Multi-Attribute Bias Mitigation via Representation Learning

Rajeev Ranjan Dwivedi[a,*,1], **Ankur Kumar**[a,1] **and Vinod Kumar Kurmi**[a]

[a]Indian Institute of Science Education and Research (IISER) Bhopal

Abstract.

Real-world images frequently exhibit multiple overlapping biases, including textures, watermarks, gendered makeup, scene-object pairings, etc. These biases collectively impair the performance of modern vision models, undermining both their robustness and fairness. Addressing these biases individually proves inadequate, as mitigating one bias often permits or intensifies others. We tackle this multi-bias problem with Generalized Multi-Bias Mitigation (GMBM), a lean two-stage framework that needs group labels only while training and minimizes bias at test time. First, Adaptive Bias-Integrated Learning (ABIL) deliberately identifies the influence of known shortcuts by training encoders for each attribute and integrating them with the main backbone, compelling the classifier to explicitly recognize these biases. Then Gradient-Suppression Fine-Tuning prunes those very bias directions from the backbone's gradients, leaving a single compact network that ignores all the shortcuts it just learned to recognize. Moreover we find that existing bias metrics break under subgroup imbalance and train–test distribution shifts, so we introduce Scaled Bias Amplification (SBA): a test-time measure that disentangles model-induced bias amplification from distributional differences. We validate GMBM on FB-CMNIST, CelebA, and COCO, where we boost worst-group accuracy, halve multi-attribute bias amplification, and set a new low in SBA—even as bias complexity and distribution shifts intensify—making GMBM the first practical, end-to-end multi-bias solution for visual recognition.

1 Introduction

In recent years, the remarkable success of machine learning models in image classification has been tempered by growing evidence of their vulnerability to biases in training data. Typically, these biases take the form of shortcuts—spurious correlations or unintended cues that models exploit to boost average performance at the expense of reliability and fairness [11, 35]. For instance, classifiers might learn to associate water backgrounds with boats [32], face recognition systems can amplify gender and skin-tone biases present in their training sets [3], and in medical imaging, COVID-19 diagnostic models have been shown to rely on dataset-specific artifacts—such as hospital tags or watermarks—leading to performance degradation when these cues are absent [6].

To counter these issues, a range of bias-mitigation approaches have been developed. Most of these techniques operate under the assumption of a single, known spurious cue and often demand privileged group annotations during training or validation [29, 26]. However, recent studies demonstrate that this simplification fails to capture the complexity of real-world datasets—such as ImageNet and popular facial attribute benchmarks—which typically harbor multiple simultaneously; unknown biases like watermarks, textures, and latent correlations. As a result, models trained under these conditions may exploit various spurious features, leading to unpredictable failures when any one of these cues is removed or altered. Although in-processing strategies like distributionally robust optimization (DRO) and last-layer retraining can improve worst-group performance for a solitary bias, they falter in multi-bias scenarios [29, 16]. Likewise, unsupervised methods that leverage training dynamics can uncover a dominant bias direction but lack the capacity to disentangle more than one bias simultaneously [15, 36].

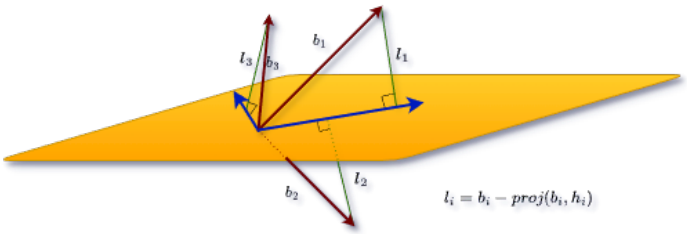

Figure 1. Gradient suppression fine tuning

In addition to this, real-world datasets often contain multiple overlapping biases that jointly erode model robustness [17, 10]. The problem becomes even more challenging when the efforts to suppress one shortcut shifts the reliance onto alternative spurious cues and even amplify it, giving rise to a "Whac-A-Mole" dilemma in bias mitigation as stated by Li *et al.* [21].

Despite practical significance, effective strategies for mitigating multiple interacting biases remain underexplored in the literature. To address multi-bias robustness in vision models, we propose Generalized Multi-Bias Mitigation (GMBM), a hyperparameter-light, two-stage framework that uses group annotations only during training to learn and then suppress multiple spurious-feature representations. Crucially, at inference time, GMBM relies *solely* on the debiased image feature—without bias labels, extra modules, or architectural changes. GMBM performs debiasing in two stages:

❶ **Adaptive Bias–Integrated Learning (ABIL).** For each known bias attribute j, we train a encoder whose penultimate-layer output captures that spurious signal. In parallel, the backbone's penultimate-layer output represents the core image feature. We

* Corresponding Author. Email: rajeev22@iiserb.ac.in.

[1] Equal contribution.

compute attention weights by applying a softmax over the cosine similarities between the image feature and each bias feature, then fuse the representation by adding the image feature to the weighted sum of bias features. Feeding this fused feature into the classification head forces the model to identify and explicitly discount each spurious channel, disentangling spurious features from task-relevant cues.

❷ **Gradient-Suppression Fine-Tuning.** We discard the bias encoders and fine-tune the backbone on clean image feature alone. To eliminate residual bias influence, we first project each bias feature onto the subspace orthogonal to the image feature, obtaining an orthogonal residual, as shown in Figure 1. We then use the standard cross-entropy loss with a penalty term that penalizes squared gradient component along each orthogonal residual, scaled by a penalty strength lambda. This enforces invariance to all known biases while preserving legitimate semantic information.

