Men Also Do Laundry: Multi-Attribute Bias Amplification

Dora Zhao [1] **Jerone T. A. Andrews** [2] **Alice Xiang** [1]

Abstract

The phenomenon of *bias amplification* occurs when models amplify training set biases at test time. Existing metrics measure bias amplification with respect to single annotated attributes (e.g., `computer`). However, large-scale datasets typically consist of instances with multiple attribute annotations (e.g., {`computer`, `keyboard`}). We demonstrate models can learn to exploit correlations with respect to multiple attributes, which are not accounted for by current metrics. Moreover, we show that current metrics can give the erroneous impression that little to no bias amplification has occurred as they aggregate positive and negative bias scores. Further, these metrics lack an ideal value, making them difficult to interpret. To address these shortcomings, we propose a new metric: *Multi-Attribute Bias Amplification*. We validate our metric's utility through a bias amplification analysis on the COCO, imSitu, and CelebA datasets. Finally, we benchmark bias mitigation methods using our proposed metric, suggesting possible avenues for future bias mitigation efforts.

1. Introduction

Despite their intent to faithfully depict the world, visual datasets are undeniably subject to historical and representational biases (Suresh & Guttag, 2021; Jo & Gebru, 2020). Left unchecked, dataset biases are invariably learned by models, especially when they are sources of efficient features for supervised learning on a given dataset (Gebru et al., 2021). For example, an image captioning model can learn to generate gendered captions by exploiting contextual cues without ever "looking" at the person in the image (Hendricks et al., 2018). Reliance on spurious correlations is undesirable since these learned associations do not always hold (Sagawa et al., 2019; Geirhos et al., 2020). More significantly, these associations risk not only perpetuating harmful social biases but also *amplifying* them (Zhao et al., 2017).

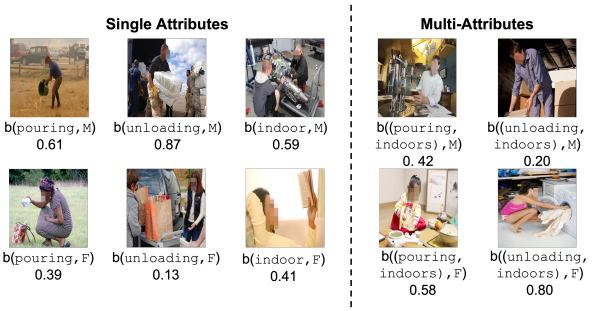

Figure 1. Bias scores (i.e., gender ratios) of the verbs `pouring` and `unloading` as well as location `indoors` in imSitu. While imSitu is skewed male (`M`) for the single attributes, the multi-attributes (e.g., {`pouring`, `indoors`}) are skewed female (`F`). Face pixelization is employed for privacy purposes.

The phenomenon of *bias amplification* refers to when a model compounds the inherent biases of its training set at test time (Zhao et al., 2017). Bias amplification has been studied across many tasks (Zhao et al., 2017; Ramaswamy et al., 2021; Wang et al., 2020b; Choi et al., 2020; Jia et al., 2020; Leino et al., 2019; Wang & Russakovsky, 2021; Hirota et al., 2022; Wang et al., 2019; Renduchintala et al., 2021). Following Zhao et al. (2017), we focus on multi-label classification. While there are metrics (Zhao et al., 2017; Wang & Russakovsky, 2021; Wang et al., 2019) that measure bias amplification in multi-label classification, they only consider the amplification between a single annotated attribute (e.g., `computer`) and a group (e.g., `female`). However, existing large-scale visual datasets often have multiple attributes per image (e.g., {`computer`, `keyboard`}). For example, 78.8% of images in the Common Objects with Context (COCO) (Lin et al., 2014) dataset contain more than one annotated attribute.

More importantly, considering multiple attributes can reveal additional nuances not present when considering only single attributes. In imSitu (Yatskar et al., 2016), individually the verb `unloading` and location `indoors` are skewed `male` (Fig. 1). However, when considering {`unloading`, `indoors`} in conjunction, imSitu is skewed `female`. Sig-

[1]Sony AI, New York [2]Sony AI, Tokyo. Correspondence to: Dora Zhao <dora.zhao@sony.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

nificantly, men tend to be pictured unloading *packages* outdoors whereas women are pictured unloading *laundry* or *dishes* indoors. Even when men are pictured indoors, they are unloading *boxes* or *equipment* as opposed to *laundry* or *dishes*. Models can similarly leverage correlations between a group and either single or multiple attributes simultaneously.

Multi-Attribute Bias Amplification Metric. We propose *multi-attribute bias amplification*, extending two previously proposed metrics (Zhao et al., 2017; Wang & Russakovsky, 2021).[1] Our new metric evaluates bias amplification arising from single and multiple attributes. We are the first to study multi-attribute bias amplification, highlighting that models exploit correlations between multiple attributes and group labels. We also address the issue that aggregated bias amplification metrics include summing positive and negative values. These values can cancel each other out, ostensibly presenting a smaller amount of amplification than what exists. Finally, as opposed to prior metrics which lack a clear ideal value, our metric is more interpretable.

Empirical Bias Amplification Analysis. Using our proposed metric, we compare the performance of multi-label classifiers trained on COCO (Lin et al., 2014), imSitu, and CelebA (Liu et al., 2015), standard benchmarks for bias amplification metrics (Zhao et al., 2017; Wang et al., 2019; Wang & Russakovsky, 2021; Hirota et al., 2022; Ramaswamy et al., 2021). As case studies, we consider bias amplification w.r.t. gender expression (COCO and imSitu), as well as hair color (CelebA) which represents the first bias amplification benchmark analysis for non-binary groups. We empirically demonstrate that on average bias amplification arising from single attributes is smaller than from multi-attributes. Thus, if we were only to consider individual attributes, not only would we obscure the nuance of understanding that multi-attributes provide but also potentially understate bias amplification.

Benchmarking Bias Mitigation Methods. Finally, we benchmark different bias mitigation methods (Zhao et al., 2017; Wang et al., 2020b; Agarwal et al., 2020) on our metric and existing bias amplification metrics. While prior works have demonstrated that models will learn to exploit different spurious correlations if one is mitigated (Li et al., 2023; 2022) we are the first to demonstrate that mitigation methods for single attribute bias can actually increase multi-attribute bias amplification. This further emphasizes the importance of our new metric as the magnitude of bias amplification is likely being underreported using single attribute metrics.

[1]Our code is available at `https://github.com/SonyResearch/multi_bias_amp`.

2. Related Work

Dataset Bias. Dataset bias is a well-studied problem in computer vision (Torralba & Efros, 2011; DeVries et al., 2019; Buolamwini & Gebru, 2018; Zhu et al., 2014; Birhane & Prabhu, 2021). Datasets are particularly predisposed to biases reflecting social inequities and disparities between true distributions and their digitized representations (Suresh & Guttag, 2021; Jo & Gebru, 2020). For example, datasets have been found to be demographically imbalanced (Buolamwini & Gebru, 2018; Yang et al., 2020; Zhao et al., 2021; Dulhanty & Wong, 2019; Paullada et al., 2021), with a particular lack of representation of females and individuals with darker skin tones. Further, there are visual differences in the way individuals from different groups are represented or interact with objects in the images (Wang et al., 2020a; Zhao et al., 2017; 2021). In our work, we focus on how dataset biases are compounded by trained models.

Measuring Fairness. A perfectly accurate model is unable to satisfy certain fairness metrics e.g., statistical parity (Dwork et al., 2012) as fairness constraints are not met in the ground-truth data (Wang & Russakovsky, 2021). In contrast, bias amplification metrics do not make recommendations as to the ideal underlying data distribution. By taking into account the correlations in the training set, bias amplification metrics instead capture additional biases introduced by a model. Bias amplification metrics can thus distinguish between cases when learned correlations are being over- or under-predicted, unlike fairness metrics such as TPR difference and FNR difference (Hardt et al., 2016).

Measuring Bias Amplification. Beyond reproducing biases, machine learning models have also been found to amplify them (Zhao et al., 2017). Hall et al. (2022) empirically show that bias amplification is influenced by dataset bias, model capacity, and training schema. To quantify bias amplification, Zhao et al. (2017) measure the difference in ground-truth object-group co-occurrences in the training set and test set co-occurrences predicted by a model. Building on this, Wang & Russakovsky (2021) propose a *directional* bias amplifcation metric to disentangle bias arising from attribute prediction versus group prediction. Others focus on *leakage* (Wang et al., 2019; Hirota et al., 2022), i.e., the change in a classifier's ability to predict group membership from ground-truth labels versus a model's predictions. Our work extends co-occurrence-based metrics (Zhao et al., 2017; Wang & Russakovsky, 2021). We are the first to consider how multi-attributes impact bias amplification and propose a novel metric to quantify this phenomenon.

Mitigating Bias Amplification. Bias amplification mitigation methods focus on either the dataset or model. Dataset-level mitigation strategies tend to center on the use of genera-

tive adversarial networks (GANs) (Ramaswamy et al., 2021; Sattigeri et al., 2019; Sharmanska et al., 2020) and counterfactuals (Kaushik et al., 2019; Wang & Culotta, 2021) to augment training sets.

More recent work (Agarwal et al., 2022) instead propose resampling strategies to address spurious correlations. Model-level strategies include corpus-level constraints (Zhao et al., 2017), adversarial debiasing (Wang et al., 2019), and domain independent training (Wang et al., 2020b). Similar to previous works (Shrestha et al., 2022; Zhao et al., 2022), we benchmark different mitigation methods. We demonstrate that existing strategies do not mitigate bias amplification from multiple attributes.

3. Multi-Attribute Bias Amplification

In this section, we introduce our multi-attribute bias amplification metric. Throughout this paper, we scale all metrics by a factor of 100.

