GRADTUNE: LAST-LAYER FINE-TUNING FOR GROUP ROBUSTNESS WITHOUT GROUP ANNOTATION

Patrik Kenfack[†§], **Ulrich Aïvodji**[†§] **& Samira Ebrahimi Kahou**[‡§*]
[†]ÉTS Montréal, [‡]University of Calgary, [§]Mila, [*]CIFAR

ABSTRACT

This work addresses the limitations of deep neural networks (DNNs) in generalizing beyond training data due to spurious correlations. Recent research has demonstrated that models trained with empirical risk minimization learn both core and spurious features, often upweighting spurious ones in the final classification, which can frequently lead to poor performance on minority groups. Deep Feature Reweighting alleviates this issue by retraining the model's last classification layer using a group-balanced held-out validation set. However, relying on spurious feature labels during training or validation limits practical application, as spurious features are not always known or costly to annotate. Our preliminary experiments reveal that ERM-trained models exhibit higher gradient norms on minority group samples in the hold-out dataset. Leveraging these insights, we propose an alternative approach called GradTune, which fine-tunes the last classification layer using high-gradient norm samples. Our results on four well-established benchmarks demonstrate that the proposed method can achieve competitive performance compared to existing methods without requiring group labels during training or validation.

1 INTRODUCTION

Despite the impressive ability of deep neural networks to achieve human-level performance on complex vision and language tasks, their dependence on the quality of training data makes them fail to generalize well on a group of data points that do align with the trend in training data. More specifically, the data used to train neural networks might contain patterns that spuriously correlate with the target task (Ye et al., 2024). For instance, the background of an image might spuriously correlate with the class label (Wah et al., 2011). Models trained using the classical Empirical Risk Minimization (ERM) can excessively rely on spurious features for prediction and fail to capture the intended core feature, which often leads to poor performance on minority groups of samples where the spurious correlation does not apply (Steinmann et al., 2024).

Kirichenko et al. (2023) have demonstrated that models trained with ERM still capture the core features in the learned representation in addition to spurious features, and the latter is overweighted in the last layer of the model. Moreover, they demonstrated that simply retraining the last classification layer of the model with a small proportion of a group-balanced held-out set (i.e., data where the spurious correlation does not hold) can mitigate the spurious correlation and achieve state-of-art performance robustness benchmark (Kirichenko et al., 2023). Moreover, even if the held-out set used for last-layer retraining contains a smaller proportion of the worst-group data, the resulting last-layer retrained model still significantly outperforms the ERM model (LaBonte et al., 2024). Furthermore, classical group robustness methods, such as group distributionally robust optimization (Group DRO (Sagawa et al., 2019)), do not necessarily learn a better representation compared to ERM but somewhat better weight the core feature in the last classification layer (Izmailov et al., 2022).

However, most existing methods addressing spurious correlation require access to spurious feature labels (group labels) for training or validation (Sagawa et al., 2019; Liu et al., 2021; Kirichenko et al., 2023; Qiu et al., 2023). This limits the practical adoption of the technique as the spurious features are generally unknown, and even when they are, labeling the data can be costly (Kenfack et al., 2024). In this work, we revise last-layer retraining to alleviate the need for group labels during training or validation. In preliminary experiments, we observe that samples in the held-out set where

the ERM-trained model does not generalize have a higher gradient magnitude. In contrast, samples on which the model performs well have smaller gradient norms. Samples with higher gradient magnitudes are mainly worst-group data that do not exhibit a spurious correlation. These results align with related works Ahn et al. (2023); Kenfack et al. (2022); Bagdasaryan et al. (2019), showing that minority groups can have a higher gradient magnitude than samples from majority groups.

Building on this observation, we propose GradTune, a method for mitigating spurious correlation without groups label by simply *fine-tuning* the last classification layer of the ERM-model using the top-k gradient norm samples. Our intensive experiments on several datasets demonstrate that GradTune can substantially improve worst-group accuracy and achieve group-robust performance comparable to state-of-the-art methods without using the group labels.

2 Problem Setup

We consider a setting where the training data $\mathcal{D}_{\text{tr}}$ contains triplets $\{(x_i, y_i, a_i)\}_{i=1}^{N}$ where $x_i \in \mathcal{X}$ is a data point, $y_i \in \mathcal{Y}$ its class label and $a_i \in \mathcal{A}$ an unknown spurious feature. When the attribute a spuriously correlates with the target y, ERM trained models to minimize the average cross-entropy loss w.r.t y can strongly rely on the spurious feature and fail to generalize on the test $\mathcal{D}_{\text{test}}$ where the spurious correlation does not apply (Ye et al., 2024). The reliance on the spurious feature can be more pronounced when it is easier to learn than the core features (Nam et al., 2020). More specifically, the training data $\mathcal{D}_{\text{tr}}$ can be represented by different subgroups in $g_i \in \mathcal{G}$, where subgroups $\{g_i\}_{i=1}^{M}$ are formed based on the cartesian product of class labels and the spurious attributes, i.e., $\mathcal{G} = \mathcal{Y} \times \mathcal{A}$. The spurious correlation means an attribute value a and a label y commonly appear at the same time. Specifically, we denote as *bias-aligned samples* (also majority group) the group where the spurious features and the label match, i.e., $a = y$, and *bias-conflicting samples* (minority group) as the group where $a \neq y$ (Ye et al., 2024).