GMBM is evaluated on both a synthetic dual-shortcut benchmark (with controlled foreground/background color cues) and a real-world CelebA [24] and COCO [22] datasets. At inference time, only the debiased backbone feature is passed to the classifier for efficient deployment without further overhead. Across both settings, GMBM consistently outperforms single-bias baselines and prior multi-bias methods, improving worst-group accuracy by up to 8% and halving spurious bias amplification.

Our key contributions can be summarized as follows:

(1) We formalize multi-bias mitigation in vision models and critically analyze the limitations inherent in single-bias approaches. (2) We introduce GMBM, the first end-to-end two-stage framework—comprising ABIL and gradient suppression—that adaptively integrates and subsequently suppresses multiple bias representations. (3) We create multi-bias evaluation benchmark designed to capture realistic scenarios of intersecting biases. (4) We provide empirical evidence demonstrating that GMBM establishes a new state-of-the-art in robustness by significantly improving both unbiased and bias-conflicting accuracy while simultaneously decreasing spurious bias amplification.

2 Related Work

Bias Identification and Discovery. A key prerequisite for any mitigation strategy is understanding what biases a model has learned. Recent work has focused on uncovering spurious correlations and underperforming subgroups without explicit bias labels. For example, Eyuboglu *et al.* [10] leverages cross-modal embeddings and an error-aware model to pinpoint under-represented subgroups, while Singla *et al.* [32] use activation maps to generate Salient ImageNet, a dataset of core vs. spurious feature masks. GSCLIP [41] offers a training-free, dataset-level shift explanation using CLIP [27] embeddings, and generative approaches have been employed to discover unknown biases via latent space manipulation[19, 18]. Jain *et al.* [15] further distill failure modes with linear probes and CLIP[27] captions to explain model errors. While these methods excel at identifying single "shortcuts", they typically do not scale to settings where multiple, intersecting biases co-occur.

Debiasing in Single-Bias Scenarios.

A large body of work targets spurious correlations between a primary attribute and one secondary (bias) attribute. When bias labels are accessible, several strategies have been developed. These include robust optimization, which re-weights groups based on their performance [29]; adversarial training, which aims to suppress bias-related signals [7, 12]; and contrastive objectives, which explicitly work to separate examples where the bias aligns with the primary attribute from those where it conflicts [37, 34]. In situations where bias labels are not available, various label-free methods have emerged. These approaches, such as Learned-Mixin (LAD) [5], Environment Inference (EIIL) [4], Just Train Twice (JTT) [23], Learning from Failure (LfF) [25], CosfairNet [8], and contrastive-based debiasing techniques [9], aim to discover or approximate bias groups by leveraging model-based heuristics. While these methods have proven effective in addressing single spurious correlations, they typically operate under the assumption of *at most one dominant bias*. Consequently, their effectiveness is limited when multiple independent biases interact within the data [18].

Multi-Attribute Bias Mitigation via Representation Learning. Methods specifically designed to address the challenge of handling simultaneous, unknown biases have only recently begun to emerge. For example, Li *et al.* [20] proposes an iterative approach that assigns pseudo-labels to discover multiple biases and subsequently trains deconfounded models. Similarly, Whac-A-Mole [21] employs targeted augmentations to simulate a variety of bias types. However, both of these methods are primarily tailored to synthetic image benchmarks and rely on hand-crafted bias generators. In contrast, representation learning techniques offer a promising avenue for addressing an arbitrary number and variety of biases in natural settings. These techniques include invariant feature extraction, information-theoretic regularizers, and multi-view contrastive learning.

In this work, we address these challenges and introduce a unified representation learning-based framework that aims to mitigate multiple overlapping biases by a two-stage network. We present our novel method in the following section 3.

3 Methodology

Problem Formulation: We consider an N-way classification dataset

$$\mathcal{D} \;=\; \big\{\big(x^{(i)},\, y^{(i)},\, b_1^{(i)},\ldots,b_k^{(i)}\big)\big\}_{i=1}^{n},$$

where n is the total number of samples in the dataset, each input $x^{(i)}$ carries a ground-truth label $y^{(i)} \in \{1,\ldots,N\}$, also $b_1^{(i)},\ldots,b_k^{(i)}$ are k known bias attributes. Let $h^{(i)} = f_{\text{pen}}(x^{(i)}) \in \mathbb{R}^d$ denote the penultimate ("pen") representation extracted by the backbone. We seek a classifier

$$g : \mathbb{R}^d \;\to\; \{1,\ldots,N\}$$

that predicts $y^{(i)}$ accurately without exploiting any spurious attribute b_j. Concretely, we call b_j *spurious* if $H\big(Y \mid B_j\big) \approx 0$, where

$$H\big(Y \mid B_j\big) = -\mathbb{E}_{B_j}\Big[\sum_{y=1}^{N} P\big(Y=y \mid B_j\big)\,\log P\big(Y=y \mid B_j\big)\Big]$$

is the Shannon conditional entropy of the random label Y given attribute B_j. In settings with multiple such biases, our goal is to ensure that the model's decision remains correct even when each b_j is removed or contradicted, thus guaranteeing robustness across all known bias dimensions.

Building on this formulation, we now introduce our two-stage debiasing framework. First, in Section 3.1 we describe *Adaptive Bias–Integrated Learning (ABIL)*, which exposes and weights each known bias cue via a soft-attention mechanism to challenge the

model to discount spurious features. Then, in Section 3.2, we detail *Inference-Time Gradient Orthogonalization*, which fine-tunes the backbone to enforce invariance to any residual bias directions.

3.1 Adaptive Bias–Integrated Learning (ABIL)

Our goal in ABIL is to *isolate* and then *attenuate* spurious bias cues, while preserving and emphasizing task-relevant information. We achieve this via a two-stream architecture and a dynamic fusion mechanism, justified as follows:

(1) Bias Encoders. We allocate an encoder per known bias attribute j, trained to predict $b_j^{(i)}$ from X_i denoted as f_{pen}^j. By dedicating separate parameters to each bias, we encourage the network to carve out *distinct* subspaces in $\mathbb{R}^d$ that specialize in capturing that particular spurious signal. This explicit disentanglement simplifies later suppression.

