4±2.5
	biased	94.9±0.2	<b>86.4±1.4</b>	<b>53.8±4.1</b>
	ablation	94.1±1.2	79.3±11.8	45.4±17.9

concept banks raises two practical considerations. First, our bias-identification stage produces class-conditional concept banks, meaning each decomposition is tailored to one class and is therefore not directly suitable for full downstream prediction. We address this by merging class-wise concept banks into a model-wide bank  $W_{\text{merged}}$  and clustering highly similar concepts (cosine similarity above 0.95). Second, projection-based mitigation requires a neutral value to which the suppressed concept is set. Since NMF matrices are naturally sparse, we use 0 as a neutral value for each concept, avoiding the need to estimate concept-specific baselines from additional data. For a test input  $x$ , we compute its representation  $a = g(x)$  and concept coefficients  $u_{\text{merged}}(a)$ , and remove the components along bias concepts  $B$  before rescaling the result to the original activation norm and performing the classification:

$$a_{\text{sup}} = a - \sum_{k \in B} u_{\text{merged}}(a)_k w_{\text{merged},k}, \quad a_{\text{rsc}} = a_{\text{sup}} \frac{\|a\|_2}{\|a_{\text{sup}}\|_2} \quad \tilde{f}(x) = h(a_{\text{rsc}}),$$

where  $a_{\text{sup}}$  is the bias-suppressed representation and  $a_{\text{rsc}}$  is its rescaling to the original activation norm  $\|a\|_2$ , which compensates for the magnitude lost during suppression. Based on the previous experiments, concepts with bias scores above 0.55 can reliably be treated as bias-aligned. Accordingly, we define  $B = \bigcup_y \{k : S_{y,k} > \tau\}$  with  $\tau = 0.55$ . We test this approach on all three datasets. For CMNIST and Waterbirds, we choose hyperparameters based on the previous studies; for CelebA, we reuse the Waterbirds settings because it uses a similar architecture, allowing us to test whether the hyperparameters transfer across datasets. To ensure that effects are linked to our identification, we also run an ablation that randomly suppresses the same number of concepts in the merged bank for each model. These experiments are reported as *ablation* in the results.

**Results.** [Table 2](#) reports overall, worst-class, and worst-group accuracy. On Waterbirds, suppression improves overall accuracy by +4.3, worst-class accuracy by +8.5, and worst-group accuracy by +17.9, clearly outperforming random ablation despite class imbalance. On CelebA, worst-class and worst-group accuracy

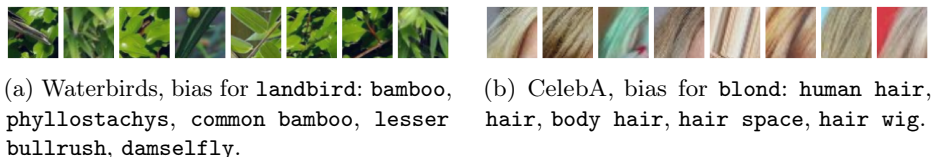


Fig. 5: Top-activating patches of representative bias concepts identified by our method, with top-5 generated labels from NOVIC [1]. **(a)** Landbird bias concept (vegetation background). **(b)** Blond-hair bias concept (hair shape/length). Extended galleries for all concepts are in [Appendix E](#).

improve by +5.1 and +10.4 with only a small drop in overall accuracy and no dataset-specific tuning (Waterbirds settings were reused); the random-suppression ablation ( $45.4 \pm 17.9$  worst-group) is consistent with the effect being concept-specific rather than a pure capacity reduction. Despite the weak correlation with the annotated gender attribute ([Sec. 4.3](#)), suppression improves fairness metrics—the qualitative analysis in [Sec. 4.5](#) suggests the identified directions capture decision-relevant spurious information that only partially coincides with gender. On CMNIST, suppression yields mixed results ([Appendix C](#)). A detailed comparison with supervised baselines (JTT, Group DRO, DFR) is provided in [Appendix D](#).

#### 4.5 Bias Concept Interpretation

Each identified concept comes with its most-activating audit-set patches, which can be inspected qualitatively. To reduce reliance on bias-prone manual inspection, for each bias concept we generate the 10 labels most associated with its top-100 patches (weighted by rank) using NOVIC [1], an open-vocabulary image classifier; we treat the labels as a qualitative summary rather than ground truth. Representative patch galleries are shown in [Fig. 5](#); per-dataset extended galleries and discussion are in [Appendix E](#) ([Figs. 8 to 10](#)).

On Waterbirds ([Fig. 5a](#)), the top labels align cleanly with the ground-truth bias: plant-related terms (**bamboo**, **phyllostachys**, **rain tree**) for **landbird** and marine terms (**wave**, **blue whale**, **sea**) for **waterbird**. On CelebA ([Fig. 5b](#)), the top labels cluster around hair (**hair**, **human hair**, **hair space**, ...); qualitative inspection of the patches suggests the direction captures hair *shape and length* rather than color. This is consistent with the low MCC against the annotated gender attribute ([Sec. 4.3](#)): gender on CelebA is a spatially distributed cue that a patch-based decomposition cannot encode as a single direction, but localized proxies such as hair length, eyeshadow, or lips carry partial gender information, which explains why suppressing these concepts still improves worst-group performance ([Sec. 4.4](#)).