Identifying bias. We denote by $\mathcal{G} = \{g_1, \ldots, g_t\}$ and $\mathcal{A} = \{a_1, \ldots, a_n\}$ a set of t group membership labels and n attribute labels, resp. Let $\mathcal{M} = \{m_1, \ldots, m_\ell\}$ denote a set of ℓ sets, containing all possible combinations of attributes, where m_i is a set of attributes and $|m_i| \in \{1, \ldots, n\}$.[2] Note $m \in \mathcal{M} \iff \text{co-occur}(m, g) \geq 1$ in both the ground-truth training set and test set, where co-occur(m, g) is the number of times m and g co-occur. We extend Zhao et al. (2017)'s single-attribute bias score to a multi-attribute setting such that the bias score of $m \in \mathcal{M}$ w.r.t. $g \in \mathcal{G}$ is defined as:

$$\text{bias}_{\text{train}}(m, g) = \frac{\text{co-occur}(m, g)}{\sum_{g' \in \mathcal{G}} \text{co-occur}(m, g')}, \tag{1}$$

where co-occur(m, g) denotes the number of times the combination of attributes m and group membership label g co-occur in the training set.

Evaluating undirected multi-attribute bias amplification. We define our undirected multi-attribute bias amplification metric as:

$$\text{Multi}_{\text{MALS}} = X, \text{Var}(\Delta_{gm}) \tag{2}$$

where

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

and

$$\Delta_{gm} = \mathbb{1}\left[\text{bias}_{\text{train}}(m, g) > |\mathcal{G}|^{-1}\right] \cdot \left(\text{bias}_{\text{test}}(m, g) - \text{bias}_{\text{train}}(m, g)\right).$$

[2]For example, if $\mathcal{A} = \{a_1, a_2, a_3\}$, then $\mathcal{M} = \{\{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}\}$.

Here $\text{bias}_{\text{test}}(m, g)$ denotes the bias score using the test set predictions of m and g. $\text{Multi}_{\text{MALS}}$ measures both the mean and variance over the change in bias score, Δ, from the training set ground truths to test set predictions. By definition, $\text{Multi}_{\text{MALS}}$ only captures group membership labels that are positively correlated with a set of attributes, i.e., due to the constraint that $\text{bias}_{\text{train}}(m, g) > |\mathcal{G}|^{-1}$.

Evaluating directional multi-attribute bias amplification. Let $\hat{m}$ and $\hat{g}$ denote a model's prediction for attribute group, m, and group membership, g, resp. Further, without loss of generality, let $P_{\text{train}}(m = 1)$ denote the probability of a training set sample containing the attribute set m, and $P_{\text{train}}(g = 1)$ denote the probability of a training set sample containing the group membership label g. We define our directional multi-attribute bias amplification metric as:

$$\text{Multi}_{\rightarrow} = X, \text{Var}(\Delta_{gm}) \tag{3}$$

where

$$X = \frac{1}{|\mathcal{G}||\mathcal{M}|} \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} y_{gm} \left|\Delta_{gm}\right| + (1 - y_{gm}) \left|-\Delta_{gm}\right|,$$

$$y_{gm} = \mathbb{1}[P_{\text{train}}(g = 1, m = 1) > P_{\text{train}}(g = 1) P_{\text{train}}(m = 1)],$$

and

$$\Delta_{gm} = \begin{cases} P_{\text{test}}(\hat{m} = 1 | g = 1) - P_{\text{train}}(m = 1 | g = 1) \\ \text{if measuring } G \rightarrow M \\ P_{\text{test}}(\hat{g} = 1 | m = 1) - P_{\text{train}}(g = 1 | m = 1) \\ \text{if measuring } M \rightarrow G \end{cases}$$

Unlike $\text{Multi}_{\text{MALS}}$, $\text{Multi}_{\rightarrow}$ captures both positive and negative correlations, i.e., regardless of whether $\text{bias}_{\text{train}}(m, g) > |\mathcal{G}|^{-1}$. Moreover, similar to the directional bias amplification metric proposed by Wang & Russakovsky (2021), $\text{Multi}_{\rightarrow}$ takes into account group membership base rates and disentangles bias amplification arising from the group influencing the attribute(s) prediction (i.e., $\text{Multi}_{G \rightarrow M}$), versus amplification from the attribute(s) influencing the group prediction (i.e., $\text{Multi}_{M \rightarrow G}$). See Appendix B for details on the relationship between the metrics.

Relation to existing single-attribute metrics. In Eqs. (2) and (3), if we only report X with $|\pm\Delta_{gm}| = \text{sgn}(\pm\Delta_{gm})|\pm\Delta_{gm}|$ and $m_i \in \mathcal{M} \iff |m_i| = 1$, then our metric $\text{Multi}_{\text{MALS}}$ reduces to Zhao et al. (2017)'s undirected bias amplification metric $\text{BiasAmp}_{\text{MALS}}$ and $\text{Multi}_{\rightarrow}$ reduces to Wang & Russakovsky (2021)'s directional bias amplification metric $\text{BiasAmp}_{\rightarrow}$. Refer to Appendix A for formal definitions of $\text{BiasAmp}_{\text{MALS}}$ and $\text{BiasAmp}_{\rightarrow}$.

Table 1. Multi-attribute, m, bias scores w.r.t. group $g = 1$ using ground-truth, $\text{bias}_{\text{train}}(m, g)$, and predicted, $\text{bias}_{\text{test}}(m, g)$, labels. We report the 95% confidence interval over five random assignments of g.

m	$\text{bias}_{\text{train}}(m, g)$	$\text{bias}_{\text{test}}(m, g)$
$\{a_1, a_2\}$	0.80	0.92 ± 0.1
$\{a_1, a_3\}$	0.49	0.50 ± 0.0
$\{a_2, a_3\}$	0.80	0.99 ± 0.0
$\{a_1, a_2, a_3\}$	0.94	1.00 ± 0.0

4. Metric Advantages

We compare our multi-attribute bias amplification metric with single-attribute metrics (Zhao et al., 2017; Wang & Russakovsky, 2021), using the MNIST dataset (Deng, 2012) to underscore the three main advantages of our metric: (1) our metric accounts for co-occurrences with multiple attributes; (2) negative and positive values do not cancel each other out; and (3) our metric is more interpretable.

Setup. We perform multi-label classification on a synthetically manipulated MNIST. For simplicity, we convert the task to binary classification (Hall et al., 2022) such that half of the classes are arbitrarily assigned to group $g = 0$ or $g = 1$. For the attributes, per image, we set a combination of three corner pixels (a_1: top left, a_2: bottom left, a_3: top right) to white. Thus, each image has a corresponding label $y = [g, a_1, a_2, a_3]$, where $a_i \in \{0, 1\}$ corresponds to a pixel being colored black (0) or white (1). We train a LeNet-5 (LeCun et al., 1989) for 50 epochs using SGD with batch size 32, momentum 0.9, and learning rate 10^{-3}. We average over five random group assignments and report the 95% confidence interval.

Advantage 1: Our metric accounts for co-occurrences with multiple attributes. If a model, for example, learns the combination of a_1 and a_2, denoted $\{a_1, a_2\}$, are correlated with g, it can exploit this correlation, potentially leading to bias amplification. By limiting the measurement to single attribute co-occurrences, $\text{BiasAmp}_{\text{MALS}}$ and $\text{BiasAmp}_{\rightarrow}$ do not account for amplification arising from co-occurrences with multiple attributes.

To illustrate this, we manipulate MNIST so that the dataset is perfectly balanced w.r.t. to single attributes, i.e., $\text{bias}_{\text{train}}(a_i, g) = 0.5$ ($\forall i \in [1, 2, 3]$), but skewed for multiple attributes. For example, although $\text{bias}_{\text{train}}(a_1, g) = \text{bias}_{\text{train}}(a_2, g) = 0.5$, the bias score for the combination of $\{a_1, a_2\}$ and g_1 is $\text{bias}_{\text{train}}(\{a_1, a_2\}, g_1) = 0.8$. Tbl. 1 shows the results using our trained models which achieve an mAP of 89.0 ± 2.6. Bias amplification is 0.0 ± 0.0 for all three single-attribute metrics. However, as shown in Tbl. 1, the bias scores calculated w.r.t. multiple attributes has increased. Therefore, bias has been amplified but is not being captured by existing metrics. Significantly, by iterating over all $m \in \mathcal{M}$ our proposed metric accounts for amplification from both single attributes (i.e., $|m| = 1$) and multiple attributes (i.e., $|m| > 1$). Thus, we capture previously unidentified attributes exhibiting amplification. While existing metrics report amplification values close to 0, our multi-attribute metric returns 9.2 ± 2.2, 0.3 ± 0.1, 0.2 ± 0.1 for $\text{Multi}_{\text{MALS}}$, $\text{Multi}_{G \rightarrow M}$, and $\text{Multi}_{M \rightarrow G}$, resp.

Advantage 2: Negative and positive values do not cancel each other out. Existing metrics aggregate over the difference in bias scores for each individual attribute. Suppose there is a dataset with two annotated attributes a_1 and a_2. It is possible that $\Delta_{ga_1} \approx -\Delta_{ga_2}$ for $\text{BiasAmp}_{\text{MALS}}$ or equivalently the difference in bias scores have opposite signs for $\text{BiasAmp}_{\rightarrow}$. In such cases, bias amplification would be approximately 0, which gives the impression little to no bias amplification has occurred.

To give a concrete example, we arbitrarily set a_1, a_2, and a_3 in MNIST to 0 with a probability of 0.7, 0.2, and 0.4, resp. The model achieves an mAP of 85.2 ± 9.9. One of the models results in $\text{BiasAmp}_{\text{MALS}} \approx 0.0$, suggesting no bias amplification has occurred. However, upon closer inspection, the bias scores for individual attributes are $\Delta_{ga_1} = 0.61$ and $\Delta_{ga_2} = -0.60$. Wang & Russakovsky (2021) recognize this limitation and suggest returning disaggregated results for each attribute-group pair. However, disaggregated values are difficult to interpret and make comparing models cumbersome if not infeasible. In contrast, our metric uses the absolute values of differences. Doing so ensures positive and negative bias amplifications over all attribute-group pairs do not cancel each other out. This allows us to report a single aggregated value, which is easier to interpret than disaggregated values per attribute-group pair.