(2) Penultimate-Layer Representations: Both the main backbone and each bias encoder output features from their penultimate (pen) layer:

$$h^{(i)} = f_{\text{pen}}\big(x^{(i)}\big), \quad b_j^{(i)} = f_{\text{pen}}^j\big(x^{(i)}\big),$$

We use the penultimate layer activations, as they capture a rich, high-dimensional embedding of the input abstractions enough to encode semantic concepts [28, 40, 33, 14], not yet collapsed into class scores. This makes h_i and each b_i^j suitable for measuring alignment and for controlled fusion, without the interference of the final decision boundary.

(3) Soft-Attention over Bias Channels. Instead of naively concatenating or summing all b_i^j, we compute:

$$\alpha_j^{(i)} = \frac{\exp\big(\cos(h_i, b_i^j)\big)}{\sum_{m=1}^{k} \exp\big(\cos(h_i, b_i^m)\big)}, \qquad \cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

This ensures that bias encoders whose representations align more strongly with the current image feature h_i receive higher weight, reflecting which spurious cues are *most likely* to influence the backbone. Furthermore, the use of softmax ensures that the attention mechanism remains fully differentiable, which enables a co-adaptive learning process during training. In this process, the backbone learns to produce features that *de-emphasize* subspaces that are subsequently down-weighted. Additionally, employing softmax has two key benefits: it prevents unbounded amplification of bias vectors and yields an interpretable probability distribution over spurious factors.

(4) Bias-Modulated Fusion. We form the composite feature

$$h_i' = h_i + \sum_{j=1}^{k} \alpha_j^{(i)}\, b_i^j \quad \longrightarrow \mathcal{L}_{\text{ABIL}} = \text{CE}\big(g(h_i'),\, y_i\big)$$

via a residual additive connection. In training we then optimize the standard *cross-entropy* (CE) loss over our composite features. The backbone's original feature h_i remains intact, while bias vectors act as *perturbations* highlighting spurious directions. By presenting classifier with both clean and bias-accentuated signals, the network naturally learns to reward reliance on h_i's invariant components and penalize shortcuts through b_i^j.

(5) Training with h_i'. The fused feature h_i' is fed into the final classification head during training. This is done in an adversarial framing where the model must solve the task *in the presence of explicit bias cues*. Consequently, the model internalizes robust features in h_i that remain predictive even when spurious channels are subsequently suppressed (in our inference-time orthogonalization).

Together, with this algorithmic setup, ABIL does not merely hide bias information, but systematically *identifies*, *weights*, and then *challenges* it leading to a backbone representation that cleanly separates task-relevant structure from known spurious factors.

3.2 Gradient-Suppression Fine-Tuning

After ABIL has equipped the backbone with bias-aware features, we perform a brief fine-tuning step to *guarantee* that no residual spurious components influence the final prediction. We reuse the bias encoders trained in ABIL to extract b_i^j alongside the backbone feature h_i. The residual vectors

$$l_i^j = b_i^j - \frac{h_i^\top b_i^j}{\|h_i\|^2} h_i \tag{1}$$

capture the pure bias directions that ABIL exposed. By penalizing

$$\sum_{j=1}^{k} (\nabla_{h_i} L_{ce} \cdot l_i^j)^2$$

we suppress any gradient component that would steer h_i back into bias subspaces, thereby enforcing *provable invariance* to all known spurious channels. Details of our two-stage algorithm (including full pseudocode) and all network/backbone architectures plus hyperparameter settings are provided in the Supplementary Material.

Overall Objective. Our framework consists of two sequential stages—adaptive bias integration followed by gradient-based fine-tuning (Fig. 2).

Stage ①: Adaptive Bias Integration. During training, each image feature is augmented by a weighted sum of bias-specific features. We compute attention weights by comparing the image feature to each bias feature using cosine similarity, then normalize these scores via softmax. The resulting fused representation is optimized with standard cross-entropy loss against the true labels.

Stage ②: Gradient-Based Fine-Tuning. In the second phase, we drop all bias encoders and fine-tune using only the original image feature. To eliminate any remaining bias influence, each bias feature is projected onto the space orthogonal to the image feature, and we add a regularization term that penalizes any gradient component aligned with these projections. A fixed penalty weight of 0.01 ensures a balance between bias suppression and classification accuracy, yielding a single, robust model.

Having described our two-stage debiasing, we now turn to how we choose and control bias attributes in our evaluation. In Section 4.1 we summarize three benchmark datasets and specific bias attributes used and give the detailed algorithm 1 for our method.

4 Experiments & Evaluation Metric

4.1 Datasets

We evaluate our method on diverse set of multi-attribute bias datasets, each designed to probe model robustness under complex spurious correlations:

- **FB-CMNIST** [30]: This is a synthetic extension of the Colored MNIST dataset[1], where each digit is modified with two spurious biases—background color and foreground (digit) color—enabling the study of models under multiple simultaneous correlations.