## 5 Related Work

Shortcut learning is a pervasive failure mode where models rely on spurious correlations that break under distribution shift [12,43]. **Bias-aware mitigation.** When the spurious attribute is known, methods reduce reliance via regularization [5], adversarial training [22], representation erasure [6], concept-supervised probing [9,21], or directional interventions [2,10]; post-training approaches include saliency-masked fine-tuning [4] and subspace projection [20]. **Bias discovery without bias labels.** Related work upweights high-loss examples [33,30] or discovers latent failure slices via environment inference and pseudo-labeling [38,45,44,16,34], though these generally do not yield a class-conditional direction that can be ranked and targeted [26,40,14]. Unsupervised concept discovery extracts interpretable directions from activations [15,11,28] and has been applied to robustness [3], concept ranking [25], and post-hoc bias proposal via explanation maps or vision-language models [8,23,35], though often without providing a concrete direction for intervention [7,36,37]. **Post-hoc mitigation with frozen models.** Existing approaches reduce spurious reliance through last-layer retraining on a reweighted split [24], sub-network extraction [27], activation erasure [18], data pruning [32], or inference-time debiasing [13,19], but still require optimization, group labels, or model edits. In contrast, we identify class-conditional spurious concept directions in a frozen model without bias labels, providing interpretable handles for downstream intervention without updating parameters. An extended discussion is provided in [Appendix F](#).

## 6 Conclusion

We presented a label-free, post-hoc method for identifying and mitigating spurious concepts in frozen vision classifiers that requires only standard class labels from a held-out audit set. Non-negative concept decomposition reliably recovers spurious attribute directions across hyperparameters on CMNIST and Waterbirds, and our gradient-concept interaction score ranks these directions without any bias annotations. Suppressing the top-ranked concepts at inference time improves worst-group accuracy by up to 17.9 percentage points on Waterbirds and 10.4 on CelebA, with no retraining or parameter updates. On CelebA, the identified concepts improve fairness metrics despite correlating only weakly with the annotated gender attribute, indicating that the method captures decision-relevant spurious information that extends beyond pre-defined bias categories. Together, these results demonstrate that unsupervised concept discovery combined with gradient probing can serve both as an interpretable auditing tool and as a practical debiasing handle for deployed models.

*Limitations and future work.* Our approach assumes access to a post-hoc bias-audit set and is most informative when enough misclassified examples are available; in low-error regimes, targeted stress tests or distribution shifts may be needed to surface failures. The single-scale patch decomposition limits the ability to

capture bias attributes that are spatially distributed (as observed on CelebA); multi-scale decompositions [15] may address this. Bias information may also be entangled across multiple concept directions, leading to collateral suppression of task-relevant features; recursive or hierarchical decompositions could help disentangle these and are a promising direction for future work.

## References

1. Allgeuer, P., Ahrens, K., Wermter, S.: Unconstrained open vocabulary image classification: Zero-shot transfer from text to image via clip inversion. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8217–8228. IEEE (2025)
2. Anders, C.J., Weber, L., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *Information Fusion* **77**, 261–295 (2022)
3. Arefin, M.R., Zhang, Y., Baratin, A., Locatello, F., Rish, I., Liu, D., Kawaguchi, K.: Unsupervised Concept Discovery Mitigates Spurious Correlations. In: Proceedings of the 41st International Conference on Machine Learning. pp. 1672–1688. PMLR (2024)
4. Asgari, S., Khani, A., Khani, F., Gholami, A., Tran, L., Mahdavi Amiri, A., Hamarneh, G.: MaskTune: Mitigating Spurious Correlations by Forcing to Explore. *Advances in Neural Information Processing Systems* **35**, 23284–23296 (2022)
5. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning De-biased Representations with Biased Representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 528–539. PMLR (2020)
6. Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., Biderman, S.: LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems* **36**, 66044–66063 (2023)
7. Bhusal, D., Clifford, M., Rampazzi, S., Rastogi, N.: FACE: Faithful Automatic Concept Extraction. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
8. Chakraborty, R., Sletten, A., Kampffmeyer, M.C.: ExMap: Leveraging Explainability Heatmaps for Unsupervised Group Robustness to Spurious Correlations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12017–12026 (2024)
9. Correa, R., Pahwa, K., Patel, B., Vachon, C.M., Gichoya, J.W., Banerjee, I.: Efficient adversarial debiasing with concept activation vector — Medical image case-studies. *Journal of Biomedical Informatics* **149**, 104548 (2024)
10. Dreyer, M., Pahde, F., Anders, C.J., Samek, W., Lapuschkin, S.: From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(19), 21046–21054 (2024)
11. Fel, T., Picard, A., Béthune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., Serre, T.: CRAFT: Concept Recursive Activation FacTorization for Explainability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2711–2721 (2023)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)

