FINE-GRAINED CLASS-CONDITIONAL DISTRIBUTION BALANCING FOR DEBIASED LEARNING

Miaoyun Zhao
Key Laboratory of Social Computing and Cognitive Intelligence
Dalian University of Technology
Liaoning, China

Qiang Zhang *
Key Laboratory of Social Computing and Cognitive Intelligence
Dalian University of Technology
Liaoning, China

ABSTRACT

Achieving group-robust generalization in the presence of spurious correlations remains a significant challenge, particularly when bias annotations are unavailable. Recent studies on Class-Conditional Distribution Balancing (CCDB) reveal that spurious correlations often stem from mismatches between the class-conditional and marginal distributions of bias attributes. They achieve promising results by addressing this issue through simple distribution matching in a bias-agnostic manner. However, CCDB approximates each distribution using a single Gaussian, which is overly simplistic and rarely holds in real-world applications. To address this limitation, we propose a novel Multi-stage data-Selective reTraining strategy (MST), which describes each distribution in greater detail using the hard confusion matrix. Building on these finer descriptions, we propose a fine-grained variant of CCDB, termed FG-CCDB, which enhances distribution matching through more precise confusion-cell-wise reweighting. FG-CCDB learns sample weights from a global perspective, effectively mitigating spurious correlations without incurring substantial storage or computational overhead. Extensive experiments demonstrate that MST serves as a reliable proxy for ground-truth bias annotations and can be seamlessly integrated with bias-supervised methods. Moreover, when combined with FG-CCDB, our method performs on par with bias-supervised approaches on binary classification tasks and significantly outperforms them in highly biased multi-class and multi-shortcut scenarios.

1 INTRODUCTION

Neural networks trained with standard Empirical risk minimization (ERM) Vapnik (1998) often suffer from spurious correlations: shortcuts that are predictive of the target class in the training data but irrelevant to the true underlying classification function LaBonte et al. (2023b). Samples exhibiting such spurious correlations typically dominate the training distribution and form the majority groups, while samples with different or conflicting correlations constitute the minority groups Radford et al. (2021). This imbalance across groups is also referred to as biased data, which results in poor ERM performance on the minority ones, sometimes even no better than random guessing Shah et al. (2020). Spurious correlations are prevalent in many high-stakes applications, including toxic comments identification Borkan et al. (2019), medical diagnosis Castro et al. (2020), and autonomous driving Pourkeshavarz et al. (2024), where both robustness and fairness are critical but overlooked by conventional methods. Take the traffic sign classification task as a vivid example Liu et al. (2023), in which the training data exhibits a strong bias: 99% of stop signs appear in red, whereas stop signs of other colors are rare and constitute a minority group Beery et al. (2018). Consequently, the classifier relies on the red color as a shortcut for recognizing stop signs, ignoring

*Corresponding author.

the textual "stop" features. This leads to biased predictions and poor generalization when the color cue is absent or misleading. Arjovsky et al. (2019); Geirhos et al. (2020); Beery et al. (2018). These challenges underscore the urgent need to develop classification methods that remain reliable across diverse data subgroups, especially in the presence of spurious correlations.

One of the most effective strategies for improving robustness against spurious correlations is to retrain models using group-balanced subsets derived from bias annotations Kirichenko et al. (2023). However, given the massive scale of modern datasets, manually labeling bias attributes is often prohibitively expensive, which motivates the development of annotation-free alternatives. Recent studies have shown that models trained with naïve ERM tend to favor biased solutions, which generalize poorly to minority groups — offering a "free lunch" for bias modeling Pezeshki et al. (2024); Puli et al. (2023). Accordingly, various methods have been developed to identify misclassified samples as belonging to minority groups. These approaches either explicitly highlight such samples or implicitly simulate group-balancing during the debiasing process to enhance group robustness LaBonte et al. (2023a); Pezeshki et al. (2024); Li et al. (2023a); Liu et al. (2021). However, they often rely on empirically chosen hyperparameters to control the upweighting of minority groups, which can easily lead to overemphasis on these groups and, in turn, degrade performance on the majority ones. As a result, held-out annotations are often required for effective hyperparameter tuning. Recent research on Class-conditional distribution balancing (CCDB) Zhao et al. (2025) reveals that spurious correlations arise from the mismatches between class-conditional and marginal distributions (usually caused by bias cues), and addresses it by reweighting samples to minimize the mutual information between bias cues and class labels without hyperparameter searching. However, CCDB performs coarse distribution matching by treating each distribution as a single Gaussian, which rarely holds in real-world applications. In practice, instances within the same class often exhibit multi-modal distributions due to hidden bias cues. Thus, this coarse matching fails to capture intra-class variations, leaving residual spurious correlations unaddressed.

To resolve these limitations, we propose a fine-grained distribution matching technique based on CCDB, termed Fine-Grained Class-Conditional Distribution Balancing (FG-CCDB), which achieves stronger mitigation of spurious correlations without relying on bias annotations. Our approach is developed from two key perspectives: (i) **Fine-grained distribution description.** Inspired by the "free lunch" phenomenon in ERM — where models tend to overfit to spurious correlations — we introduce a Multi-stage data-Selective reTraining strategy (MST) for bias characterization, which capable of tackling multi-shortcuts by relate the hard confusion matrix to bias-aligning and conflicting partitions, and employing a multi-stage, data-selective retraining strategy to enhance the reliability of these partition assignments, which iteratively refines predictions from the overfitted model. This process yields a confusion matrix that approximates the ground-truth group partition when spurious correlations arise from a single shortcut. (ii) **Fine-grained distribution matching.** Building on the confusion matrix identified by MST, we extend CCDB into a fine-grained formulation, termed Fine-Grained Class-Conditional Distribution Balancing (FG-CCDB). It provides a discrete multi-modal approximation of both class-conditional and marginal distributions, enabling precise mode-wise alignment and thus more thorough mitigation of spurious correlations than the original CCDB. **The main contributions of this work are as follows:** (i) We propose an annotation-free bias exploration method with multi-stage refinement, based on model overfitting, which generalizes beyond singular shortcut and serves as a reliable alternative to human annotations. (ii) We introduce FG-CCDB, a lightweight and scalable debiasing method that enables fine-grained mode-wise reweighting and is well-suited for multi-class classification and multi-shortcut mitigation. (iii) Extensive experiments show that our method matches or surpasses bias-supervised baselines, achieving strong performance without requiring bias annotations.

2 Related work

The related work is structured around the two core aspects of our contribution.

2.1 Bias exploration

Primary approaches define bias as texture Bahng et al. (2020), background Venkataramani et al. (2024), or image style Li et al. (2025)—features presumed irrelevant to class labels. These methods often rely on tailored architectures or training schemes to detect specific bias cues Hong & Yang

(2021), but generalize poorly to unknown biases. To overcome this, recent data-driven strategies interpret bias as group imbalance or latent substructures. Some methods, like JTT Liu et al. (2021), LfF Nam et al. (2020), and RIDGE Pezeshki et al. (2024), identify bias via consistently misclassified (hard) samples under ERM. Others rely on model disagreement, *e.g.,* DebiAN Li et al. (2022b) iteratively trains a bias "discoverer" alongside a main classifier, XRM Pezeshki et al. (2024) uses a pair of biased auxiliary models to generate pseudo group labels across the training set, DDB Ciranni et al. (2025) utilizes a diffusion model to generate bias-aligned data, which amplifies the bias reliance. Other methods, such as GEORGE Sohoni et al. (2020), apply unsupervised feature clustering to decompose each class into latent subgroups. Few of these methods conduct a thorough evaluation on the quality of bias prediction. Another trend leverages vision-language models (e.g., CLIP Radford et al. (2021)) to infer explainable bias attributes Jain et al. (2023); Kim et al. (2024); Wiles et al. (2022), though they are constrained by predefined vocabularies and may miss unexpected biases.

2.2 Bias mitigation

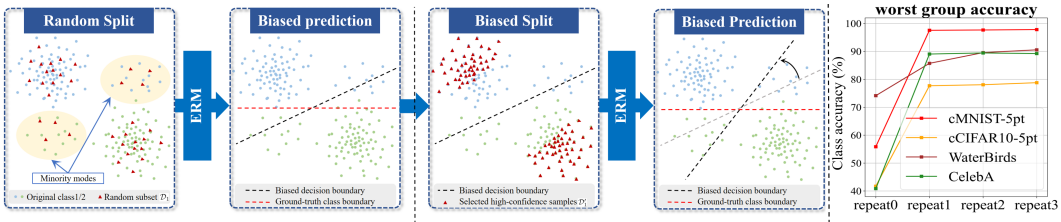
Bias annotation dependent. With the assistance of bias annotations, a variety of methods have been developed to mitigate spurious correlations. GroupDRO Sagawa et al. (2020) groups data based on class and bias annotations and optimizes for the worst-group performance. DFR Kirichenko et al. (2023) improves robustness by retraining only the last layer using a small, balanced validation set. MAPLE Zhou et al. (2022) uses a measure based on validation set with explicit bias annotations to reweight training samples. LISA Yao et al. (2022) utilizes data augmentation technique to encourage bias-invariant features. Though effective, relying on costly bias annotations limits their scalability in real applications.

Bias-conflicting samples dependent. To mitigate spurious correlations without manual annotation, recent studies often leverage disagreements among auxiliary models to identify bias-conflicting samples and focus learning on them.Nam et al. (2020); Liu et al. (2023); Chu et al. (2021); Liu et al. (2021). To better identify bias-conflicting samples, SELF LaBonte et al. (2023b) proposes to split the training data and applying early stopping for effective bias-conflicting detection. uLA Tsirigotis et al. (2023b) leverages pretrained self-supervised models to extract bias-relevant information. DPR Han et al. (2024) uses the Generalized cross-entropy loss Nam et al. (2020) to amplify model bias. However, these methods rely on empirically tuned parameters—often requiring a split of annotated subsets—and their simple binary partitioning into bias-aligned and bias-conflicting samples is insufficient to fully capture the structure of bias, ultimately limiting generalization.