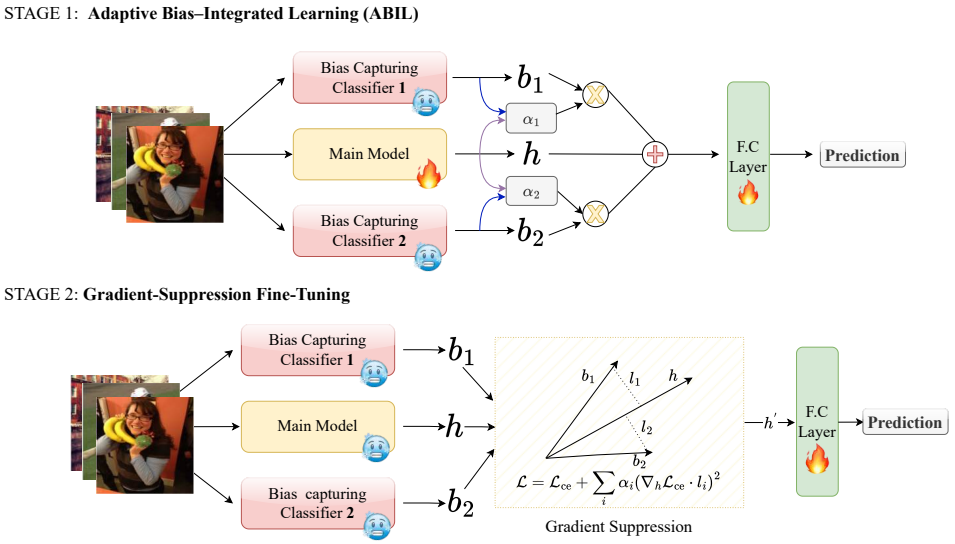

Figure 2. Overview of the Generalized Multi-Bias Mitigation (GMBM) Framework. Stage 1 (ABIL)—Multiple bias encoders are trained alongside the main model to explicitly capture known spurious features. Their outputs are integrated with the backbone feature, forcing the classifier to learn to discount bias-aligned directions. Stage 2 (GSFT)—after discarding the bias encoders, the model is fine-tuned with a gradient-penalty term that suppresses residual alignment with bias directions. This results in a compact, bias-invariant backbone used at inference time.

- **CelebA** [24]: CelebA comprises real-world face images annotated with 40 binary attributes, including gender, hair color, age, and various facial accessories. We employ this dataset to study multi-attribute bias in a binary gender classification setting. Concretely, we take the Male attribute as our target label and select *Wearing_Lipstick* and *Heavy_Makeup* as spurious attributes since they are known to correlate strongly with the female gender and can introduce significant bias [18].
- **COCO** [22]: We construct a custom dataset based on the COCO2017 dataset [22] to study gender and object-based biases following the work in [38]. Gender labels are derived from image captions by scanning for gender-indicative keywords. To introduce bias labels, we define two object bias categories using COCO instance annotations. **Bias Category 1** includes various sports and outdoor objects objects while **Bias Category 2** includes various indoor and kitchen objects.

Implementation Details: For FB-CMNIST, a simple 7-layer convolutional neural network was used as both the model backbone and the bias-capturing classifier. For the CelebA and COCO datasets, the standard ResNet-18 architecture was employed. The CMNIST model was trained for 80 epochs, followed by 10 epochs of fine-tuning. Similarly, the ResNet-18 models were trained for 20 epochs and fine-tuned for an additional 10 epochs. In both cases, we used an initial learning rate of 10^{-3} for the main training phase and 10^{-4} for fine-tuning, both steps were employed with the Adam optimizer. A batch size of 128 was used across all experiments. Table 1 summarizes the hyperparameters used.

4.2 Evaluation Metrics

In our problem setting, "multiple bias" attributes may be correlated with one another, jointly influencing the model's predictions and amplifying bias. Prior work has largely relied on two key metrics—unbiased precision and bias-conflicting accuracy—to evaluate the effectiveness of bias mitigation algorithms. While useful in simple scenarios dominated by a "single bias", these metrics fall short in capturing the nuanced effects of multiple, inter-dependent biases. In such settings, evaluating individual unbiased or bias-conflicting accuracies may overlook how well the model disentangles and mitigates the joint influence of correlated biases.

Table 1. Key hyper-parameters used throughout all experiments.

Symbol	Meaning	Default
T_1	ABIL epochs	6
T_2	Suppression epochs	3
β	Bias–loss weight	0.2
λ_{supp}	Gradient penalty	10^{-2}
d	Embedding width	128

Multi-Attribute Bias Amplification (MABA). Zhao *et al.* [39] observe that multiple bias attributes can interact to skew predictions in ways that single-bias metrics cannot account for. MABA addresses this by examining every combination of bias attributes $m \in \mathcal{M}$ together with each target label $g \in \mathcal{G}$. For each pair (m, g), one counts the number of training samples in which m and g co-occur, denoted $\text{co-occur}_{\text{train}}(m, g)$, and normalizes over all labels to obtain the training bias

$$\mathcal{B}_{train} = \text{bias}_{\text{train}}(m, g) = \frac{\text{co-occur}_{\text{train}}(m, g)}{\sum_{g' \in \mathcal{G}} \text{co-occur}_{\text{train}}(m, g')}.$$

An analogous procedure on the model's predicted labels yields the test-time bias distribution $\text{bias}_{\text{test}}(m, g)$. To focus on meaningful spurious associations, Δ_{gm} is defined as:

$$\Delta_{gm} = \mathbb{1}\{\mathcal{B}_{train} > 1/|\mathcal{G}|\} \left(\mathcal{B}_{test} - \mathcal{B}_{train}\right),$$

Δ_{gm} thereby ignores the attribute–label pairs whose training frequency does not exceed the uniform prior. The overall amplification is then summarized by the average absolute shift

$$\mathcal{X} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{g \in \mathcal{G}} |\Delta_{gm}|$$

Algorithm 1 Generalized Multi-Bias Mitigation (GMBM)

Require: Dataset $\mathcal{D} = \{(x, y, b_1, \ldots, b_k)\}$; backbone f with penultimate hook f_{pen}; classifier g; bias encoders $f_{\text{pen}}^{1..k}$; bias-classifier heads $\hat{b}_{1..k}$; weights $\beta, \lambda_{\text{supp}}$

