This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MAVias: Mitigate any Visual Bias

Ioannis Sarridis[1,2] Christos Koutlis[1] Symeon Papadopoulos[1] Christos Diou[2]
[1]Information Technologies Institute, CERTH, Greece
[2]Department of Informatics and Telematics, Harokopio University of Athens, Greece
{gsarridis, ckoutlis, papadop}@iti.gr {isarridis, cdiou}@hua.gr

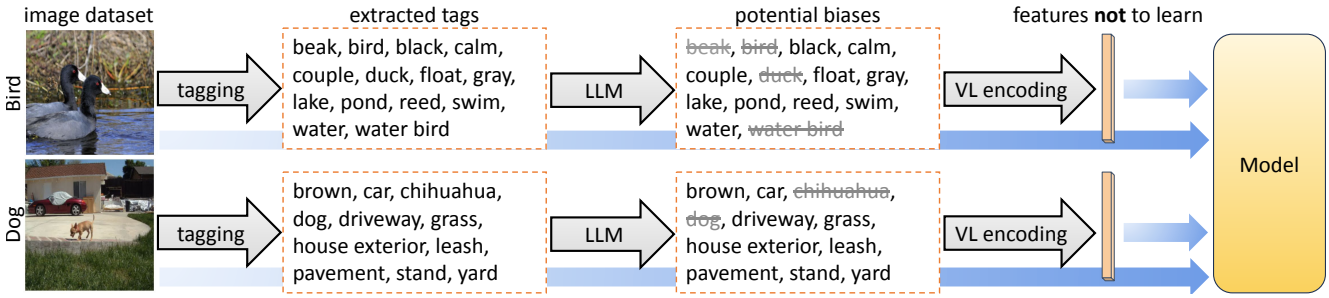


Figure 1. MAVias identifies instance-level potential visual biases through foundational models that extract tags representing visual features and assess relevance to the target class. Then, MAVias encodes these features within the vision-language space and integrates them into a bias-aware framework to train a model that is invariant to such visual biases.

Abstract

Mitigating biases in computer vision models is an essential step towards trustworthy artificial intelligence systems. Existing bias mitigation methods are limited to predefined biases, preventing their use in visual datasets where multiple, possibly unknown biases exist. To address this limitation, we introduce MAVias, an open-set bias mitigation approach that leverages foundation models to discover spurious associations between visual attributes and target classes. MAVias first captures a wide variety of visual features in natural language via a foundation image tagging model, and then leverages a large language model to select visual features that define the target class, resulting in a set of language-coded potential visual biases. It then translates these biases into vision-language embeddings and introduces an in-processing bias mitigation approach to prevent the model from encoding information related to them. Experiments on diverse datasets, including CelebA, Waterbirds, ImageNet, and UrbanCars, show that MAVias effectively detects and mitigates a wide range of biases in visual recognition tasks, outperforming current state-of-the-art.

1. Introduction

Computer Vision (CV) progress has been largely driven by Deep Learning (DL) advances [9, 15] and large-scale datasets [6, 22, 26], enabling models to learn complex patterns and visual features with impressive accuracy. However, alongside this progress, concerns have emerged about biases embedded in these models [11, 30, 33, 40, 43, 47] – often stemming from unintended correlations present in the training data [13, 23, 28, 31]. The correlated attributes that are irrelevant act as "shortcuts" and can significantly impact the model's reliability and generalization [12, 37, 38]. It is important to note that in the context of this paper, we define as visual bias any characteristic that does not contribute to defining the target class, which we refer to as "irrelevant". To address this, several methodologies have been developed. Broadly, these fall into two categories: Bias label-Aware (BA) and label-Unaware (BU) methods. BA methods leverage the annotations of attributes introducing the biases to address them [2, 20, 44, 48]. BU methods focus on extracting bias pseudo-labels in cases of extreme biases where a bias-proxy model (or bias-capturing classifier) can be trained using the task's target labels that closely align with the bias labels [16, 29, 32, 41].

While effective in certain contexts, both BA and BU methods have limited applicability when there are multiple, complex, and possibly unknown biases. Common challenging scenarios include:

- **Unknown biases**: In large general-purpose CV datasets, such as ImageNet, biases are difficult to identify and

largely remain unknown, as they vary widely across different classes and are not prominent enough to allow training of bias proxy models. For instance, in the ImageNet9 example of Tab. 1, a sample labeled as a *dog* could introduce biases related to the background scene (e.g., *armchair*, *couch*, *pillow*, *red*), the color of the dog (e.g., *black* and *white*), or accessories like a *neckband*.

- **Potential biases beyond a predefined set**: The CelebA example in Tab. 1, shows that beyond the *hair color*, additional biases may be present, such as clothing styles (e.g., *business suit*, *tie*).
- **Poor representation of predefined bias**: In some cases, biases are reduced to single labels, such as "rural background" in the UrbanCars dataset. However, as shown in Tab. 1, more nuanced descriptors at the instance level (e.g., *path*, *tree*, *wood*, *forest*, *hydrant*, *red*, etc.) provide a richer representation of bias.