Advantage 3: Our metric is more interpretable. There is a lack of intuition as to the *ideal* bias amplification value. One interpretation is that smaller values are more desirable. This becomes less clear when values are negative, as occurs in several bias mitigation works (Wang et al., 2020b; Ramaswamy et al., 2021). Negative bias amplification indicates bias in the predictions is in the opposite direction than that in the training set. However, this is not always ideal. First, there often exists a trade-off between performance and smaller bias amplification values. Second, high magnitude negative bias amplification may lead to erasure of certain groups. For example, in imSitu, $\text{bias}_{\text{train}}(\texttt{typing, F}) = 0.52$. Negative bias amplification signifies that the model underpredicts (`typing, F`), which could reinforce negative gender stereotypes (Zhao et al., 2017).

Instead, we may want to minimize the distance between the bias amplification value and 0. This interpretation offers the

advantage that large negative values are also not desirable. However, a potential dilemma occurs when interpreting two values with the same magnitude but opposite signs, which is a value-laden decision and depends on the system's context. Additionally, under this alternative interpretation, Adv. 2 becomes more pressing as this suggests we are interpreting models as less biased than they are in practice.

Our proposed metric is easy to interpret. Since we use absolute differences, the ideal value is unambiguously 0. Further, reporting variance provides intuition as to whether amplification is uniform across all attribute-group pairs or if particular pairs are more amplified.

5. Bias Amplification Analysis

We now analyze the advantages of our proposed metric by evaluating bias amplification when group membership is balanced w.r.t. to single attributes.

5.1. Setup

In our experiments, we focus on COCO, imSitu, and CelebA, which contain multiple attributes per image and are frequently used in bias amplification analyses (Zhao et al., 2017; Wang et al., 2019; Ramaswamy et al., 2021). First, for COCO, group membership is binary gender expression, i.e., $\{\texttt{female}, \texttt{male}\}$, and attributes correspond to objects. To obtain the group labels, we follow Zhao et al. (2017) and use the provided captions. We only consider objects occurring > 100 times with either group, leading to 52 objects in total. Second, for imSitu, group membership is binary gender expression, and attributes correspond to verbs and location. We derive group labels from the gendered agent terms. We only consider the 361 verbs that occur > 5 with each group and have a binary location label (i.e., $\{\texttt{indoor}, \texttt{outdoor}\}$). Finally, for CelebA, group membership is non-binary hair color, i.e., $\{\texttt{black hair}, \texttt{blond hair}, \texttt{brown hair}\}$, and attributes correspond to other annotated physical characteristics in the dataset. We only consider the 23 physical characteristics that occur > 500 times with each group.

To balance the datasets w.r.t. single attributes, $\forall(a, g) \in \mathcal{A} \times \mathcal{G}$ we greedily oversample, until the bias score $\text{bias}_{\text{train}}(a, g) \in [|\mathcal{G}|^{-1} \pm \epsilon]$.[3] We train "balanced" models on each of the three datasets using ResNet-50 architectures (He et al., 2016) initialized from weights learned on ImageNet (Russakovsky et al., 2015), where the classification layer has been replaced such that it jointly predicts group membership and attributes. Refer to Appendix C for more details.

[3]For COCO and imSitu, $\epsilon = 0.025$. For CelebA, $\epsilon = 0.07$, as we must balance across a larger number of groups.

Table 2. We report multi-attribute bias amplification with the variance in brackets when varying $|m_i|$, the minimum number of attributes in a combination. $|m_i| \geq 1$ includes biases from single and multi-attributes. We report 95% confidence interval over five models trained using random seeds for COCO (a), imSitu (b), and CelebA (c).

(a) COCO	$\lvert m_i\rvert \geq 2$	$\lvert m_i\rvert \geq 1$
$\text{Multi}_{\text{MALS}}$	$22.3 \pm 0.7, [4.6 \pm 0.1]$	$21.9 \pm 0.2, [4.5 \pm 0.1]$
$\text{Multi}_{M \to G}$	$22.7 \pm 0.3, [12.9 \pm 0.2]$	$22.2 \pm 0.3, [13.0 \pm 0.0]$
$\text{Multi}_{G \to M}$	$0.3 \pm 0.0, [0.0 \pm 0.0]$	$0.3 \pm 0.0, [0.0 \pm 0.0]$
(b) imSitu	$\lvert m_i\rvert \geq 2$	$\lvert m_i\rvert \geq 1$
$\text{Multi}_{\text{MALS}}$	$18.0 \pm 0.3, [3.0 \pm 0.1]$	$9.4 \pm 0.2, [1.6 \pm 0.1]$
$\text{Multi}_{M \to G}$	$14.5 \pm 0.2, [4.1 \pm 0.2]$	$13.0 \pm 0.1, [3.2 \pm 0.1]$
$\text{Multi}_{G \to M}$	$0.1 \pm 0.0, [0.0 \pm 0.0]$	$0.1 \pm 0.0, [0.0 \pm 0.0]$
(c) CelebA	$\lvert m_i\rvert \geq 2$	$\lvert m_i\rvert \geq 1$
$\text{Multi}_{\text{MALS}}$	$23.2 \pm 0.4, [2.3 \pm 0.1]$	$23.1 \pm 0.4, [2.3 \pm 0.1]$
$\text{Multi}_{M \to G}$	$5.5 \pm 0.0, [0.0 \pm 0.0]$	$5.5 \pm 0.0, [0.0 \pm .0.0]$
$\text{Multi}_{G \to M}$	$0.6 \pm 0.0, [0.1 \pm 0.0]$	$0.6 \pm 0.0, [0.1 \pm 0.0]$

5.2. Experimental Analysis

We analyze the different bias amplification metrics when using balanced models. First, we examine the effect of including multiple attributes. We then analyze the metrics when taking the absolute value versus the raw differences. Finally, we take a closer look into the variance of bias amplification scores per attribute.

More bias arises from multiple attributes. To examine the effect of including multi-attributes, we vary the minimum number of attributes in a combination (i.e., $|m|$). In Tbl. 2, we consider all possible combinations of attributes for $|m| \geq 1$ (i.e., all attribute sets with at least a single attribute) or $|m| \geq 2$ (i.e., only attribute sets with at least two attributes).

Across all datasets and metrics, the bias amplification when $|m| \geq 2$ is greater than or equal to that when $|m| \geq 1$, underscoring the importance of considering multiple attribute groups as detailed in Adv. 1. For COCO and imSitu, $\text{Multi}_{\text{MALS}}$ is greater when $|m| \geq 2$ than when $|m| \geq 1$, which implies that the mean bias amplification arising from single attributes is lower than that from multiple attributes. We note for CelebA, bias amplification is approximately equal when $|m| \geq 1$ and $|m| \geq 2$, indicating the amount of bias being amplified due to single attributes is negligible compared to the amount of bias being amplified due to multiple attributes (i.e., $|m| \geq 2$).

When considering disaggregated values, there is considerably higher bias arising from certain singular verbs, such as `constructing`, `reading`, and `vacuuming`, for $M \to G$ prediction. Finally, we note for single attributes, larger bias amplification occurs with attributes that co-occur with males, but more for females when considering multiple attributes. See Appendix G for more qualitative results.

Table 3. Performance, existing metrics, and our proposed multi-attribute metrics on balanced versions of COCO, imSitu and CelebA. There are three versions of the metrics: calculated using raw differences from training to test set (a), absolute value of differences (b), and the variances of these differences (c). We report the 95% confidence interval over five models trained using random seeds.

(a) Raw	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
COCO	53.8 ± 0.1	-1.9 ± 0.2	-16.3 ± 0.2	-1.5 ± 0.2	-1.4 ± 0.4	-0.0 ± 0.0	-0.1 ± 0.0
imSitu	67.0 ± 0.1	0.3 ± 0.1	-2.4 ± 0.3	0.7 ± 0.1	-0.5 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
CelebA	78.2 ± 0.1	0.6 ± 0.1	-8.4 ± 0.6	-0.3 ± 0.0	0.5 ± 0.1	0.6 ± 0.0	0.0 ± 0.0
(b) Absolute		$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
COCO		5.5 ± 0.2	21.9 ± 0.2	5.0 ± 0.2	22.2 ± 0.3	0.3 ± 0.0	0.3 ± 0.0
imSitu		1.3 ± 0.0	9.4 ± 0.2	11.9 ± 0.1	13.0 ± 0.1	0.1 ± 0.0	0.1 ± 0.0
CelebA		27.4 ± 1.1	23.1 ± 0.4	0.8 ± 0.0	5.5 ± 0.0	1.1 ± 0.0	0.6 ± 0.0
(c) Variance		$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
COCO		0.2 ± 0.0	4.5 ± 0.1	0.4 ± 0.0	13.0 ± 0.2	0.0 ± 0.0	0.0 ± 0.0
imSitu		0.1 ± 0.0	1.6 ± 0.1	2.5 ± 0.0	3.2 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
CelebA		0.0 ± 0.0	2.7 ± 0.1	0.0 ± 0.0	2.2 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Raw differences collapse individual attributes. Tbl. 3 shows how each metric differs when using raw or absolute differences. We find that the magnitude of mean bias amplification is significantly higher when using absolute differences versus raw differences. For example, on imSitu, $\text{BiasAmp}_{M\to G}$ is 0.7 ± 0.1 for raw differences, but when using absolute differences $\text{BiasAmp}_{M\to G}$ is 11.9 ± 0.1. Further, this effect is particularly evident for $G \to M$ where amplification appears to be approximately 0.0 when in fact amplification does occur as evidenced by the absolute scores. This emphasizes the fact that negative and positive values for individual attributes are collapsed to zero when using raw differences, highlighting the advantage of summing over the absolute differences (see Adv. 2). Therefore, when using raw differences, we arrive at the following erroneous conclusion: models are minimally amplifying biases.