As the average accuracy does not fully capture the robustness of the model to spurious correlation, we use the worst-group accuracy (WGA) to measure the model's reliance on spurious correlation for predictions. Furthermore, in the presence of spurious features, a model f parametrized by θ is optimized to minimize the loss of the worst-performing subgroup, i.e.,

$$\arg\min_{\theta} \max_{g \in \mathcal{G}} \frac{1}{|g|} \sum_{i \in g}^{|g|} \mathcal{L}(f(x_i; \theta), y_i) \tag{1}$$

3 Related work

We categorize baseline methods for comparison into two types depending on whether they require group labels during the training and/or validation.

Methods requiring spurious features during training or validation When group labels are available in training data, Group DRO (Sagawa et al., 2019) can be applied to optimize for the worst-performing subgroups. Group DRO also requires a group-labeled validation set for model selection and hyperparameter tuning. Spurious correlation can also be mitigated using reweighting and subsampling to ensure group-balanced training data (Sagawa et al., 2020) or using synthetic data to augment the minority group (Goel et al., 2020). However, group information can be costly to collect or unavailable due to privacy restrictions (Kenfack et al., 2024)

Some existing methods require spurious features only in the validation set for hyperparameter tuning or model selection. Notable examples include *Just Train Twice*(JTT) (Liu et al., 2021). This approach first trains a biased model and then upweights misclassified samples to train a second model, aiming to improve the performance of the worst-performing subgroup. (Kirichenko et al., 2023) proposed *Deep Feature Reweighting* (DFR), a method that trains an ERM model and fine-tunes the last layer using a held-out, group-balanced validation set. Similarly *Selective last-layer fine-tuning* (SELF) (LaBonte et al., 2024): SELF fine-tunes the last layer using a fine-tuning set consisting of samples with higher disagreement between the outputs of ERM and an early-stoped models (LaBonte et al., 2024). It only requires group labels in the validation set for model selection. *Automatic Feature Reweighting* (AFR) (Qiu et al., 2023) retrains the last layer of an ERM-trained model using a weighted loss that

emphasizes examples where the ERM model performs poorly. Unlike these methods, GradTune can improve worst-group performance without intensive hyperparameter tuning on a group-labeled validation.

Methods that do not utilize any spurious label information An example is bias-unsupervised logit adjustment (uLA) (Tsirigotis et al., 2024), which employs self-supervised feature representation learning combined with a classifier layer trained using ERM. It fine-tunes the last layer using logit adjustment (Liu et al., 2022) to mitigate biases learned by the ERM-trained classifier. Additionally, to highlight the importance of data quality for fine-tuning, we consider an ERM baseline where the last layer is retrained using random data points from a held-out validation set; we name it *random finetuning*. Results comparing the proposed method against the presented baselines are provided in Section 5.

We focus on the case where the group information is unknown during the training and validation, and we only consider it the test set for evaluation.