13. Gerych, W., Zhang, H., Hamidieh, K., Pan, E., Sharma, M., Hartvigsen, T., Ghassemi, M.: BendVLM: Test-Time Debiasing of Vision-Language Embeddings. *Advances in Neural Information Processing Systems* **37**, 62481–62502 (2024)
14. Ghaznavi, M., Asadollahzadeh, H., Noohdani, F.H., Tabar, S.V., Hasani, H., Alvanagh, T.A., Rohban, M.H., Baghshah, M.S.: Exploiting What Trained Models Learn for Making Them Robust to Spurious Correlations without Group Annotations. In: *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions* (2025)
15. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards Automatic Concept-based Explanations. In: *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019)
16. Ghosh, S., Syed, R., Wang, C., Choudhary, V., Li, B., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. In: *Findings of the Association for Computational Linguistics: ACL 2025*. pp. 22935–22970. Association for Computational Linguistics (2025)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
18. He, Q., Xu, K., Yao, A.: EvA: Erasing Spurious Correlations with Activations. In: *The Thirteenth International Conference on Learning Representations* (2024)
19. Hirota, Y., Chen, M.H., Wang, C.Y., Nakashima, Y., Wang, Y.C.F., Hachiuma, R.: SANER: Annotation-free Societal Attribute Neutralizer for Debiasing CLIP. In: *The Thirteenth International Conference on Learning Representations* (2024)
20. Holstege, F., Wouters, B., Giersbergen, N.V., Diks, C.: Removing Spurious Concepts from Neural Network Representations via Joint Subspace Estimation. In: *Proceedings of the 41st International Conference on Machine Learning*. pp. 18568–18610. PMLR (2024)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: *Proceedings of the 35th International Conference on Machine Learning*. pp. 2668–2677. PMLR (2018)
22. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning Not to Learn: Training Deep Neural Networks With Biased Data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9012–9020 (2019)
23. Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., Shin, J.: Discovering and Mitigating Visual Biases through Keyword Explanation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11082–11092 (2024)
24. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. In: *The Eleventh International Conference on Learning Representations (ICLR)* (2023)
25. Kowal, M., Dave, A., Ambrus, R., Gaidon, A., Derpanis, K.G., Tokmakov, P.: Understanding Video Transformers via Universal Concept Discovery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10946–10956 (2024)
26. LaBonte, T., Hill, J.C., Zhang, X., Muthukumar, V., Kumar, A.: The Group Robustness is in the Details: Revisiting Finetuning under Spurious Correlations. *Advances in Neural Information Processing Systems* **37**, 121598–121629 (2024)
27. Le, P.Q., Schlötterer, J., Seifert, C.: Out of Spuriousity: Improving Robustness to Spurious Correlations without Group Annotations. *Transactions on Machine Learning Research* (2024)

28. Lee, J.H., Mikriukov, G., Schwalbe, G., Wermter, S., Wolter, D.: Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go? In: *Computer Vision – ECCV 2024 Workshops*. pp. 266–287. Springer Nature Switzerland (2024)
29. Lee, J., Lee, J., Jung, S., Choo, J.: Improving Evaluation of Debiasing in Image Classification. arXiv preprint arXiv:2206.03680 (2023)
30. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. pp. 6781–6792 (2021)
31. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015)
32. Mulchandani, V., Kim, J.E.: Severing Spurious Correlations with Data Pruning. In: *The Thirteenth International Conference on Learning Representations* (2024)
33. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 33, pp. 20673–20684 (2020)
34. Olesen, V., Weng, N., Feragen, A., Petersen, E.: Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods. In: *Ethics and Fairness in Medical Imaging*. pp. 3–13. Springer Nature Switzerland (2025)
35. Paduraru, C.D., Barbalau, A., Filipescu, R., Nicolicioiu, A.L., Burceanu, E.: Concept-Drift: Uncovering Biases through the Lens of Foundation Models. In: *Interpretable AI: Past, Present and Future* (2024)
36. Pahde, F., Dreyer, M., Weckbecker, M., Weber, L., Anders, C.J., Wiegand, T., Samek, W., Lapuschkin, S.: Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Directional Divergence. In: *The Thirteenth International Conference on Learning Representations* (2024)
37. Panousis, K.P., Ienco, D., Marcos, D.: Coarse-to-Fine Concept Bottleneck Models. *Advances in Neural Information Processing Systems* **37**, 105171–105199 (2024)
38. Pezeshki, M., Bouchacourt, D., Ibrahim, M., Ballas, N., Vincent, P., Lopez-Paz, D.: Discovering Environments with XRM. In: *Proceedings of the 41st International Conference on Machine Learning*. pp. 40551–40569. PMLR (2024)
39. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks. In: *International Conference on Learning Representations* (2019)
40. Tsigotis, C., Monteiro, J., Rodriguez, P., Vazquez, D., Courville, A.C.: Group Robust Classification Without Any Group Information. *Advances in Neural Information Processing Systems* **36**, 56553–56575 (2023)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. 2010-001, California Institute of Technology (2011)
42. Wu, H., Bezold, G., Gunther, M., Boulton, T., King, M.C., Bowyer, K.W.: Consistency and Accuracy of CelebA Attribute Values . In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. pp. 3258–3266. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023)
43. Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Porter, C.W., Coffee, M., Zeng, Z., Li, S., Wang, Z., Huang, T.H.K., Rehg, J.M., Kautz, H., Zhang, A.: The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning. *Transactions on Machine Learning Research* (2025)

44. Zare, S., Nguyen, H.V.: Frustratingly Easy Environment Discovery for Invariant Learning. *Computer Sciences & Mathematics Forum* **9**(1), 2 (2024)
45. Zhang, Z., Feng, M., Li, Z., Xu, C.: Discover and Mitigate Multiple Biased Subgroups in Image Classifiers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10906–10915 (2024)
46. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(6), 1452–1464 (2018)

## A Algorithm

The bias identification procedure is summarized in Alg. 1.