Bias-agnostic. Beyond bias-aware techniques, several bias-agnostic approaches have emerged, motivated by diverse perspectivesPuli et al. (2023); Jain et al. (2024). MASKTUNE Asgari et al. (2022), ExMap Chakraborty et al. (2024), and DaC Noohdani et al. (2024) reduce reliance on spurious features by identifying bias-related regions via heatmaps, which restricts their applicability to the image domain. Stable learning approachesZhang et al. (2021); Yu et al. (2023) treat spurious correlations as effects of unknown confounders and attempt to mitigate them by decorrelating features, though this is difficult to achieve in practice. GERNE Asaad et al. (2025) leverages the gradient differences between two batches to identify a debiasing direction, along which the model is optimized. CCDB Zhao et al. (2025) seeks to mitigate spurious correlations by minimizing the mutual information between spurious features and class labels via distribution matching. Although effective, its coarse matching strategy limits generalization performance.

3 Our method

Our work builds on the existing method CCDB, which attributes spurious correlations to distribution mismatches and addresses them through sample reweighting without requiring bias annotations. However, CCDB performs distribution matching in a relatively coarse manner by modeling each distribution as a single Gaussian. To enable more accurate alignment–and thereby more effective spurious correlations elimination–we propose a fine-grained extension. Specifically, we introduce a multi-stage data-selective retraining strategy (MST) that characterizes bias structure via the hard confusion matrix, allowing for a discrete multi-modal description of each distribution. Based on these multi-modal distributions, we develop Fine-Grained Class-Conditional Distribution Balancing (FG-CCDB), which performs alignment at the mode level.



(a) Initial biased learning (b) Bias enhancement learning (multiple times) (c)WGA along the repeatation of (b)

Figure 1: The MST framework progresses from an initial partition with limited minority group coverage (a) to a more complete identification in stage (b). (c) WGA under different bias-capturing qualities.

We consider the task of predicting a label $y \in \mathcal{Y}, \mathcal{Y} = \{1, \ldots, C\}$ based on an input $\boldsymbol{x} \in \mathcal{X}, \mathcal{X} \subset \mathbb{R}^d$. Following prior work, we define a *shortcut* as an explainable attribute (*e.g.,* color, background) that is spuriously correlated with class labels and highly predictive, and focus on a more general setting in which each data point $(\boldsymbol{x}, y)$ may be associated with one or more shortcuts. Motivated by Tsirigotis et al. (2023a), we use an auxiliary biased model to predict the bias label $s \in \mathcal{S}$, which share the same label space as y, *i.e.,* $|\mathcal{S}| = |\mathcal{Y}|$. Note that our goal is for s to capture general and harmful bias information that humans may not preconceive Li et al. (2022b), rather than only physically interpretable attributes. The value s represents spurious signals that an ERM model prefers over core features and that consequently cause evaluation failures. $s = i$ denotes all spurious cues that cause samples from other classes to be misclassified as class i. These cues may correspond to interpretable shortcuts, combinations of multiple shortcuts, or entangled, uninterpretable patterns. By combining s and y, we partition the dataset into modes $\mathcal{M} = \mathcal{S} \times \mathcal{Y}$, which exactly corresponde to the hard confusion matrix. When the bias corresponds to a single shortcut, this reduces to conventional group partitions. To distinguish our data partitioning approach from traditional group-based methods, we refer to the partitions derived from the confusion matrix as **modes**. Accordingly, diagonal entries represent majority (bias-aligning) modes, and off-diagonal entries correspond to minority (bias-conflicting) modes. With the confusion matrix, one can infer a discrete multi-modal approximation of both the class-conditional and marginal distributions over bias information. The goal is twofold: (i) to train a biased model that can effectively explore the underlying bias cues; (ii) to train a debiased model that invariant to bias information and achieves uniform performance across all modes.

3.1 Bias Exploration Through Overfitting

In this section, we introduce the proposed multi-stage data-selective retraining (MST) technique and demonstrate its compatibility with existing bias-supervised methods. It is well established that, in the presence of spurious correlations, ERM tends to overfit to majority groups in training data, leading to an over-reliance on bias cues and poor generalization to minority groups. Recent studies LaBonte et al. (2023b); Tsirigotis et al. (2023b) have made preliminary attempts to exploit this overfitting behavior to mitigate spurious correlations, revealing that the predictions of overfitted models are strongly aligned with bias cues. Furthermore, Lee et al. (2023); Ciranni et al. (2025) find that removing bias-conflicting samples improves bias prediction and point out that, in principle, if all bias-conflicting samples were removed, one could train a bias-capturing model that provides ideal learning signals for debiasing. Inspired by these insights, we propose a multi-stage framework for refined bias prediction, which further leverages model overfitting and serves as an approximate substitute for human annotations. The overall framework consists of two basic stages (Figure 1(a)(b)): initial bias learning and bias enhancement learning. The first stage extracts primary bias patterns, while the second amplifies them in the model's predictions, yielding a reliable bias predictor.

Initial bias learning. Given an accessible train dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of N samples and C classes. Following prior works Zhao et al. (2025); LaBonte et al. (2023b); Pezeshki et al. (2024), we explore bias information by randomly splitting $\mathcal{D}$ into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1$ contains a fraction γ of the original data (Figure 1(a), left). We then perform naïve ERM on $\mathcal{D}_1$ to train a biased model $f_{\boldsymbol{\theta}_1}$, which typically performs well on majority groups but poorly on minority groups. Unlike prior worksLaBonte et al. (2023b); Pezeshki et al. (2024), which use $95\%/50\%$ of the data for biased training, our goal is to maximize the model's alignment with bias cues to better reveal underlying spurious correlations. As demonstrated in our experiments (Figure 3(right)), a smaller γ proves more effective for bias exploration, with $\gamma = 10\%$ emerging as a sweet spot. Since the data is

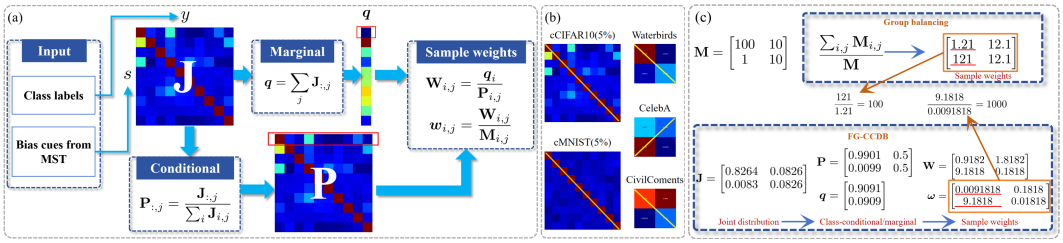


Figure 2: (a) The framework of our FG-CCDB (b) The joint distribution $\mathbf{J}$ of bias and class labels estimated by our method. (c) Toy example to show how FG-CCDB differs from group balancing.

randomly split, some samples from minority modes inevitably participate in training, which weakens the model's tendency to align its predictions with bias cues (Figure 1(a) right). To counteract this effect, we introduce a subsequent amplification stage.

Bias enhancement learning. Amplifying bias in model predictions is non-trivial. Our key idea is to guide the bias prediction model to focus exclusively on majority modes, *i.e.,* to construct a training subset $\mathcal{D}_1$ that contains little to no samples from minority modes. This idea of removing bias-conflicting samples has been shown to effectively amplify bias in prior work Lee et al. (2023); Ciranni et al. (2025). Such a setup forces the model to overfit to the majority modes and align more strongly with the corresponding bias cues, thereby behaving like a bias predictor and exhibiting near-zero generalization ability on minority modes. To achieve this, we introduce a data selection procedure based on the predictions of $f_{\boldsymbol{\theta}_1}$, forming an extremely biased subset $\mathcal{D}_1'$, on which a more biased model is trained. Specifically, for each sample $(\boldsymbol{x}_i, y_i)$ in $\mathcal{D}$, we infer the softmax output as $\boldsymbol{h}_i = f_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i)$. Within each class, we select the top β fraction of samples ($\beta \in [0, 1]$) with the highest prediction confidence (measured by $\boldsymbol{h}_i$), and aggregate them to form $\mathcal{D}_1'$. In our experiments, we find that $\beta = 50\%$ offers a stable and reliable choice. Since $f_{\boldsymbol{\theta}_1}$ is biased toward majority modes, the high-confidence samples are more likely to come from those modes. Consequently, $\mathcal{D}_1'$ filters out most minority mode instances and is thus more biased than $\mathcal{D}_1$. (Figure 1(b) left). We then train a new biased model $f_{\boldsymbol{\theta}_2}$ using naïve ERM on $\mathcal{D}_1'$. The resulting model serves as the final bias predictor to produce bias labels for $\mathcal{D}$. Combined with the target class labels, the resulting hard confusion matrix yields estimated mode partitions over the space $|\mathcal{S}| \times |\mathcal{Y}|$, which can serve as a proxy for group annotations in bias-supervised methods (Figure 1(b) right).

Notably, the "Bias enhancement learning" stage can be repeated to further improve bias prediction accuracy. Only the biased model from the final repetition is used to generate bias labels. As shown in the experiments(Figure1(c)), a single iteration already achieves performance comparable to existing methods, while further iterations lead to gradually converging performance with diminishing gains.

3.2 Fine-grained class-conditional distribution balancing

In this section, we present the Fine-grained Class-Conditional Distribution Balancing (FG-CCDB) approach. With the hard confusion matrix obtained via MST, FG-CCDB improves both the quality of distribution matching and the efficiency of sample reweighting.