4 GradTune: Gradient-based fine-tuning for mitigating spurious correlation

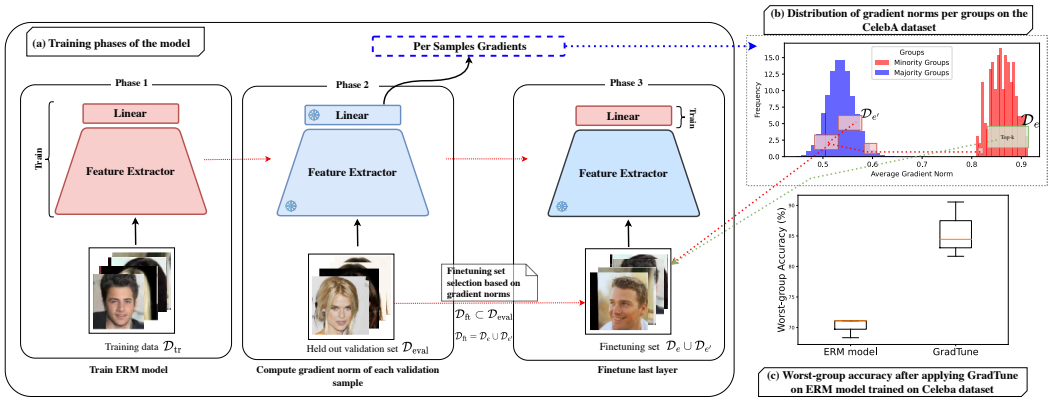

Figure 1: Overview of the proposed method on the CelebA dataset, where the hair color spuriously correlates with gender. Fig. (a) showcases the three-phase training on the proposed method: (i) during the first phase, we train the ERM model using the training data (ii) in the second phase, we use the pretrained ERM model to compute the gradient norm of each data point in the held-out validation set (iii) we derive the fine-tuning set by sampling top-k ($\mathcal{D}_e$) gradient norms data points, and we sample the small proportion of data point at random ($\mathcal{D}'_e$). Finally, we finetune the last classification layer using the fine-tuning set. Fig. (b) shows that on the CelebA dataset, the minority group has a higher gradient magnitude than the majority group. Fig. (c) shows after the three phases of our debiasing mechanism, the performance of the minority group significantly improved compared to the ERM model.

In this section, we describe the training process of our proposed method, named GradTune, for mitigating spurious correlation without group labels. We begin by providing an overview of the three phases of the training process and then dive into a detailed analysis of fine-tuning based on gradient norms as a means to identify and mitigate spurious.

4.1 Training phases of GradTune

The training process of GradTune can subdivided into three phases: (i) ERM training, (iii) sample gradient norm computation, and (iii) last layer fine-tuning. We first train the ERM model and then fine-tune its last layer using a subset of samples from the held-out validation set with high gradient norms. This subset with higher gradient norms mainly consists of samples the ERM-trained model failed to learn, and we hypothesize that using them to fine-tune the last layer of a pretrained ERM-model can mitigate the spurious correlation learned by the model. GradTune improves on Ahn et al.

(2023), which first trains a bias amplified model to weight the training data points and then trains the debiased model using the weighted training dataset. The weights of each sample in the training dataset are proportional to their gradient norms; the per-sample gradient vector is computed using the bias-amplified model trained with generalized cross-entropy loss (Ahn et al., 2023; Liu et al., 2021). The advantages of GradTune are twofold: First, the ERM model with the classical cross-entropy loss alone can learn spurious features, which is reflected in sample gradients disparity across subgroups (Section 4.2), and sample gradients are only computed only the validation set instead of the entire training set. Second, recent work (Izmailov et al., 2022; Kirichenko et al., 2023; Tsirigotis et al., 2024) demonstrated that core features are learned by the ERM model and focusing on the last classification can substantially improve group robustness and reduce reliance on spurious features. Figure 1(a) shows an overview of each phase of GradTune.

ERM training. The first step of GradTune consists of ERM training the model using the training set. As the training data might contain spurious features that are not predictive of the target label, the ERM-trained model will capture the spurious features and achieve low classification error on the group of samples exhibiting the spurious correlation while having a higher error on samples where the spuriousness does not hold. The sensitivity of the ERM-trained model to spuriousness can guide the identification of worst-performing subgroups. For example, Liu et al. (2021); Nam et al. (2020); LaBonte et al. (2024) uses the misclassification of the ERM model to derive a set of samples that the model needs to improve, hypothesizing that this set mainly contains samples from the worst-performing subgroups. Instead, we consider in GradTune the gradient norm of each sample's loss with the hypothesis that the samples from the worst-performing subgroups exhibit a higher gradient norm for updating the model.

Reweighting set based on sample gradient norms. After training the ERM model on the training set, we compute the per-sample gradient norm on the held-out validation set without updating the model parameters. More specifically, we compute the classification loss for each sample and evaluate the norm of the gradient's loss w.r.t the model parameters. Given the computation cost of evaluating samples' gradient norm across all network layers, we only compute the gradient norms of the last classification layer. This is a frequently used technique for reducing the computational complexity of computing sample gradients (Killamsetty et al., 2021; Ahn et al., 2023; Kenfack et al., 2022; Mirzasoleiman et al., 2020). More specifically, we compute the gradient norm of each sample in the validation set as follows:

$$h(x, y) = \|\nabla_{\theta_{\text{fc}}} \mathcal{L}(f_\theta(x; \theta), y)\|_2, \forall (x, y) \in \mathcal{D}_{\text{eval}} \tag{2}$$

Where θ_{fc} are the parameters of the last classification layer, $\mathcal{L}$ the cross-entropy loss, f the ERM-trained model, and $\mathcal{D}_{\text{eval}}$ the held-out validation set. Following Ahn et al. (2022), we use the L_2 norm, which has demonstrated better performance in identifying out-of-distribution data based on the gradient vectors, while another type of norm can be considered (Huang et al., 2021).