Bias amplification is not uniform across attribute groups. Considering the variance of the metrics, we observe the variance of multi-attribute $\text{Multi}_{\text{MALS}}$ is an order of magnitude higher than single-attribute $\text{BiasAmp}_{\text{MALS}}$. We find that this occurs as there is a small set of multi-attributes groups where more bias amplification is arising (see Appendix E for bias score distribution plots and Appendix G for qualitative analysis of top attributes that contribute to bias amplification). In contrast, bias arising from single attributes is more uniform. As discussed in Adv. 3, having this insight can help guide potential interventions such that we can target the more problematic groups of attributes.

5.3. Understanding Factors of Bias Amplification

We examine three factors that can influence the magnitude of bias amplified, i.e., group salience, attribute group size, and model performance.

$G \to M$ bias amplification depends on person salience. Amplification arising from group to attribute prediction ($\text{BiasAmp}_{G\to M}$, $\text{Multi}_{G\to M}$) is consistently low for COCO and imSitu. Although spurious correlations with gender exist in many parts of the image (Meister et al., 2022), the person is the main source of gender cues. For datasets such as COCO and imSitu where the person may not be the image's focal point, it is likely group membership is less easily inferred and thus has a smaller impact on attribute prediction. To validate our intuition, we evaluate $\text{BiasAmp}_{G\to M}$ and $\text{Multi}_{G\to M}$ with images from COCO containing person bounding boxes of varying sizes. There is a strong positive correlation between bias amplification and the size of the person bounding box—i.e., Pearson's r of 0.89 and 0.94 for $\text{BiasAmp}_{G\to M}$ and $\text{Multi}_{G\to M}$, resp. This is in line with Hall et al. (2022)'s findings, which suggest that harder to recognize groups can result in lower bias amplification. In contrast, for CelebA where the person is the focal point of the image, we find group to attribute prediction is larger (1.1 ± 0.0 for $\text{BiasAmp}_{G\to M}$ and 0.6 ± 0.0 for $\text{Multi}_{G\to M}$).

Amplification occurs across all attribute group sizes. While previous bias amplification metrics only consider amplification at $|m_i| = 1$, we provide results when varying attribute group sizes (see Fig. 2). Across all attribute group sizes, we observe that amplification occurs. Further, for $\text{Multi}_{\text{MALS}}$ on all datasets and $\text{Multi}_{G\to M}$ on COCO and imSitu, bias amplification increases as we increase $|m_i|$. Although, $\text{Multi}_{M\to G}$ is smaller in magnitude, indicating there is not much bias arising from the attributes on group prediction, amplification exists (i.e., > 0) and remains constant across the various attribute group sizes $|m_i|$.

Performance and bias amplification are negatively correlated. We calculate two performance metrics: mAP, the average precision over all individual attributes, and mAP*, a modified version of mAP that takes into account the different combinations of attributes in the ground-truth training set. As shown in Tbl. 4, we find generally there is a negative correlation between bias amplification and performance metrics. For example, the Pearson's r between mAP* and

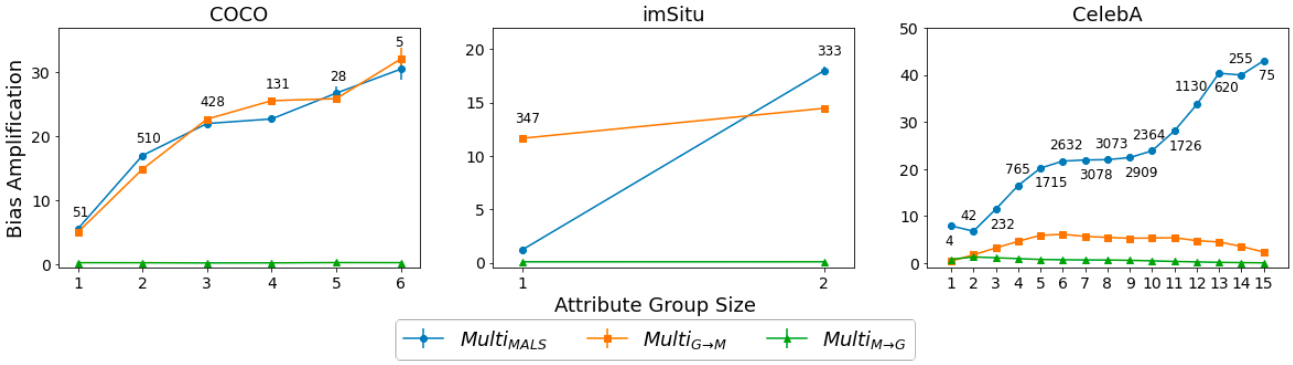

Figure 2. Visualization of bias amplification when varying the size of the attribute group, $|m|$. We plot the mean amplification score over five runs with random seeds of the model. Above each point, we include the number of attributes with size $|m|$. Error bars represent the standard deviation over runs.

Table 4. mAP*, a modified version of precision over all combinations of attributes, and the Pearson's r between mAP* at various combination sizes with the respective multi-bias amplification metrics on balanced versions of COCO , imSitu, and CelebA. We do not report the correlations for imSitu as there are only two attribute group sizes.

	COCO	imSitu	CelebA
mAP*	66.6 ± 0.1	31.4 ± 0.0	61.6 ± 0.0
corr(mAP*, $\text{Multi}_{\text{MALS}}$)	-0.1 ± 0.0	-	-0.9 ± 0.0
corr(mAP*, $\text{Multi}_{M \to G}$)	-0.5 ± 0.1	-	1.0 ± 0.0
corr(mAP*, $\text{Multi}_{G \to M}$)	-0.2 ± 0.0	-	-0.4 ± 0.0

the $\text{Multi}_{\text{MALS}}$ are -0.1 ± 0.0 and -0.9 ± 0.0 for COCO for and CelebA respectively. This occurs since mAP* generally decreases as the attribute group size increases whereas amplification increases.

6. Benchmarking Bias Mitigation Methods

We now benchmark the performance of previously proposed bias mitigation methods (Wang et al., 2019; Zhao et al., 2017; Wang et al., 2020b; Agarwal et al., 2022) using our multi-attribute bias amplification metrics.

6.1. Setup

Datasets. As the mitigation methods we benchmark were originally proposed for gender bias mitigation, we focus on COCO and imSitu datasets. See Appendix F for results on CelebA. We obtain gender and attribute labels using the same process as described in Sec. 5. The key difference here is that we do *not* oversample to balance attribute-group co-occurrence pairs. As a result, for COCO, we have 18,177, 4,545, and 10,795 images for the train, validation, and test splits, where 30.9% of the instances are labeled female. For imSitu, we have 10,240, 6,175 and 24,698 images for the train, validation, and test splits, where 40.7% of the instances are labeled female.

Mitigation Methods. Following Wang et al. (2020b), we benchmark five mitigation methods. As a simple baseline, we use oversampling, the method from Sec. 5, which greedily samples to balance w.r.t. single attributes. We also select four popular strategies that have been evaluated for single-attribute bias amplification metrics: corpus constraints (RBA) (Zhao et al., 2017), adversarial de-biasing (ADV) (Wang et al., 2019), data repair (Agarwal et al., 2022)[4], and domain independent training (DOMIND) (Wang et al., 2020b).[5]

For each mitigation method, we set all parameters as proposed in respective papers. As a baseline, we train a ResNet-50 without any mitigation techniques using the same training protocol from Sec 5. We refer to this model as ORIGINAL.

6.2. Experimental Analysis

We now analyze the performance of different mitigation methods. Here we start by comparing their performance on single-attribute versus multi-attribute metrics. We conclude by discussing the trade-offs between different methods. All results are in Tbl. 5.

Mitigating single-attribute bias amplification is not enough. Even when mitigation methods decrease bias am-

[4]For DATA REPAIR, the models are trained on a smaller number of instances since the method involves subsampling the dataset.

[5]Wang et al. (2020b) propose four inference methods: conditioning on the known group, choosing the maximum over the groups, summing the probability of different groups, and summing the scores of the different groups. We report results on the first as it minimizes bias amplification in their work. See Appendix F for results on the other methods.

Table 5. Comparison of five mitigation methods—OVERSAMPLING, RBA (Zhao et al., 2017), ADV (Wang et al., 2019), DOMIND (Wang et al., 2020b), and DATA REPAIR (Agarwal et al., 2022)—against the baseline ORIGINAL. We compare mAP, single attribute bias amplification, and multi-attribute bias amplification on COCO (a) and imSitu (b). We report the 95% confidence interval over five models trained using random seeds. The bold values indicate the best performing method: distance to 0 for single-attribute and smallest value for multi-attribute metrics.