The original CCDB proposes to mitigate spurious correlations by directly minimizing the mutual information between bias cues and target classes, which is achieved by aligning each class-conditional distribution with the marginal distribution, while simultaneously balancing class proportions — a generalization to traditional class balancing technique. Specifically, the objective is to minimize:

$$\mathcal{L}_{\boldsymbol{\omega}} = I(\boxed{\boldsymbol{z}}, y) - H(y) = \mathbb{E}_{p_{\boldsymbol{\omega}}(y)} D_{\mathrm{KL}}[p_{\boldsymbol{\omega}}(\boxed{\boldsymbol{z}}|y) \| p(\boxed{\boldsymbol{z}})] + \mathbb{E}_{p_{\boldsymbol{\omega}}(y)} \log p_{\boldsymbol{\omega}}(y) \tag{1}$$

where $\boxed{\boldsymbol{z}}$ denotes the latent feature (with gradients detached) extracted by the biased model prior to the fully connected layer, which predominantly captures bias cues. $\boldsymbol{\omega}$ denotes the sample weights to be optimized, and $D_{\mathrm{KL}}[\cdot\|\cdot]$ refers to the Kullback–Leibler divergence Kullback & Leibler (1951). Since the true distributions associated with $\boldsymbol{z}$ and y are unknown, CCDB approximates them using single Gaussian, which is insufficient for complex data with inherently multi-modal structures. Moreover, CCDB's sample-level reweighting requires storing and processing feature representations for the entire dataset, incurring additional computational cost.

Our work adopts the same objective as Equation 1. To overcome the aforementioned limitations, we derive a discrete multi-modal approximation of both class-conditional and marginal distributions from the hard confusion matrix, which enables localized, mode-wise distribution matching and leads

to more accurate and scalable reweighting. As shown in Figure 2 (a), we represent the confusion matrix as $\mathbf{M} \in \mathbb{R}^{C\times C}$, where $\mathbf{M}_{i,j}$ denotes the number of samples belonging to mode $(s, y) = (i, j)$, Thus, the joint distribution over $(\boldsymbol{z}, y)$ is approximated with a discretized version over modes (s, y), which is characterized by matrix $\mathbf{J} \in \mathbb{R}^{C\times C}$ with $\mathbf{J}_{i,j} = \frac{\mathbf{M}_{i,j}}{N}$ represents the probability of a sample belonging to mode $(s, y) = (i, j)$, N is the total number of training samples. By design, we define a class-conditional distribution matrix $\mathbf{P} \in \mathbb{R}^{C\times C}$ such that the j-th column $\mathbf{P}_{:,j}$ encodes $p(\boldsymbol{z}|y = j)$, and a marginal distribution vector $\boldsymbol{q} \in \mathbb{R}^C$ that captures $p(\boldsymbol{z})$. Both $\mathbf{P}$ and $\boldsymbol{q}$ are computed directly from $\mathbf{J}$ as follows:

$$p(\boldsymbol{z}|y = j) \overset{\text{def}}{\approx} \mathbf{P}_{:,j} = \frac{\mathbf{J}_{:,j}}{\sum_i \mathbf{J}_{i,j}}, \qquad p(\boldsymbol{z}) \overset{\text{def}}{\approx} \boldsymbol{q} = \sum_j \mathbf{J}_{:,j} \tag{2}$$

Figure 2 (b) shows the joint distribution matrix $\mathbf{J}$ estimated by our MST across four datasets. Clear spurious correlations are observed, as evidenced by the strong diagonal elements (aligned along the yellow line), which indicate a high dependency between labels and bias cues. To eliminate these spurious correlations and minimize equation1, we introduce mode-level weighting parameter $\mathbf{W} \in \mathbb{R}^{C\times C}$ to adjust each class-conditional distribution so that it aligns with the marginal distribution. A straightforward solution for $\mathbf{W}$ is,

$$\mathbf{W}_{i,j} = \frac{\boldsymbol{q}_i}{\mathbf{P}_{i,j}}, \qquad \text{for } i, j = 1, \cdots, C \tag{3}$$

Note, equation 3 achieves exact distribution matching, *i.e.,* $\mathbf{W}_{:,j} \odot \mathbf{P}_{:,j} = \boldsymbol{q}$, meaning that all class-conditional distributions are reweighted to align with the same marginal distribution $\boldsymbol{q}$, where $\odot$ denotes the Hadamard product. For a given mode $(s, y) = (i, j)$, assuming uniform contribution from its samples, the corresponding sample weight is,

$$\boldsymbol{w}_{i,j} = \frac{\mathbf{W}_{i,j}}{\mathbf{M}_{i,j}}, \qquad \text{for } i, j = 1, \cdots, C \tag{4}$$

Note that beyond distribution matching, Equation 4 inherently solved the class imbalance issue: the mode with more samples gets smaller weights. As a result, FG-CCDB simultaneously minimizes both terms in Equation 1. These weights are subsequently used during debiasing to reweight training samples according to their mode identities.

It is worth noting that *our distribution matching fundamentally differs from conventional group balancing* (see example in Figure2(c)): (i) Unlike group balancing, which aims to reduce differences across all entries in the mode matrix, our method focuses solely on aligning the class-conditional distributions—i.e., reducing the variation among columns in $\mathbf{P}$—while preserving intra-column imbalance. This allows for more flexible training by merely minimizing mutual information rather than enforcing strict equality. (ii) By minimizing the divergence between conditional and marginal distributions, our method and CCDB implicitly achieve "covariate balance" from the view of causal inference, specifically, by finding a reweighting that makes the confounder (bias) independent of the treatment (core feature), ultimately forcing the statistical model to rely solely on core features for inference Neal (2020). (iii) Simple scale balancing between majority and minority modes is insufficient for generalization, as majority modes typically exhibit greater diversity. Our method applies a more aggressive reweighting strategy. For example, the ratio between the largest and smallest mode weights in FG-CCDB reaches 1000, compared to just 100 in conventional group balancing. *Compared to CCDB, our sample reweighting approach offers several key advantages*: (i) It performs distribution matching across multiple localized regions defined by the confusion matrix, enabling more precise alignment and more thorough removal of spurious correlations; (ii) The sample weights are computed efficiently in closed form, without requiring any iterative optimization; (iii) Instead of assigning weights individually to each sample, FG-CCDB assigns a shared weight to samples within the same mode, resulting in negligible computational and memory overhead.

After completing MST and FG-CCDB, we train a debiased model $f_{\boldsymbol{\phi}}$ by incorporating sample weights into the data sampling process using PyTorch's "*torch.utils.data.WeightedRandomSampler*" following Zhao et al. (2025). Unless otherwise specified, we refer to the entire procedure as FG-CCDB for brevity. A full algorithm of the proposed method is provided in Appendix A.

4 Experimental results

In this section, we demonstrate the effectiveness of our method from five perspectives: (i) We conduct experiments on real-world binary classification benchmarks with either single or multiple short-

Table 1: Classification performance on real-world datasets. We report the average test accuracy(%) and std.dev. over 5 random seeds. Best bias-agnostic results in bold.

Methods	Bias label		Waterbirds		CelebA		CivilComments	
	Train	Val	i.i.d.	WGA	i.i.d.	WGA	i.i.d.	WGA
GroupDRO	Yes	Yes	93.50	91.40	92.90	88.90	84.2	73.7
DFR	Yes	Yes	94.20±0.4	92.90±0.2	91.30±0.3	88.30±1.1	87.2±0.3	70.1±0.8
LfF	No	Yes	97.50	75.20	86.00	77.20	68.2	50.3
JTT	No	Yes	93.60	86.00	88.00	81.10	83.3	64.3
LC	No	Yes	-	90.50±1.1	-	88.10±0.8	-	70.30±1.2
SELF	No	Yes	-	93.00±0.3	-	83.90±0.9	-	79.10±2.1
DaC	No	Yes	95.3±0.4	92.3±0.4	91.4±1.1	81.9±0.7	-	-
ERM	No	No	**97.30**	72.60	**95.60**	47.20	81.6	66.7
MASKTUNE	No	No	93.00±0.7	86.40±1.9	91.30±0.1	78.00±1.2	-	-
uLA	No	No	91.50±0.7	86.10±1.5	93.90±0.2	86.50±3.7	-	-
XRM	No	No	90.60	86.10	91.0	88.5	83.5	70.1
DebiAN	No	No	90.80	78.19	84.0	52.9	-	-
DDB	No	No	-	90.34	-	-	-	-
GERNE	No	No	-	89.88±0.67	-	74.24±2.51	-	63.10±0.22
CCDB	No	No	92.59±0.10	90.48±0.28	90.08±0.19	85.27±0.28	83.60±0.21	75.00±0.26
FG-CCDB	No	No	92.50±0.52	**90.56**±0.24	89.71±0.54	**89.22**±0.19	**86.99**±0.14	**78.52**±0.42

Table 2: Results on UrbanCars.

Methods	Bias label		I.D.	Gap due to shortcuts(↑)		
	Train	Val	Acc	BG	CoObj	BG+CoObj
GroupDRO	Yes	Yes	91.6	-10.9	-3.6	-16.4
JTT	No	Yes	95.9	-8.1	-13.3	-40.1
DaC	No	No	98.17	-3.78	-9.78	-58.58
ERM	No	No	97.6	-15.3	-11.2	-69.2
ExMap	No	No	-	-5.9	-9.9	-30.7
DebiAN	No	No	98.0	-14.9	-10.5	-69.0
DDB	No	No	86.39	**-1.85**	**-0.52**	**-0.12**
FG-CCDB	No	No	92.98	-4.17	-7.37	-4.9

Table 3: Ablation study on four datasets.