After computing the gradient of each sample. i.e., $H = \{h(x, y) \mid (x, y) \in \mathcal{D}_{\text{eval}}\}$, we derive the fine-tuning set (also called the reweighting set) $\mathcal{D}_{\text{ft}} \subset \mathcal{D}_{\text{eval}}$ that will be used in the next step to mitigate spurious correlations learned by ERM model f_θ. Our fine-tuning set $\mathcal{D}_{\text{ft}}$ is constructed such that it mainly contains samples with higher gradient norms ($\mathcal{D}_e$) along with a smaller proportion of the samples sampled at random proportionally to their gradient norms ($\mathcal{D}_{e'}$). For example, Figure 1(b) shows, on the CelebA dataset, the gradient distribution of minority groups (i.e., blond males) against the majority subgroups (i.e., blond females); gradients norms are computed on ERM models trained with three independent random seeds; as can be seen, the top-k gradients norm might only contain blond male images, which is why we add a smaller proportion of samples to the fine-tuning set at random. More specifically, the fine-tuning data set consists of two subsets: (1) $\mathcal{D}_e$ containing samples in the top-k gradient norms and (2) $\mathcal{D}_{e'}$ samples not in the top-k sampled in the remaining validation set ($\mathcal{D}_{\text{eval}} \setminus \mathcal{D}_e$), with a sampling probability proportional to their gradient norms. More formally, $\mathcal{D}_{\text{ft}} = \mathcal{D}_e \cup \mathcal{D}_{e'}$ where $\mathcal{D}_e$ and $\mathcal{D}_{e'}$ are defined as follows: $\mathcal{D}_e = \{(x_i, y_i) \mid h(x_i, y_i) \in \text{Top-}k(H)\}$ and $\mathcal{D}_{e'} \sim \text{RandomSampler}(\mathcal{D}_{\text{eval}} \setminus \mathcal{D}_e, M' - k)$

Where M' is the size of the fine-tuning set, and k are hyperparameters. RandomSampler sample data points with sampling probability proportional to their gradient norm. We fixed M' to only 500

samples and following LaBonte et al. (2024), and we recommend using a higher proportion of M' for setting k to select enough minority samples.

Last layer fine-tuning The last step of our method focuses on fine-tuning the last classification layer using the fine-tuning set ($\mathcal{D}_{\text{ft}}$) obtained in the previous step. Here, we fine-tune the last layer, without resetting the model's parameters, of the ERM model using the fine-tuning set. In other words, we fine-tune the ERM model in a continual learning fashion using a small proportion of the validation set, consisting mainly of samples from the worst-performing subgroups and a smaller proportion of samples from the best-performing subgroups to avoid catastrophic forgetting in the ERM model. This means our fine-tuning set is close to, but not perfectly group-balanced, and recent literature (LaBonte et al., 2024; Jain et al., 2024), in the context of biased training data, has demonstrated that we can substantially improve group robustness without necessarily relying on a group-balanced dataset. According to LaBonte et al. (2024), worst-group accuracy may be affected by characteristics of the reweighting dataset other than group balance. Recent studies positioned class-balance training as a solid baseline for mitigating spurious correlation without group labels (Idrissi et al., 2022). Therefore, we perform the fine-tuning step with class-balance sampling during training to account for the imperfect group imbalance in the fine-tuning set. Our results show that class-balance sampling during fine-tuning is an important aspect of GradTune for better improving WGA. As can be seen in Figure 1(c), we report on the CelebA dataset, the average WGA of a Resnset-50 model trained with ERM before and after applying GradTune; the model is trained for three random seeds, and the WGA is averaged across seeds, results shows that GradTune can improve the ERM-model's WGA by up to 30%. This suggests the gradient norm disparity across samples in the held-out validation set provides a strong signal about the data impacting the worst-group accuracy. In the following subsection, we will see on several benchmarks that the gradient norms of the samples from the worst-performing subgroups are higher than those from the best-performing subgroups.

4.2 Does worst-performing groups experience a higher gradient norm?

In this subsection, we demonstrate on several datasets that samples from the worst-performing subgroups receive higher gradient norms from ERM-trained models. We trained ERM models across three independent random seeds and measured the per-sample gradient norms on the held-out validation set.

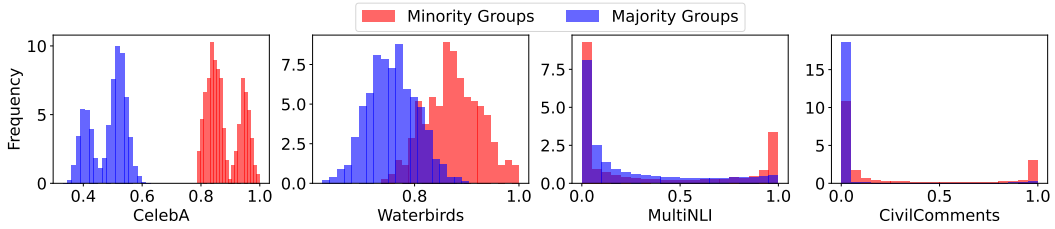

Figure 2: Distribution of the gradient norm of samples from different subgroups across different datasets. We trained ERM models and computed the gradient norm of samples in the held-out validation dataset. The gradient norms are average over three seeds.

Figure 2 shows, across datasets, that ERM-trained models provide higher gradient norms to data points where spurious correlation applies. The next section shows how leveraging these insights and applying GradTune on ERM models improves their robustness to spurious correlation without knowing any information about the spurious features.