(a) COCO	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
ORIGINAL	53.4 ± 0.2	-0.6 ± 0.3	14.5 ± 0.6	2.2 ± 0.4	12.5 ± 0.2	$\mathbf{-0.0 \pm 0.0}$	0.4 ± 0.0
OVERSAMPLING	51.5 ± 0.1	1.1 ± 0.1	14.0 ± 0.4	-3.4 ± 0.2	12.5 ± 0.3	-0.2 ± 0.0	$\mathbf{0.3 \pm 0.0}$
RBA	50.7 ± 1.1	3.8 ± 1.7	14.9 ± 1.1	-6.3 ± 3.5	17.3 ± 2.2	0.1 ± 01	0.4 ± 0.0
ADV	$\mathbf{59.0 \pm 0.1}$	-0.7 ± 0.9	17.1 ± 0.4	7.0 ± 0.6	14.7 ± 0.6	0.1 ± 0.0	0.3 ± 0.0
DOMIND	56.1 ± 0.3	0.4 ± 0.6	$\mathbf{12.6 \pm 0.8}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	0.3 ± 0.0	0.3 ± 0.0
DATA REPAIR	48.5 ± 0.1	$\mathbf{0.3 \pm 0.1}$	17.2 ± 0.3	1.9 ± 0.3	11.7 ± 0.2	-0.0 ± 0.0	0.4 ± 0.0
(b) imSitu	**mAP**	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
ORIGINAL	67.1 ± 0.1	2.5 ± 0.1	37.5 ± 0.1	-0.3 ± 0.1	20.6 ± 0.1	0.0 ± 0.0	0.2 ± 0.0
OVERSAMPLING	66.3 ± 0.1	-4.5 ± 0.2	35.8 ± 0.1	-2.4 ± 0.1	20.1 ± 0.1	-0.0 ± 0.0	0.2 ± 0.0
RBA	54.7 ± 0.5	$\mathbf{-1.4 \pm 0.3}$	35.4 ± 0.3	-6.2 ± 0.3	40.7 ± 0.5	-0.1 ± 0.0	0.3 ± 0.0
ADV	58.1 ± 0.1	4.1 ± 0.3	38.7 ± 0.3	0.6 ± 0.4	28.1 ± 0.3	-0.0 ± 0.0	0.2 ± 0.0
DOMIND	$\mathbf{69.6 \pm 0.1}$	10.2 ± 0.9	37.5 ± 0.4	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	0.1 ± 0.0	0.2 ± 0.0
DATA REPAIR	62.3 ± 0.1	-1.8 ± 0.1	$\mathbf{16.2 \pm 0.1}$	-0.1 ± 0.1	24.2 ± 0.1	$\mathbf{-0.0 \pm 0.0}$	$\mathbf{0.1 \pm 0.0}$

plification for single attributes, this does not always translate to the multi-attribute case. For example, on COCO, while DATA REPAIR outperforms all methods on $\text{BiasAmp}_{\text{MALS}}$, reducing bias amplification from -0.6 ± 0.3 on ORIGINAL to 0.3 ± 0.1, it increases $\text{Multi}_{\text{MALS}}$ from 14.5 ± 0.6 to 17.2 ± 0.3. In fact, DOMIND on COCO is the only method able to successfully mitigate amplification across all three multi-attribute metrics. Therefore, although current mitigation methods may "work" for single attributes, they increase multi-attribute bias amplification. Most significantly, this demonstrates that the overall amount of bias amplification may not actually be decreasing. Our finding underscores the need for mitigation methods that address bias amplification for both single and multi-attributes.

The best mitigation method is dataset dependent. Across the datasets, no mitigation method clearly outperforms another. For example, DOMIND outperforms ORIGINAL for COCO on all metrics except $\text{BiasAmp}_{G\to M}$ and $\text{Multi}_{G\to M}$; however, the method fares worse on imSitu. This is likely because DOMIND attempts to distinguish between group membership within attributes (e.g., woman with computer versus man with computer). Since imSitu attributes (e.g., indoor or outdoor location) have more diverse appearances compared to objects in COCO, it may be more difficult to learn these boundaries. A potential avenue for inquiry is developing training methods that work well for more general attributes like those found in imSitu.

The best mitigation method is metric dependent. The relative ordering of the mitigation methods changes across metrics. For example, DOMIND fares well for $M \to G$, consistently achieving bias amplification of 0.0 ± 0.0. For DOMIND, the prediction of attributes is conditioned on the group at test time, making it unlikely that the attributes would affect group prediction. However, for $G \to M$ amplification, DOMIND either increases or does not significantly change the amount of bias amplification. In contrast, we have the case as with DATA REPAIR on imSitu where $\text{Multi}_{G\to M}$ and $\text{Multi}_{\text{MALS}}$ decrease but $\text{Multi}_{M\to G}$ increases w.r.t. the baseline. DATA REPAIR is ill-suited to scenarios where group amplification is of particular concern. Thus, only benchmarking on one metric can lead to the wrong conclusion that bias has been mitigated when in fact it has not changed or even increased. Overall, these results underscore the importance of considering which type of bias amplification is most relevant when deciding upon a mitigation strategy as the best-suited method may vary.

7. Discussion

7.1. Implications for Mitigation Strategies

Wang et al. (2019) illustrated that balancing datasets only w.r.t. group membership is insufficient. We implement more sophisticated balancing strategies: balancing each group with individual attributes and using Agarwal et al. (2020)'s proposed data repair method. While balancing can reduce bias from single attributes, it does not work for multiple attributes. Multi-attribute bias amplification highlights the futility of balancing datasets. In particular, as balancing each group with all possible combinations of attributes is likely to be infeasible given the large number of attributes represented in datasets.

A potential avenue is to augment training datasets with synthetic images. While prior works (Sharmanska et al., 2020; Ramaswamy et al., 2021; Sattigeri et al., 2019) use GANs to mitigate bias amplification, these efforts have focused on face-centric datasets, such as CelebA and Diversity in Faces (Merler et al., 2019), as opposed to more complex real-world image scenes. Using generative methods we can manipulate scenes to change the attributes pictured or alter

a model's perception of group membership.[6] This would permit us to account for both single and multi-attributes when balancing datasets.

7.2. Multi-Attribute Bias Amplification Limitations

Reliance on annotations. As with existing co-occurrence metrics (Zhao et al., 2017; Wang & Russakovsky, 2021), our proposed metric only measures bias amplification w.r.t. annotated attributes. While we capture multi-attribute bias amplification, we cannot account for amplification that occurs due to unlabeled attributes. In addition, due to the lack of self-reported demographic annotations, we rely on third-party judgement of group membership. We acknowledge that relying on proxy judgements reifies the incorrect notion that, e.g., gender identity can be visually inferred and reducing gender to a binary is a harmful practice (Hamidi et al., 2018; Keyes, 2018). Such reliance is a problem that many fairness researchers face (Zhao et al., 2021; Buolamwini & Gebru, 2018; Wilson et al., 2019) due to a lack of self-reported demographic information.

Entanglement between metrics. Taking the absolute value of differences conflates two values of the same magnitude but with different signs. Both values are equal contributors to bias amplification. Some may argue that a bias amplification score of $-c$ is more desirable than $+c$ since "bias" has decreased. The preference between a positive or negative amplification is a value-laden decision that ultimately depends on the attribute and group. To illustrate, revisiting the example from Sec. 4, negative bias amplification for $\{\texttt{typing, F}\}$ can contribute to erasure. Conversely, positive bias amplification for $\{\texttt{baking, F}\}$ (which has a positive bias score of 0.6 in imSitu's train set) reproduces and amplifies a harmful social stereotype. In both cases, regardless of the direction, bias amplification is undesirable.

Rather than making an overarching prescription, we propose using a multiplicity of fairness metrics. For example, reporting both raw and absolute differences permits a finer-grained analysis of where amplification is arising.

Equal weighting of attribute groups. Similar to single-attribute metrics, we weight all attribute combinations equally. However, depending on the context, one may want to place more weight on socially salient attribute combinations. For example, if a classifier trained on COCO is deployed in an athletics setting, amplification w.r.t. sports attributes such as `frisbee` or `tennis racket` may be more important. Alternatively, if the classifier is deployed in a classroom setting, these attributes may no longer be significant. We suggest metric users consider the deployment context of their system and adjust the weights accordingly.

[6] We do not advocate visual manipulations of group membership, given the ethical concerns of changing people's appearances, in particular for data collected without informed consent.

8. Conclusion

Our proposed metric, Multi-Attribute Bias Amplification, illustrates the need to consider multiple attributes when measuring bias amplification. For perfectly "balanced" datasets, we find bias amplification occurs wrt multi-attributes, regardless of whether we use raw or absolute differences. Further, we are the first to show that methods which mitigate single attribute bias can inadvertently increase multi-attribute bias amplification. Overall, multi-attribute bias amplification provides a better characterization of the extent to which a model introduces bias from training to prediction.

Acknowledgements. This work was funded by Sony Research. We thank William Thong and Julienne LaChance for their helpful comments and suggestions.

References

Agarwal, S., Muku, S., Anand, S., and Arora, C. Does data repair lead to fair models? curating contextually fair data to reduce model bias. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

Agarwal, V., Shetty, R., and Fritz, M. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Birhane, A. and Prabhu, V. U. Large image datasets: A pyrrhic win for computer vision? In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2018.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. Fair generative modeling via weak supervision. In *International Conference on Machine Learning (ICML)*, 2020.

Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

DeVries, T., Misra, I., Wang, C., and van der Maaten, L. Does object recognition work for everyone? In *CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision*, 2019.

Dulhanty, C. and Wong, A. Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. *CVPR Workshop on Fairness, Accountability, Transparency and Ethics in Computer Vision*, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Hall, M., van der Maaten, L., Gustafson, L., and Adcock, A. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Hamidi, F., Scheuerman, M. K., and Branham, S. M. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Conference on Human Factors in Computing Systems (CHI)*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, 2018.

Hirota, Y., Nakashima, Y., and Garcia, N. Quantifying societal bias amplification in image captioning. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jia, S., Meng, T., Zhao, J., and Chang, K.-W. Mitigating gender bias amplification in distribution by posterior regularization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jo, E. S. and Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

Kaushik, D., Hovy, E., and Lipton, Z. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2019.

Keyes, O. The misgendering machines: Trans/HCI implications of automatic gender recognition. *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2015.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems (NeurIPS)*, 1989.

Leino, K., Fredrikson, M., Black, E., Sen, S., and Datta, A. Feature-wise bias amplification. In *International Conference on Learning Representations (ICLR)*, 2019.

Li, Z., Hoogs, A., and Xu, C. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., Xu, C., and Ibrahim, M. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

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

Meister, N., Zhao, D., Wang, A., Ramaswamy, V. V., Fong, R., and Russakovsky, O. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191*, 2022.

