

TLDR: Text Based Last-layer Retraining for Debiasing Image Classifiers

Juhyeon Park^{1*}, Seokhyeon Jeong^{2*}, and Taesup Moon^{1, 2, 3†}

{parkjh9229, sh102201, tsmoon}@snu.ac.kr

¹ IPAI, Seoul National University ² ECE, Seoul National University

³ ASRI / INMC / AIIS, Seoul National University

Abstract

An image classifier may depend on incidental features stemming from a strong correlation between the feature and the classification target in the training dataset. Recently, Last Layer Retraining (LLR) with group-balanced datasets is shown to be efficient in mitigating the spurious correlation of classifiers. However, the acquisition of image-based group-balanced datasets is costly, which hinders the general applicability of the LLR method. In this work, we propose to perform LLR based on text datasets built with large language models to debias a general image classifier. To that end, we demonstrate that text can generally be a proxy for its corresponding image beyond the image-text joint embedding space, which is achieved with a linear projector that ensures orthogonality between its weight and the modality gap of the joint embedding space. In addition, we propose a systematic validation procedure that checks whether the generated words are compatible with the embedding space of CLIP and the image classifier, which is shown to be effective for improving debiasing performance. We dub these procedures as **TLDR (Text-based Last layer retraining for Debiasing image classifieRs)** and show our method achieves the performance that is competitive with the LLR methods that require group-balanced image dataset for retraining. Furthermore, TLDR outperforms other baselines that involve training the last layer without any group annotated dataset. Codes: <https://github.com/beotborry/TLDR>

1. Introduction

An image classifier may grant excessive importance to an inconsequential attribute of an input image as a result of detecting a strong correlation between the target and the attribute discovered in the training dataset. Such *spurious* correlations can present a substantial problem in domains

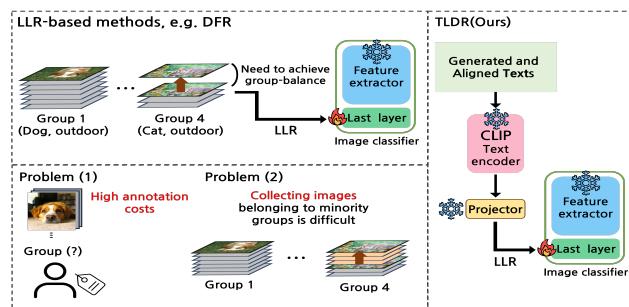


Figure 1. (Left) Typical procedure of LLR-based methods for debiasing an image classifier, which still suffers from high annotation cost for collecting group-balanced holdout image dataset, particularly when the number of classes and groups increases. (Right) In contrast, TLDR effectively constructs a sufficient group-balanced set using only with *texts*, by generating a substantial amount of data through LLMs with ease, to debias *any* general image classifier. More details are depicted in Fig. 2.

such as medical AI and autonomous driving, in which classification errors can cause severe consequences for humans.

To that end, numerous methods have been proposed to reduce the classifier’s reliance on the spurious features, *i.e.*, to *debias* the classifier [11, 15, 19, 21, 25, 32]. Among those, DFR [15] was recently proposed — it suggests collecting a small holdout dataset that has a balanced number of data samples for each group, which stands for a sub-dataset with the same spurious attribute and target labels, and only retrain the last linear classification layer of a biased classifier based on the collected group-balanced holdout data. Such a procedure is often dubbed as Last Layer Re-training (LLR) and is shown to be very effective in debiasing a classifier. However, such an approach of collecting a group-balanced holdout dataset has a few critical limitations as shown in Fig. 1. Namely, the size of the holdout set still needs to be considerable to ensure the debiasing (namely, cannot be in the regime of *few-shot* setting), hence, the annotation cost becomes expensive. Moreover, collecting a group-balanced dataset becomes significantly more difficult as data pertaining to the minority group is often scarce and hard to collect in the wild. Furthermore, as the number of classes or

*Equal contribution

†Corresponding author

spurious attributes increases, the cost of collecting a group-balanced image dataset increases exponentially.

In order to lift the burden of the requirement of collecting additional group-balanced datasets, DrML [33] recently proposed a way to retrain the last linear layer of an image classifier only with *text* data, thanks to the availability of the powerful joint embedding space produced by a multimodal pre-trained encoder, such as CLIP [23]. Namely, for an image classifier that consists of a linear layer operating on top of the CLIP image embedding space, they showed that it is possible to achieve *cross-modal transferability* — *i.e.*, the text embedding can be used as a proxy for its corresponding image embedding in CLIP embedding space. Based on this finding, they used the text data to retrain the biased linear layer and showed the resulting debiased classifier still can work well with the image embeddings.

However, we argue their result was limited since the proposed method is only applicable to the image classifiers that are defined on top of the joint embedding space. To the best of our knowledge, whether the cross-modal transferability can be extended to a more general embedding space, *e.g.*, image embedding space obtained by a ResNet classifier, and hence, whether the text can be used to debias a general image classifier has not been investigated yet. In addition, DrML solely relied on the additional metadata for each benchmark dataset to generate the text dataset, which again limits the applicability to a general setting in which such metadata is not available. Furthermore, using only the metadata to generate the text dataset may not be sufficient to achieve optimal debiasing performance due to its limited quantity or incompatibility with the embedding space of CLIP.

To that end, we aim to address the limitations of both DFR [15] and DrML [33] to develop a framework for debiasing a *general* image classifier solely using text. Namely, unlike DFR, we remove the requirement of collecting annotated group-balanced *image* data by developing general mechanisms for generating text data, without using any metadata. Moreover, unlike DrML, we achieve cross-modal transferability across different embedding spaces so that our method can be applied to debiasing general image classifiers. Specifically, we first develop a linear projector that can project a CLIP embedding to another image embedding space while preserving the cross-modal transferability, considering the concept of the modality gap [18] of CLIP joint embedding space. Then, only with the knowledge of the existence and the type of spurious correlations, we *generate* group-balanced text data using publicly available Large Language Models (LLM), such as GPT [1] or LLaMA [28]. The generated texts, which are not confined to some limited metadata, are then followed by a systematic procedure to only select those that are appropriate for debiasing. Finally, we retrain the last layer of the biased image classi-

fier using the projected CLIP text embeddings of the generated group-balanced text data. In our experimental results, we show our method, dubbed as TLDR (**T**ext-based **L**ast layer retraining for **D**ebiasing image classifie**R**s), outperforms other state-of-the-art LLR-based methods that do not require group-annotated data as ours and is competitive with DFR [15], which additionally uses the expensive group-annotated, group-balanced *image* dataset.