Methods	Waterbirds		CelebA		cMNIST	cCIFAR10
	i.i.d.	WGA	i.i.d.	WGA	i.i.d.	i.i.d.
GroupDRO	93.50	91.40	92.90	88.90	84.20	57.32
GroupDRO -MST	90.82±0.05	88.47±0.35	88.69±0.15	85.21±0.02	84.07±0.22	55.73±0.54
DFR	**94.20**±0.4	**92.90**±0.2	91.30±0.3	88.30±1.1	-	-
DFR-MST	92.53±0.50	91.49±0.72	88.80±0.20	85.87±0.29	-	-
FG-CCDB	92.50±0.52	90.56±0.24	89.71±0.54	**89.22**±0.19	98.21±0.02	78.06±0.30
FG-CCDB -sup	91.54±0.11	91.76±0.13	**93.14**±0.16	89.09±0.12	**98.26**±0.21	**78.53**±0.37

cuts, such as Waterbirds Zhou et al. (2022), CelebA Zhou et al. (2022), CivilComments Koh et al. (2021), and UrbanCars Li et al. (2023b) to validate the overall effectiveness of our method; (ii) We further evaluate our method on challenging multi-class datasets, including cMNIST Li et al. (2022a) and cCIFAR10 Hendrycks & Dietterich (2018) to assess its robustness under highly biased conditions; (iii) To evaluate the reliability of the bias cues explored by MST, we compare them with ground-truth bias annotations and analyze the effects of repeating the "bias enhancement learning" procedure; (iv) we conduct an ablation study to demonstrate that each technical component (MST and FG-CCDB) makes a distinct and independent contribution to the final performance. (v) Finally, we analyze the effects of hyperparameters (γ and β) on the performance of MST.

For all datasets, we adopt the same train-validation-test split following Liu et al. (2021); Tsirigotis et al. (2023b) for fair comparison. Results are averaged over 5 random seeds, and for each seed, the best-performing model (the one with the highest worst-class accuracy on the validation set) is selected Tsirigotis et al. (2023b). Unless otherwise stated, we repeat the "bias enhancement learning" process three times for FG-CCDB. See the appendix for the detailed experimental setup.

Compared methods. To demonstrate the superiority of our method in addressing spurious correlations and its potential to serve as an approximate substitute for bias-supervised methods, we compare it with both bias-supervised and bias-agnostic techniques. GroupDRO Sagawa et al. (2020) and DFR Kirichenko et al. (2023) are fully bias-supervised during both training and validation, and serve as strong baselines. LfF Nam et al. (2020), JTT Liu et al. (2021), LC Liu et al. (2023), DaC Noohdani et al. (2024), and SELF LaBonte et al. (2023b) rely on pseudo-bias supervision during training, but still require bias annotations during validation to achieve optimal performance. In contrast, ERM, uLA Tsirigotis et al. (2023b), MASKTUNE Asgari et al. (2022), XRM Pezeshki et al. (2024), DebiAN, ExMap, DDB Ciranni et al. (2025), GERNE Asaad et al. (2025), and CCDB Zhao et al. (2025), similar to our method, are entirely bias-agnostic throughout both training and validation.

4.1 Binary classification with a single or multiple shortcuts

The results on real-world binary classification with a single shortcut are shown in Table1. Although i.i.d. performance reflects overall accuracy, it can mask disparities across groups. In contrast, worst-group accuracy (WGA) directly measures robustness by focusing on the most challenging subpopulations. With bias annotations available during both training and validation, GroupDRO and DFR

Table 4: Results on cMNIST and cCIFAR10 with various bias-conflicting ratios in the training set. The test accuracy(%) is averaged over 5 random seeds.The best results are indicated in bold.

Methods	Bias label		cMNIST				cCIFAR10			
	Train	Val	0.5%	1%	2%	5%	0.5%	1%	2%	5%
GroupDRO	Yes	Yes	63.12	68.78	76.30	84.20	33.44	38.30	45.81	57.32
LfF	No	Yes	52.50±2.43	61.89±4.97	71.03±2.44	80.57±3.84	28.57±1.30	33.07±0.77	39.91±0.30	50.27±1.56
LC	No	Yes	71.25±3.17	82.25±2.11	86.21±1.02	91.16±0.97	34.56±0.69	37.34±0.69	47.81±2.00	54.55±1.26
DaC	No	Yes	53.24	75.02	87.60	94.70	21.01	28.01	36.56	51.06
ERM	No	No	35.19±3.49	52.09±2.88	65.86±3.59	82.17±0.74	23.08±1.25	25.82±0.33	30.06±0.71	39.42±0.64
uLA	No	No	75.13±0.78	81.80±1.41	84.79±1.10	92.79±0.85	34.39±1.14	62.49±0.74	63.88±1.07	74.49±0.58
GERNE	No	No	77.25±0.17	83.98±0.26	87.41±0.31	90.98±0.13	39.90±0.48	45.60±0.23	50.19±0.18	56.53±0.32
CCDB	No	No	83.20±2.17	87.95±1.59	91.02±0.28	96.37±0.25	55.07±0.85	63.28±0.46	67.78±0.78	74.64±0.34
FG-CCDB	No	No	**89.02**±0.45	**94.93**±0.17	**96.18**±0.19	**98.21**±0.02	**55.28**±0.54	**64.66**±0.48	**71.69**±0.31	**78.06**±0.30

demonstrate strong generalization performance on the worst group, serving as a challenging upper bound. In contrast, methods that only use bias annotations during validation show a bit inferior performance. The situation becomes more challenging when access to bias annotations is not permitted. In this case, existing bias-agnostic methods consistently fall short of the supervised upper bound on at least one of the datasets. Remarkably, SELF, CCDB and our method surpass the supervised upper bound on CivilComments by a large margin. This is because they apply stronger upweighting to the minority groups/modes. Among all compared methods, including those with full supervision, our method consistently achieves the best or competitive WGA across all three datasets, highlighting its effectiveness in eliminating the need for human annotations.

Table2 presents the results on UrbanCars with multiple shortcuts: background (BG) and co-occurring object (CoObj). The in-distribution accuracy(I.D. Acc) and gap-related metrics are adopted from Li et al. (2023b)(See appendix for details). The BG/CoObj/BG+CoObj Gap is the drop in accuracy between mean and cases when only the BG/CoObj/BG+CoObj is uncommon. A smaller drop indicates better generalization. On average, BG+CoObj is the most challenging one and most compared methods suffer a significant drop on it. GroupDRO can mitigate multiple shortcuts; however, they require access to labels of both shortcuts. Although DDB shows the smallest overall drops across all bias-conflicting scenarios, its base I.D. Acc is the lowest among all compared methods. Overall, our method consistently achieves the best balance between high I.D. Acc and small drops compared to other bias-agnostic methods (particularly on the challenging BG+CoObj generalization). It performs comparably to, or better than, methods that rely on bias annotations. These results confirm that our approach provides a general framework for handling multi-shortcut scenarios. Please refer to Appendix D for more details.

4.2 Multi-class classification under extreme spurious correlations

In this section, we use the synthetic datasets cMNIST and cCIFAR10 to evaluate the effectiveness of our method under challenging multi-class settings with extreme spurious correlations. For each dataset, we vary the ratio of bias-conflicting samples in the training set to control the strength of spurious correlations and evaluate performance on a completely unbiased test set. Following Tsirigotis et al. (2023b), the bias-conflicting ratios are set to $\{0.5\%, 1\%, 2\%, 5\%\}$ for both datasets, where 0.5% indicates an extremely biased scenario. The generalization accuracies are reported in Table 4. We observe that: (i) On both datasets, our method consistently achieves the best performance. In particular, it outperforms the second-best method by a large margin on cMNIST; (ii) On cCIFAR10, the improvements become more pronounced as the bias-conflicting ratio increases (i.e., at 2% and 5%).

A comparison of the results in Table 1 and Table 4 reveals a phenomenon similar to that reported in Zhao et al. (2025): bias-supervised methods tend to perform well on basic binary classification tasks, whereas bias-agnostic methods are relatively more effective in complex multi-class classification scenarios. In contrast, CCDB demonstrates strong performance across both scenarios. With fine-grained distribution matching, FG-CCDB further boosts performance over CCDB by a significant margin, highlighting the effectiveness of more thorough spurious correlations elimination.

To demonstrate the effectiveness of FG-CCDB in mitigating spurious correlations, we analyze how sample reweighting influences the correlation between feature dimensions and class/bias information, as shown in Figure 3(left). We compute the correlation of each feature dimension with class/bias information, and visualize their distributions using box plots Zhao et al. (2025). Before

sample reweighting, strong spurious correlations in the training data lead the biased model f_{θ_2} to rely heavily on bias-related features, with most dimensions exhibiting high correlation with bias and low correlation with class. After applying FG-CCDB weights on features from f_{θ_2}, the correlation with bias drops significantly, while the correlation with class increases. Moreover, after debiasing training on the reweighted data, this shift toward class-relevant features is further amplified, confirming that FG-CCDB effectively reduces the model's reliance on spurious features.

4.3 The quality of bias exploration

In this section, we evaluate the effectiveness of MST by measuring its mode-prediction F1-score, precision, and recall against the ground-truth annotations. Results regarding the smallest-mode are shown in Figure4(b). Since our method progressively filters out bias-conflicting samples, it retains far fewer such samples than XRM and JTT, achieving the highest F1-score across the four datasets. This confirms the principle that removing bias-conflicting samples improves bias prediction. JTT misidentifies a large number of majority samples as belonging to the smallest-mode(low precision). In contrast, XRM tends to misidentify minority-modes samples as majority modes(low recall). With multi-stage refinement, our method achieving the best overall performance. As discussed in Section 3, repeating the "bias enhancement learning" process can further improve both bias prediction accuracy and consequently mode prediction accuracy. To validate this claim, we conduct experiments with different numbers of repetitions. The mode prediction performance across varying repetition counts are shown in Figure 4(a). The dashed lines represent the standard accuracy across all modes, while the solid lines show the recall for each individual mode. We observe that repetition has a particularly strong effect on minority groups (highlighted in bold), as evidenced by the significant improvement in their recall with more repetitions. Please refer to Appendix Figure 8 for convergence results with additional repetitions.

Figure 1(c) shows the final WGA for classification as repetition increases. Notably, performance improves substantially after the first repetition and then plateaus, especially on cMNIST and CelebA, suggesting that a single repetition is often sufficient to achieve satisfactory performance.