---

**Algorithm 1** Post-hoc bias identification.

---

**Input:** Frozen classifier  $f = h \circ g$ , bias-audit set  $\mathcal{D}$ , patch size  $s$ , #concepts  $r$ , step size  $d$

**Output:** Per-class ranked concept sets  $\mathcal{R}$

- 1: Initialize per-class ranked concept sets  $\mathcal{R} \leftarrow \emptyset$
- 2: **for** class  $y \in \{1, \dots, C\}$  **do**
- 3: Collect  $\mathcal{X}_y = \{x_i : \hat{y}(x_i) = y\}$  and patches  $\mathcal{P}_y = \{\pi_s(x) : x \in \mathcal{X}_y\}$
- 4: Compute  $A_y = g(\mathcal{P}_y)$  and NMF concepts  $W_y$  via Eq. (1)
- 5: Compute  $\text{FN}_y$  and  $\text{FP}_y$  on  $\mathcal{D}$
- 6: **for**  $k = 1, \dots, r$  **do**
- 7: **for all**  $x \in \text{FN}_y \cup \text{FP}_y$  **do**
- 8: Form  $a = g(x)$  and  $a' = a - d \nabla_a L(h(a), y_i)$
- 9: Solve NNLS in Eq. (2) to obtain  $u_{y,k}(a)$  and  $u_{y,k}(a')$ , then  $I_{y,k}(a) = \mathbf{1}[u_{y,k}(a) > 0]$
- 10: **end for**
- 11: Compute  $S_{y,k}$  via Eq. (4)
- 12: **end for**
- 13: Add class- $y$  ranking to  $\mathcal{R}$ : concepts  $w_{y,k}$  (columns of  $W_y$ ) ranked by  $S_{y,k}$  with their top-activating patches
- 14: **end for**
- 15: **return**  $\mathcal{R}$

---

## B Dataset Construction Details

*CMNIST*. In the biased training set each digit label is assigned a “bias color” that matches the class label with probability 0.95. The *unbiased* variant makes all color-label combinations equally likely, while the *shifted-bias* variant uses a different color-to-label assignment. A key advantage of CMNIST is the ability to modulate the bias correlation coefficient and generate biased versions of any sample by modifying its color, allowing for controlled experiments with known ground-truth bias directions.

*Waterbirds*. Overlays birds from the Caltech-UCSD Birds (CUB) dataset [41] on Places [46] backgrounds. The water/land background is correlated with the label with probability 0.95 in training and 0.5 in testing.

*CelebA*. CelebA [31] contains aligned celebrity face images with 40 binary attribute annotations. In the spurious-correlation setting it is used as a binary blond vs. not-blond prediction task with gender as the spurious attribute [42]; a biased training split is formed by subsampling so that target and spurious attribute are correlated at 0.95. Evaluation reports group-wise and worst-group

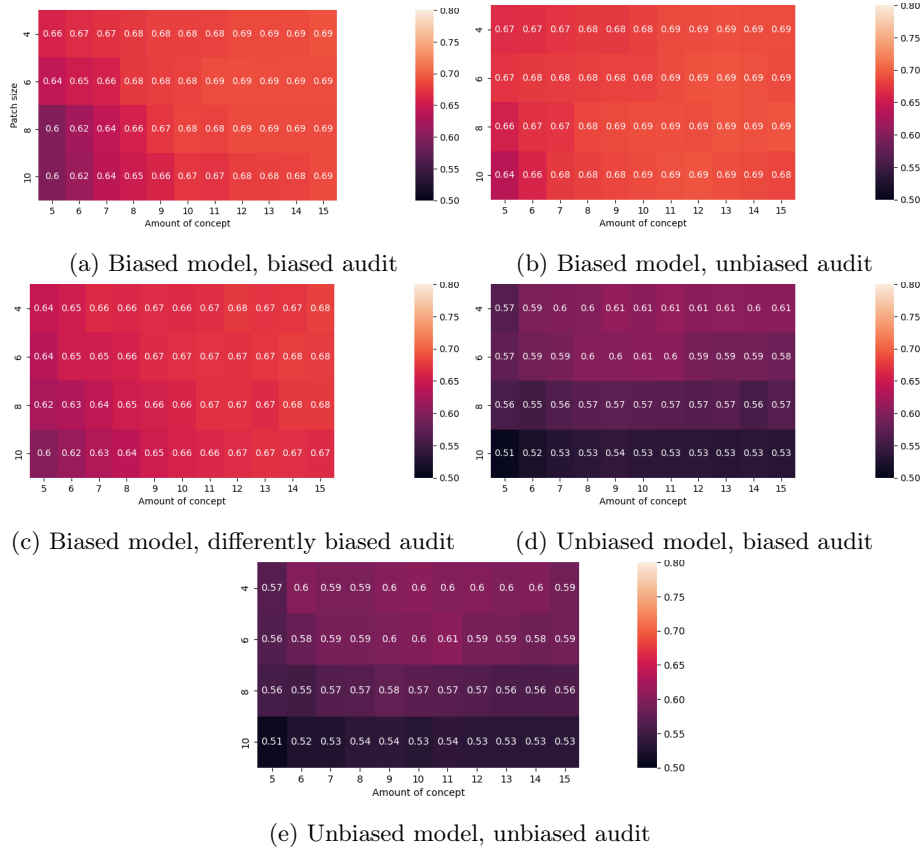


Fig. 6: CMNIST: cosine similarity between the estimated color-bias direction and the most aligned concept vector as a function of patch size and number of concepts. Values are computed over 10 runs on all 10 classes.

accuracy over the four (blond/not-blond) $\times$ (female/male) combinations. The test split keeps a strong imbalance (e.g., only 180 blond men against 9767 not-blond women), which makes overall accuracy an unreliable standalone metric.