Ensure: Debiased backbone f_{debiased}

Stage 1: Adaptive Bias-Integrated Learning (ABIL)

for all minibatch $(X, Y, B_1, \ldots, B_k)$ **do**
  $H \leftarrow f_{\text{pen}}(X)$
  for $j = 1$ **to** k **do** ▷ bias representations
    $B_j \leftarrow f_{\text{pen}}^{j}(X)$
  end for
  $\alpha_j \leftarrow \mathrm{softmax}\big(\cos(H, B_j)\big)$
  $H' \leftarrow H + \sum_{j=1}^{k} \alpha_j B_j$
  $\mathcal{L}_{\text{main}} \leftarrow \mathrm{CE}\big(g(H'), Y\big)$
  $\mathcal{L}_{\text{bias}} \leftarrow \sum_{j=1}^{k} \mathrm{CE}\big(\hat{b}_j(B_j), B_j\big)$
  Update $f, g, f_{\text{pen}}^{j}, \hat{b}_j$ w.r.t. $\mathcal{L}_{\text{main}} + \beta\, \mathcal{L}_{\text{bias}}$
end for
Freeze $f_{\text{pen}}^{1..k}$; Discard $\hat{b}_{1..k}$

Stage 2: Gradient-Suppression Fine-Tuning

for all minibatch (X, Y) **do**
  $H \leftarrow f_{\text{pen}}(X)$
  for $j = 1$ **to** k **do**
    $B_j \leftarrow f_{\text{pen}}^{j}(X)$
    $L_j \leftarrow B_j - \dfrac{\langle H, B_j \rangle}{\|H\|^2} H$
  end for
  $\mathcal{L}_{\text{ce}} \leftarrow \mathrm{CE}\big(g(H), Y\big)$
  $\mathcal{L}_{\text{grad}} \leftarrow \sum_{j=1}^{k} \big(\nabla_H \mathcal{L}_{\text{ce}} \cdot L_j\big)^2$
  Update f_{pen}, g w.r.t. $\mathcal{L}_{\text{ce}} + \lambda_{\text{supp}}\, \mathcal{L}_{\text{grad}}$
end forreturn $f_{\text{debiased}} \leftarrow f$

and the variance $\mathrm{Var}(\Delta_{gm})$, together forming the Multi$_{\text{MALS}}$ [39] metric. A well-debiased model will produce test-time bias distributions that closely mirror the training distributions, resulting in low $\mathcal{X}$ (minimal average amplification) and low variance (consistent mitigation across all attribute combinations). By capturing joint, distributional shifts rather than isolated biases, MABA provides a nuanced measure of how spurious correlations propagate through the model.

Problems with MABA: A key limitation of the MABA metric is that it treats all group-attribute pairs equally, irrespective of their frequency within the training dataset. This becomes problematic when certain combinations are severely underrepresented. In such cases, the bias amplification estimate is unstable and sensitive to noise due to the small sample size, and these rare, practically insignificant combinations can disproportionately influence the metric. This equal weighting scheme results in skewed interpretations, especially in imbalanced datasets. To overcome this limitation, we give two variants of the MABA metric:

❶ **Min-Support MABA:** Exclude the group-attribute pairs that do not meet a minimum support threshold (τ) in the training data. This ensures that amplification is computed only for statistically reliable combinations, resulting in more robust and interpretable bias estimates.

$$\Delta_{gm}^{\text{min-supp}} = \mathbb{1}\left\{\text{co\_occur}_{train}(g, m) > \tau\right\} \cdot (\mathcal{B}_{test} - \mathcal{B}_{train})$$

$$\text{Min-Support MABA} = \frac{1}{|\mathcal{M}|} \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} \left|\Delta_{gm}^{\text{min-supp}}\right|$$

❷ **Weighted MABA:** Here we introduce a frequency-based weighting scheme, where each group-attribute pair is weighted in proportion to its occurrence in the training set. This ensures that more representative groups have greater influence on the final amplification score, leading to a more balanced and realistic measurement of model-induced bias.

$$\Delta_{gm}^{\text{weighted}} = w_{gm} \cdot [\mathcal{B}_{\text{test}}(g, m) - \mathcal{B}_{\text{train}}(g, m)]$$

$$M_w = \frac{1}{|\mathcal{M}|} \sum_{g,m} |\Delta_{gm}^{\text{weighted}}|, \quad w_{gm} = \frac{c_{gm}}{\sum_{g',m'} c_{g'm'}}$$

where $c_{gm} = \text{co\_occur}_{train}(g, m)$.

While the Min-Support MABA variant addresses challenges related to severely under-represented subgroups, and Weighted MABA tackles both under- and over-representation, a critical limitation remains: the original MABA metric and the proposed variants falter in the presence of distribution shift between training and test datasets. If the joint distribution of attributes and groups differs between training and test sets, Δ_{gm} reflects not just model bias but also dataset shift. The MABA metric may combine distributional shifts between the training and test sets with model-induced bias amplification. When the underlying joint distribution of attributes m and groups g differs between the training set ($P_{\text{train}}(m, g)$) and the test set ground truth ($P_{\text{test, actual}}(m, g)$), the difference $\Delta_{gm} = \text{bias}_{\text{test}}(m, g) - \text{bias}_{\text{train}}(m, g)$ may reflect variations in dataset distributions rather than solely the model's tendency to amplify training biases, diminishing the metric's ability to isolate effects specific to the model.