4.4 Ablation study

Our method comprises two core technical modules: MST and FG-CCDB, which together demonstrate superior performance. In this section, we integrate these modules with existing methods and observe the resulting performance improvements to verify the effectiveness and versatility of our approach, as detailed below.

(i) To assess the effectiveness of MST, we replace ground-truth annotations in bias-supervised methods, *i.e.,* GroupDRO and DFR, with bias predictions generated by MST. This results in their unsupervised counterparts, denoted as GroupDRO-MST and DFR-MST, respectively. The results are reported at the top of Table 3. Remarkably, the generalization performance of these unsupervised variants is comparable to their supervised versions using human annotations. Although the bias predictions are not perfect, they are sufficiently accurate to identify most minority modes, confirming the effectiveness of our MST as an approximate substitute for human bias annotations.

(ii) To evaluate the effectiveness of FG-CCDB independently of MST, we replace the predicted bias with human annotations, resulting in a supervised version, FG-CCDB-sup. The results are reported at the bottom of Table 3. When bias annotations are available, FG-CCDB-sup further boosts performance, achieving results comparable to existing supervised methods on Waterbirds,

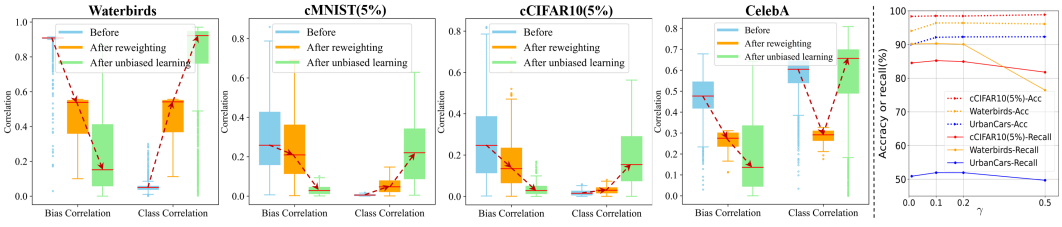


Figure 3: **Left:** effect of FG-CCDB sample reweighting in reshaping the data distribution and mitigating spurious correlations. **Right:** mode prediction accuracy and Smallest-mode recall along γ.

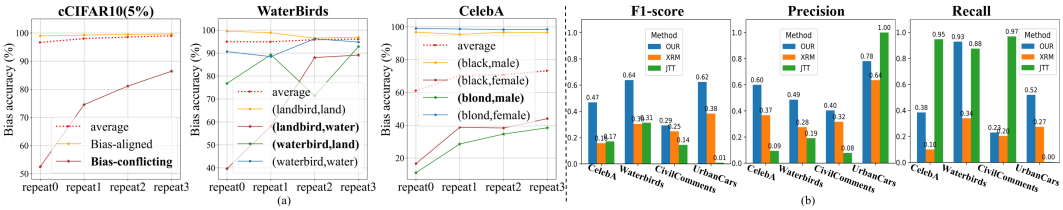


Figure 4: (a) The mode prediction accuracy along the repeating of the "bias enhancement learning" procedure; (b) Smallest-mode F1-score, precision, and recall compare with existing methods.

Table 5: The F1-score of the smallest-mode prediction under different top high-confidence ratio β.

	cCIFAR10(5%)	Waterbirds	CelebA	UrbanCars
Bias-align ratio	95.00%	94.97%	51.72%	90.25%
$\beta = 30\%$	0.65	0.53	0.32	0.47
$\beta = 50\%$	0.72	0.64	**0.47**	0.62
$\beta = 70\%$	**0.79**	**0.67**	0.40	**0.64**

and outperforming them on the others, especially in multi-class settings. *This justified our statement on a more aggressive reweighting and indicates that FG-CCDB is a more effective strategy than naïve group balancing for handling spurious correlations*. Moreover, the performance gap between FG-CCDB-sup and the original FG-CCDB is marginal, further confirming the effectiveness of our method in reducing reliance on human bias annotations.

4.5 Hyperparameter Analysis

In this section, we evaluate the effect of the hyperparameters γ for "Initial Bias Learning" and β for selecting top high-confidence samples on MST's final performance. The results are presented in Figure 3(right) and Table 5.

The hyperparameter γ controls the proportion of samples selected for training the initial bias model. Intuitively, a smaller γ leads to stronger overfitting to bias cues and thus greater reliance on them. As expected, the results in Figure 3(right) show that when $\gamma \leq 0.2$, both prediction accuracy and smallest-mode recall remain high. However, as γ increases to 0.5, the performance drops significantly. We find that $\gamma = 0.1$ serves as a sweet spot, while also saving computation compared to $\gamma = 0.2$.

F1-score with $\beta \in \{30\%, 50\%, 70\%\}$ are reported in Table 5. The hyperparameter β controls the proportion of top high-confidence samples selected to filter out bias-conflicting samples and amplify the model's bias. Intuitively, this value relates to the smallest bias-aligned ratio across classes, as shown in the first row of Table 5. Except for CelebA, whose ratio is slightly above 50%, all other datasets have ratios exceeding 90%. Accordingly, $\beta = 50\%$ serves as a reasonable middle-ground choice. For CelebA, which has a relatively low bias-aligned ratio, $\beta = 50\%$ achieves the best performance; whereas for datasets with ratios exceeding 90%, both $\beta = 50\%$ and $\beta = 70\%$ yield high F1-scores, with $\beta = 70\%$ performing the best. Intuitively, when bias annotations are unavailable, selecting the top 50% high-confidence samples is likely to capture the bias-aligned subset while excluding bias-conflicting samples.

5 Conclusions

In this paper, we address the challenge of robust group generalization under spurious correlations without requiring bias annotations. Following the distribution matching paradigm, we propose a method that integrates a reliable bias prediction module with fine-grained class-conditional distribution matching. Our approach demonstrates strong performance on real-world datasets with single or multiple shortcuts, as well as highly biased multi-class datasets, often matching or outperforming methods that rely on human-provided group annotations. By leveraging the model's overfitting behavior, our method offers a novel alternative to traditional group balancing strategies and effectively reduces reliance on manual supervision. However, its effectiveness may be limited in scenarios where the overfitting signal fails to capture bias cues—for example, in CelebA, which has only one minority group, or in CivilComments, where majority groups dominate one class while minority groups appear in another. These settings present different spurious correlation patterns that weaken the overfitting signal used for bias prediction. Addressing this limitation remains an important direction for future research.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

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

Ihab Asaad, Maha Shadaydeh, and Joachim Denzler. Gradient extrapolation for debiased representation learning. In *ICCV*, 2025. URL https://arxiv.org/abs/2503.13236.

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23284–23296. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/93be245fce00a9bb2333c17ceae4b732-Paper-Conference.pdf.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

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

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Rwiddhi Chakraborty, Adrian Sletten, and Michael C Kampffmeyer. Exmap: Leveraging explainability heatmaps for unsupervised group robustness to spurious correlations. In *CVPR*, pp. 12017–12026, 2024.

Sanghyeok Chu, Dongwan Kim, and Bohyung Han. Learning debiased and disentangled representations for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 8355–8366, 2021.

Massimiliano Ciranni, Vito Paolo Pastore, Roberto Di Via, Enzo Tartaglione, Francesca Odone, and Vittorio Murino. Diffusing debias: Synthetic bias amplification for model debiasing. In *NeurIPS*, 2025.

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.

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.

Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. Mitigating spurious correlations via disagreement probability. *Advances in Neural Information Processing Systems*, 37:74363–74382, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *ICLR*, 2023.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Improving subgroup robustness via data selection. *Advances in Neural Information Processing Systems*, 37:94490–94511, 2024.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. 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.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11552–11579. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/265bee74aee86df77e8e36d25e786ab5-Paper-Conference.pdf.

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, 2023b.

Jungsoo Lee, Jeonghoon Park, Daeyoung Kim, Juyoung Lee, Edward Choi, and Jaegul Choo. Revisiting the importance of amplifying bias for debiasing. In *37th AAAI Conference on Artificial Intelligence, AAAI 2023*, pp. 14974–14981. AAAI Press, 2023.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022a.

Gaotang Li, Jiarui Liu, and Wei Hu. Bias amplification enhances minority group performance. *CoRR*, abs/2309.06717, 2023a. URL https://doi.org/10.48550/arXiv.2309.06717.

Ruimeng Li, Yuanhao Pu, Zhaoyi Li, Chenwang Wu, Hong Xie, and Defu Lian. Invariant representation learning via decoupling style and spurious features. *Machine Learning*, 114(2):37, 2025.

Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pp. 270–288. Springer, 2022b.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *CVPR*, pp. 20071–20082, 2023b.

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. PMLR, 2021.

Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. In *The Eleventh International Conference on Learning Representations*, 2023.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Brady Neal. Introduction to causal inference. *Course lecture notes (draft)*, 132, 2020.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *CVPR*, pp. 27662–27671, June 2024.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with XRM. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gPStP3FSY9.

Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14874–14884, June 2024.

Aahlad Manas Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *Advances in Neural Information Processing Systems*, 36:71874–71910, 2023.

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

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 *ICLR*, 2020.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

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.

Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 56553–56575. Curran Associates, Inc., 2023a.

Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. *Advances in Neural Information Processing Systems*, 36:56553–56575, 2023b.

V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

Rahul Venkataramani, Parag Dutta, Vikram Melapudi, and Ambedkar Dukkipati. Causal feature alignment: Learning to ignore spurious background features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4666–4674, January 2024.

Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. In *NeurIPS ML Safety Workshop*, 2022.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Han Yu, Peng Cui, Yue He, Zheyan Shen, Yong Lin, Renzhe Xu, and Xingxuan Zhang. Stable learning via sparse variable independence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10998–11006, 2023.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Miaoyun Zhao, Qiang Zhang, and Chenrong Li. Class-conditional distribution balancing for group robust classification. *arXiv preprint arXiv:2504.17314*, 2025.

Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*, pp. 27203–27221. PMLR, 2022.

A THE ALGORITHM OF OUR PROPOSED METHOD

The complete procedure of our proposed FG-CCDB is summarized in Algorithm 1. It consists of three stages: bias exploration, sample weight inference, and unbiased classifier training.

Algorithm 1 Fine-grained class-conditional distribution balancing (FG-CCDB)

Input: Randomly initialized network $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ for bias prediction, $f_{\boldsymbol{\phi}}$ for unbiased classification; training set $\mathcal{D}$, validation set $\mathcal{D}_v$.
Output: unbiased classifier $f_{\boldsymbol{\phi}}$.
#Stage1: bias exploration via multi-stage data-selective retraining
1: Randomly sample a subset $\mathcal{D}_1$ from $\mathcal{D}$ with proportion γ ($\gamma = 10\%$).
2: Train $f_{\boldsymbol{\theta}_1}$ on $\mathcal{D}_1$ using ERM.
3: Select the top β ($\beta = 50\%$) most biased samples from $\mathcal{D}$ to form an extremely biased subset $\mathcal{D}_1'$.
4: Train $f_{\boldsymbol{\theta}_2}$ on $\mathcal{D}_1'$ using ERM.
5: use $f_{\boldsymbol{\theta}_2}$ to infer bias labels for all samples in $\mathcal{D}$, and modeling joint distribution via hard confusion matrix.
 # Stage2: Sample weight inference
6: Infer class-conditional and marginal distribution over the bias cues using equation2.
7: Compute sample weights using Equation3, and 4 from the main manuscript.
 # Stage 3: Unbiased classifier training
8: Train classifier $f_{\boldsymbol{\phi}}$ on reweighted samples using standard ERM.
9: Select the best-performing $f_{\boldsymbol{\phi}}$ based on the highest worst-class accuracy on the validation set $\mathcal{D}_v$.

B EXPERIMENTAL SETUP

Datasets. The experiments are conducted on five benchmark datasets known to exhibit spurious correlations. Waterbirds, CelebA, CivilComments, and UrbanCars are real-world datasets in which each class is spuriously correlated with background, gender, certain demographic identities, or a combination of multiple shortcuts respectively. cMNIST and cCIFAR10 are synthetic ten-way classification tasks, where each class is spuriously linked to a specific color or noise pattern. For all datasets, we adopt the same train-validation-test split following Liu et al. (2021); Tsirigotis et al. (2023b) for fair comparison.

Training setup. For fair comparison, we adopt model architectures following Tsirigotis et al. (2023b); LaBonte et al. (2023b): a 3-hidden layer MLP for cMNIST, ResNet18 He et al. (2016) For cCIFAR10, ResNet50 He et al. (2016) for Waterbirds and CelebA, and BERT Devlin et al. (2019) for CivilComments. ResNet18 and ResNet50 are pretrained on ImageNet-1K, and BERT is pretrained on Book Corpus and English Wikipedia. No data augmentation is applied to cMNIST and CivilComments, while simple augmentations (random cropping and horizontal flipping) are used to the remaining datasets, following Ahuja et al. (2021). This ensures that the improvements we observed are attributed to the proposed methodology, rather than to data augmentations that could potentially nullify the bias attribute. For our method, both the initial bias learning and the bias enhancement learning span 20 epochs, and the final unbiased learning involves 5000 iterations across all datasets. Results are averaged over 5 random seeds, and for each seed, the best-performing model (the one with the highest worst-class accuracy on the validation set Tsirigotis et al. (2023b)) is selected. Unless otherwise stated, we repeat the "bias enhancement learning" process three times for FG-CCDB.

On Hyperparameters. All experiments were conducted on a single NVIDIA A40 GPU. The hyperparameters and optimization settings for the MST and FG-CCDB modules on each dataset are summarized in Table 6. Both modules share the same batch size, scheduler, optimizer, and optimizer hyperparameters. For cMNIST and CivilComments, no data augmentation is applied to either module, while for the remaining datasets, simple data augmentations (*i.e.,* ResizedCrop and HorizontalFlip) are applied only for the FG-CCDB module. All stages in MST are trained with the same *Epoch* number. In contrast to CCDB, our MST framework consists of at least two stages: the first stage provides an initial bias prediction, which is further refined by the subsequent stages.

The experimental results in the main manuscript (see Figure3(right)) show that selecting the parameter γ within the range of $1\% \leq \gamma \leq 20\%$ has a negligible impact on the final performance. Accordingly, we set $\gamma = 10\%$ across all datasets to ensure strong performance while maintaining low computational cost.

Table 6: The optimization setup for our FG-CCDB.

Dataset	Optimizer	Scheduler	LR	Batch size	Weight decay	{Epoch,Iter}	γ	Augmentation
cMNIST	Adam	None	1×10^{-2}	256	1×10^{-4}	$\{20, 5000\}$	0.1	None
cCIFAR10	Adam	None	1×10^{-5}	256	1×10^{-4}	$\{20, 5000\}$	0.1	ResizedCrop, HorizontalFlip
Waterbirds	Adam	None	1×10^{-5}	256	1×10^{-4}	$\{20, 5000\}$	0.1	ResizedCrop, HorizontalFlip
CelebA	Adam	None	1×10^{-5}	256	1×10^{-4}	$\{20, 5000\}$	0.1	ResizedCrop, HorizontalFlip
CivilComments	AdamW	Linear	1×10^{-5}	16	1×10^{-4}	$\{20, 5000\}$	0.1	None

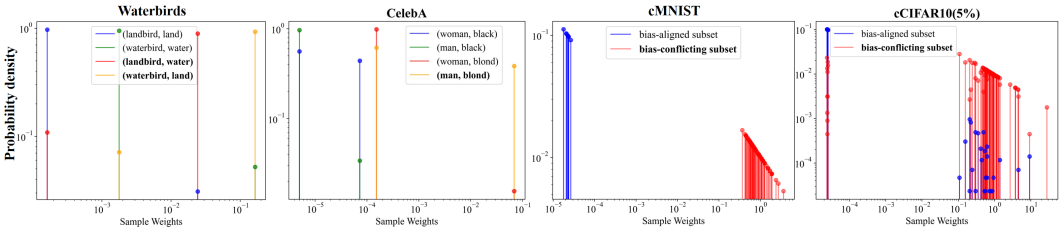


Figure 5: The distribution of the sample weights assigned by FG-CCDB within each mode on four datasets.

C THE SAMPLE WEIGHTS INFERRED BY OUR FG-CCDB

To assess whether our distribution-matching approach, FG-CCDB, effectively distinguishes minority modes from majority ones and assigns appropriate sample weights in the singular shortcut case, we analyze the distribution of inferred sample weights across different modes. The results on four datasets are summarized in Figure 5, with the minority modes highlighted in bold. As expected, samples from the majority modes are assigned low weights, typically concentrated below 0.01, while samples from minority modes receive significantly higher weights, clustered around 1. These results demonstrate that FG-CCDB successfully differentiates between majority and minority modes, and up-weights the latter in a balanced manner, aligning both class-conditional and marginal distributions.

D ADDITIONAL EXPERIMENTAL RESULTS AND DETAILS FOR URBANCARS

For UrbanCars, the class label corresponds to the car type (country or urban), while the spurious attributes consist of two shortcuts: the background (BG) and the co-occurring object (CoObj), both of which are also labeled as country or urban. The ground-truth group partition of the training data is shown in Figure6. The majority groups contain urban car images combined with urban backgrounds (*e.g.,* alleys) and urban co-occurring objects (*e.g.,* fire plugs), and vice versa for country car images. The remaining combinations constitute the minority groups. As shown in Li et al. (2023b), mitigating spurious correlations in datasets with multiple shortcuts presents a Whac-A-Mole dilemma: mitigating one shortcut often amplifies the model's reliance on the others.

Evaluation Metrics for the UrbanCars Dataset. Compared to datasets with a single shortcut, four new metrics are proposed for multi-shortcut scenarios to better evaluate performance across different shortcut combinations.

(i) In-Distribution Accuracy (I.D. Acc): This metric computes the weighted average of per-group accuracies, where the weights are proportional to each group's frequency in the training set (*i.e.,* its correlation strength, as shown in Figure6). Following the "average accuracy" definition in Sagawa et al. (2020), it reflects model performance when no group shift occurs.

Table 7: Classification performance on multi-shortcuts UrbanCars. In addition to our worst-group accuracy, the measurements following Li et al. (2023b) are also provided.

Methods	Given Condition	I.D. Acc	Gap due to shortcut			Urbancar(BG)		Urbancar(CoObj)		Urbancar	
			BG	CoObj	BG+CoObj	Mean	WGA	Mean	WGA	Mean	WGA
LfF	Yes	97.2	-11.6	-18.4	-63.2	-	-	-	-	-	-
JTT	Yes	95.9	-8.1	-13.3	-40.1	-	-	-	-	-	-
DebiAN	No	98.0	-14.9	-10.5	-69.0	-	-	-	-	-	-
ExMap	No	-	-5.9	-9.9	-30.7	93.2	71.4	93.2	79.2	-	-
FG-CCDB	None	92.98	**-4.17**	**-7.37**	**-4.9**	91.04$_{\pm 0.04}$	87.84$_{\pm 0.2.12}$	93.08$_{\pm 0.14}$	90.24$_{\pm 0.28}$	88.56$_{\pm 0.30}$	81.28$_{\pm 3.9}$

(ii) BG Gap: The drop in accuracy from the I.D. Acc to the accuracy on groups where the background (BG) is uncommon but the co-occurring object (CoObj) remains common (*cf.* the first yellow column in Figure6).

(iii) CoObj Gap: The drop in accuracy from the I.D. Acc to the accuracy on groups where the CoObj is uncommon but the BG remains common (*cf.* the second yellow column in Figure6).