Existing BA and BU methods are not designed to mitigate such biases, leading to models that are biased and/or do not achieve their optimal generalization potential. To address this, Mitigate Any Visual bias (MAVias) introduces a flexible and scalable solution, capable of identifying and mitigating open-set biases in CV datasets. MAVias begins by extracting descriptive tags that capture various visual features, such as general-purpose objects, actions, scenes, and visual attributes. Recent advancements in image tagging [53] cover large and comprehensive vocabularies (i.e., $> 4,000$ tags), effectively meeting the open-set requirements of MAVias. Then, these tags are processed by a Large Language Model (LLM) to identify which ones are irrelevant to each of the target classes, leading to a rich set of *language-encoded visual biases*, text descriptions of visual characteristics that are irrelevant to the classification task at hand. MAVias translates these biases into vision-language embeddings, projects them to the main model's backbone space and then to the classification layer. This projection layer is trained simultaneously alongside the main model, and during training, the output logits are a linear combination of the logits of the main model and those of the projection layer that captures visual biases. This setup allows the main model to be exposed to the biased features – those representing irrelevant information to the target class – in a controlled way that leads to bias-invariant representations. Overall, MAVias provides an effective, end-to-end solution for identifying and mitigating biases in open-set scenarios. We evaluate the proposed method on several datasets involving single-attribute biases (CelebA, Waterbirds) as well as multi-attribute predefined biases (UrbanCars), demonstrating state-of-the-art performance. Furthermore, experiments were conducted on ImageNet9, involving unknown biases, where the suggested approach demonstrates significant gains (from 5.24% to 10.21%) in terms of accuracy compared to existing competitive approaches.

The main contributions of this paper are the following:

- A framework for identifying instance-specific open-set potential visual biases in CV datasets.
- A learning strategy that exploits foundation models to learn bias-invariant representations that force the under-training model to avoid encoding any number of identified potential biases per training sample.
- An extensive evaluation study including 4 thematically diverse datasets demonstrating the effectiveness and general applicability of MAVias, which outperforms the state-of-the-art.

MAVias implementation is available as part of the VB-Mitigator library [39].

2. Related Work

Bias identification. Several recent methods leverage text (such as captions, keywords, or tags) for bias detection, highlighting the potential of this approach in this domain. For instance, Say My Name (SaMyNa) [4] is an explainability method that tries to discover model biases through text-based pipelines. Similarly, the Bias-to-Text (B2T) [21] and Language-guided (Lg) [54] frameworks discover biases by extracting common keywords from the captions of misclassified images. These methods aim to identify the main source of bias rather than discovering potentially biased visual characteristics in an open-set setting. Furthermore, OpenBias [8] is a framework for detecting biases in text-to-image generative models by leveraging LLMs to propose potential biases from captions. Although operating in a different domain (text-to-image generation), similarly to MAVias OpenBias acknowledges the need for discovering biases in an open-set setting. However, it provides an LLM with text-only image descriptions asking for potential types of bias thus missing visual grounding of the depicted information, which could offer essential context and semantics. In contrast, MAVias takes a more structured and systematic approach; it evaluates each descriptive tag by querying the LLM to determine whether it is directly relevant to the target class using a detailed prompt (specified in the supp. material).

Furthermore, bias identification methods typically use a vanilla model and validation data to infer specific biases, while MAVias focuses on defining irrelevant visual features *a priori*, and leveraging them in model training to mitigate them. While some open-set identification methods can provide bias labels to define subgroups for use by existing bias mitigation approaches [4], it is known that mitigation methods struggle with multi-attribute biases and cannot scale beyond single-attribute subgrouping [25]. In contrast, MAVias uses instance-level irrelevant visual features without relying on dataset statistics (i.e., neither bias labels nor subgroups w.r.t. them are defined), allowing it to handle complex biases effectively.

Table 1. Examples of extracted tags for various datasets. Red color indicates the irrelevant tags ($\mathcal{B}^{(i)}$).

Dataset	ImageNet9	Waterbirds	UrbanCars	CelebA
Target Class	dog	bird species	car type	hair color
Bias Type	unknown	background	background and objects	gender
Sample				
Extracted Tags	armchair, black, chair, couch, dog, neckband, pillow, red, sit, white	bamboo, bamboo forest, bird, blue, branch, green, hide, parrot, perch, sit, stand, stem, tree, yellow	car, path, forest, hydrant, lush, park, red, road, sedan, silver, SUV, tree, white, wood	black, business suit, dress, dress shirt, man, stand, stare, suit, tie, wear

Bias mitigation. Recent bias mitigation methods include those with direct access to the labels of attributes introducing bias (i.e., BA methods) [2, 16, 20, 25, 36, 38, 44, 48] and those that do not take advantage of such labels but instead rely on deriving pseudo-labels (i.e., BU methods) [1, 3, 32, 41, 46, 49, 51]. For instance, Learning Not to Learn (LNL) [20] is a BA method that discourages the model from predicting the attribute introducing bias, while Bias Contrastive-Bias Balance (BC-BB) [16] and FairKL [2] rely on the bias labels to enforce bias-neutral pairwise similarities between the samples using contrastive learning. Some methods have indirect access to the bias labels by utilizing bias-capturing classifiers trained on different datasets offering bias-related information explicitly [41, 42]. Finally, other methods infer pseudo-labels from the biased vanilla model to identify the biases [1, 5, 32]. While these methods have been effective in mitigating biases, they primarily depend on predefined bias labels or pseudo-labels derived from biased models. On the contrary, this work explores open-set scenarios and introduces a flexible bias mitigation approach that can discover and handle multiple, diverse biases without requiring a bias-labeling system.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ be a dataset consisting of N images, where $\mathbf{x}^{(i)} \in \mathcal{X}$ represents an input image, and $y^{(i)} \in \mathcal{Y}$ is the corresponding target label. The goal is to train a DL model $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$, parameterized by $\boldsymbol{\theta}$, to predict the target label $y^{(i)}$ given an image $\mathbf{x}^{(i)}$, while mitigating a set of potential biases, $\mathcal{B}^{(i)}$, present in $\mathbf{x}^{(i)}$ that may lead to biased predictions. In this context, the term "potential biases" refers to all visual attributes present in $\mathcal{X}$ that are irrelevant to the target class, such as elements of the background of images or other so-called "shortcuts".