4.3 Scaled Bias Amplification (SBA)

To quantify how much our model amplifies existing group–attribute biases on unseen data (test set), we compute SBA using only test-set counts. Let $C_{g,m}^{\text{pred}}$ and $C_{g,m}^{\text{actual}}$ be the number of test instances of group g with attribute m predicted by the model and observed in the ground truth, respectively. We first convert these into proportions:

$$\text{pred}_{g,m} = \frac{C_{g,m}^{\text{pred}}}{\sum_{g'} C_{g',m}^{\text{pred}}}, \qquad \text{act}_{g,m} = \frac{C_{g,m}^{\text{actual}}}{\sum_{g'} C_{g',m}^{\text{actual}}},$$

and define the amplification gap

$$\Delta_{g,m} = \text{pred}_{g,m} - \text{act}_{g,m}.$$

To prevent noisy estimates for rare attributes from dominating the score, we weight each gap by

$$\omega_{g,m} = \frac{1}{\sqrt{\sum_{g'} C_{g',m}^{\text{actual}}} + \epsilon},$$

where:

- The $\sqrt{\cdot}$ yields sublinear scaling so that rarer attributes receive higher weight, but not excessively so—balancing sensitivity to true bias amplification against variance from small-sample noise.
- $\epsilon > 0$ ensures numerical stability (avoiding division by zero) and caps the maximum weight when counts are extremely low.

Finally, SBA is the average weighted absolute gap over all groups G and attributes M:

$$\text{SBA} = \frac{1}{|G|\,|M|} \sum_{g \in G} \sum_{m \in M} \omega_{g,m} \left|\Delta_{g,m}\right|.$$

This formulation yields a single, interpretable scalar: sensitive enough to detect bias amplification yet stable under rare-group noise.

Key Benefits of SBA: SBA delivers a more robust and interpretable measure of model-induced bias amplification by relying exclusively on test-set comparisons and a subgroup-aware weighting scheme.

- **Robustness to Distribution Shifts.** SBA uses only test-set ground truth, avoiding the train-test co-occurrence mismatches that destabilize MABA. As the bias ratio q increases from 0.90 to 0.99, SBA for ERM grows steadily (0.265→0.611→1.077), whereas MABA's variance explodes (>850 at $q = 0.99$). In contrast, SBA for BAdd and GMBM remains low and stable across all q (Table 6).
- **Interpretability and Subgroup Sensitivity.** A test-set scaling factor ω_{gm} weights each group–label pair by its frequency, preventing majority-group dominance and ensuring rare, bias-conflicting instances contribute proportionately to the final score.
- **Stable Variance Profiles.** SBA's variance across subgroups rises with ERM's increasing bias (0.233→1.000 →2.748) but stays low for BAdd and GMBM ($\leq$ 0.258) even under extreme training skew (Table 2), confirming its resilience to imbalance.

This combination of test-set centric evaluation and subgroup-aware weighting makes SBA a more dependable fairness metric for real-world, imbalanced, or multi-attribute datasets, overcoming key limitations of prior metrics. Additional comparisons and ablations appear in the Supplementary Material.

Table 2. Variance of SBA (↓ better) on FB-CMNIST test set demonstrating GMBM's stability under increasing spurious correlation.

Model	$q = 0.90$	$q = 0.95$	$q = 0.99$
ERM	0.233	1.000	2.748
BAdd	0.013	0.048	**0.256**
GMBM	**0.010**	**0.042**	0.258

5 Results

We evaluate GMBM against multiple baselines on three benchmarks: FB-CMNIST with controlled bias ratios ($q = 0.90, 0.95, 0.99$), CelebA (male vs. gender-correlated attributes), and a custom COCO split (sports/outdoor vs. kitchen/indoor object biases). We report (i) unbiased and bias-conflicting accuracies, (ii) Multi-Attribute Bias Amplification (MABA) variants, and (iii) the proposed Scaled Bias Amplification (SBA) metric.

5.1 Unbiased & Bias-Conflicting Accuracy

Table 3 compares results of our method on FB-CMNIST dataset with varying bias ratios of 0.90, 0.95, and 0.99. GMBM maintains the highest unbiased accuracy across all bias ratios, achieving 96.1% ($q = 0.90$), 91.5% ($q = 0.95$), and 74.6% ($q = 0.99$). This outperforms the strongest baseline (**BAdd** [30]) by +0.5, +2.5, and +5.1%, demonstrating robustness even under extreme bias ratio (Table 3).

On CelebA dataset, for male classification with *Wearing_Lipstick* and *Heavy_Makeup* as spurious attributes, GMBM attains the best unbiased accuracies (96.7% and 95.5%) and bias-conflicting accuracies (94.5% and 92.0%), improving over highest baseline by up to +1.5% in conflict cases (Table 4). On the COCO dataset, with sports-object and kitchen-object biases, GMBM achieves unbiased accuracies of 83.78% and 83.19%, and bias-conflicting accuracies of 83.85% and 82.35%, surpassing all baselines.(Table 5).

Table 3. Unbiased accuracy on the FB-CMNIST test set for various methods under bias ratios q, highlighting GMBM's superior robustness compared to single and multi-bias baselines. (↑ is better)

Method	$q = 0.90$	$q = 0.95$	$q = 0.99$
Vanilla	82.5	57.9	25.5
BC-BB [13]	80.9	66.0	40.9
EnD [34]	82.5	57.5	25.7
FLAC [31]	84.4	63.1	32.4
FairKL [2]	87.6	61.6	42.0
BAdd [30]	95.6	89.0	69.5
GMBM	**96.1**	**91.5**	**74.6**

Table 4. Unbiased and bias-conflicting accuracies on a 30% CelebA test split for gender classification. (↑ is better)