(iv) BG+CoObj Gap: The drop in accuracy from the I.D. Acc to the accuracy on groups where both BG and CoObj are uncommon (*cf.* the red column in Figure6).

BG Gap and CoObj Gap measure the model's robustness to distribution shifts caused by each individual shortcut. BG+CoObj Gap evaluates robustness in the most challenging scenario, where both shortcuts are absent.

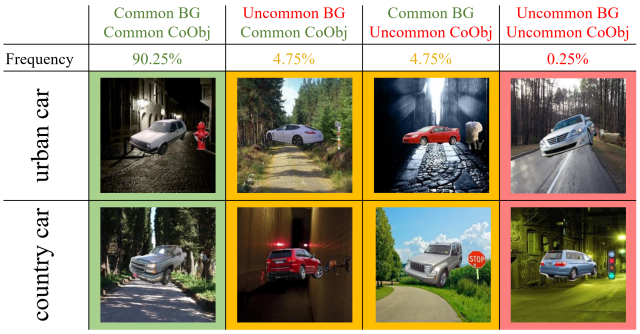


Figure 6: Unbalanced groups in the UrbanCars training set based on two shortcuts: background and co-occurring object (the figure is adopted from Li et al. (2023b))

Following Chakraborty et al. (2024), two variants of UrbanCars are constructed: (i) UrbanCars (BG), where only the background object serves as the spurious attribute; (ii) UrbanCars (CoObj), where only the co-occurring object serves as the spurious attribute.

We compare the worst-group accuracy (WGA) on these two variants plus the original one, as shown in Table7. Our method achieves significantly higher WGA than ExMap on both variants, further confirming our claim that FG-CCDB captures bias information through mode partitioning in a more general manner. This makes it applicable to both singular and multiple shortcut scenarios.

E Additional Discussions

R1W1: How iterative bias amplification improves minority-mode recall

In addition to our experimental results, the validation of MST is supported by the following research findings: (i) Easy-to-learn property of bias attributes Nam et al. (2020). ERM tend to overfit spurious correlations only when they are "easier" to learn than the desired core features. This property has been successfully exploited in many debiasing methods Nam et al. (2020); Pezeshki et al. (2024); LaBonte et al. (2023b); Zhao et al. (2025); Lee et al. (2023) to detect and highlight underrepresented bias-conflicting samples. Thus, the initial step of MST is well motivated. (ii) Removing bias-conflicting samples improves bias prediction. Prior works Lee et al. (2023); Ciranni et al. (2025) show that even a small number of bias-conflicting samples can severely degrade the estimation of bias-aligned vs. bias-conflicting partitions. In principle, if all bias-conflicting samples were removed, one could train a bias-capturing model that provides ideal learning signals for debiasing. These methods obtain a bias-amplified model either by explicitly removing bias-conflicting samples

Table 8: The F1-score of the smallest-mode prediction under different top high-confidence ratio β.

	cCIFAR10(5%)	Waterbirds	CelebA	UrbanCars
Bias-align ratio	95.00%	94.97%	51.72%	90.25%
$\beta = 30\%$	0.65	0.53	0.32	0.47
$\beta = 50\%$	0.72	0.64	**0.47**	0.62
$\beta = 70\%$	**0.79**	**0.67**	0.40	**0.64**
Adaptive	0.76	0.69	0.43	0.66

or by generating only bias-aligning samples. Our MST shares the same core insight but adopts a different mechanism: we use a multi-stage bias amplification process that progressively filters out bias-conflicting samples by selecting those with the highest confidence. (iii) Bias-aligned samples tend to have higher confidence. As revealed in Lee et al. (2023), bias attributes are easier to learn than intrinsic attributes; thus, ERM model assigns higher predicted probabilities to bias-aligned samples. This phenomenon has also been effectively used in works on GCE Zhang & Sabuncu (2018). Therefore, selecting top-confidence samples at each stage in MST is an effective strategy for filtering out bias-conflicting samples.

R1Q1: Why fix the top-50% high-confidence samples per-class for bias enhancement?

We denote by β the ratio used to select the top high-confidence samples for brevity. Our choice of $\beta = 50\%$ is based on a practical and widely observed property of spurious-correlation datasets. In typical settings, within each class, the bias-aligned partition is larger than the bias-conflicting partition; otherwise, spurious correlations would not arise, as pointed out in Ciranni et al. (2025). This implies that the bias-aligned partition occupies more than 50% of the samples in that class. Table 8 summarizes the smallest bias-aligned ratio across classes for each dataset. Except for CelebA, which has a value only slightly above 50%, the other datasets have ratios exceeding 90%. Therefore, when bias annotations are unavailable, selecting the top 50% high-confidence samples is highly likely to capture the bias-aligned partition while excluding bias-conflicting samples. We emphasize that this is an empirical principle rather than a strict theoretical guarantee. However, it is consistently supported by prior works on spurious correlations and by our empirical results.

To further address potential concerns regarding $\beta = 50\%$, we conduct experiments with alternative proportions (30% and 70%) and an adaptive version based on class-wise confidence distributions (assigning higher β to classes with higher average confidence). The F1-scores are shown in Table 8. Clearly, $\beta = 50\%$ represents a reasonable middle-ground option. For CelebA, which has a low bias-aligned ratio, $\beta = 50\%$ performs best, whereas for datasets with bias-aligned ratios above 90%, $\beta = 70\%$ yields the best performance. The adaptive strategy is primarily effective when the data exhibits noticeable class imbalance. We consider further exploration of this approach as promising future work.

R2W1: How iterative bias amplification improves minority-mode recall

Please refer to R1W1.

R2W2: comparison with recent label-free debiasing methods

We incorporate comparisons with recent label-free debiasing methods: DDB Ciranni et al. (2025), DaC Noohdani et al. (2024), and GERNE Asaad et al. (2025). DDB utilizes a diffusion model to generate bias-aligned data, which amplifies the bias reliance of the bias model and provides useful information for the debiasing process. DaC identifies the causal components of images using class activation maps from models trained with ERM. It then intervenes on the images by combining these components and retrains the model on the augmented data. Both DDB and DaC are specifically designed for image data. GERNE assumes that the difference between the gradients of two batches captures a debiasing direction and optimizes the model along this direction. The results are summarized in Table1, Table2 and Table4. Although DaC uses bias annotations during validation, its performance on CelebA remains significantly lower than ours. Our method demonstrates substantial advantages over GERNE and DDB across CelebA, CivilComments, and the multi-shortcut UrbanCars dataset. Notably, on UrbanCars, while DDB exhibits the smallest overall drops across different bias-conflicting scenarios, its base I.D. accuracy is the lowest among all compared methods.

R2W3: The performance on multi-bias scenarios The experiments on multi-bias (multi-shortcut) scenarios may have been overlooked. We conducted experiments on the UrbanCars dataset, which

contains multiple shortcuts (i.e., background and co-occurring objects). The corresponding results and discussion can be found in Section 4.1 and Table 2.

Overall, our method consistently achieves the best balance between high I.D. accuracy and minimal drops compared to other bias-agnostic methods, particularly on the challenging BG+CoObj generalization. It performs comparably to — or better than — methods that rely on bias annotations. These results confirm that our approach provides a general framework for handling multi-shortcut scenarios.

R2W4: whether FG-CCDB can compensate for imperfect bias predictions

We have shown the performance of FG-CCDB under different mode partition qualities in Figure 1(c), which may have been overlooked. By observing Figure 4(a), we find that repetition has a particularly strong effect on minority groups: performance increases significantly after the first repetition and then gradually converges. Accordingly, in Figure 1(c), the WGA obtained by subsequent FG-CCDB shows a similar trend: it jumps from a relatively low accuracy after the first repetition and then gradually converges to a stable value. We conclude that: (i) When MST provides poor mode partitioning ("repeat0"), the errors are significant, and FG-CCDB is affected by these errors, resulting in relatively low WGA. (ii) When MST provides acceptable mode partitioning (with a repetition count of 1 or higher), the WGA of FG-CCDB increases and shows only marginal improvement with further repetitions, even though the mode partition quality continues to improve. This indicates that FG-CCDB can compensate for imperfect mode partitions once the partition quality is sufficiently high.

R2Q1: how well the MST matches human labels? performance comparison results with the latest methods

Please refer to R3W3 and R3W4 for a quantitative evaluation of MST's performance. Please refer to R2W2 for a comparison with the latest label-free and generative model-based methods.

R3W1: Definition of 'mode' and whether major biases are captured by MST

We define the "mode" (s, y) as a black-box concept because our goal is for s to capture general and harmful bias information that humans may not preconceive Li et al. (2022b), rather than only physically interpretable attributes. The value s represents spurious signals that an ERM model prefers over core features and that consequently cause evaluation failures. We do not aim to model spurious attributes are not preferred by ERM and therefore do not lead to generalization errors. In this sense, model mistakes serve as indicators of harmful spurious correlations. Regarding the type of bias we focus on, we clarify that **our model is unlikely to fail to capture such harmful bias cues**. The reasons are as follows.

First, extensive prior works Nam et al. (2020); Pezeshki et al. (2024); LaBonte et al. (2023b); Zhao et al. (2025); Lee et al. (2023) operate under the widely accepted assumption that naive ERM tends to misclassify or produce low-confidence predictions on bias-conflicting samples. These studies demonstrate that ERM naturally learns spurious correlations, providing reliable learning signals for debiasing.

Second, for stronger theoretical grounding, we connect our idea to the Equal Opportunity Fairness (EOF) criterion Li et al. (2022b); Hardt et al. (2016) and show that our method is equivalent to find the bias cues that cause a classifier's predictions to strongly violate this fairness criterion, as detailed below.