Merler, M., Ratha, N., Feris, R. S., and Smith, J. R. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Ramaswamy, V. V., Kim, S. S. Y., and Russakovsky, O. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Renduchintala, A., Díaz, D., Heafield, K., Li, X., and Diab, M. Gender bias amplification during speed-quality optimization in neural machine translation. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference for Learning Representations (ICLR)*, 2019.

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., and Varshney, K. R. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Sharmanska, V., Hendricks, L. A., Darrell, T., and Quadrianto, N. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020.

Shrestha, R., Kafle, K., and Kanan, C. An investigation of critical issues in bias mitigation techniques. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. 2021.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Wang, A. and Russakovsky, O. Directional bias amplification. In *International Conference on Machine Learning (ICML)*, 2021.

Wang, A., Narayanan, A., and Russakovsky, O. REVISE: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision (ECCV)*, 2020a.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *International Conference on Computer Vision (ICCV)*, 2019.

Wang, Z. and Culotta, A. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Wang, Z., Qinami, K., Karakozis, I., Genova, K., Nair, P., Hata, K., and Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Wilson, B., Hoffman, J., and Morgenstern, J. Predictive inequity in object detection. In *CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision*, 2019.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *ACM Conference on Fairness, Acountability and Transparency (FAccT)*, 2020.

Yatskar, M., Zettlemoyer, L., and Farhadi, A. Situation recognition: Visual semantic role labeling for image understanding. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Zhao, D., Wang, A., and Russakovsky, O. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021.

Zhao, E., Huang, D.-A., Liu, H., Yu, Z., Liu, A., Russakovsky, O., and Anandkumar, A. Scaling fair learning to hundreds of intersectional groups. 2022.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Zhu, X., Anguelov, D., and Ramanan, D. Capturing longtail distributions of object subcategories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. Existing metrics definition

Using the notation from Sec. 3, we provide formal definitions for $\text{BiasAmp}_{\text{MALS}}$ and $\text{BiasAmp}_{\rightarrow}$. We first provide the definition for undirected bias amplification (Zhao et al., 2017):

$$\text{BiasAmp}_{\text{MALS}} = \frac{1}{|\mathcal{A}|} \sum_{g \in \mathcal{G}} \sum_{a \in \mathcal{A}} \mathbb{1}\left[\text{bias}_{\text{train}}(a, g) > |\mathcal{G}|^{-1}\right] \cdot \left(\text{bias}_{\text{test}}(a, g) - \text{bias}_{\text{train}}(a, g)\right). \tag{4}$$

Note $\text{bias}_{\text{test}}(a, g)$ is the bias score from the attribute and group label test set predictions, whereas $\text{bias}_{\text{train}}(a, g)$ is the bias score from the attribute and group label training set ground truths.

Let $\hat{a}$ and $\hat{g}$ denote a model's prediction for attribute, a, and group membership, g, resp. Further, without loss of generality, let $P_{\text{train}}(a = 1)$ denote the probability of a training set sample containing the attribute a, and $P_{\text{train}}(g = 1)$ denote the probability of a training set sample containing the group membership label g. We now provide the definition for directional bias amplification (Wang & Russakovsky, 2021):

$$\text{BiasAmp}_{\rightarrow} = \frac{1}{|\mathcal{G}||\mathcal{A}|} \sum_{g \in \mathcal{G}} \sum_{a \in \mathcal{A}} y_{ga} \Delta_{ga} + (1 - y_{ga})(-\Delta_{ga}), \tag{5}$$

where

$$y_{ga} = \mathbb{1}\left[P_{\text{train}}(g = 1, a = 1) > P_{\text{train}}(a = 1) P_{\text{train}}(g = 1)\right]$$

$$\Delta_{ga} = \begin{cases} P_{\text{test}}(\hat{a} = 1 | g = 1) - P_{\text{train}}(a = 1 | g = 1) \\ \text{if measuring } G \rightarrow A \\ P_{\text{test}}(\hat{g} = 1 | a = 1) - P_{\text{train}}(g = 1 | a = 1) \\ \text{if measuring } A \rightarrow G \end{cases} \tag{6}$$

B. Relationship between metrics

We provide clarification on the relationship between the three multi-attribute bias amplification metrics. Importantly, we do not expect the metrics to be correlated. First, $\text{Multi}_{\text{MALS}}$ only captures positive correlations while directional metrics capture positive and negative. Second, the two directional metrics capture different phenomena: $M \rightarrow G$ measures the influence of attributes on group prediction and $G \rightarrow M$ measures the influence of group on attribute prediction.

To illustrate this, we use the MNIST setup from Sec. 3 but apply a Gaussian blur (radius 5), making group prediction hard. As shown in Fig. 3, after blurring the digits, $\text{Multi}_{\text{MALS}}$and $\text{Multi}_{G \rightarrow M}$ decrease. $\text{Multi}_{M \rightarrow G}$ increases from 0.3 to 1.3. This is expected as the classifier relies on other attributes to predict the group. Conversely, since the group is difficult to recognize, the classifier is unlikely to rely on this feature when predicting attributes, causing $\text{Multi}_{G \rightarrow M}$ to decrease.

Example Image		
	Original	Blurred
$\text{Multi}_{\text{MALS}}$	5.0 ± 4.5	1.8 ± 0.6
$\text{Multi}_{M \rightarrow G}$	0.3 ± 0.2	1.3 ± 0.2
$\text{Multi}_{G \rightarrow M}$	8.9 ± 5.3	0.0 ± 0.0

Figure 3. Comparison of multi-attribute metrics for the original and blurred MNIST images. The example images have attributes $a_1 = 1$ and $a_2, a_3 = 0$.

C. Dataset preprocessing

The greedy oversampling procedure results in $45{,}657$, $12{,}351$, and $27{,}499$ images respectively in the training, validation, and test sets for COCO. The splits for imSitu are $40{,}470$, $10{,}668$, and $17{,}036$ images. Finally, the splits for CelebA are $379{,}661$, $65{,}895$, and $52{,}397$ images.

The model is trained for 50 epochs using an Adam optimizer (Kingma & Ba, 2015) with L2 weight decay of 10^{-6}, batch size of 32, and a learning rate of 10^{-5}. Based on the known variance in fairness metrics (Wang & Russakovsky, 2021), we train five models with random seeds and report the 95% confidence interval.

We provide additional details on how the datasets are preprocessed.

C.1. COCO

Gender expression labels. We follow Zhao et al. (2021) and use an expanded word set to automatically derive perceived gender expression from the captions provided in the COCO (Lin et al., 2014) dataset. When searching for gendered words, we first convert the captions to lowercase. We then use the following set of keywords to query the captions:

- Male: {"male", "boy", "man", "gentleman", "boys", "men", "males", "gentlemen", "father", "boyfriend"}
- Female: {"female", "girl", "woman", "lady", "girls", "women", "females", "ladies", "mother", "girlfriend"}

One of the five COCO captions must contain a gendered term. Further, if both a male and female gendered term appear in the captions, the instance was discarded. This methodology matches that of prior works (Zhao et al., 2017; Wang et al., 2019; Agarwal et al., 2022)

Attribute labels. We only include attributes that have occurred more than 100 times with both male and female instances. This leaves us with the following 52 attributes: {"person", "bicycle", "car", "motorcycle", "bus", "truck", "traffic light", "bench", "cat", "dog", "horse", "backpack", "umbrella", "handbag", "tie", "suitcase", "frisbee", "skis", "sports ball", "kite", "surfboard", "tennis racket", "bottle", "wine glass", "cup", "fork", "knife", "spoon", "bowl", "banana", "sandwich", "pizza", "donut", "cake", "chair", "couch", "potted plant", "bed", "dining table", "tv", "laptop", "remote", "cell phone", "microwave", "oven", "sink", "refrigerator", "book", "clock", "vase", "teddy bear", "toothbrush"}. For attribute groups, we do not use "person" as this redundantly appears in all groups.

C.2. imSitu

Gender expression labels. To derive gender expression labels for imSitu, we use the agent annotations associated with each instance. We use the same set of keywords that we used for COCO to query male and female instances.

Action and location labels. From the 504 annotated verbs in imSitu, we consider only those that co-occur with a gendered agent. Further, we only include actions that have occurred more than 5 times with both male and female instances. This leaves us with 361 verbs in total.

We turn location into a binary prediction between indoor and outdoor. To derive labels, we consider the top 50 place annotations in imSitu. We manually annotate each location as either "indoor" or "outdoor," discarding locations that were ambiguous. This results in 25 location annotations that are divided into indoor and outdoor as follows:

- Outdoors: {"outdoors", "outside", "field", "street", "road", "sidewalk", "beach", "farm", "forest", "yard"}
- Indoors: {"room", "inside", "kitchen", "office", "gymnasium", "shop", "house", "hospital", "classroom", "workshop", "stage", "bed", "classroom", "living room", "bathroom"}

C.3. CelebA

Hair color labels. We use hair color as the group. To do so, we use three of the annotations for hair color included with CelebA: "black hair," "blond hair," "brown hair."