Method	WearingLipstick		HeavyMakeup	
	Unbiased	Bias-conflicting	Unbiased	Bias-conflicting
Vanilla	93.2	89.1	92.0	84.7
FairKL [2]	82.7	74.7	84.4	77.9
BC-BB [13]	91.6	85.8	89.7	81.8
EnD [34]	95.1	91.0	92.3	85.3
FLAC [31]	95.4	91.6	93.2	87.2
BAdd [30]	95.8	93.0	94.9	91.0
GMBM	**96.7**	**94.5**	**95.5**	**92.0**

Table 5. Unbiased and bias-conflicting accuracies on the COCO validation set for gender classification with sports/outdoor vs. kitchen/indoor object biases, showing that GMBM outperforms prior methods on both majority and minority bias-conflicting groups. (↑ is better)

Method	Sports Object		Kitchen Object	
	Unbiased	Bias-conflicting	Unbiased	Bias-conflicting
Vanilla	70.81	64.61	73.20	67.36
FairKL [2]	76.32	67.11	74.35	76.90
EnD [34]	77.11	70.97	82.38	77.34
FLAC [31]	80.02	77.31	80.22	79.95
BAdd [30]	81.28	77.81	82.91	**83.05**
GMBM	**83.78**	**83.85**	**83.19**	82.35

5.2 Multi-Attribute Bias Amplification (MABA) Variants

To assess how models amplify existing group–attribute biases, we compute two MABA variants: *Min-Support* and *Weighted* MABA.

On FB-CMNIST, under distribution shift between train/test, MABA metrics exhibit high variance even for debiased models (variance $>$ 850 at $q = 0.99$), highlighting their instability when co-occurrence statistics diverge (Table 6). On the CelebA, GMBM achieves the lowest Min-Support MABA mean (0.67) and variance (0.61), improving over ERM (mean 0.74, var 0.85) and BAdd (mean 0.90) (Table 7). Similarly, on COCO dataset, GMBM again yields the most reliable bias estimates: Min-Support MABA mean 0.73 (vs. 8.53 for ERM) and variance 1.66 (vs. 96.27), demonstrating consistent bias suppression (Table 8). Note that f*or weighted MABA, the variance is not reported, since group-wise scaling distorts the interpretability of variance.*

Table 6. Comparison of Base MABA, Min-Support and Weighted MABA means and variances for ERM, BAdd [30], and GMBM on FB-CMNIST across bias ratios q, illustrating the high volatility of traditional MABA metrics under distribution shift and GMBM's relative consistency.

q	MABA Variant	Metric	ERM	BAdd [30]	GMBM
0.90	Base MABA	Mean	14.34	16.78	16.69
		Variance	393.01	519.94	513.69
	Min Support	Mean	14.24	16.67	16.68
		Variance	393.90	513.88	513.42
	Weighted	Mean	10.98	12.69	12.69
		Variance	-	-	-
0.95	Base MABA	Mean	14.41	12.41	12.38
		Variance	384.42	381.58	388.78
	Min Support	Mean	14.34	12.43	12.29
		Variance	378.11	382.13	379.09
	Weighted	Mean	10.71	7.75	7.79
		Variance	-	-	-
0.99	Base MABA	Mean	14.54	22.82	26.66
		Variance	388.18	679.15	857.12
	Min Support	Mean	14.69	22.82	26.64
		Variance	385.48	679.72	853.92
	Weighted	Mean	10.35	16.62	18.84
		Variance	-	-	-

Table 7. Performance of MABA variants on CelebA 30% test split

MABA Variant	Metric	ERM	BAdd [30]	GMBM
Base MABA	Mean	0.74	0.90	0.67
	Variance	0.85	1.05	0.61
Min support	Mean	0.74	0.90	0.67
	Variance	0.85	1.05	0.61
Weighted	Mean	0.86	1.03	0.70
	Variance	-	-	-

Table 8. MABA metrics on COCO-validation set

MABA Variant	Metric	ERM	BAdd [30]	GMBM
Base MABA	Mean	8.549	9.46	7.88
	Variance	628.50	634.83	625.08
Min support	Mean	8.53	3.53	0.73
	Variance	96.27	14.65	1.66
Weighted	Mean	8.84	2.60	0.35
	Variance	-	-	-

Despite these improvements, the high variability of MABA metric under distribution shift motivates our SBA metric.

5.3 Scaled Bias Amplification (SBA)

SBA quantifies how much a model's predictions exaggerate the correlation between different groups and attributes when evaluated on a test set. Unlike some other metrics, it is calculated using only the test data, which makes it more reliable when the biases appearing in the training data are different from the test data. It also uses a weighting system to ensure that rare combinations of groups and attributes are not ignored, providing a more balanced view of bias across different subgroups.

Under ERM, SBA increases sharply with bias ratio ($0.26 \rightarrow 0.61 \rightarrow 1.07$), whereas GMBM remains low and stable ($0.05 \rightarrow 0.11 \rightarrow 0.32$), confirming effective mitigation of spurious amplification on the FB-CMNIST dataset (Table 9). On CelebA, GMBM achieves SBA 0.35 vs. ERM 0.37 and BAdd 1.39; on COCO, SBA is 0.10 for GMBM vs. 0.84 (ERM) and 0.15 (BAdd). Across all settings, GMBM consistently yields the lowest, consistent, and interpretable SBA scores, underscoring its superior ability to prevent bias amplification in unseen data (Table 9).