Our contributions are summarized as follows.

- We propose a sufficient condition for preserving cross-modal transferability within the general image classifier’s embedding space when linear alignment between embedding spaces is possible. Moreover, we present a closed-form solution of a linear projector between embedding spaces with the condition, which eliminates the need for gradient-based optimization to learn the projector.
- We introduce a systematic process to construct a sufficient number of text data for LLR and select only those that are well-aligned with the embedding spaces of CLIP and the given image classifier.
- We experimentally demonstrate that our TLDR achieves competitive performance in debiasing a general image classifier and show it is particularly effective when the minority group has a considerably low data proportion.

2. Preliminaries

2.1. Problem Setting

Our work shares the group robustness problem setting introduced in [25]. The data distribution can be specified by the group $\mathcal{G}$ defined as the Cartesian product of a set of labels $\mathcal{Y}$ and a set of spurious attributes $\mathcal{A}$, *i.e.*, $\mathcal{G} := \mathcal{Y} \times \mathcal{A}$. For example, in the Waterbirds dataset [25], the label indicates whether a bird in an image is a landbird or waterbird, and the spurious attribute is the background (BG) of the image. Thus, the group can be specified as $\mathcal{G} = \{\text{landbirds, waterbirds}\} \times \{\text{land BGs, water BGs}\}$. Due to the prevalence of waterbirds with water BGs as well as landbirds with land BGs, the minority groups are (landbirds, water BGs) and (waterbirds, land BGs). The reliance of the classifier on the spurious features is typically evaluated by the Worst Group Accuracy (WGA).

2.2. Prior Works on Group Robustness

Plenty of works have been suggested to mitigate spurious correlation problems in classification [11, 15, 19, 21, 25, 32]. Existing works can be categorized with the level of assumption on the group information. For instance, when the group information on the whole dataset is fully available, we can exploit such information to achieve group robustness, as in Group DRO [25], by minimizing the worst group loss. As mentioned in the Introduction, our work is closely aligned with studies that use LLR (or fine-tuning) with holdout

dataset, often dubbed as *reweighting dataset*. We describe the most directly related work below, and provide a more comprehensive review in Sec. A of the Supplementary Materials (S.M.).

DFR [15] reveals that deep neural networks often persist in learning core features despite the existence of spurious correlations in the training dataset. Consequently, they demonstrated that simple LLR with a group-balanced dataset alone can achieve group robustness. This method is cost-effective and less complex as it does not require retraining of the entire classifier. However, as pointed out earlier, collecting the group-annotated and group-balanced image dataset can often become costly, particularly when the number of groups increases and the samples from the minority group are hard to collect.

AFR [22] merges the concept of the LLR and the inference of group information of the data [19, 21, 32]. They proposed LLR of the Empirical Risk Minimization (ERM) model with a weighted loss function that assigns higher importance to instances where the ERM model exhibits poor predictions, thereby prioritizing the minority group. However, their approach necessitates a split of the original training dataset to create a reweighting dataset for LLR, potentially requiring additional training of the ERM model even when a pre-trained model is available. Furthermore, as the proportion of the minority group decreases, the performance may significantly drop as the reweighting dataset may not contain enough minority examples.

SELF [16] leverages training checkpoints akin to [19] to infer data belonging to a minority group in a holdout dataset, constructed from half of the validation set. Specifically, they select data where the prediction discrepancies between the fully trained model and the early stopped model are substantial, then these selected data constitute of the reweighting dataset used for fine-tuning the last layer. This process involves additional training of the entire model to store checkpoints of the early stopped model and to perform class-balanced ERM, which cannot utilize the ERM-trained model. Furthermore, if the proportion of minority examples in the class-balanced ERM training phase is low, the performance of the model may decline because the disagreement-based inference of the group information can be inaccurate.

2.3. Notations

We denote the feature extractor and the last linear layer of a general image classifier by f_θ and h_ϕ , with parameters θ and ϕ respectively. We mainly consider two embedding spaces — a *joint* embedding space generated by image-text contrastive learning-based models, *e.g.*, CLIP [23] or ALIGN [12], and an image embedding space generated by the penultimate layer of a *general* image classifier. The representation vector for each space and data is explicitly denoted by $z_{\text{modality}}^{\text{space}}$; *e.g.*, an embedding of text T

residing in the CLIP’s embedding space is represented as $z_T^{\text{CLIP}} \in \mathbb{R}^{d_{\text{CLIP}}}$ and an embedding of an image I obtained by f_θ is denoted by $z_I^{f_\theta} \in \mathbb{R}^{d_{f_\theta}}$. Note d_{CLIP} and d_{f_θ} denote the dimension of each embedding space. We mainly consider CLIP [23] as a representative joint embedding space throughout the paper.

2.4. Cross-modal Transferability

In [18], the existence of the modality gap, which refers to a constant gap between image and text embeddings in the joint embedding space, was first discovered. Following [33], we consider the instance-wise modality gap, of which the definition is restated below.

Definition 1 (Modality gap in CLIP) *Let (I, T) denote an image-text pair. Then, the modality gap g in the joint embedding space of CLIP is defined as*

$$g := z_I^{\text{CLIP}} - z_T^{\text{CLIP}}.$$

Empirically, g can be approximated as a constant regardless of (I, T) .

Remark: Although Definition 1 is with respect to CLIP space, g can be defined in any joint embedding space where two modalities are well aligned [18].