5 Experiments

This section provides empirical results demonstrating the superiority of GradTune against existing state-of-the-art methods. We first describe the experimental setup, followed by the baseline methods used for comparison, and then present the results and discussion. The source code is available in the following repository: `https://github.com/patrikken/GradTune`.

Table 1: Comparison to other baseline methods. We report the average and standard deviation across three independent runs. Bolded and underlined represent the best and second best values, respectively.

Method	Group Labels		Waterbirds		CelebA		MultiNLI		CivilComments	
	train	val	Average	WGA	Average	WGA	Average	WGA	Average	WGA
Group DRO	✓	✓	93.5	91.4	92.9	88.9	81.4	77.1	88.9	69.9
JTT	✗	✓	93.3	$85.6_{\pm 0.2}$	88.0	81.1	78.6	72.6	92.6	69.3
DFR	✗	✓	$94.2_{\pm 0.5}$	$91.9_{\pm 1.0}$	$92.7_{\pm 0.5}$	$87.6_{\pm 2.2}$	$81.0_{\pm 0.1}$	$70.2_{\pm 0.4}$	$86.0_{\pm 0.0}$	$76.1_{\pm 0.2}$
SELF	✗	✓	$94.4_{\pm 0.5}$	$91.2_{\pm 1.1}$	$92.7_{\pm 2.1}$	$68.3_{\pm 11.0}$	$69.4_{\pm 10.3}$	$50.3_{\pm 22.7}$	$65.6_{\pm 28.3}$	$56.3_{\pm 24.1}$
AFR	✗	✓	$94.2_{\pm 1.2}$	$90.4_{\pm 1.1}$	$91.3_{\pm 0.3}$	$82.0_{\pm 0.5}$	$81.4_{\pm 0.2}$	$73.1_{\pm 0.6}$	$89.8_{\pm 0.6}$	$68.7_{\pm 0.6}$
ERM	✗	✗	$87.4_{\pm 1.1}$	$73.2_{\pm 1.0}$	$\underline{93.5}_{\pm 0.2}$	$71.5_{\pm 1.9}$	$\mathbf{81.8}_{\pm 0.2}$	$\underline{62.6}_{\pm 1.6}$	$\mathbf{90.1}_{\pm 0.1}$	$70.5_{\pm 0.8}$
Random	✗	✗	$90.5_{\pm 1.2}$	$80.9_{\pm 2.9}$	$92.4_{\pm 0.8}$	$81.7_{\pm 3.9}$	$80.8_{\pm 0.8}$	$56.8_{\pm 8.8}$	$85.9_{\pm 0.9}$	$71.2_{\pm 1.2}$
uLA	✗	✗	$91.5_{\pm 0.7}$	$\underline{86.1}_{\pm 1.5}$	$\mathbf{93.9}_{\pm 0.2}$	$\mathbf{86.5}_{\pm 3.7}$	-	-	-	-
GradTune	✗	✗	$\mathbf{94.3}_{\pm 0.2}$	$\mathbf{91.0}_{\pm 0.6}$	$90.7_{\pm 0.6}$	$\underline{85.6}_{\pm 3.8}$	$\underline{81.6}_{\pm 0.4}$	$\mathbf{64.2}_{\pm 4.3}$	$\underline{89.9}_{\pm 0.2}$	$\mathbf{78.6}_{\pm 1.3}$

5.1 SETUP

Datasets We consider four datasets commonly used for spurious correlation studies (Kirichenko et al., 2023; Ahn et al., 2023; Izmailov et al., 2022; LaBonte et al., 2024) across vision and language tasks: Waterbirds, CelebA dataset (Liu et al., 2015), MultiNLi dataset (Williams et al., 2017), and CivilComments (Koh et al., 2021). Details about the datasets can be found in Appendix A

Models. For the vision tasks, we use the Restnet-50 model pretrained (He et al., 2016) on ImageNet-1k (Russakovsky et al., 2015) and BERT model pretrained on Book Corpus (Kenton & Toutanova, 2019) for language tasks. For fair comparison to previous work (LaBonte et al., 2024; Qiu et al., 2023; Kirichenko et al., 2023), we use half of the validation set for the fine-tuning set and keep all the hyperparameters, i.e., we do not perform model selection using the other half as in LaBonte et al. (2024); Qiu et al. (2023). More details about the hyperparameters can be found in Appendix B. For applying GradTune to fine-tuning the last layer, we fix the size of the fine-tuning set $M' = 500$ following LaBonte et al. (2024) and use 80% of M' for the top-k, i.e., $k = 400$ and provide ablation on different values in Section 5.1.