3.2. Method

3.2.1. Language-driven Bias Modeling

In many general-purpose CV datasets, biases manifest through visual information that can be described in text. MAVias, first, utilizes an image tagging model to extract tags that describe the visual information present in an image. Formally, for each image $\mathbf{x}^{(i)}$, we derive a set of tags $\mathcal{T}^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \ldots, t_{m_i}^{(i)}\}$, where m_i represents the total number of tags for i-th sample. These capture various visual attributes, including colors, objects, backgrounds, and other features that can either describe the target class or potentially introduce bias.

The next step is to filter out tags that should not influence the decisions of an unbiased classifier. To achieve this, we leverage an LLM to identify which tags are irrelevant to the target class $y^{(i)}$. We denote this subset of potential biases as $\mathcal{B}^{(i)} \subseteq \mathcal{T}^{(i)}$. These encapsulate visual features that could lead to biased predictions if considered by the model. To ensure the LLM correctly identifies these irrelevant tags, we carefully design the prompt used for this task by providing precise instructions on what tag should be considered relevant (e.g., physical components, defining features, inherent characteristics, etc.) or not (e.g., background details, lighting, textures, other objects, etc.). Details on the prompt formulation process are provided in the supp. material. Table 1 shows several examples of $\mathcal{B}^{(i)}$ for samples belonging to different datasets.

For each set of irrelevant tags $\mathcal{B}^{(i)}$ associated with an image $\mathbf{x}^{(i)}$, we employ a vision-language model to generate a single embedding $\mathbf{e}^{(i)} \in \mathbb{R}^d$, where d is the dimension of the embedding. This is produced using the prompt "a photo of $t_1^{(i)}, t_2^{(i)}, \ldots, t_{k_i}^{(i)}$", where $k_i \leq m_i$ is the number

of irrelevant tags for the i-th sample. This aggregate embedding captures the combined information from all irrelevant tags, providing an instance-level representation of the biased features that could affect the model's behavior. Comparison with alternative embedding approaches is provided in Sec. 4.6.

3.2.2. Bias Mitigation

We define the main model $f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$, a DL classifier composed of the following: (i) a backbone $f_{\boldsymbol{\theta}_{bb}}(\mathbf{x}^{(i)})$, which extracts feature representations $\mathbf{h}^{(i)} \in \mathbb{R}^r$, where r is the feature vector size; (ii) a classification head $f_{\boldsymbol{\theta}_c}(\mathbf{x}^{(i)})$ outputting the logits $\mathbf{z}^{(i)}_{\text{main}} = f_{\boldsymbol{\theta}_c}(\mathbf{h}^{(i)}) \in \mathbb{R}^p$, where $p = |\mathcal{Y}|$. The overall main model is expressed as $f_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = f_{\boldsymbol{\theta}_c}(f_{\boldsymbol{\theta}_{bb}}(\mathbf{x}^{(i)}))$. In parallel, we introduce a projection layer $g_{\boldsymbol{\phi}}$, parameterized by $\boldsymbol{\phi}$, which takes the visual bias embeddings $\mathbf{e}^{(i)}$ as input and outputs embeddings $\mathbf{b}^{(i)} \in \mathbb{R}^p$. Note that $g_{\boldsymbol{\phi}}$ is employed to project $\mathbf{e}^{(i)}$ to the feature space of $\mathbf{h}^{(i)}$. Then the corresponding logits are derived through $\mathbf{z}^{(i)}_{\text{tag}} = f_{\boldsymbol{\theta}_c}(\mathbf{b}^{(i)}) \in \mathbb{R}^p$. The final logits $\mathbf{z}^{(i)}$ for each sample are the addition of the main model logits $\mathbf{z}^{(i)}_{\text{main}}$ and the visual bias logits $\mathbf{z}^{(i)}_{\text{tag}}$:

$$\mathbf{z}^{(i)} = \mathbf{z}^{(i)}_{\text{main}} + \mathbf{z}^{(i)}_{\text{tag}}. \tag{1}$$

It is worth noting that bias often arises during DL model training because biased attributes in the training data are easier to learn and thus dominate gradient updates [35]. To counteract this, MAVias incorporates visual bias logits into the main model's logits, ensuring that as the bias in a sample increases, its impact on gradient updates is reduced. The intuition behind this mechanism is that for bias-aligned samples, the value of $\mathbf{z}_{tag}$ is high, effectively reducing the magnitude of $\mathbf{z}_{main}$ and its contribution to the total logits $\mathbf{z}$. This leads to significantly reduced gradients for these samples. This is supported both empirically in Sec. 4.5 and theoretically in the supp. material. In other words, $\mathbf{z}_{tag}$ assists in decoupling the learning of biased features from the actual task at hand, which allows the main model to focus on the relevant features for the target prediction.