The existence of the modality gap does not affect the achievement of cross-modal transferability, which enables using embeddings from different modalities interchangeably. Namely, for an image-text pair (I, T) , [33] showed that $h(z_I^{\text{CLIP}}) \approx h(z_T^{\text{CLIP}})$, in which h is a linear classifier on top of the CLIP embedding space. However, the limitation of such cross-modal transferability is that it could be only achieved on the joint embedding space like CLIP and not on other general embedding spaces, *e.g.*, an embedding space obtained by a general feature extractor f_θ .

In this paper, we aim to address such a limitation and enable the cross-modal transferability beyond of the joint embedding space. To that end, we consider a *linear* projector $\Pi : \mathbb{R}^{d_{\text{CLIP}}} \rightarrow \mathbb{R}^{d_{f_\theta}}$ that projects z_I^{CLIP} to $z_I^{f_\theta}$. Namely, we denote (W, b) as the linear matrix and bias vector that defines Π . Then, we make the following (rough) assumption, which we believe is sensible because both z_I^{CLIP} and $z_I^{f_\theta}$ reside in linearly separable spaces in the sense that the high average accuracy can be achieved with a linear layer.

Assumption 1 *There exists a linear projector Π that makes $\Pi(z_I^{\text{CLIP}}) \approx z_I^{f_\theta}$. More specifically, we assume $\|\Pi(z_I^{\text{CLIP}}) - z_I^{f_\theta}\|_2^2 < \|z_{I_1}^{f_\theta} - z_{I_2}^{f_\theta}\|_2^2$ for any arbitrary images I_1 and I_2 that belong to the same class as I .*

The discussion on whether the assumption is valid is presented in Sec. C.4 of the S.M. Now, for the cross-modal transferability on the embedding space of f_θ , we would like to achieve $h_\phi(z_I^{f_\theta}) \approx h_\phi(\Pi(z_T^{\text{CLIP}}))$ for a pair (I, T) ,

namely, we would like to use the projected embedding of CLIP text embedding interchangeably for the image embedding on f_θ . We can then derive a *sufficient condition* for the projector Π under our assumption by examining the equations below:

$$h_\phi(\Pi(z_T^{\text{CLIP}})) = h_\phi(W^\top z_T^{\text{CLIP}} + b) \quad (1)$$

$$= h_\phi(W^\top (z_I^{\text{CLIP}} - g) + b) \quad (2)$$

$$= h_\phi(\Pi(z_I^{\text{CLIP}}) - W^\top g) \approx h_\phi(z_I^{f_\theta} - W^\top g), \quad (3)$$

in which Eq. (2) follows from Definition 1, and Eq. (3) follows from the assumption and the continuity of h_ϕ . Thus, from the above equations, we can easily deduce the sufficient condition for the cross-modal transferability is $W^\top g = 0$; i.e., the modality gap g should lie in the nullspace of $W^\top$. Guided by this sufficient condition, we obtain our linear projector Π and introduce our method in detail in the next section.

3. Method: TLDR

Our TLDR is summarized as follows and in Fig. 2.

1. Estimating (W, b) of Π which connects embedding spaces generated by the CLIP and a trained image classifier. (Sec. 3.1, Fig. 2(a))
2. Generating synonyms for category names of classes and spurious attributes with LLM and validating whether the generated words are well aligned with the embedding space of CLIP and the trained image classifier. (Sec. 3.2, Fig. 2(b))
3. Text-based LLR with the projected embedding of the text prompts, which are obtained from the words validated by Step 2. (Sec. 3.3, Fig. 2(c))

3.1. Closed-Form Estimation of the Projector

In line with Lemma 1, we impose the constraint $W^\top g = 0$ on Π to extend the cross-modal transferability in an embedding space of an arbitrary image classifier. We simply estimate g by sampling image-text pairs from the COCO-Caption dataset [2] and averaging their gaps, i.e., $\hat{g} = \frac{1}{N} \sum_{i=1}^N (z_{I_i}^{\text{CLIP}} - z_{T_i}^{\text{CLIP}})$. We emphasize that the gap estimates can be easily obtained from open-sourced image-text paired dataset [2, 26] and the estimated gap is independent of the dataset used to train the image classifier, f_θ . With the estimated gap, we solve the constrained ridge regression problem to obtain (W, b) of Π . Even if there is a discrepancy between $\hat{g}$ and each $(z_{I_i}^{\text{CLIP}} - z_{T_i}^{\text{CLIP}})$, the impact is minimal in our method, as detailed in Sec. C.5 of the S.M.

Lemma 1 (Constrained ridge regression estimate of Π)

Let $X \in \mathbb{R}^{n \times d_{\text{CLIP}}}$ be a matrix of CLIP embeddings of n training images, $Y \in \mathbb{R}^{n \times d_{f_\theta}}$ be a matrix of embeddings generated by f_θ for the corresponding images. Moreover, let $W \in \mathbb{R}^{d_{\text{CLIP}} \times d_{f_\theta}}$, $b \in \mathbb{R}^{d_{f_\theta}}$ define the linear projector Π

and $g \in \mathbb{R}^{d_{\text{CLIP}}}$ be a modality gap. Following Assumption 1, we assume the following linear relationship between X and Y :

$$Y \sim XW + b + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma). \quad (4)$$

Then, the ridge regression estimate of (W, b) with the constraint $W^\top g = 0$ is

$$W^* = \tilde{W} - (X^\top X + \lambda I)^{-1} g (g^\top (X^\top X + \lambda I)^{-1} g)^{-1} g^\top \tilde{W} \quad (5)$$

$$b^* = \frac{1}{n} (Y - XW^*)^\top \mathbf{1} \quad (6)$$

in which $\tilde{W} = (X^\top X + \lambda I)^{-1} X^\top Y$, and $\mathbf{1}$ is a all-ones vector. (We also assume Σ is diagonal.)