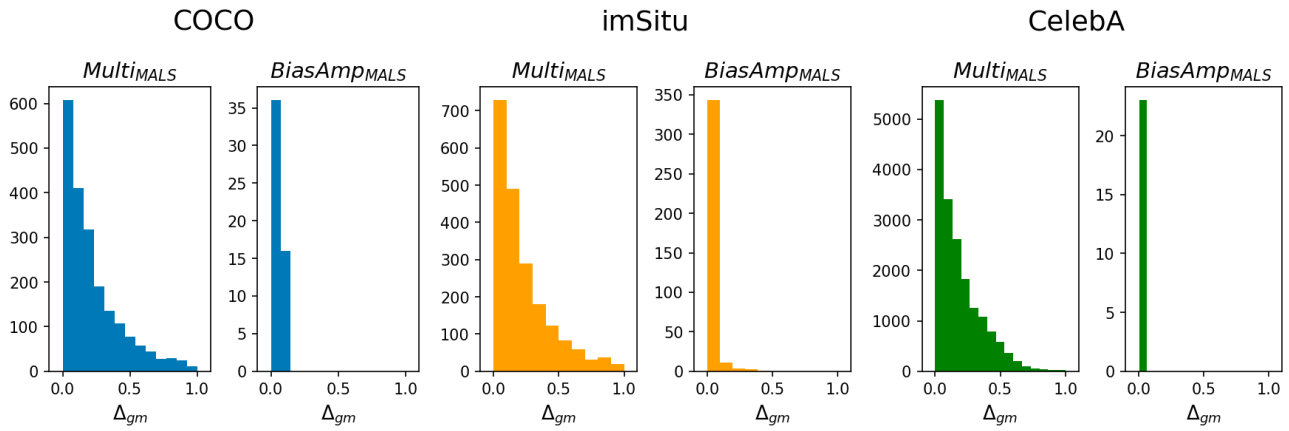

Figure 4. Visualization of the absolute change in bias scores, $|\Delta_{gm}|$ from the ground-truth training set to the predicted labels distribution for undirected multiple-attribute and single-attribute bias amplification across COCO, imSitu, and CelebA.

Physical attribute labels. For the attributes in CelebA, we only include those that have occurred more than 500 times with each hair color respectively. This leaves us with the following 23 attributes: {"Arched Eyebrows", "Bags Under Eyes", "Bangs", "Big Nose", "Eyeglasses", "Heavy Makeup", "High Cheekbones", "Male", "Mouth Slightly Open", "Narrow Eyes", "No Beard", "Oval Face", "Pale Skin", "Pointy Nose", "Receding Hairline", "Rosy Cheeks", "Smiling", "Straight Hair", "Wavy Hair", "Wearing Earrings", "Wearing Lipstick", "Wearing Necklace", "Young"}.

D. Training details

All models in this work were developed using PyTorch. The models are trained and evaluated on 1 NVIDIA T4 Tensor Core GPU with 64 GB of GPU memory and 2.5 GHz Cascade Lake 24C processors. The operating system is Linux 64-bit Ubuntu 18.04.

E. Bias scores

In Sec. 5.2, we observe that the variance for undirected multi-attribute bias amplification is greater than that for undirected single attribute metrics. We visualize the disaggregated changes in bias score (Δ_{gm}) for single and multi-attribute metrics across the three datasets (see Fig. 4). For all datasets, we see the distribution is right-skewed for Multi$_{\text{MALS}}$, meaning there is a small subset of multi-attribute groups where more amplification is occurring.

F. Bias mitigation methods

F.1. Method descriptions

We provide a description of the five mitigation strategies used in Sec. 6. We evaluate on the same test set consisting of 10,795, 10,240, and 11,351 images for COCO, imSitu, and CelebA respectively.

Oversampling. We use a greedy oversampling process to balance group membership for each single attribute. $\forall(a, g) \in \mathcal{A} \times \mathcal{G}$, the sampling process terminates when the bias score $\text{bias}_{\text{train}}(a, g) \in [|\mathcal{G}|^{-1} \pm \epsilon]$, where $\epsilon = 0.025$ for COCO, imSitu and $\epsilon = 0.07$ for CelebA. This results in 45,657, 40,470, and 379,661 training images for COCO and imSitu respectively. We train a ResNet-50 for 50 epochs using an Adam optimizer with batch size 32, learning rate of 10^{-5}, and weight decay of 10^{-6}.

RBA. Reducing Bias Amplification (Zhao et al., 2017) is employed after training. Here, the method uses corpus level constraints so that the predictions match a specified distribution. To ensure that the algorithm converges, the method uses Lagrangian relaxation. We use RBA to optimize the prediction distribution so it matches that of the training set.

Table 6. Comparison of three inference methods for DOMIND (Wang et al., 2020b): (1) choosing the maximum over groups, (2) summing the probability of different groups, and (3) summing the sores of different groups. We compare mAP, single attribute bias, and multi-attribute bias amplification on COCO, imSitu, and CelebA. For notation, y, g, and x refer to the attributes, group, and image respectively. $|G|$ indicates the number of groups.

COCO	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
$\|G\|N$sigm, max $P(y\|g,x)$	54.7	-61.4	59.7	-3.1	39.6	-0.2	0.3
$\|G\|N$sigm $\sum_g P(y\|g,x)$	54.7	-8.4	31.3	-6.1	31.2	-0.1	0.3
$\|G\|N$sigm $\sum_g s(y\|g,x)$	54.8	-61.4	59.9	-3.1	39.6	-0.1	0.3
imSitu	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
$\|G\|N$sigm, max $P(y\|g,x)$	49.9	-64.7	59.7	-3.1	39.6	-0.2	0.3
$\|G\|N$sigm $\sum_g P(y\|g,x)$	51.1	-5.5	36.3	-5.0	37.1	-0.0	0.3
$\|G\|N$sigm $\sum_g s(y\|g,x)$	49.4	-64.7	32.8	-13.0	70.3	-0.0	0.5
CelebA	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
$\|G\|N$sigm, max $P(y\|g,x)$	37.4	-99.0	40.6	-13.6	34.2	-3.9	0.7
$\|G\|N$sigm $\sum_g P(y\|g,x)$	38.0	-3.6	8.1	-4.0	27.5	30.7	0.8
$\|G\|N$sigm $\sum_g s(y\|g,x)$	37.8	-99.0	-13.6	-3.8	43.2	34.2	0.7

Adversarial Debiasing. This method removes group information from the intermediate representation so that the model is not influenced by the group when making predictions. The goal during training is thus to improve the classifier's ability to predict attributes while making it difficult for the adversary to predict group membership.

We remove group information from the final convolutional layer of a ResNet-50 (i.e., **adv @ conv5**. To train our adversarial debiasing method, we follow the procedure from Wang et al. (2019). We train with the adversarial loss using a learning rate of 5×10^{-6} for 60 epochs.

Domain Independent. As opposed to adversarial debiasing which promotes *fairness through blindness*, Wang et al. (2020b) claim that domain independent training promotes *fairness through awareness*. Concretely, domain independent training attempts to learn to differentiate between the same attribute for different groups. For example, in the case of gender bias amplification on COCO, the classifier will attempt to learn the boundary between (`computer, F`) and (`computer, M`). We train the ResNet-50 for 50 epochs using an Adam optimizer with batch size 32, learning rate of 10^{-5}, and weight decay of 10^{-6}.

At inference time, there are gn predictions (i.e., n predictions for each group). To select which predictions we use, we take the predictions associated with the ground-truth group membership for the instance. We also provide results in Tbl. 6 for the three other inference methods introduced by Wang et al. (2020b).

Data Repair. Agarwal et al. (2022) attempt to repair the existing dataset using a fair selection process. Concretely, this method curates a dataset via subsampling until co-occurring attributes are well represented. In the supervised setting, the method uses a greedy approach that aims to minimize c_v, which is equal to standard deviation divided by the mean number of images per attribute. We use five random seeds, resulting in a mean coefficient variation of 1.78, 5.83, and 0.65 for COCO, imSitu, and CelebA. The subsampled train and validation set size $3{,}000 \times g$ and $500 \times g$ where $g = 2$ for COCO, imSitu and $g = 3$ for CelebA.

F.2. Additional mitigation results

We report the result of applying the mitigation methods on the CelebA dataset in Tbl. 7. It is important to note that the mitigation methods (e.g., RBA, ADV, DOMIND, and DATA REPAIR) were originally proposed for gender bias and only benchmarked on a binary group. Here, we extend the mitigation methods for the non-binary group of hair color. For single-attribute metrics, proposed methods can help mitigate bias amplification for specific metrics. For example, ADV reduces $\text{BiasAmp}_{\text{MALS}}$ from 2.0 ± 0.3 to 1.0 ± 1.2. However, overall, we see that mitigation methods are unable to reduce bias amplification, even increasing amplification in some instances, especially with respect to the multi-attribute metrics.

Table 7. Comparison of five mitigation methods—OVERSAMPLING, RBA (Zhao et al., 2017), ADV (Wang et al., 2019), DOMIND (Wang et al., 2020b), and DATA REPAIR (Agarwal et al., 2022)—against the baseline ORIGINAL. We compare mAP, single attribute bias amplification, and multi-attribute bias amplification on CelebA. We report the 95% confidence interval over five models trained using random seeds. The bold values indicate the best performing method: distance to 0 for single-attribute and smallest value for multi-attribute metrics.

CelebA	mAP	$\text{BiasAmp}_{\text{MALS}}$	$\text{Multi}_{\text{MALS}}$	$\text{BiasAmp}_{M\to G}$	$\text{Multi}_{M\to G}$	$\text{BiasAmp}_{G\to M}$	$\text{Multi}_{G\to M}$
ORIGINAL	79.0 ± 0.2	2.0 ± 0.3	$\mathbf{17.5 \pm 0.8}$	$\mathbf{-0.1 \pm 0.1}$	2.3 ± 0.2	$\mathbf{0.1 \pm 0.1}$	$\mathbf{0.2 \pm 0.0}$
OVERSAMPLING	73.9 ± 0.1	-1.1 ± 0.1	19.2 ± 0.5	-0.9 ± 0.0	4.0 ± 0.0	-0.2 ± 0.1	0.5 ± 0.0
RBA	71.5 ± 2.7	25.8 ± 12.9	38.8 ± 11.2	-8.8 ± 4.0	21.0 ± 10.5	-0.8 ± 0.3	0.2 ± 0.1
ADV	73.9 ± 0.1	$\mathbf{1.0 \pm 1.2}$	27.4 ± 2.6	4.9 ± 0.8	22.3 ± 2.1	-3.3 ± 0.1	0.4 ± 0.0
DOMIND	$\mathbf{82.7 \pm 0.1}$	3.6 ± 0.3	18.7 ± 0.9	-2.1 ± 0.0	$\mathbf{0.0 \pm 0.0}$	0.3 ± 0.1	0.3 ± 0.0
DATA REPAIR	82.2 ± 0.2	-10.6 ± 0.3	27.4 ± 0.7	-0.5 ± 0.3	3.3 ± 0.2	-0.5 ± 0.1	0.4 ± 0.0

Table 8. Top three groups of attributes with the largest contribution to bias amplification when training and evaluating on a perfectly "balanced" dataset and the difference in bias score for the respective attribute group. The ranking is calculated by averaging over the deltas with random seed assignments. We report the mean difference in bias score and the 95% confidence interval. Bolded values indicate the bias is amplifying for male gender expression for COCO and imSitu. Bolded values indicate bias is amplifying for black hair color and italic values indicate bias is amplifying for blonde hair.