Formally, a classifier f satisfies EOF criterion if:

$$\Pr\{\hat{y} = k | s = 0, y = k\} = \Pr\{\hat{y} = k | s = 1, y = k\} \tag{5}$$

where the LHS and RHS are the true positive rates (TPR) for negative ($s = 0$) and positive ($s = 1$) groups in target class $k \in \{1...K\}$. As noted in Li et al. (2022b), a significant TPR discrepancy between groups indicates that classifier f contains bias regarding s.

In our setting without bias annotations, we train an overfitted ERM and use its predictions s as a general bias cues. Specifically, given a dataset $\mathcal{D}$ with spurious correlations, where minority groups are non-empty and target labels are correct, we train an ERM model f on a small random subset of $\mathcal{D}$ and evaluate it on the full dataset, obtaining accuracy a.

- If $a = 100\%$, TPRs for each (s, y) pair resemble Figure 7(a). This implies that bias cues are not preferred and f likely relies exclusively on core features. No debiasing is needed.
- If $a < 100\%$, overfitting occurs, though to different degrees. The TPRs within each class show severe violations of the EOF criterion (*e.g.,* Figure 7(b) for class $k = 0$, $\Pr\{\hat{y} = 0|s = 0, y = 0\} \gg \Pr\{\hat{y} = 0|s \neq 0, y = 0\}$), indicating that f indeed captures and relies on the bias encoded in s.

Thus, in principle, as long as $a < 100\%$, our method leveraging ERM overfitting reliably captures harmful implicit bias cues. Unlike Li et al. (2022b), our approach directly identifies cues that maximally violate EOF without requiring interleaving optimization.

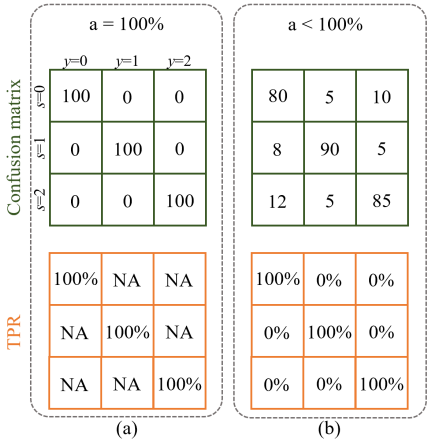


Figure 7: The hard confusion matrix and the TPRs. Take a 3-class classification task as an example, with each class contains 100 samples.

R3W2: The hyperparameter choices in MST

In fact, we have conducted ablation studies on γ in Figure 3(right) and discussed it in Section 4.4, which may have been overlooked. To further validate its robustness across datasets and bias strengths, we include additional results on UrbanCars. These results consistently show that $\gamma = 10\%$ serves as a sweet spot for maximizing the smallest-mode recall. Please refer to R1Q1 for our discussion regarding the use of the top 50% high-confidence samples.

R3W3: On MST's ability to capture complex biases in multi-shortcut scenarios

As we have pointed out in R3W1, we focus only on biases that are harmful — i.e., those that cause ERM models to overfit and make incorrect predictions — and our goal is to correct them. If the model overfits to "noise or irrelevant features" rather than physically interpretable biases, we treat such noise or irrelevant features as harmful bias and aim to balance them to improve ERM performance.

As demonstrated in Line 156 of the main manuscript, our model captures spurious cues that lead to overfitting and, consequently, incorrect predictions. These cues may correspond to interpretable shortcuts, combinations of multiple shortcuts, or entangled, uninterpretable patterns. Therefore, when multiple competing biases exist, MST can reveal the full bias structure, representing multiple competing biases within a single bias cue.

We have conducted experiments in Section 4.1 (Table 2) to demonstrate the effectiveness of our method in complex multi-shortcut scenarios, which may have been overlooked. For example, in UrbanCars, there are two competing shortcuts (background and co-occurring objects) and our method exhibits substantially less bias towards any specific background, co-object, or their combination, even outperforming methods that rely on multiple shortcut annotations.

Additionally, we compare the Recall of bias-conflicting modes on UrbanCars obtained by XRM, JTT, and our MST in Table 9. The results show that even under multi-shortcut conditions, our method successfully identifies bias-conflicting samples covering all minority groups, whereas XRM fails to capture group $(0, 1, 1)$, and JTT fails to capture groups $(0, 1, 0)$, $(0, 1, 1)$, and $(1, 1, 0)$.

Table 9: Recall of minority groups in UrbanCars predictions by MST, XRM, and JTT. Group (e_1, e_2, e_3): $e_1 = 0/1$ indicates urban/country car, $e_2 = 0/1$ indicates urban/country object, and $e_3 = 0/1$ indicates urban/country background.

	(0,0,1)	(0,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,1,0)
MST	**45.79%**	**58.42%**	**70.00%**	**100.00%**	**64.55%**	**28.57%**
XRM	41.05%	30.51%	0.00%	60.00%	10.12%	14.06%
JTT	0.53%	0.00%	0.00%	10.00%	0.53%	0.00%

R3W4: Quantitative Comparison of MST with XRM and DebiAN.

The comparison with XRM and DebiAN on final debiasing performance was already provided in Table 1 and Table 2 of the main manuscript (Section 4.1), which may have been overlooked. Similarly, the comparison with XRM and JTT on bias capturing was already presented in Figure4(b) and discussed in section 4.3, which also may have been overlooked. Theoretically, XRM trains its biased model on a random half of the training data, which contains far more bias-conflicting samples than ours, resulting in lower precision and recall on the smallest-mode. A similar explanation accounts for JTT's poor recall. DebiAN uses an alternating training scheme, where the classifier gradually mitigates biases during the discovery phase, making it difficult for its discoverer to reliably predict biases; therefore, we do not include DebiAN in the bias-capturing comparison.

To provide a more comprehensive evaluation, we have additionally included the F1-score in Figure 4(b). Our method consistently achieves the highest F1-score.

R3W5: F1-score to evaluate MST mode partitions.

Please refer to R3W4 for quantitative evaluation of MST-generated mode partitions.

The effect of the MST's prediction quality on the subsequent FG-CCDB was already provided in Figure1(c) and may have been overlooked. Please refer to R2W4 for a detailed discussion.

R3Q1: on further subdivision within modes or continuous weights.

We appreciate the reviewer's insightful suggestion. While a mode may contain potential substructures, our assumption of intra-mode homogeneity is not a theoretical requirement but a practical approximation, motivated by the following: (i) the "mode" definition is conditioned on both the predicted bias and the label (s, y). The auxiliary bias model partitions data according to the most dominant spurious patterns revealed by ERM overfitting. This ensures that samples assigned to the same mode share the most influential bias cues, which is sufficient for effective reweighting. In practice, these dominant bias cues account for the majority of generalization errors, while finer-grained variations within a mode have only marginal influence. (ii) Empirically, uniform per-mode weighting is stable and effective. We experimented with an alternative design (i.e., entropy-based intra-mode splitting that divides each mode into high-entropy and low-entropy subsets) but found it introduced noise and degrades performance. For example, on Waterbirds, WGA drops from 90.56% to 89.90%, and on cCIFAR10 with an extremely small bias-conflicting portion (the smallest group contains only 19 samples), performance drops from 55.28% to 50.18%. This suggests that finer intra-mode partitioning requires additional sub-bias cues to correctly guide matching, which are unavailable under the current setting.

FG-CCDB focuses on mode-level bias amplification guided by dominant shortcuts. Incorporating a more detailed internal structure is beyond the scope of this work. We therefore consider mode-level homogeneity a reasonable and empirically validated design trade-off, with finer-grained mode modeling left as future work.

R3Q2: Performance curves over additional iterations to demonstrate MST convergence.

We provide the mode partition and WGA results with additional repetition counts in Figure 8. The results show that when the number of repetitions exceeds 3, the improvement in mode-partition accuracy slows down and eventually converges to a stable point. Correspondingly, the WGA remains nearly unchanged once the repetition count is greater than 2.

This behavior is expected. As repetitions progresses, bias-conflicting samples are gradually filtered out, causing the bias-aligned ratio of the selected training subset to increase and eventually stabilize.

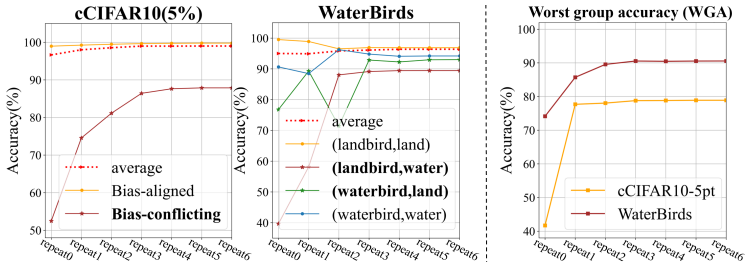


Figure 8: Mode prediction recall (left) and WGA under varying mode prediction quality (right) across repetitions of the "bias enhancement learning" procedure..

Table 10: The computation cost of compared methods on cCIFAR10.

	Our	ERM	uLA
Bias discovery	80 epochs	NA	500 epoch
Debiasing	5000 iters≈28 epochs	300 epoch	500 epochs

Table 11: The running time (hour) of MST, evaluated on a single NVIDIA A40 GPU.

	cCIFAR10	Waterbirds	CelebA	UrbanCars
MST	0.27h	0.35h	1.26h	0.14h

Once the learned bias model reaches a stable level of bias reliance, further top-confidence selection no longer changes the bias-aligned ratio, and the mode partition consequently remains unchanged.

R3Q3: Computational cost of MST.

To avoid misunderstanding, uLA is also a two stage method. Compared to single-stage training methods like ERM, the additional training time mainly comes from MST. However, this overhead is acceptable in practical applications for the following reasons: (i) In each MST stage, we use only 10%, 50%, 50%, and 50% of the training data, which significantly reduces the computational burden. (ii) We observe that the ERM model already exhibits strong bias reliance in the early training phase — a phenomenon widely reported in prior works. Therefore, we set a small number of epochs for each MST stage. The main computation cost are compared in Table 10, and the running-time of MST is summarize in Table 11.