The proof of the lemma is in Sec. D of the S.M. λ is a hyperparameter for the ℓ_2 regularization and searched based on the NMSE [14] criterion, which evaluates the normalized ℓ_2 distance between original embedding and its prediction, i.e., $\text{NMSE}(z, \hat{z}) := \frac{\|z - \hat{z}\|_2^2}{\|z\|_2^2}$. Throughout the paper, (W^*, b^*) are computed with the training set of each dataset, and λ is searched with the corresponding validation set, following the same procedure as training the image classifier. In addition, we do not perform ℓ_2 normalization on the CLIP embeddings and g turns out to be almost constant without any normalization; refer to Sec. C.2 of the S.M.

It is important to note that our proposed algorithm has a clear advantage in that it eliminates the need for gradient descent-based optimization for learning the Π and additional tedious search of hyperparameters. (e.g., learning rate, weight decay, etc.) Moreover, it turns out that training Π to minimize $\|z_I^{f_\theta} - \Pi z_T^{\text{CLIP}}\|_2^2$ is not as effective as our approach since it does not sufficiently consider the domain in which the (f_θ, h_ϕ) are trained. The detailed analysis on this is discussed in Sec. 4.4. Besides, there are alternative methods for mitigating the impact of g such as adding $\hat{g}$ to z_T^{CLIP} to recover z_I^{CLIP} , or methods suggested by [8, 17]. However, we found that these are less effective than ours and have some limitations; refer to Sec. C.6 of the S.M.

3.2. Text-based Dataset Construction

Given the category names of the class label and spurious attribute in $\mathcal{Y}$ and $\mathcal{A}$, respectively, we generate diverse words corresponding to those names by prompting GPT-3.5 [1] to generate synonyms of them. Such diversity is essential in carrying out LLR as we show in our experiments. We denote the set of generated words for the y -th element in $\mathcal{Y}$ as $\mathcal{T}^y$ and the i -th generated word is denoted as $t_i^y \in \mathcal{T}^y$. $t_i^a \in \mathcal{T}^a$ is defined in the same way for $\mathcal{A}$. The full list of generated words is given in Sec. E.8 of the S.M.

Despite the impressive generation capability of LLMs, it is well-known that they can exhibit hallucinations and

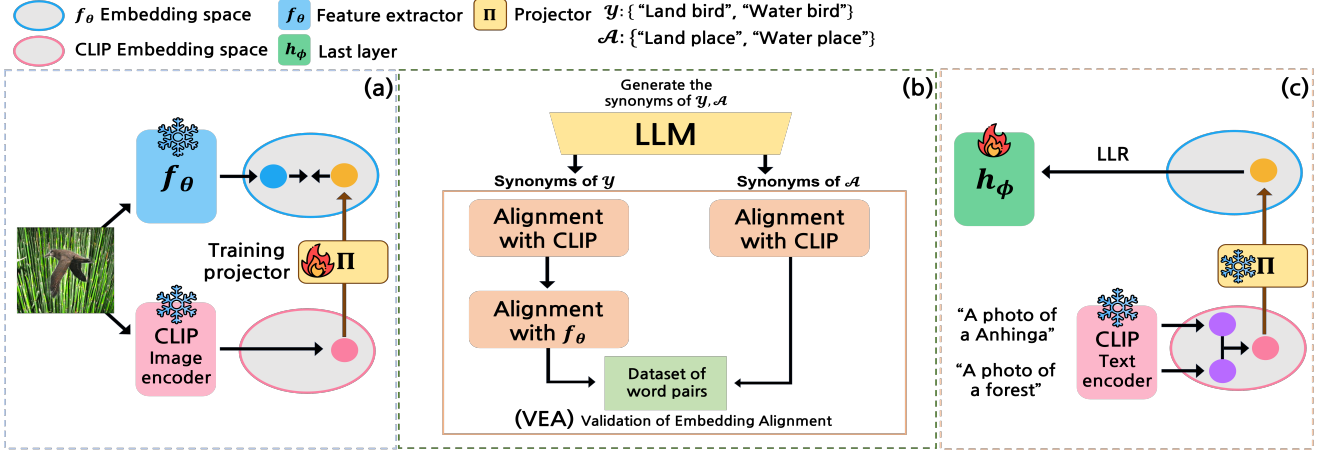


Figure 2. (a) Diagram of obtaining the projector Π . The same image is fed to the encoder of f_θ and CLIP image encoder. Then, (W^*, b^*) of Π minimizes the distance between the embedding of f_θ and the projected embedding from CLIP. (b) Diagram of the validation of the embedding alignment (VEA) of the generated words. (c) Diagram of the text-based LLR. Text embeddings for each class and spurious attribute are averaged in the embedding space of CLIP and projected to the embedding space of f_θ with Π . Consequently, the projected embedding is fed to h_ϕ , which is used for retraining.

provide inaccurate information [10, 35]. Furthermore, the generated texts and their corresponding projected embeddings may be out-of-distribution (OOD) examples for CLIP and f_θ respectively, which can diminish the performance of LLR. Hence, we implement validation of embedding alignment (VEA) for the generated words to only keep those that are compatible with embedding spaces of CLIP and f_θ .

[VEA for CLIP] Since we expect the CLIP embeddings of the generated words in $\mathcal{T}^y$ for each $y \in \mathcal{Y}$ to be close to z_y^{CLIP} , we employ the following simple and effective rule to remove words that are misaligned with CLIP — we only select t_i^y of which cosine similarity with its corresponding class y in the CLIP text embedding space is the largest. Namely, t_i^y is selected only when

$$\arg \max_{y' \in \mathcal{Y}} \cos(z_{P_1(t_i^y)}^{\text{CLIP}}, z_{P_1(y')}^{\text{CLIP}}) = y \quad (7)$$

holds. It is done for the generated words for $\mathcal{T}^a$ analogously.