Baselines. We considered five baseline methods for comparison and them depending on whether they use group labels during the training and/or validation: this includes Group DRO (Sagawa et al., 2019); *Just Train Twice* (JTT) (Liu et al., 2021); *Deep Feature Reweighting* (DFR) (Kirichenko et al., 2023); *Selective Last-layer Fine-tuning* (SELF) (LaBonte et al., 2024); *Automated Feature Reweighting*(AFR) (Qiu et al., 2023); *Bias-Unsupervised Logit Adjustment* (uLA) (Tsirigotis et al., 2024); ERM model trained with class-balance[1]; *Random finetuning* fine-tuning last-layer with random samples from $\mathcal{D}_{\text{ft}}$. More details about each baseline and related work can be found in the Appendix 3.

5.2 RESULTS AND DISCUSSION.

Table 1 summarizes the comparison to other baselines across the four datasets considered. We report the average accuracy and WGA obtained across three independent random seeds. Since we do not conduct hyperparameter tuning for these experiments, we also do not use group annotations for model selection. The results show that GradTune achieves competitive performance with methods requiring group information during training or validation. Notably, GradTune outperforms uLA on the waterbirds dataset and achieves comparable performance on the CelebA dataset. Note that for the CelebA dataset, the worst-performing subgroup of $85.6\%_{\pm 3.8}$ comes from the majority group (non-blond female) while the minority group's (blond-male) accuracy has improved from 73.2% to 89%, this suggests the fine-tuning step targetted improvement on the minority group. These results show that we can fine-tune ERM trained on high gradient samples and substantially improve WGA without intensive hyperparameter tuning, as in existing methods, or applying early stopping to worst-group validation accuracy. Surprisingly, we can also see that fine-tuning the ERM model with random samples from the held-out validation set improves WGA. A similar observation was

[1]Throughout the paper, the performance of the ERM reported is the ERM model trained with class balance, which is a strong baseline for group robustness without group annotations (Idrissi et al., 2022; LaBonte et al., 2024).

made in (LaBonte et al., 2024) while the extent of improvement in WGA depends on the contribution of the selected data points in the fine-tuning set. Our results posit gradient norms across samples as reliable selection criteria for fine-tuning sets when the group information is unknown. However, we observed that the WGA improvement of our methods is not substantial on the MultiNLI dataset. Our analysis revealed that the validation set in MultiNLI does contain enough data points from minority groups to improve substantially beyond the ERM model.

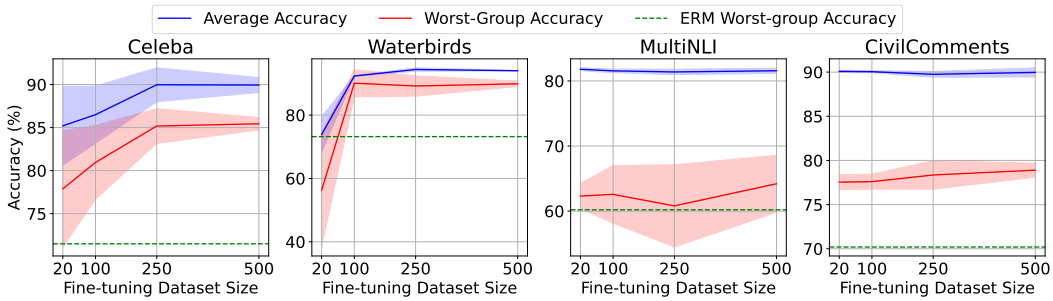

Figure 3: Study of the size of the fine-tuning set. Even with as little as 20 samples in the fine-tuning set, GradTune improves WGA over ERM, and the performance gets better as the size of the fine-tuning set increases.

On Figure 3, we plot the worst-group accuracy against different sizes of the fine-tuning step ($\mathcal{D}_{\text{ft}}$), i.e., 20, 100, 250, and 500. As can be seen in the Figure, even only 20 data points in the fine-tuning set, GradTune improve the WGA of the ERM-trained model, and the performance gets better as more data points are included in the fine-tuning set, especially for the Celeba and Waterbirds data, while the WGA is not much impacted in the MultiNLI and CivilComments datasets that require more training data for significant improvement.

6 Conclusion

In this paper, we present a novel method called GradTune for identifying and mitigating spurious correlations without using group labels. We demonstrate that ERM-trained models exhibit higher gradient norms on samples from the minority group in the hold-out dataset. The central intuition of this work is that fine-tuning the last classification layer with these high-gradient norm samples can substantially reduce the spurious correlation learned by the model and effectively emphasize the core features. Through various experiments and ablation studies, we show the effectiveness of the proposed methods, with competitive performance with existing methods, while not using group labels during training and validation for model selection. Developing model selection techniques without group labels remains an important and open research direction.

Acknowledgement

The authors thank the Digital Research Alliance of Canada and Denvr for computing resources. SEK is supported by CIFAR and NSERC DG (2021-4086) and UA by NSERC DG (2022-04006).

References

Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. Mitigating dataset bias by using per-sample gradient. In *The Eleventh International Conference on Learning Representations*, 2022.