## C CMNIST Results

This section presents the full CMNIST results that complement the Waterbirds and CelebA results in the main text. These results were moved to the appendix because the simple and artificial nature of CMNIST makes it better suited as a controlled toy setting than as a headline benchmark.

### C.1 Concept–Bias Alignment

We sweep the patch size  $s$  and number of concepts  $r$ . On CMNIST (Figs. 6a to 6c), biased models on average always yield a strongly aligned concept across hyperparameters (minimum cosine similarity 0.6 with low variance below 0.03), and moderate settings ( $r \geq 9$ ,  $s \leq 8$ ) work best. While too few concepts hurt the emergence of a bias-aligned concept, too many concepts do not substantially improve the decomposition, plausibly because NMF then distributes bias information across multiple vectors rather than concentrating it in a single direction. Smaller patch sizes are generally preferred, likely because they focus patches on local evidence. The nature of the audit dataset has little influence on alignment overall, with effects mainly visible at low numbers of concepts. This suggests that non-negative concept decomposition can be used across different audit-set distributions. We note, however, that samples containing the bias attribute must still be present, even if the attribute is not correlated with the target label; otherwise, the decomposition cannot build patches that expose the bias attribute.

When the concept bank is built from an unbiased dataset (Figs. 6d and 6e), concepts remain weakly aligned with the bias attribute. This confirms that the unbiased model does not rely on bias, as no strong bias direction emerges in its representation layer. It also suggests that non-negative concept decomposition is not prone to hallucinating spurious concepts when none are present.

*Gradient step size.* Sweeping  $d$  on this setup, we find that the bias score becomes more reliable as the step size grows and plateaus without degrading at larger values; we therefore fix  $d = 2 \times 10^4$  for all remaining experiments.

### C.2 Bias Score Validation

On CMNIST, we examine the relation between bias score and bias alignment (Fig. 7). On biased models (Figs. 7a to 7c), alignment is widely spread at low bias scores, but this spread narrows as the score increases. Around a bias score of 0.7, concepts consistently exceed 0.4 in alignment, and the average alignment curve (green) confirms that concepts above this threshold have mean alignment above our 0.55 criterion. This pattern holds across audit sets, although with weaker confidence when the audit set shares the same bias as the model, since informative error examples (false negatives without the spurious cue and false positives with the spurious cue) become rarer.

On unbiased models (Figs. 7d and 7e), there is no strong correlation between the bias score and bias alignment, as expected: the model has not learned a spurious bias, so errors stem from random noise rather than systematic spurious attributes. No concept exceeds a bias score of 0.7, reducing the risk of false bias identification. This holds even when the audit dataset itself is biased, indicating that the method is resilient to distributional properties of the audit set.

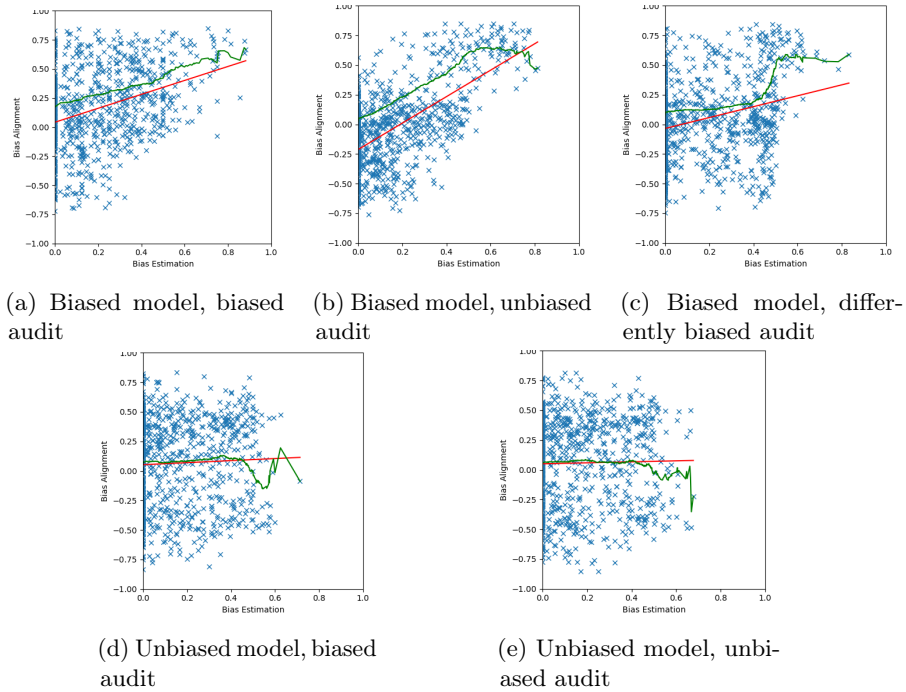


Fig. 7: CMNIST: bias score against bias-alignment score. Blue crosses represent concepts, red curves are linear regressions, and green curves show the average bias-alignment score of concepts above the current bias score threshold.

### C.3 Correlation with Ground-Truth Bias

Table 3 reports the correlation between identified bias concepts and ground-truth bias labels for CMNIST. On biased models, identified concepts show significantly higher MCC with the color bias than other concept–bias pairs ( $p < 0.001$ ). On unbiased models, no such pattern holds, confirming that the method does not hallucinate bias where none was learned.