[VEA for f_θ] Once the VEA for CLIP is done, the remaining words will be semantically well-aligned with the CLIP embedding space. However, particularly for the remaining words in $\mathcal{T}^y$, they still may not be well-aligned with the f_θ 's embedding space since their “projected” CLIP embeddings may be OOD examples within the embedding space of f_θ . Such misalignment may cause additional confusion when building a new classification boundary via LLR. To that end, we implement a logit-based VEA step for f_θ to further remove the misaligned words in $\mathcal{T}^y$ — namely, we only select t_i^y that satisfies

$$\arg \max_{y' \in \mathcal{Y}} h_\phi(\Pi(z_{P_1(t_i^y)}^{\text{CLIP}}))_{y'} = y, \quad (8)$$

in which Π is the projector defined by (W^*, b^*) , $h_\phi(\cdot)_{y'}$

denotes the y' -th element of $h_\phi(\cdot)$, and P_1 is a prompt template with $P_1(t) = \text{"A photo of a } \{t\} \text{"}$. We note that this VEA step is *only* done for the words in $\mathcal{T}^y$.

[Constructing Dataset of Word Pairs] We construct a dataset of word pairs based on the remaining words after the VEA. We simply consider all possible combinations for each group, *i.e.*, $\mathcal{T}^y \times \mathcal{T}^a$ for group (y, a) where $\mathcal{T}^y, \mathcal{T}^a$ now include only validated words.

We emphasize that our proposed word generation and VEA procedure enables text-based LLR in a situation where appropriate words are not given in the metadata of a dataset, which was not sufficiently discussed in DrML [33]. In addition, we effectively improve the performance of LLR by using a sufficient amount of text data compared to DrML, which is discussed in detail in Sec. 4.5.

3.3. Text-based Last Layer Retraining

We use the word pairs to compute the CLIP text embeddings, which are projected onto the embedding space of f_θ using (W^*, b^*) . When computing the CLIP text embedding for a sample in group (y, a) , we do not simply concatenate t_i^y and t_j^a using a single text prompt as in [33]. Instead, we average the two text embeddings computed for each t_i^y and t_j^a , *i.e.*, compute $\frac{1}{2}(z_{P_1(t_i^y)}^{\text{CLIP}} + z_{P_1(t_j^a)}^{\text{CLIP}})$ as the embedding for (y, a) . We empirically find that this way better reflects the individual embeddings of t_i^y and t_j^a . Furthermore, to represent the more diverse nature of images with texts, we adopt the 80 CLIP prompt templates used in zero-shot classification in [23]. That is, when each (t_i^y, t_j^a) pair is fetched, we compute $\frac{1}{2}(z_{P_k(t_i^y)}^{\text{CLIP}} + z_{P_k(t_j^a)}^{\text{CLIP}})$, in which $P_k(\cdot)$ is randomly selected among the 80 prompt templates.

The retraining of the last layer is done by mini-batch optimization as in [16]. Moreover, we sample group-balanced

training sets every epoch to maximally utilize available data. By default, we adopt early stopping based on validation WGA as in [16, 19, 22, 25].

4. Experimental Results

4.1. Dataset Description

Waterbirds [25] is a synthetic dataset constructed with CUB [30] and Places [36]. Minor groups together make up about 5% of the training dataset.

CelebA [20] is a large collection of celebrity faces along with 40 attribute annotations per image. In line with [25], we consider a classification problem where $\mathcal{Y} = \{\text{dark hair, blond hair}\}$ and $\mathcal{A} = \{\text{woman, man}\}$. The minority group, "blond men," makes up about 1% of the training dataset.

SpuCoAnimals [13] is constructed based on ImageNet-1K [4] with $\mathcal{Y} = \{\text{landbirds, waterbirds, small dogs, large dogs}\}$ and $\mathcal{A} = \{\text{land BGs, water BGs, indoor BGs, outdoor BGs}\}$. Bird classes are spuriously correlated with land/water BGs and dog classes with indoor/outdoor BGs. This dataset provides a more realistic setting beyond the binary $\mathcal{Y}, \mathcal{A}$. The total proportion of minority groups is about 5%.

4.2. Experimental Setup

[Model Architecture] We used ImageNet pre-trained ResNet-50 [9] as the (f_θ, h_ϕ) , with the exception of SpuCoAnimals. In the case of SpuCoAnimals, given that it originated from ImageNet, we used a randomly initialized model to prevent any potential information leakage. Additionally, an experiment was conducted with ViT [6] based (f_θ, h_ϕ) to verify the applicability of TLDR to more generalized vision models, as elaborated in Sec. B.2 of the S.M. We employed CLIP ViT-B/32 throughout our experiments.

[Number of Words Generated] We generated 200 synonyms for each category name, except for the "large dog" class in SpuCoAnimals where GPT-3.5 cannot generate more than 100 unique breeds due to its limitations.

[Estimation of the g] The modality gap was estimated from 1K image-text pairs sampled from the validation set of COCO-Caption dataset [2] unless otherwise stated.

[ReLU on the Projected Embedding] Since all embeddings of ResNet undergo the ReLU operation, we applied ReLU to each projected embedding $\Pi(z_T^{\text{CLIP}})$ to minimize the discrepancy. If $z_I^{f_\theta}$ possesses real values rather than non-negative values, the ReLU operation can be omitted.

4.3. Main Experimental Results

[Main results] We present a comparison of TLDR with other baselines in Tab. 1. It is noteworthy that TLDR exhibits competitive performance with DFR, which explicitly uses a group-balanced image dataset for LLR, on the Waterbirds and CelebA. Furthermore, TLDR even demonstrates superior performance on SpuCoAnimals compared to DFR.

In addition, TLDR outperforms AFR and SELF across all datasets. The lower performance of DFR and SELF on SpuCoAnimals may be attributed to the limited data for each group in the group-balanced dataset. The small number may affect both retraining and the evaluation of the validation WGA since DFR and SELF randomly halve the validation set for WGA evaluation, leading to the suboptimal hyperparameter search. In contrast, TLDR avoids this issue by neither splitting the validation set nor using it for LLR.

[Post Hoc Utilization of Baselines] As previously mentioned, utilizing TLDR provides a notable advantage in employing a pre-trained model. Since the usual development process involves training a model with ERM, identifying weaknesses in the model, and subsequently addressing them, having the capability to leverage an already trained model constitutes a significant practical benefit. To assess the performance of AFR and SELF under circumstances where additional training of the entire model is not feasible due to computational costs, we evaluate their performances in a post hoc manner. We used half of the validation set as a reweighting dataset for AFR and excluded class-balanced ERM for SELF.