Sumyeong Ahn, SeongYoon Kim, and Seyoung Yun. Mitigating dataset bias by using per-sample gradient. In *Eleventh International Conference on Learning Representations, ICLR 2023*. International Conference on Learning Representations, 2023.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

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*, pp. 770–778, 2016.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Improving subgroup robustness via data selection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Patrik Joslin Kenfack, Kamil Sabbagh, Adín Ramírez Rivera, and Adil Khan. Repfair-gan: Mitigating representation bias in gans using gradient clipping. *arXiv preprint arXiv:2207.10653*, 2022.

Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=3HE4vPNIfX`.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. *arXiv preprint arXiv:2212.01433*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

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.

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*, pp. 28448–28467. PMLR, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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. *International Conference on Learning Representations*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

David Steinmann, Felix Divo, Maurice Kraus, Antonia Wüst, Lukas Struppek, Felix Friedrich, and Kristian Kersting. Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *arXiv preprint arXiv:2412.05152*, 2024.

Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. *Advances in Neural Information Processing Systems*, 36, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

A DATASETS

We evaluate the worst-case performance of the proposed method on three classification tasks: two from the vision domain (Waterbirds and CelebA) and two from the language domain (CivilComments and MultiNLI)

- **Waterbirds** (Sagawa et al., 2019; Liu et al., 2021) is a dataset of birds derived from Caltech-UCSD Birds (CUB) (Wah et al., 2011) by synthetically creating a spurious correlation between bird species and the background. In particular, the class label is the type of bird appearing in the image (waterbirds and landbirds), and the background landscape (water, land) spuriously correlates with the bird type. Here, the minority subgroups represent images with the background landscape not aligned with the bird type, i.e., {`waterbird`, `land background`} and {`landbird`, `water background`}.
- **CelebA** (Liu et al., 2015) dataset contains images of celebrities with 40 facial attributes. In this dataset, the attribute `hair color` is spuriously correlated `gender`. We consider hair color {`blond`, `non-blond`} as the class label and gender {`male`, `female`} as group information.
- **CivilComments** (Koh et al., 2021) is a textual dataset collected from online comments. The task is to predict whether a comment is `toxic` or `non-toxic`. The label is spuriously correlated with comments related to some demographic subgroups such as gender (male, female), race (white, black), and sexual orientation (LGBT). We consider a binary indicator of comments related to these demographic subgroups as spurious group information.
- **MultiNLI** (Williams et al., 2017) is a language dataset that classifies pairs of sentences as contradiction, entailment, or neither. The spurious feature is the presence of negation in the second sentence; the presence of negation words ("no", "never"...) is correlated with the contradiction class and serves as a spurious feature.

Table 2: Statistics of the datasets considered. Class probabilities exhibit significant variations when conditioned on spurious features. It's important to note that Waterbirds is the sole dataset with a distribution shift, while MultiNLI is the only inherently class-balanced dataset. The minority groups within each class are in italics. Due to rounding, the probabilities may not total exactly 1.

Dataset	Group g		Training distribution $\hat{p}$			Data quantity		
	Class y	Spurious s	$\hat{p}(y)$	$\hat{p}(g)$	$\hat{p}(y \mid s)$	Train	Val	Test
Waterbirds	landbird	land	.768	.730	.984	3498	467	2225
	landbird	*water*		.038	.148	184	466	2225
	waterbird	*land*	.232	.012	.016	56	133	642
	waterbird	water		.220	.852	1057	133	642
CelebA	non-blond	female	.851	.440	.758	71629	8535	9767
	non-blond	*male*		.411	.980	66874	8276	7535
	blond	female	.149	.141	.242	22880	2874	2480
	blond	*male*		.009	.020	1387	182	180
CivilComments	neutral	no identity	.887	.551	.921	148186	25159	74780
	neutral	*identity*		.336	.836	90337	14966	43778
	toxic	*no identity*	.113	.047	.079	12731	2111	6455
	toxic	identity		.066	.164	17784	2944	8769
MultiNLI	contradiction	no negation	.333	.279	.300	57498	22814	34597
	contradiction	*negation*		.054	.761	11158	4634	6655
	entailment	no negation	.334	.327	.352	67376	26949	40496
	entailment	*negation*		.007	.104	1521	613	886
	neither	no negation	.333	.323	.348	66630	26655	39930
	neither	*negation*		.010	.136	1992	797	1148