### C.4 Inference-Time Bias Mitigation

Table 4 reports the effect of suppressing identified bias concepts on CMNIST. On biased CMNIST, suppression does not consistently improve metrics and in some configurations decreases worst-group accuracy, particularly with the unbiased audit set, which is unexpected given this configuration’s strong performance in the bias score validation. The ablation rows confirm this is not an artefact of removing concept capacity in general: random suppression degrades metrics further than targeted suppression in all conditions, so the identified directions remain preferentially bias-aligned even when their removal does not translate into worst-group gains. A plausible explanation is that on CMNIST the color attribute

Table 3: CMNIST: correlation between identified bias concepts and ground-truth bias labels (cf. Table 1 for Waterbirds and CelebA).

Train dataset	Audit bias	#bias	$f_{\text{bias}}$	$f_{\text{other}}$	$\bar{\Phi}_{\text{bias}}$	$\bar{\Phi}_{\text{other}}$	$p$
Biased CMNIST	biased	$4.9 \pm 1.8$	0.959	0.919	$0.418 \pm 0.184$	$0.261 \pm 0.204$	0.001
	unbiased	$5.7 \pm 1.0$	1.000	0.880	$0.399 \pm 0.154$	$0.156 \pm 0.114$	0.001
	shifted bias	$2.7 \pm 1.1$	0.963	0.856	$0.479 \pm 0.260$	$0.140 \pm 0.136$	0.001
Unbiased CMNIST	biased	$1.4 \pm 1.1$	0.027	0.004	$0.027 \pm 0.004$	$0.046 \pm 0.027$	0.857
	unbiased	$2.1 \pm 1.7$	0.381	0.305	$0.038 \pm 0.017$	$0.043 \pm 0.025$	0.724

Table 4: CMNIST: effect of suppressing identified bias concepts at inference time (cf. Table 2 for Waterbirds and CelebA). “-” under *Audit bias* denotes the base frozen model without mitigation.

Train dataset	Audit bias	Accuracy	Worst-class acc	Worst-group acc
Biased CMNIST	-	$76.7 \pm 0.5$	$64.6 \pm 1.8$	$31.2 \pm 5.1$
	biased	$76.4 \pm 2.7$	$61.9 \pm 11.2$	$23.0 \pm 12.5$
	biased ablation	$72.2 \pm 3.1$	$54.6 \pm 11.3$	$11.1 \pm 9.5$
	unbiased	$74.1 \pm 2.4$	$52.0 \pm 5.5$	$15.0 \pm 12.8$
	unbiased ablation	$67.8 \pm 4.9$	$36.5 \pm 12.5$	$4.6 \pm 7.0$
	shifted bias	$76.9 \pm 0.7$	$64.2 \pm 2.4$	$30.4 \pm 7.1$
	shifted ablation	$74.1 \pm 2.2$	$55.3 \pm 8.8$	$19.1 \pm 10.4$
Unbiased CMNIST	-	$95.8 \pm 0.1$	$93.3 \pm 0.5$	$88.3 \pm 1.3$
	biased	$94.8 \pm 1.6$	$88.8 \pm 7.1$	$82.4 \pm 8.1$
	biased ablation	$95.0 \pm 0.8$	$87.7 \pm 6.7$	$79.3 \pm 10.4$
	unbiased	$94.0 \pm 2.7$	$81.6 \pm 20.0$	$75.5 \pm 20.2$
	unbiased ablation	$95.0 \pm 0.7$	$89.4 \pm 5.1$	$81.5 \pm 5.8$

is almost linearly separable and tightly entangled with digit shape at our chosen rank, so removing the color-aligned direction also removes class-discriminative signal; this is consistent with the lower best-case worst-group accuracy observed relative to Waterbirds. Investigating multi-scale or recursive decompositions to disentangle these directions is left to future work. On unbiased CMNIST, suppressing identified concepts expectedly decreases all metrics, as the model did not rely on a spurious direction to begin with.

## D Comparison with Supervised Baselines

Table 5 contextualizes our inference-time mitigation with supervised methods that retrain the model using group annotations. All reference numbers are from [24] (ResNet-50); our Waterbirds model is a ResNet-18, our CelebA model is a ResNet-50.

Table 5: Worst-group accuracy (%) of debiasing methods under different supervision and retraining requirements. ERM = empirical risk minimization (standard training). <sup>†</sup>ResNet-50 numbers from [24]; our Waterbirds model uses a ResNet-18.

Method	Bias labels	Retraining	Waterbirds	CelebA
ERM (ours)	none	none	45.9±1.5	43.4±2.5
<b>Ours</b>	<b>none</b>	<b>none</b>	<b>63.8±4.0</b>	<b>53.8±4.1</b>
JTT <sup>†</sup> [30]	val	full	86.7	81.1
Group DRO <sup>†</sup> [39]	train+val	full	91.4	88.9
DFR <sup>†</sup> [24]	val	last layer	92.9±0.2	88.3±1.1

Our method is the only one that requires neither bias labels nor parameter updates, operating entirely at inference time on a frozen model. The gap to supervised baselines reflects this strict constraint; methods that retrain with group-balanced data achieve higher absolute worst-group accuracy but presuppose label collection and model access that may not be available for deployed systems. Our approach provides a complementary, zero-cost diagnostic and mitigation handle that can be applied before deciding whether more invasive retraining is warranted.