The results are marked with $\star$ in Tab. 1. While AFR's overall performance improves due to the benefit of training the ERM model with the full training dataset, TLDR still outperforms. In contrast, the performance of SELF drops significantly because the model is not trained in a class-balanced manner, which implies that class-balanced ERM is critical to maintaining the performance of SELF, necessitating the additional training of the entire model. On the other hand, TLDR can be applied to the pre-trained model without any additional training of the whole model, which is efficient and practical for use.

[Adjustment the Proportion of Minority Examples]

We compared TLDR with other baselines by adjusting the proportion of minority groups from both the training and validation sets. Unlike the experiments conducted by [15, 16, 22] where only the proportion of the validation set is controlled, our experimental setting is more realistic; usually training and validation sets are split from an entire dataset, so the proportion changes in both. We excluded SpuCoAnimals since we observed a case where the validation WGAs are all zero for all hyperparameters as the ratio becomes smaller, making a fair comparison between the baselines challenging. For a fair comparison, we avoided early stopping for all methods as discussed in [22].

Fig. 3a illustrates that TLDR consistently outperforms AFR and SELF across all cases and even has superior performance to DFR on Waterbirds. The performance of all baselines is inevitably influenced by the ratio, considering the diminished size of minority groups in their reweighting dataset. Moreover, the reduced ratio may impact SELF's ability to identify minority data, as the scarcity of minority

Table 1. Test WGA & average accuracy for each dataset. The reported experimental results of each baseline are given in S.M., and we did our best to reproduce the reported results of each method. $\checkmark$ denotes utilizing group annotated image validation set for training the last layer, on the other hand, a single $\checkmark$ denotes using the set for only hyperparameter searching or model selection. $\star$ denotes the setting where ERM model is used, *i.e.* no additional split of training dataset or class-balancing. Refer to Sec. 4.3 for a detailed experimental setting. We calculated average accuracy in the same way with [25]. All numbers are averaged from 4 random seeds and the highest WGAs are bolded among the last three rows that share the same settings.

Method	Group Info	Post-hoc	Waterbirds		CelebA		SpuCoAnimals	
			Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)
ERM	$\star / \star$	-	72.2 $\pm$ 0.7	98.1 $\pm$ 1.1	47.6 $\pm$ 3.1	95.2 $\pm$ 0.1	6.3 $\pm$ 1.6	81.3 $\pm$ 0.9
Group-DRO [25]	$\checkmark / \checkmark$	$\star$	88.2 $\pm$ 0.5	93.3 $\pm$ 0.7	90.3 $\pm$ 0.3	92.3 $\pm$ 1.9	39.5 $\pm$ 4.8	47.1 $\pm$ 3.6
DFR _{Tr} ^{Val} [15]	$\star / \checkmark \checkmark$	$\checkmark$	92.5 $\pm$ 0.7	94.8 $\pm$ 0.3	86.6 $\pm$ 1.1	90.3 $\pm$ 0.2	22.4 $\pm$ 2.4	68.4 $\pm$ 1.1
AFR [22]	$\star / \checkmark$	$\star$	87.4 $\pm$ 0.8	90.4 $\pm$ 0.6	79.4 $\pm$ 0.8	91.7 $\pm$ 0.3	16.2 $\pm$ 6.1	59.9 $\pm$ 3.5
SELF [16]	$\star / \checkmark$	$\star$	91.4 $\pm$ 2.1	94.5 $\pm$ 1.6	79.4 $\pm$ 3.2	91.9 $\pm$ 0.7	7.3 $\pm$ 2.9	86.7 $\pm$ 1.0
*AFR	$\star / \checkmark$	$\checkmark$	86.3 $\pm$ 1.9	91.7 $\pm$ 0.4	80.1 $\pm$ 3.3	90.6 $\pm$ 1.0	22.3 $\pm$ 3.4	53.8 $\pm$ 7.7
*SELF	$\star / \checkmark$	$\checkmark$	91.2 $\pm$ 0.7	96.0 $\pm$ 0.7	56.9 $\pm$ 4.9	95.0 $\pm$ 0.1	6.9 $\pm$ 0.6	77.2 $\pm$ 2.2
TLDR	$\star / \checkmark$	$\checkmark$	92.1 $\pm$ 0.3	95.2 $\pm$ 0.8	85.4 $\pm$ 1.2	89.0 $\pm$ 0.9	36.2 $\pm$ 1.7	55.8 $\pm$ 2.9

groups in the training data might not provide sufficient information to discern differences in predictions between the early-stopped model and the fully trained model. In contrast, TLDR has fairly robust performance across varying ratios, making it well-suited for situations where images belonging to minority groups are rare.

4.4. Analyses on the Projector

[Direct Alignment of Πz_T^{CLIP} and $z_I^{f_\theta}$] One may suggest that direct alignment of Πz_T^{CLIP} and $z_I^{f_\theta}$ would work using the image-text pairs in COCO-Caption. However, we experimentally found that approach to be ineffective since it does not consider the dataset on which the (f_θ, h_ϕ) were trained. This oversight can impact both the VEA as well as the LLR process. Since the goal is to debias f_θ within its trained domain, it’s crucial for Π to perform well within that domain. Estimating Π with extensive (I, T) pairs may inundate it with redundant information, failing to precisely capture the data distribution relevant to (f_θ, h_ϕ) .

We estimated Π_{COCO} by minimizing $\|z_I^{f_\theta} - \Pi z_T^{\text{CLIP}}\|_2^2$ with the validation split of COCO, 80% for training ($\# \approx 160\text{K}$) and 20% for validation. As shown in Tab. 2 (first row), the aforementioned issue negatively impacts the VEA process for f_θ and significantly undermines performance on CelebA. In addition, Tab. 2 (second row) demonstrates that using the Π_{COCO} for LLR remains ineffective, even if the VEA is still done with Π_{ours} which is obtained with Lemma 1 and the datasets used for training (f_θ, h_ϕ) . Hence, estimating Π with the same data used for training of (f_θ, h_ϕ) becomes imperative, thereby validating the novelty and effectiveness of our approach.