Furthermore, since both the main model and the projection layer are trained concurrently, it is essential to ensure the stability of the training process. To achieve this, we introduce a loss function that combines the classification loss with a logit alignment term. The classification loss ensures that the combined logits $\mathbf{z}^{(i)}$ accurately predict the target label $y^{(i)}$, while the alignment term controls the relative magnitudes of the main model's logits $\mathbf{z}^{(i)}_{\text{main}}$ and the visual bias logits $\mathbf{z}^{(i)}_{\text{tag}}$. By doing so, we prevent $g_{\boldsymbol{\phi}}$ from dominating or being overshadowed by $h_{\boldsymbol{\theta}}$. Specifically, the loss is computed as: $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_{cls}(\mathbf{z}^{(i)}, y^{(i)}) + \alpha \cdot \mathcal{L}_{align}(\mathbf{z}^{(i)}_{\text{main}}, \mathbf{z}^{(i)}_{\text{tag}})$, where: (i) $\mathcal{L}_{cls}(\mathbf{z}^{(i)}, y^{(i)})$ is the classification loss (e.g., cross-entropy loss) between the final logits $\mathbf{z}^{(i)}$ and the ground truth label $y^{(i)}$; (ii) $\mathcal{L}_{align}(\mathbf{z}^{(i)}_{\text{main}}, \mathbf{z}^{(i)}_{\text{tag}})$ is the logit alignment term for the norm of the logits, calculated as:

$$\mathcal{L}^{(i)}_{align} = \frac{1}{2} \left\| \|\mathbf{z}^{(i)}_{\text{main}}\| - \lambda \cdot \|\mathbf{z}^{(i)}_{\text{tag}}\| \right\|^2 \tag{2}$$

where $\|\cdot\|$ denotes the l^2-norm, $\lambda \in (0, 1)$ is a scaling factor and $\alpha \in (0, 1)$ is a weighting factor that balances the influence of the $\mathcal{L}_{align}(\mathbf{z}^{(i)}_{\text{main}}, \mathbf{z}^{(i)}_{\text{tag}})$ in the total loss function. Typically, the greater the bias in the data, the smaller the value of λ should be, as this leads to smaller $\mathbf{z}_{main}$ values and reduced gradient updates for the bias-aligned samples. Details on the effects of hyperparameters λ and α are provided in the supp. material. The overview of the proposed framework is illustrated in Fig. 2.

4. Experiments

4.1. Datasets

We use Biased-CelebA [16], Waterbirds [38], UrbanCars [25], and Imagenet9 [50]. Biased-CelebA is a subset of the CelebA dataset, containing facial images annotated with 40 binary attributes. In this subset, *BlondHair* is the target attribute, while the *gender* attribute introduces bias with a 90% correlation. Waterbirds features a 95% co-occurrence between waterbirds (or landbirds) and aquatic (or terrestrial) backgrounds. UrbanCars is an artificially generated dataset with a 95% co-occurrence between car body types and relevant urban or rural backgrounds. ImageNet9 consists of 9 coarse ImageNet classes.

4.2. Model Architectures

For comparability purposes, we employ the same network architectures as those used in other works [2, 16, 41, 42]. In particular, for CelebA, we adopt the ResNet-18 architecture [14], for Waterbirds, UrbanCars, and ImageNet9 datasets, we use ResNet-50 networks. In all experiments, the projection layer is a dense layer that gets vision-language embeddings as input and its output size is aligned with the feature size of the main model.

4.3. Implementation Details

The SGD optimizer is employed for all datasets except for CelebA, where Adam optimizer is used. We use an initial learning rate of 0.001, which is divided by 10 every 1/3 of the training epochs. The weight decay is set to 10^{-4}. The batch size is 128 for CelebA and 64 for Waterbirds, UrbanCars, and ImageNet9. Following previous works [16, 25, 38], we train the models for 40, 100, 300, and 40 for CelebA, Waterbirds, UrbanCars, and ImageNet9, respectively. For Waterbirds and UrbanCars, we do not use a learning rate scheduler. The parameters (α, λ) are (0.01, 0.5), (0.05, 0.6), (0.01, 0.4), and (0.001, 0.7) for CelebA,

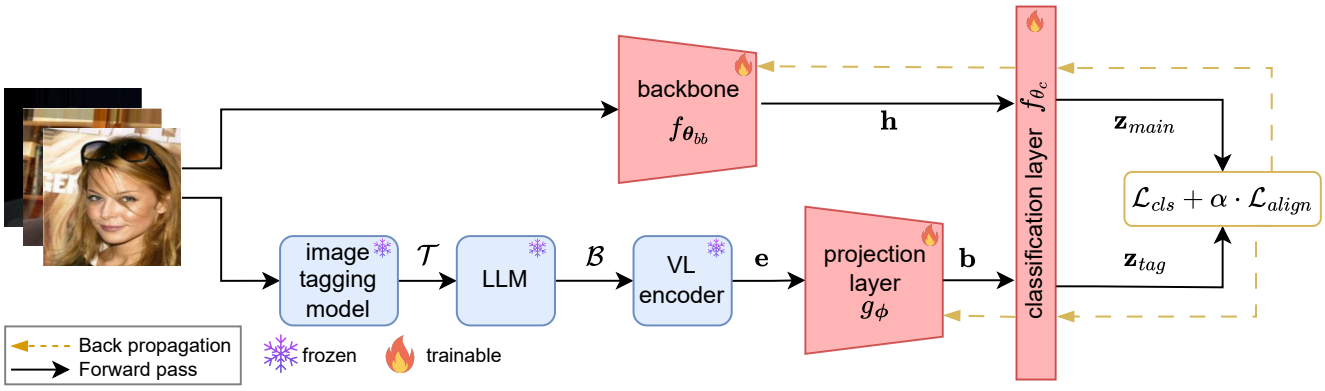


Figure 2. Illustration of the proposed framework for mitigating any visual bias during model training. For inference, only the backbone and the classification layer are considered (i.e., f_θ).