## E Bias Concept Interpretation

This appendix presents the full patch galleries and automatically generated labels for all bias concepts identified by our method on CMNIST, Waterbirds and CelebA. For each identified concept, we retrieve its top-100 most-activating patches from the bias-audit set and generate the 10 labels most associated with these patches (weighted by activation rank) using NOVIC [1], an open-vocabulary image classifier. Each subfigure shows a representative subset of the top-activating patches alongside the generated labels. Common concepts appear consistently across random seeds, while uncommon concepts emerge in only a fraction of seeds. Representative examples from both categories are included to illustrate the range of identified bias directions.

*CMNIST.* CMNIST bias concepts (Fig. 8) align closely with ground-truth color bias. Regardless of the audit nature, as long as the model is biased, we observe that the ground-truth color bias label of each class appears systematically in the top-5 most appearing labels for bias concept of that class. One notable exception is class nine with the bias *grey* for which no bias concept was ever recorded in any of the 10 experiments nor audit type. This seems to show that the *grey* number is treated as a default by the model and therefore not strongly encoded. Interestingly, when the audit had unaligned bias, label three biased with *yellow* in the model and *red* in the audit also never displayed any bias concept.

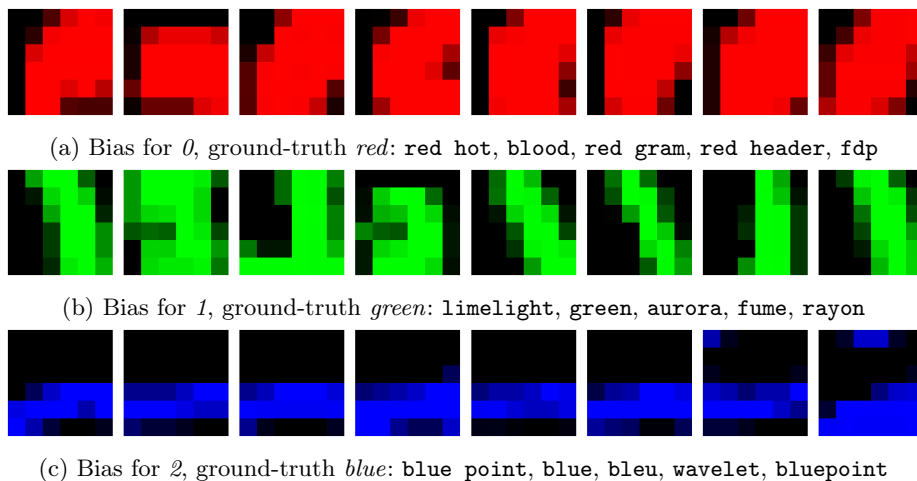


Fig. 8: Top-activating patches of bias concepts for CMNIST with corresponding top-5 generated labels. Computed using biased model and similarly biased audit.

*Waterbirds*. The Waterbirds bias concepts (Fig. 9) align closely with the ground-truth background bias. The two common concepts capture the dominant spurious cues: dense vegetation for *landbird* (a) and open water for *seabird* (c). The uncommon concepts reveal subtler patterns: autumnal foliage (b) and coastal boundaries (d). The foliage concept appears in about half of the seeds and occasionally includes a bird wing whose camouflage pattern resembles a leaf texture. The coastal concept appears in roughly one-fifth of seeds and often confuses the bird–sea boundary for a coastline. These results confirm that the NMF decomposition reliably separates background-related directions from bird-related ones.

*CelebA*. The CelebA bias concepts (Fig. 10) are more diverse and harder to interpret, reflecting the distributed nature of the gender cue in this dataset. Unlike Waterbirds, where the bias is spatially localized in the background, gender-related information on CelebA is spread across multiple facial regions. The identified concepts span background color and darkness (a), lower-face features such as lips and facial hair (b), hair texture and color for **blond** (c) and **not blond** (d), eye-region features including make-up cues (e), and forehead and hairline structure (f). While no single concept correlates strongly with the binary gender attribute, collectively these directions capture decision-relevant spurious information: suppressing them improves worst-group accuracy by 10.4 percentage points (Table 2). This pattern is consistent with the observation in Sec. 4.3 that gender on CelebA cannot be represented as a single concept direction under a patch-based decomposition.



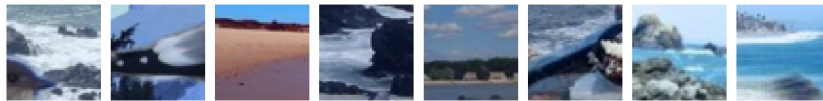
(a) Bias for *landbird* (common): bamboo, fishpole bamboo, phyllostachys, web, common bamboo, lesser bullrush, genus griselinia, damselfly, cue, yellow berry



(b) Bias for *landbird* (uncommon): fall, rain tree, birds eye maple, web, silverbush, black oak, blossom, brown oak, canopy, snow. Appears in about half of the seeds; the fifth patch is a bird wing whose camouflage fits the “leaf” pattern.



(c) Bias for *seabird* (common): wavelet, wave, gray whale, blue, blue whale, sea, rippling, harbor seal, oceanic abyss, pilot whale

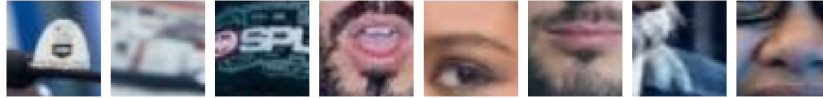


(d) Bias for *seabird* (uncommon): coast, sea coast, wave, hill, mountain, beach, coast boykinia, gray whale, breakwater, tidal wave. Appears in about 1/5th of the seeds; the bird-sea boundary is often confused for a coast.