Rank	COCO Single	COCO Multi
1	{bus} 14.6 ± 0.8	**{bus, banana}** 100.0 ± 0.0
2	{dog} 13.3 ± 1.0	{bottle,cup,cake,dining table,microwave,refrigerator} 100.0 ± 0.0
3	**{surfboard}** 12.5 ± 1.6	{dog,pizza,couch} 100.0 ± 0.0
Rank	**imSitu Single**	**imSitu Multi**
1	**{packaging}** 48.3 ± 0.0	**{indoor, crying}** 86.7 ± 23.4
2	**{giggling}** 36.4 ± 3.7	**{indoors, drumming}** 85.7 ± 4.2
3	{confronting} 30.2 ± 2.0	**{indoors, flinging}** 81.8 ± 0.0
Rank	**CelebA Single**	**CelebA Multi**
1	{receding hairline} 6.5 ± 1.0	{*arched eyebrows, heavy makeup, no beard, pale skin, pointy nose, receding hairline, wearing earrings, wearing lipstick*} 100.0 ± 0.0
2	{*rosy cheeks*} 3.9 ± 0.1	{eyeglasses, high cheekbones, mouth slightly open, no beard, pointy nose, smiling, wavy hair, wearing earrings, wearing lipstick} 100.0 ± 0.0
3	{*narrow eyes*} 3.7 ± 1.0	**{heavy makeup, high cheekbones, narrow eyes, no beard, receding hairline, smiling, straight hair, wearing lipstick, wearing necklace, young}** 100.0 ± 0.0

G. Qualitative results

We consider the top attributes or groups of attributes that contribute to bias amplification. Concretely, we examine the difference in bias scores between predictions and the training set when calculating $\text{BiasAmp}_{\text{MALS}}$ and $\text{Multi}_{\text{MALS}}$. The results on the "balanced" datasets from Sec. 5 and mitigation methods from Sec. 6 are found in Tbls. 8, 9, and 10.

"Balanced" models. We find that multi-attributes surfaces a different set of attributes than when only considering for single attributes. In Tbl. 8, there is some overlap between single attributes in the multi-attribute groups (e.g., bus, dog, receding hairline) for COCO and CelebA. However, for imSitu, the multi and single attributes are disjoint.

Bias mitigation methods. Next, we look at the groups surfaced after applying mitigation techniques (Tbl. 9, 10). We note certain groups of attributes occur across many mitigation methods. For example, in COCO, {car, motorcycle, truck, handbag} occurs in RBA, ADV, and DOMIND. Similarly, in imSitu, {indoors, crying} also occurs in OVERSAMPLE and RBA as well as ORIGINAL. This indicates there are certain groups of attributes which may be more difficult to debias than others. Learning to address the bias amplification arising from these more difficult groups can be a fruitful direction for future work in this space.

Table 9. Top three groups of attributes with the largest contribution to bias amplification for bias mitigation methods and the difference in bias score for the respective attribute group. Results are reported on COCO (top) and imSitu (bottom). The ranking is calculated by averaging over the deltas for five runs with random seed assignments. We report the mean difference in bias score and the 95% confidence interval. Bolded values indicate the bias is amplifying for male gender expression.

Rank	ORIGINAL	OVERSAMPLE	RBA	ADV	DOMIND	DATA REPAIR
1	{bench, dog, cup}	{bench, dog, cup}	{sport ball, potted plant}	**{bench, umbrella, potted plant}**	**{bench, umbrella, potted plant}**	**{sports ball, kite}**
	100.0 ± 0.0	100.0 ± 0.0	80.0 ± 0.0	72.8 ± 3.9	67.0 ± 8.6	100.0 ± 0.0
2	{sports ball, surfboard}	{sports ball, surfboard}	**{car, motorcycle, truck, handbag}**	**{car, motorcycle, truck, handbag}**	**{backpack, handbag, chair, cell phone}**	{car, surfboard}
	100.0 ± 0.0	100.0 ± 0.0	76.4 ± 5.3	67.1 ± 5.3	56.8 ± 17.0	83.3 ± 0.0
3	{sports ball, potted plant}	**{handbag, pizza, chair, dining table}**	{frisbee, clock}	{car, laptop}	**{car, motorcycle, truck, handbag}**	**{bus, cell phone}**
	80.0 ± 0.0	80.9 ± 17.5	76.4 ± 5.3	63.6 ± 0.0	56.6 ± 9.8	81.3 ± 0.0
Rank	ORIGINAL	OVERSAMPLE	RBA	ADV	DOMIND	DATA REPAIR
1	**{indoors, crying}**	**{indoors, crying}**	**{indoors, crying}**	**{indoors, repairing}**	**{indoors, repairing}**	{indoors, slicing}
	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	78.5 ± 2.7	64.0 ± 28.0	100.0 ± 0.0
2	**{indoor, repairing}**	**indoors, drumming}**	**{indoor, repairing}**	{indoors, checking}	**{indoors, assembling}**	**{indoors, watering}**
	80.0 ± 0.0	100.0 ± 0.0	80.0 ± 0.0	64.1 ± 8.3	55.6 ± 0.0	100.0 ± 0.0
3	**indoors, drumming}**	**{indoors, resting}**	**{indoors, resting}**	{checking}	**{indoor, racing}**	**{indoors, pumping}**
	80.0 ± 35.1	67.1 ± 4.7	61.4 ± 13.4	61.7 ± 7.2	55.0 ± 21.5	100.0 ± 0.0

H. Runtime analysis

Naively implemented, the runtime for calculating the multi-attribute metric could be exponential as we could iterate through all possible combinations of attributes $\mathcal{A}$ of size $|m|$ for $|m| \in \{1, \dots |\mathcal{A}|\}$. This is especially costly given that $|\mathcal{A}|$ can be large for many visual datasets. However, we only consider the groups of multiple attributes that exist in the dataset; many groups of multiple attributes do not occur in either the training or the test set. Rather, our implementation can run in $\mathcal{O}(n)$ time where n is equal to the number of instances in either the training or test set.

Table 10. Top three groups of attributes with the largest contribution to bias amplification for bias mitigation methods on CelebA and the difference in bias score for the respective attribute group. The ranking is calculated by averaging over the deltas for five runs with random seed assignments. We report the mean difference in bias score and the 95% confidence interval. Bolded values indicate bias is amplifying for black hair color and italic values indicate bias is amplifying for blonde hair.

	1	2	3
ORIGINAL	{*heavy makeup,no beard,oval face,pale skin,pointy nose,receding hairline,smiling,wearing lipstick,young*}	{bags under eyes,high cheekbones,male,mouth slightly open,narrow eyes,receding hairline,smiling,wavy hair}	{bags under eyes,bangs,eyeglasses,male,narrow eyes,wavy hair}
	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
OVERSAMPLE	{bags under eyes,high cheekbones,male,mouth slightly open,narrow eyes,receding hairline,smiling,wavy hair}	{*arched eyebrows,heavy makeup,high cheekbones,mouth slightly open,no beard,oval face,receding hairline,smiling,straight hair,wearing lipstick,wearing necklace,young*}	{*heavy makeup,no beard,oval face,pale skin,pointy nose,receding hairline,smiling,wearing lipstick,young*}
	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
RBA	{*bangs,big nose,eyeglasses,high cheekbones,mouth slightly open,no beard,pointy nose,smiling,wearing lipstick,wearing necklace*}	{*bangs,big nose,eyeglasses,high cheekbones,mouth slightly open,no beard,pointy nose,smiling,wearing lipstick*}	{bags under eyes,big nose,high cheekbones,male,mouth slightly open,no beard,receding hairline,smiling,wavy hair}
	100.0 ± 0.0	100.0 ± 0.0	90.0 ± 17.5
ADV	**{arched eyebrows,bags under eyes,bangs,big nose,heavy makeup,high cheekbones,mouth slightly open,no beard,smiling,straight hair,wearing earrings,wearing lipstick}**	{bags under eyes,male,no beard,wavy hair,wearing earrings}	**{arched eyebrows,bangs,big nose,heavy makeup,no beard,straight hair,wearing earrings,wearing lipstick,young}**
	87.5 ± 0.0	83.3 ± 0.0	83.1 ± 17.5
DOMIND	{*big nose,eyeglasses,high cheekbones,mouth slightly open,no beard,receding hairline,smiling,wearing earrings,wearing necklace,young*}	{bangs,eyeglasses,heavy makeup,high cheekbones,mouth slightly open,no beard,smiling,wearing lipstick}	{*bags under eyes,big nose,eyeglasses,male,narrow eyes,no beard,smiling*}
	100.0 ± 0.0	68.6 ± 5.0	64.8 ± 30
DATA REPAIR	**{high cheekbones, male, mouth slightly open, oval face, receding hairline, wearing earrings, young}**	{arched eyebrows, big nose, heavy makeup, high cheekbones, no beard, rosy cheeks, smiling, straight hair, wavy hair, wearing lipstick, young}	{*arched eyebrows, heavy makeup, no beard, oval face, pale skin, straight hair, wavy hair, wearing lipstick, young*}
	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0