Table 9. SBA scores for ERM, BAdd [30], and GMBM across FB-CMNIS, CelebA, and COCO datasets, showcasing GMBM's effectiveness at maintaining low and stable bias amplification on unseen data. (↓ better)

Dataset	Bias Ratio	ERM	BAdd [30]	GMBM
CMNIST	0.90	0.26	0.05	**0.05**
	0.95	0.61	0.11	**0.10**
	0.99	1.07	0.32	**0.32**
CelebA	–	0.37	1.39	**0.35**
COCO	–	0.84	0.15	**0.10**

GMBM not only attains state-of-the-art unbiased and bias-conflicting accuracies across synthetic and real-world benchmarks, but also demonstrably curtails multi-attribute bias amplification—both under traditional MABA variants and our more robust SBA metric. These results validate GMBM's effectiveness in learning and suppressing multiple spurious features. Alternative attention-weighting schemes, full breakdowns by bias ratio, and latent-space diagnostics—appears in the Supplementary Material.

6 Discussion & Future Work

Integrating multiple bias representations via attention–weighted fusion steers the backbone toward the most influential shortcuts, driving SBA down to 0.05 on FB-CMNIST and 0.10 on COCO while simultaneously boosting bias-conflicting accuracy. Yet GMBM still relies on group labels for every known bias attribute at training time a reasonable assumption for curated benchmarks, but onerous in many real-world pipelines where such annotations are unavailable. A second limitation is its sensitivity to biases that are nearly inseparable from the target label: on CelebA, SBA of 0.35 suggests that gender remains partly entangled with make-up cues despite mitigation.

Future work can relax these constraints by augmenting ABIL with unsupervised bias discovery. One avenue is to cluster systematic failure modes or latent shortcut directions uncovered by linear probes, echoing the approach of Jain et al. [15]; the discovered prototypes could be fed back into the same gradient-suppression stage to remove unknown biases without extra labels. Complementarily, an online gating mechanism that attenuates any feature whose gradient persistently correlates with emerging shortcut clusters would let GMBM adapt post-deployment. These extensions would broaden applicability while preserving the label-light character of the framework.

7 Conclusion

We present GMBM, a two-stage framework that (i) learns to expose and attenuate multiple spurious cues via attention-based bias integration and (ii) enforces invariance through gradient-suppression fine-tuning. In tandem, we introduce the *Scaled Bias Amplification (SBA)*: a test-time metric that quantifies the extent to which a model exaggerates group–attribute correlations under distributional shifts, while normalizing for subgroup over- and under-representation. Our experiments on synthetic and real-world image-classification benchmarks show that GMBM delivers state-of-the-art unbiased and bias-conflicting accuracies while dramatically reducing bias amplification. By tackling multi-attribute spurious correlations in a label-light, inference-efficient way, we advance fairness in vision models and pave the way for adaptive, domain-aware debiasing strategies.

References

[1] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh. Learning de-biased representations with biased representations. In *International conference on machine learning*, pages 528–539. PMLR, 2020.

[2] C. A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, and P. Gori. Unbiased supervised contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023.

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[5] L. Darlow, S. Jastrzębski, and A. Storkey. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv preprint arXiv:2011.11486*, 2020.

[6] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021.

[7] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *ICCV*, pages 15087–15096, 2021.

[8] R. R. Dwivedi, P. Kumari, and V. K. Kurmi. Cosfairnet:a parameter-space based approach for bias free learning. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. URL https://papers.bmvc2024.org/0738.pdf.

[9] Z. et al. Contrastive adapters for foundation model group robustness. *NeurIPS*, 2022.

[10] S. Eyuboglu, M. Varma, K. Saab, J.-B. Delbrouck, C. Lee-Messer, J. Dunnmon, J. Zou, and C. Ré. Domino: Discovering systematic errors with cross-modal embeddings. *ICLR*, 2022.

[11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

[12] S. Gong, X. Liu, and A. K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 330–347. Springer, 2020.

[13] Y. Hong and E. Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

[14] E. Hosseini and E. Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930, 2023.

[15] S. Jain, H. Lawrence, A. Moitra, and A. Madry. Distilling Model Failures as Directions in Latent Space. In *International Conference on Learning Representations*, 2023.

[16] P. Kirichenko, P. Izmailov, and A. G. Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *ICLR*, 2023.

[17] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, and I. Mosseri. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[18] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

[19] Z. Li and C. Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979, 2021.

[20] Z. Li, A. Hoogs, and C. Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *ECCV*, pages 270–288. Springer, 2022.

[21] Z. Li, I. Evtimov, A. Gordo, C. Hazirbas, T. Hassner, C. C. Ferrer, C. Xu, and M. Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *CVPR*, pages 20071–20082, 2023.

[22] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[23] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *ICML*, pages 6781–6792. PMLR, 2021.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[25] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

[26] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, 2020.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[28] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[29] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*, 2020.

[30] I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou. Badd: Bias mitigation through bias addition. *arXiv preprint arXiv:2408.11439*, 2024.

[31] I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou. FLAC: Fairness-Aware Representation Learning by Suppressing Attribute-Class Associations . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(02):1148–1160, Feb. 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3487254. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2024.3487254.

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

[33] E. A. Stanley, R. Souza, M. Wilms, and N. D. Forkert. Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. *EBioMedicine*, 111, 2025.

[34] E. Tartaglione, C. A. Barbano, and M. Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021.

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

[36] C. Tsirigotis, J. Monteiro, P. Rodriguez, D. Vazquez, and A. Courville. Group robust classification without any group information. *NeurIPS*, 2023.

[37] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022.

[38] D. Zhao, A. Wang, and O. Russakovsky. Understanding and evaluating racial biases in image captioning, 2021. URL https://arxiv.org/abs/2106.08503.

[39] D. Zhao, J. Andrews, and A. Xiang. Men also do laundry: Multi-attribute bias amplification. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42000–42017. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhao23a.html.

[40] Z. Zhao, Y. Ziser, and S. B. Cohen. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.847. URL https://aclanthology.org/2024.emnlp-main.847/.

[41] Z. Zhu, W. Liang, and J. Zou. Gsclip: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*, 2022.