Waterbirds, UrbanCars, and ImageNet9, respectively. RAM [53], GPT-4o [34], OpenCLIP [18] are employed for image tagging, irrelevant tag filtering, and vision-language encoding, respectively. For all the experiments presented in the main manuscript, we run the compared methods using the code provided by the corresponding papers, using the suggested hyperparameters. Specifically, we use 50 epochs for bias discovery and 100 upweight for JTT; $\alpha = 110$ for FLAC-B; 0.1 SD coefficient; and $\alpha = 0.7$ for LfF. For the baseline methods results presented in the supplementary material (i.e., closed-set), we present performance as reported in the corresponding works to avoid potential underperformance due to reimplementation. Experiments were conducted on an NVIDIA A100 GPU. All experiments were repeated for 5 different random seeds.

4.4. Evaluation Protocol

To assess bias, we primarily use worst-group accuracy (WG Acc.), which measures the accuracy of the least-performing group within a dataset, and average accuracy (Avg. Acc.), which is the mean accuracy across all groups and reflects overall model performance. The formation of these groups determines whether the evaluation supports a closed- or open-set bias scenario. Accordingly, we implement two distinct evaluation protocols. The first one is designed to align with our approach, focusing on open-set biases, while the second protocol adheres to established evaluation standards in the literature (i.e., closed-set scenario). Furthermore, since MAVias is designed for BU scenarios, we evaluate its performance against the following widely-employed and competitive BU methods: LfF [32], JTT [27], SD [35], Debian [24], and FLAC-B [41].

Open-set. Here, we form groups based on the presence/absence of the detected open-set bias attributes (i.e., no predefined biases are considered). To achieve that, we use a vanilla model and the potential biases $\mathcal{B}$ extracted through MAVias to identify the subset of tags $\mathcal{B}' \subseteq \mathcal{B}$ that actually introduce bias to the model. We define bias as occurring when the model exhibits increased accuracy on images that include a particular tag, relative to its overall accuracy across the entire dataset. After deriving $\mathcal{B}'$ for each class within the datasets, we categorize each group of images belonging to a class into two sub-groups based on whether the

Table 2. Open-set performance comparison across CelebA, Waterbirds, and UrbanCars datasets.

Dataset	Method	WG Acc.	Avg. Acc.
CelebA	LfF [32]	14.7 ±15.2	67.1 ±4.4
	JTT [27]	31.5 ±8.0	61.6 ±8.5
	SD [35]	13.3 ±8.2	67.4 ±1.9
	Debian [24]	12.0 ±8.7	67.0 ±2.6
	FLAC-B [41]	12.0 ±8.7	65.9 ±2.4
	MAVias	**66.7** ±4.7	**81.4** ±1.8
Waterbirds	LfF [32]	30.0 ±6.8	72.7 ±1.4
	JTT [27]	64.7 ±2.4	85.2 ±4.6
	SD [35]	35.0 ±16.3	75.5 ±4.1
	Debian [24]	37.5 ±0.0	74.7 ±0.5
	FLAC-B [41]	37.5 ±8.8	75.2 ±2.7
	MAVias	**75.4** ±0.9	**87.5** ±1.2
UrbanCars	LfF [32]	34.6 ±2.6	61.0 ±1.4
	JTT [27]	69.0 ±3.3	77.8 ±0.3
	SD [35]	40.4 ±2.7	66.5 ±1.1
	Debian [24]	33.2 ±8.0	61.1 ±2.1
	FLAC-B [41]	28.5 ±4.3	57.3 ±1.7
	MAVias	**84.4** ±2.2	**89.3** ±1.3

Table 3. Open-set performance on ImageNet9 in terms of accuracy across 7 test sets.

Method	MIXED-NEXT (↑)	MIXED-RAND (↑)	NO-FG (↓)	ONLY-BG-B (↓)	ONLY-BG-T (↓)	ONLY-FG (↑)	ORIGINAL (↑)
Vanilla	82.66 ±0.1	85.06 ±0.0	64.16 ±0.1	35.18 ±0.1	44.74 ±0.1	<u>93.12</u> ±0.0	97.69±0.0
LfF [32]	78.70 ±0.1	81.47 ±0.2	61.07 ±0.1	34.82 ±0.2	44.46 ±0.0	88.99 ±0.2	94.34 ±0.2
JTT [27]	84.43 ±0.1	86.16 ±0.5	61.09 ±2.0	32.04 ±1.0	<u>36.62</u> ±4.7	92.09 ±0.5	97.71 ±0.1
Debian [24]	83.02 ±0.4	85.64 ±0.3	64.53 ±0.4	34.45 ±0.1	45.00 ±0.6	93.06 ±0.1	<u>97.89</u> ±0.1
SD [35]	<u>87.56</u> ±0.57	<u>88.92</u> ±0.74	62.60 ±1.05	31.42 ±2.93	40.81 ±3.00	**93.71** ±0.71	**98.16** ±0.06
FLAC-B [41]	84.60 ±0.46	86.62 ±0.45	<u>59.84</u> ±1.67	<u>29.71</u> ±0.53	40.38 ±1.28	92.72 ±0.73	97.89 ±0.09
MAVias	**88.26**±0.1 (+0.70)	**89.64**±0.2 (+0.72)	**53.02**±0.7 (-6.82)	**21.83**±0.4 (-7.88)	**32.48**±0.6 (-4.14)	91.90±0.4 (-1.81)	96.92±0.2 (-1.24)

samples contain at least one of the biased tags. This results in $2 \times p$ groups. Subsequently, we measure the WG Acc. and Avg. Acc. across all groups. To ensure a fair assessment, the optimal training epoch for each method is selected based on the overall accuracy, without considering any information related to the biases. An exception to this protocol is ImageNet9. In the original ImageNet9 test set, we observed that for several classes, the subgroups without any biases contain very few samples (fewer than 4), making reliable evaluation difficult. We therefore use the official seven test set variations, which allow for a more comprehensive evaluation of the model's reliance on factors beyond the target object: ORIGINAL, ONLY-BG-B (bounding box of a black object), ONLY-BG-T (bounding box of an inpainted object), NO-FG Black (segmented object removed), ONLY-FG (black background), MIXED-RAND (random background of a random class), and MIXED-NEXT (random background of the next class).