Fig. 9: Top-activating patches of common and uncommon bias concepts for Waterbirds with corresponding top-10 generated labels.



(a) Bias for *not blond*: black, sky, maroon, blue sky, gray, blue, indigo, brown ash, black ash, dark chocolate. This uncommon concept shows the model decision is influenced by background color; it correlates fairly well with the *men* bias (MCC 0.318) and matches societal stereotypes.



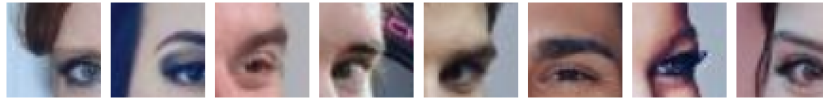
(b) Bias for *not blond*: lip, eyebrow, lip rouge, lip gloss, mouth, eyelash, brow, facial hair, sideburn, nose. Appears with variation fairly commonly and illustrates the distributed nature of the gender concept.



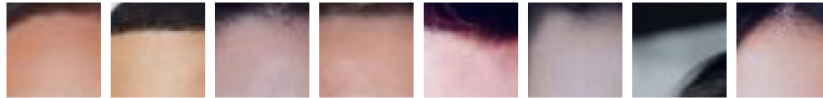
(c) Bias for *blond*: human hair, hair, body hair, hair space, hair wig, brown ash, false hair, marble wood, fiber, golden thread. A common concept featuring blond hair; the presence of long hair suggests the direction may capture shape rather than color, illustrating that color and shape information co-locate on the same patches.



(d) Bias for *not blond*: hair, human hair, hair space, black ash, fiber, black, hair wig, nebula, raven, fur. Features only dark hair; the bias-mitigation experiments indicate this concept does encode bias although the exact aspect is ambiguous.



(e) Bias for *not blond*: eyebrow, eyelash, brow, make up, lash, human eye, eye, sideburn, eyeshadow, human nose. Concentrates on the eyes; plausibly bias-aligned since eye make-up is correlated with gender in CelebA.



(f) Bias for *not blond*: hair, forehead, sideburn, eyebrow, brow, human hair, hair space, body hair, eyelash, hair wig. Features foreheads; derived from patches classified as *not blond* it contains mostly dark hair and captures hairline/forehead cues that can vary by gender.

Fig. 10: Top-activating patches of common and uncommon bias concepts for CelebA with corresponding top-10 generated labels.

## F Extended Related Work

Reliance on spurious correlations is a common failure mode in vision, where predictors perform well on the training distribution but fail to generalize when the underlying correlations differ at deployment [12]. Surveys show such shortcuts persist across model families and datasets, and are addressed with methods that vary in supervision and access to the training pipeline [43]. Prior work broadly either (i) mitigates bias given known spurious or group information, (ii) improves robustness without explicit bias supervision during training, or (iii) discovers spurious cues post-hoc when the bias is unknown.

**Bias-aware Mitigation.** When spurious attributes or groups are available, approaches reduce reliance via regularization [5], adversarial training [22], or representation-space erasure [6]. Other methods use concept supervision to probe where a bias is encoded and how it affects predictions [9,21], or intervene along a specified artifact direction using augmentation, projection, or gradient penalties [2,10]. While effective with correct bias signals, these methods assume the spurious attribute is known and typically require training-time access; in contrast, we operate post-hoc on a frozen model without bias labels. MaskTune reduces shortcut reliance via post-training fine-tuning on saliency-masked inputs [4], and subspace-projection methods such as JSE remove specified spurious concepts [20], but both still rely on optimization and/or spurious supervision to define what to suppress.

**Bias Discovery Without Bias Labels.** A related line of work discovers latent groups or failure slices without subgroup labels, aiming to explain performance gaps and support robust learning. Loss-based schemes upweight examples misclassified by an auxiliary model [33,30], and environment-discovery or subgroup-identification methods recover pseudo-groups from model behavior or representations [38,45,44]; slice-discovery methods generate pseudo-labels for downstream mitigation [16,34]. Group-free evaluation and model selection are also non-trivial [26], motivating proxy validation and pseudo-labeled validation sets [40,14]. These approaches localize failures, but generally do not yield a direct, class-conditional representation-space direction that can be ranked and targeted.

**Automatic Concept Discovery and Post-hoc Debiasing.** Concept discovery methods extract human-interpretable concepts from activations without concept annotations [28,15,11], and have been used to improve robustness by reweighting or rebalancing discovered concepts during training [3]. Related work studies factorization-based concept ranking [25], hierarchical or patch-level concept bottlenecks [37], and the faithfulness and reliability of concept directions for intervention [7,36]. Other post-hoc pipelines cluster explanation maps or use vision-language models to propose candidate biases [2,8,23,35], but often require manual validation and do not provide a concrete class-conditional direction that can be directly suppressed.

**Post-hoc Mitigation with Frozen Models.** Several methods reduce spurious reliance after ERM training without group labels, including last-layer retraining on a class-balanced split (DFR) [24], sub-network extraction [27], activation-level erasure [18], and data pruning [32]. Vision-language work similarly studies inference-time debiasing under deployment constraints [13,19]. While aligned in motivation, these approaches still perform optimization, pruning, or model edits, and typically do not expose an interpretable direction; we instead identify class-conditional spurious concept directions in a frozen model to support downstream intervention without updating parameters.