Table 2. Comparison of the estimation schemes of Π .

VEA	LLR	Waterbirds		CelebA		SpuCoAnimals	
		Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)
Π_{COCO}	Π_{COCO}	86.5 $\pm$ 0.8	97.4 $\pm$ 0.0	5.6 $\pm$ 1.2	90.4 $\pm$ 1.1	21.0 $\pm$ 3.2	70.7 $\pm$ 2.6
Π_{ours}	Π_{COCO}	86.8 $\pm$ 0.3	97.2 $\pm$ 0.1	83.8 $\pm$ 1.4	87.3 $\pm$ 0.4	26.4 $\pm$ 5.5	63.0 $\pm$ 3.1
Π_{ours}	Π_{ours}	92.1 $\pm$ 0.3	95.2 $\pm$ 0.8	85.4 $\pm$ 1.2	89.0 $\pm$ 0.9	36.2 $\pm$ 1.7	55.8 $\pm$ 2.9

[Number of Pairs for Estimation of g] We varied the number of pairs used to estimate the modality gap from 0 to 1K to check how our method is affected by the number of pairs. The results are illustrated in Fig. 3b. Notably, without gap information, the WGA and average accuracy significantly decrease. It is noteworthy that a mere 10 image-text pairs suffice to estimate the modality gap, yielding comparable performance to cases with a larger number of pairs. This suggests a remarkably low burden in terms of pair collection. However, the experiment result on SpuCoAnimals exhibits a different tendency. This is because the searched λ value is considerably larger compared to the other two datasets, resulting in a smaller $\|\mathbf{W}\|_F$ and, consequently, reduced $\|\mathbf{W}^\top \mathbf{g}\|_1$. Therefore, the efficacy of the constraint $\mathbf{W}^\top \mathbf{g} = 0$ is not adequately represented.

4.5. Analyses on Text Dataset Construction

[Comparison with DrML] We re-emphasize that DrML [33] merely uses words that are already provided in the metadata. Therefore, it cannot be seamlessly applied to CelebA, SpuCoAnimals, or other general datasets, which lack appropriate texts in metadata. In addition, using only words in the metadata may be insufficient due to their limited quantity or misalignment with the embedding spaces, which can lead to poor debiasing performance. To demonstrate the superiority of our text dataset construction procedure, we carried out experiments on Waterbirds drawing comparison with DrML. We assessed the performance on the CLIP backbone, comparing DrML’s text data against ours, both with and without the inclusion of VEA in Tab. 3. First of all, it is noteworthy that VEA has a positive impact also on the DrML’s text data, improving both the test WGA and the average accuracy. Furthermore, we observe that TLDR’s text generation scheme *with* VEA is significantly superior to that of DrML, and the VEA step is essential as relying solely on all the generated words proves to be less effective. We believe that these results clearly show the

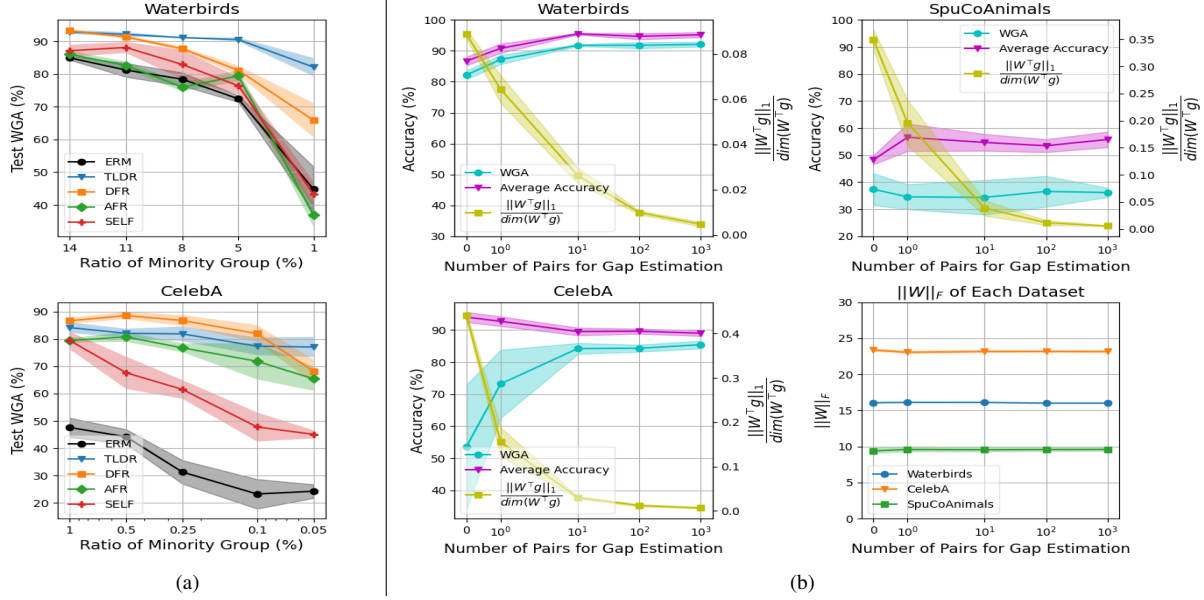


Figure 3. All numbers are averaged from 4 random seeds. (a): Experimental results of controlling the minority ratio on Waterbirds and CelebA. (b): Experimental results of controlling the number of pairs for gap estimation on all datasets.

Table 3. Comparison between DrML and TLDR for the construction of text datasets.

Method	VEA	Need Metadata?	Ratio of Remaining Words	Waterbirds	
				Worst(%)	Mean(%)
CLIP w/ Linear Classifier	-	-	-	33.5 $\pm$ 0.6	96.1 $\pm$ 0.1
DrML	$\times$	$\checkmark$	204 / 204	54.8 $\pm$ 0.3	93.3 $\pm$ 0.4
	$\checkmark$	$\checkmark$	181 / 204	59.5 $\pm$ 0.4	94.2 $\pm$ 0.2
DrML w/ TLDR's Words	$\times$	$\times$	800 / 800	47.1 $\pm$ 0.7	96.1 $\pm$ 0.0
	$\checkmark$	$\times$	545 / 800	73.1 $\pm$ 1.0	92.3 $\pm$ 0.2

Table 4. Result of ablation studies on the VEA. ES stands for early stopping.