Closed-set. Here, we follow the protocols suggested by the dataset providers and previous works for predefined biases. In particular, for CelebA we use the accuracy of the underrepresented groups (i.e., bias-conflicting accuracy) and the average accuracy across all groups (i.e., unbiased accuracy) [16, 41]. For Waterbirds, we use the WG Acc. and the Avg. Acc. across all groups. In the case of UrbanCars, we calculate the weighted average accuracy across groups, referred to as In-Distribution Accuracy (I.D. Acc), with weights determined by group representation ratios. Furthermore, I.D. Acc serves as a baseline for assessing accuracy drops related to background (BG Gap), co-occurring objects (CoObj Gap), and both background and co-occurring objects combined (BG+CoObj Gap).

4.5. Comparative Analysis

Table 2 presents the open-set performance comparison across the CelebA, Waterbirds, and UrbanCars datasets. For CelebA, MAVias significantly outperforms competing methods by achieving a 66.7% (+35.2%) WG accuracy and 81.4% (+14%) Avg. accuracy. As expected, most existing BU methods exhibit poor performance, as they struggle to handle scenarios with multiple concurrent biases [25]. For the Waterbirds, MAVias reaches a WG accuracy of 75.4% (+10.7%) and an Avg. accuracy of 87.5% (+2.3%), surpassing competing approaches by a notable margin. Similarly, in the UrbanCars dataset, MAVias achieves a WG accuracy of 84.4% (+15.4%) and an Avg. accuracy of 89.3% (+11.5%). For ImageNet9, as shown in Tab. 3, MAVias achieves accuracy improvements on the test sets with modified backgrounds, MIXED-NEXT and MIXED-RAND (+0.7% and +0.72).

Similarly, for the sets where background information is suppressed (ONLY-BG-B, ONLY-BG-T, and NO-FG) MAVias achieves improvements of +7.88%, +4.14%, and +6.82%, respectively. As observed, most competitive methods, including MAVias, underperform on ONLY-FG compared to the vanilla model. This is likely because the vanilla model can exploit biases present in the foreground (e.g., colors) to increase its accuracy. Finally, there is a 1.24% drop in performance on the ORIGINAL test set, which aligns with expectations, as MAVias is designed to rely less on shortcuts that boost overall performance. The top-10 biased tags for each Imagenet9 class are reported in Tab. 4. Moreover, Tab. 5 reports the top-10 irrelevant tags derived using MAVias for CelebA, Waterbirds, and UrbanCars datasets. For CelebA, MAVias successfully detects the primary bias (i.e., gender) while also revealing other biases associated with facial expressions (smile), shot types (selfies), and clothing items (e.g., dresses, business suits, or ties). In the case of Waterbirds, we observe that, in addition to background elements, waterbirds are predominantly depicted in flight, whereas landbirds are often observed perched on tree branches. Finally, for UrbanCars, as anticipated, the top irrelevant tags correspond to objects that are common in urban/rural environments.

For the closed-set evaluation, as reported in Tab. 6 MAVias achieves +2.7% (89.7%) Unbiased and +2.2 (87.1%) Bias-conflict accuracy compared to the second-best performing method, on CelebA. On Waterbirds, MAVias attains 93.7% WG Acc., outperforming all compared BU methods (+5%), while slightly improving the average accuracy (+0.7%). On UrbanCars, MAVias significantly enhances performance by reducing the BG, CoObj, and BG+CoObj gaps to 4.1, 2.4, and 6.7, respectively, representing absolute improvements of 4, 8.1, and 33.4. The full closed-set results are provided in the suppl. material.

In addition, by examining the logit spaces, as shown in

Table 4. Top-10 biased tags for each class of ImageNet9.

Class	Top-10 Irrelevant Tags
Bird	perch, sit, branch, tree, tree branch, water, blue, twig, brown, sky
Carnivore	stand, grass, stone, lush, fur, walk, lay, enclosure, tree, red
Dog	brown, white, stand, black, lush, stare, green, carpet, blanket, bed
Fish	sea, swim, water, underwater, aquarium, tank, coral reef, catch, man, yellow
Insect	green, plant, break, floor, sit, stem, leaf, close-up, white, yellow
Instrument	play, woman, ceiling, table, pillar, cloth, hang, band, tool, hand
Primate	tree, sit, branch, tree branch, stare, black, enclosure, brown, stone, log
Reptile	floor, stone, green, branch, tree, grass, tree branch, water, lay, sit
Vehicle	road, track, red, building, curb, equipment, travel, load, train track, railroad

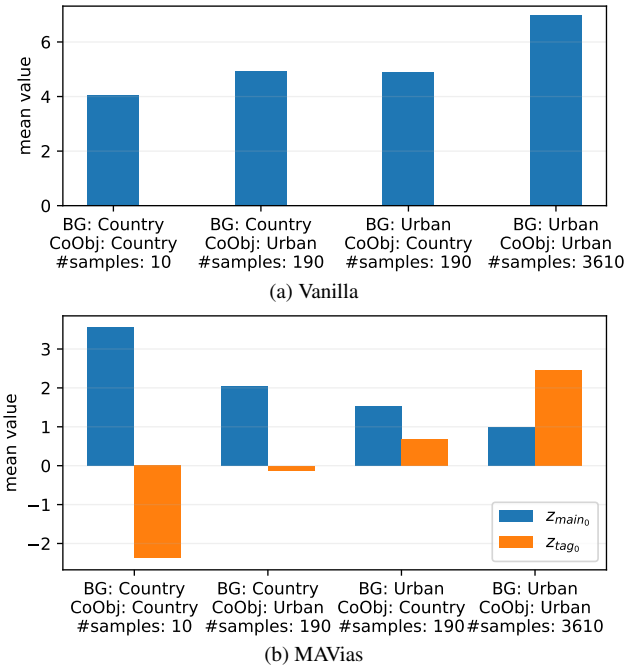
Table 5. Top-10 biased tags for each class of CelebA, Waterbirds, and UrbanCars.

$\mathcal{D}$	Class	Top-10 Irrelevant Tags
CelebA	Blonde	woman, dress, pose, smile, girl, beautiful, curl, selfie, actor, carpet
	Non-Blonde	wear, man, shirt, selfie, tie, black, stand, stare, dress shirt, business suit
Waterbirds	Landbird	stand, perch, sit, tree, yellow, branch, floor, brown, forest, green
	Waterbird	water, stand, white, sea, fly, lake, sky, ledge, pond, stone
UrbanCars	Country Car	road, rural, animal, stand, white, jump, cow, lift, bull, highway
	Urban Car	park, house exterior, house, red, building, flip, home, lift, white, jump

Fig. 3 (a), the vanilla model increases logit values proportionally with the bias, ranging from 4 to 7. This is an expected behavior, as the biases are easier to learn than the actual target and thus the model illustrates higher confidence for the bias-aligned samples. This phenomenon is referred to as Gradient Starvation [35] and describes how bias is introduced. Also, it is theoretically shown that methodologies penalizing high logit values can mitigate this phenomenon [35]. As shown in Fig. 3 (b), the interactions between the main model and the projection layer in the training procedure (defined by Eq. (1) and (2)) reduce the main model's logit values proportionally to the data bias.

Table 6. Brief closed-set performance comparison between MAVias and competitive BU methods.

Dataset	Metric	Best BU Method	MAVias
CelebA	Unbiased	87.0±0.6 [41]	**89.7**±0.6
	Bias-Conflict	84.9±2.2 [41]	**87.1**±1.7
Waterbirds	WG Acc.	88.7±0.4 [49]	**93.7**±0.4
	Avg. Acc.	93.8±0.7 [49]	**94.5**±0.4
UrbanCars	BG Gap	8.1 [27]	**4.1**±0.6
	CoObj Gap	10.5 [24]	**2.4**±1.4
	BG+CoObj Gap	40.1 [27]	**6.7**±1.4



(a) Vanilla

(b) MAVias

Figure 3. Logits for UrbanCars training samples belonging to groups defined by the *urban car* class and Background (BG) and Co-occurring Object (CoObj) biases.

4.6. Ablation Analysis

An important aspect in MAVias is how the embeddings are calculated. Table 7 compares two well-known vision-language models, i.e., OpenCLIP and SigLip, and the text-based model BERT. Furthermore, Tab. 8 compares MAVias's performance using various LLMs. GPT-4o shows the highest performance, with open-source models also performing competitively. For experiments, we use GPT-4o to minimize LLM-related errors. Table 9 presents the performance of GPT-4o in deriving relevant tags for the Waterbirds, CelebA, and UrbanCars datasets. The ground truth was manually established.

Furthermore, we investigate the impact of the different

Table 7. Impact of the employed encoders on the MAVias performance. Results pertain to the Waterbirds dataset.

Model	WG Acc.	Avg. Acc.
BERT-L [7]	91.7$\pm$1.3	92.8$\pm$1.0
SigLip-L-16-256 [52]	92.7$\pm$1.2	93.5$\pm$1.0
OpenCLIP (ViT-L-14) [18]	93.7$\pm$0.4	94.5$\pm$0.4

Table 8. Impact of the employed LLM on the MAVias performance. Results pertain to the Waterbirds dataset.

Model	#parameters	WG Acc.	Avg. Acc.
Qwen2.5 [17]	7B	91.9$\pm$2.4	94.4$\pm$ 0.5
Mistral-small [19]	22B	92.9$\pm$0.4	94.9$\pm$0.2
Gemma2 [45]	9B	93.5$\pm$0.6	94.5$\pm$0.6
Llama3.1 [10]	8B	93.6$\pm$0.6	94.4$\pm$0.4
GPT-4o [34]	>>175B	93.7$\pm$0.4	94.5$\pm$0.4

Table 9. Performance of GPT-4o.

Dataset	Precision	Recall
Waterbirds	96.1	79.0
CelebA	81.8	75.0
UrbanCars	89.2	71.7

Table 10. Impact of Eq. (2) on the MAVias performance. Results pertain to Waterbirds dataset.

Type	WG Acc.	Avg. Acc.
w/o Eq. (2)	89.4$\pm$1.6	95.2$\pm$0.4
w/ Eq. (2)	93.7$\pm$0.4	94.5$\pm$0.4

MAVias components. First, as discussed in Sec. 3, the logit alignment term in Eq. (2) is crucial for balancing the logits between the two models. Table 10 shows the impact of removing this term from the training on the effectiveness of MAVias.