ES	VEA	Waterbirds		CelebA		SpuCoAnimals	
		Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)
$\checkmark$	$\times$	91.1 $\pm$ 0.8	93.1 $\pm$ 1.3	84.1 $\pm$ 1.2	87.8 $\pm$ 1.5	38.9 $\pm$ 7.0	51.3 $\pm$ 2.5
$\checkmark$	$\checkmark$	92.1 $\pm$ 0.3	95.2 $\pm$ 0.8	85.4 $\pm$ 1.2	89.0 $\pm$ 0.9	36.2 $\pm$ 1.7	55.8 $\pm$ 2.9
$\times$	$\times$	88.1 $\pm$ 0.7	88.6 $\pm$ 0.5	76.0 $\pm$ 2.5	82.2 $\pm$ 1.2	34.7 $\pm$ 7.7	49.9 $\pm$ 3.3
$\times$	$\checkmark$	92.0 $\pm$ 0.4	95.6 $\pm$ 0.2	84.1 $\pm$ 1.9	88.0 $\pm$ 1.8	28.3 $\pm$ 2.9	52.1 $\pm$ 1.5

novelty of TLDR’s text dataset construction procedure.

[Ablation study on VEA] We present the results of ablation studies on the VEA in Tab. 4. Indeed, the test WGA improves on Waterbirds and CelebA. In addition, it significantly reduces the variance of the test WGA on SpuCoAnimals. We note that the decrease of the test WGA on SpuCoAnimals is not decisive, considering the substantial variance in the test WGA when VEA is not applied. The gain from VEA becomes larger when the early stopping is not applied, which can be practical in a situation where the ratio of minority groups is too low to allow proper early stopping.

5. Conclusion & Limitations

In this study, we demonstrate that a general image classifier can be debiased with text-based LLR. TLDR stands out by not necessitating a group-annotated, balanced image

dataset, nor requires additional training of the entire model, making it easily applicable. In addition, TLDR introduces a systematic procedure for constructing text datasets that leverages debiasing performance. Nevertheless, there are some limitations to our method. As our work is based on CLIP, we can leverage the concepts that CLIP understands. We anticipate that this limitation can be addressed with the advancements in multi-modal contrastive models across various domains, such as ConVIRT [34] for the medical domain and WikiSatNet [29] for satellite imagery. Additionally, our approach assumes prior knowledge of the model’s weakness. Considering the plenty of works focused on discovering the model’s weaknesses [5, 7, 24, 27], we anticipate that acquiring knowledge about the model’s weaknesses can be readily achieved. Lastly, there may be concerns about the inherent bias of LLMs and CLIP [3, 31], but we did not encounter any substantial negative impact on our results.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant [No.2021R1A2C2007884] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [RS-2021-II211343, RS-2021-II212068, RS-2022-II220113, RS-2022-II220959] funded by the Korean government (MSIT). It was also supported by AOARD Grant No. FA2386-23-1-4079. Juhyeon Park is grateful for financial support from Hyundai Motor Chung Mong-Koo Foundation.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, volume 33, pages 1877–1901, 2020. [2](#), [4](#)
- [2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. [4](#), [6](#)
- [3] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Bodeti. Fairerclip: Debiasing zero-shot predictions of clip in rkhss. In *International Conference on Learning Representations (ICLR)*, 2024. [8](#)
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. [6](#)
- [5] Greg d’Eon, Jason d’Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1962–1981, 2022. [8](#)
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. [6](#)
- [7] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations (ICLR)*, 2022. [8](#)
- [8] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can’t believe there’s no images! learning visual tasks using only language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2672–2683, 2023. [4](#)
- [9] 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 (CVPR)*, pages 770–778, 2016. [6](#)
- [10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. [5](#)
- [11] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the Conference on Causal Learning and Reasoning (CLeaR)*, pages 336–351, 2022. [1](#), [2](#)
- [12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 4904–4916, 2021. [3](#)
- [13] Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. *arXiv preprint arXiv:2306.11957*, 2023. [6](#)
- [14] Siwon Kim, Jinoh Oh, Sungjin Lee, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Grounding counterfactual explanation of image classifiers to textual concept space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10942–10950, 2023. [4](#)
- [15] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2022. [1](#), [2](#), [3](#), [6](#), [7](#)
- [16] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 11552–11579, 2023. [3](#), [5](#), [6](#), [7](#)
- [17] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. In *International Conference on Learning Representations (ICLR)*, 2023. [4](#)
- [18] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17612–17625, 2022. [2](#), [3](#)
- [19] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 6781–6792, 2021. [1](#), [2](#), [3](#), [6](#)
- [20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. [6](#)
- [21] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations (ICLR)*, 2021. [1](#), [2](#), [3](#)
- [22] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 28448–28467, 2023. [3](#), [6](#), [7](#)
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

- vision. In *Proceedings of the International conference on machine learning (ICML)*, pages 8748–8763, 2021. 2, 3, 5
- [24] Nazneen Rajani, Weixin Liang, Lingjiao Chen, Meg Mitchell, and James Zou. Seal: Interactive tool for systematic error analysis and labeling. *arXiv preprint arXiv:2210.05839*, 2022. 8
- [25] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 6, 7
- [26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 25278–25294, 2022. 4
- [27] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 19339–19352, 2020. 8
- [28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [29] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. Learning to interpret satellite images in global scale using wikipedia. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3620–3626, 2019. 8
- [30] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6
- [31] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024. 8
- [32] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pages 26484–26516, 2022. 1, 2, 3
- [33] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3, 5, 7
- [34] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, pages 2–25, 2022. 8
- [35] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lema Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. 5
- [36] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464, 2017. 6