In addition, let us report the accuracy of $g_\phi(\cdot)$ on the training samples of the *urban car* class from the UrbanCars dataset, which correlates the target classes with relevant backgrounds and co-occurring objects. As shown in Tab. 11, the g_ϕ shows a clear distinction in its performance: its accuracy increases from 0% for bias-conflicting samples (i.e., those with a *country* background and co-occurring object) to 98.42% for bias-aligned samples (i.e., those with an *urban* background and co-occurring object). This result confirms that our tag-based approach produces embeddings that effectively capture bias, providing a foundation for the main model $f_\theta(\cdot)$ to focus on learning unbiased features.

Furthermore, we investigate the impact of the vocabulary size employed for image tagging. The original vocabulary of RAM has a size of 4585 words. Tab. 12 reports the performance of MAVias for different portions of the original vocabulary. Notably, using 30% or more of the original vocabulary yields highly effective models.

Table 11. MAVias: $g_\theta(\cdot)$ accuracy for UrbanCars training samples belonging to subgroups of *urban car* class with different biases.

BG	CoObj	#samples	$g_\theta(\cdot)$ acc.
Country	Country	10	00.00
Country	Urban	190	54.21
Urban	Country	190	67.36
Urban	Urban	3610	98.42

Table 12. Performance of MAVias across varying portions of the original tag vocabulary. Results pertain to the Waterbirds dataset.

size	WG Acc.	Avg. Acc.
10%	78.1	87.1
20%	79.4	89.0
30%	91.1	92.0
40%	93.1	92.3
50%	93.8	93.2
60%	93.2	93.5
70%	92.9	94.2
80%	93.0	94.3
90%	93.6	94.5
100%	93.7	94.5

5. Conclusion

We presented a method to address open-set biases in CV using instance-level descriptive tags. Unlike previous methods relying on predefined biases, MAVias offers a flexible solution for identifying and mitigating unknown biases. It detects visual features that may introduce bias and reduces their impact. Extensive experiments show MAVias outperforms existing methods in detecting and mitigating complex, real-world biases. The main requirement is an image tagging model with a comprehensive tag vocabulary and an LLM suited to the application context. For example, cloud-deployed models are not ideal for applications where sensitive data is involved, e.g., medical images, images of personal identity documents, etc.

Acknowledgments

This research was supported by the EU Horizon Europe projects MAMMOth (grant no. 101070285), ELIAS (grant no. 101120237), and ELLIOT (grant no. 101214398).

References

[1] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 3

[2] Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. *arXiv preprint arXiv:2211.05568*, 2022. 1, 3, 4

[3] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 3

[4] Massimiliano Ciranni, Luca Molinaro, Carlo Alberto Barbano, Attilio Fiandrotti, Vittorio Murino, Vito Paolo Pastore, and Enzo Tartaglione. Say my name: a model's bias discovery framework. *arXiv preprint arXiv:2408.09570*, 2024. 2

[5] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 8

[8] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235, 2024. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8

[11] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223: 103552, 2022. 1

[12] Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. *Advances in neural information processing systems*, 31, 2018. 1

[13] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[16] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34: 26449–26461, 2021. 1, 3, 4, 6

[17] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. 8

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 5, 8

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 8

[20] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 1, 3

[21] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092, 2024. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[23] Haoxuan Li, Chunyuan Zheng, Sihao Ding, Peng Wu, Zhi Geng, Fuli Feng, and Xiangnan He. Be aware of the neighborhood effect: Modeling selection bias under interference for recommendation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[24] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate net-

works. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022. 5, 6, 7, 3

[25] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023. 2, 3, 4, 6

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[27] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 5, 6, 7, 3

[28] Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[29] Shenyu Lu, Yipei Wang, and Xiaoqian Wang. Debiasing attention mechanism in transformer without demographics. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021. 1

[31] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[32] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1, 3, 5, 6

[33] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356, 2020. 1

[34] OpenAI. Chatgpt-4, 2023. [Large language model]. 5, 8

[35] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 4, 5, 6, 7

[36] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pages 28448–28467. PMLR, 2023. 3

[37] Ryan Ramos, Vladan Stojnic, Giorgos Kordopatis-Zilos, Yuta Nakashima, Giorgos Tolias, and Noa Garcia. Processing and acquisition traces in visual encoders: What does clip know about your camera? In *International Conference on Computer Vision*, 2025. 1

[38] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 3, 4

[39] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Vb-mitigator: An open-source framework for evaluating and advancing visual bias mitigation. *arXiv preprint arXiv:2507.18348*. 2

[40] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Towards fair face verification: An in-depth analysis of demographic biases. In *Proceedings of the International Workshops of ECML PKDD*, 2023. 1

[41] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Flac: Fairness-aware representation learning by suppressing attribute-class associations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3, 4, 5, 6, 7

[42] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Badd: Bias mitigation through bias addition. *arXiv preprint arXiv:2408.11439*, 2024. 3, 4

[43] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Facex: Understanding face attribute classifiers through summary model explanations. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 758–766, 2024. 1

[44] Enzo Tartaglione, Carlo Alberto Barbano, and Marco 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. 1, 3

[45] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 8

[46] Rahul Venkataramani, Parag Dutta, Vikram Melapudi, and Ambedkar Dukkipati. Causal feature alignment: Learning to ignore spurious background features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4666–4674, 2024. 3

[47] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022. 1

[48] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for

bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 1, 3

[49] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, pages 37765–37786. PMLR, 2023. 3, 7

[50] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 4

[51] Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. PMLR, 2024. 3

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 8

[53] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 2, 5

[54] Zaiying Zhao, Soichiro Kumano, and Toshihiko Yamasaki. Language-guided detection and mitigation of unknown dataset bias. *arXiv preprint arXiv:2406.02889*, 2024. 2