B HYPERPARAMETERS

We use standard hyperparameters following recent literature that uses fine-tuning for spurious correlation mitigation (LaBonte et al., 2024; Kirichenko et al., 2023; Izmailov et al., 2022). For the vision tasks, we use the Restnet-50 model pretrained (He et al., 2016) on ImageNet-1k (Russakovsky et al., 2015) and BERT model pretrained on Book Corpus and English Wikipedias (Kenton & Toutanova, 2019) for language tasks. These pretrained models serve as the starting point for ERM models across the four datasets we consider. For data preprocessing, we apply standard ImageNet normalization along with typical flip and crop augmentation for the vision tasks and BERT tokenization for the language tasks. For ERM and last-layer finetuning (Section 4.1), we do not vary any hyperparameters; their fixed values are listed in Table 3. Specifically, our reweighting set is fixed to 500, containing the top-400 gradient norm samples, and the remaining 100 data points are sampled at random proportionally to their gradient norm. As in recent work on last-layer retraining methods in Kirichenko et al. (2023); LaBonte et al. (2024), the held-out dataset has a fixed size of 600 for Waterbirds, 9934 for CelebA, 22590 for CivilComments, and 41231 for MultiNLI, which correspond to half of the validation set provided in each dataset. With the ERM-trained model, we calculate the sample gradient only in these held-out sets and determine the fine-tuning set ($\mathcal{D}_{\text{ft}}$) based on gradient norms. We set the size of the fine-tuning set to just 500 samples and demonstrated in the main paper that the worst-group performance can be improved over the ERM-trained model with as few as 20 data points in the fine-tuning set.

Table 3: ERM and last-layer fine-tuning hyperparameters. For training the ERM models and fine-tuning the last layer, we used the same fixed hyperparameters as in the previous work (Kirichenko et al., 2023; LaBonte et al., 2024; Qiu et al., 2023). We increased the number of epochs for the fine-tuning step to 500 for all datasets.

Dataset	Optimizer	Initial LR	LR schedule	Batch size	Weight decay	Epochs
Waterbirds	SGD	3×10^{-3}	Cosine	32	1×10^{-4}	100
CelebA	SGD	3×10^{-3}	Cosine	100	1×10^{-4}	20
CivilComments	AdamW	1×10^{-5}	Linear	16	1×10^{-4}	10
MultiNLI	AdamW	1×10^{-5}	Linear	16	1×10^{-4}	10

C ADDITIONNAL RESULTS

Class balance fine-tuning. In the last phase of WGA, we perform last-layer fine-tuning using a class balance batch sampler; thereby, in expectation, different classes are equally represented across batches. This experiment compares the fine-tuning step of WGA with class imbalance and class balance sampling. Table 5 reports the average accuracy and the WGA of last-layer fine-tuning with class balance or imbalance sampling. Results show that while class imbalance fine-tuning improves upon the ERM model, it performs worse in terms of WGA than class balance fine-tuning, except on the CivilComments dataset where class imbalance fine-tuning has slightly better WGA. Intuitively, it is challenging to derive a perfect group-balance fine-tuning set without group labels; class-balance sampling can improve the balance across subgroups during training, which justifies the improved worst-group performance.

Table 4: Comparison between GradTune with class balance fine-tuning vs class-imbalance fine-tuning. We report the average and standard deviation over three independent runs.

Method	Waterbirds		CelebA		MultiNLI		CiviComments	
	Average	WGA	Average	WGA	Average	WGA	Average	WGA
ERM	$87.4_{\pm 1.1}$	$73.2_{\pm 1.0}$	$93.5_{\pm 0.2}$	$71.5_{\pm 1.9}$	$81.8_{\pm 0.2}$	$62.6_{\pm 1.6}$	$90.1_{\pm 0.1}$	$70.5_{\pm 0.8}$
Class imbalance	$94.8_{\pm 0.2}$	$78.3_{\pm 3.1}$	$84.6_{\pm 1.2}$	$77.9_{\pm 1.7}$	$81.1_{\pm 0.4}$	$59.7_{\pm 7.7}$	$89.3_{\pm 0.6}$	$79.1_{\pm 3.2}$
Class balance	$94.3_{\pm 0.2}$	$91.0_{\pm 0.6}$	$90.7_{\pm 0.6}$	$85.6_{\pm 3.8}$	$81.6_{\pm 0.4}$	$64.2_{\pm 4.3}$	$89.9_{\pm 0.2}$	$78.6_{\pm 1.3}$

Last layer retraining vs. fine-tuning. Table 5 compares applying GradTune with last-layer retraining vs. fine-tuning. For retraining, we reset the parameters of the last layer, while for

fine-tuning, we continue training the last layer of the ERM model without resetting the weights. Empirically, we observed that fine-tuning provides better WGA than retraining. We hypothesize this is because fine-tuning uses some of the knowledge already encoded by the initial weights while retaining, despite improving the performance, suffers from the fact the retraining set is not fully balanced. For example, on the Waterbirds, the held-out dataset is a prior group balance dataset due to the distribution shift (Liu et al., 2022), and the WGA difference between fine-tuning and retraining is 1%. More investigations are needed to fully understand the last-layer retraining/fine-tuning effect on the group robustness, which we leave for future work.

Table 5: Comparison between GradTune with last-layer fine-tuning and retraining. While retraining and fine-tuning both improve WGA, finetuning the last layer without resetting the model's weights provides better WGA. We report the average and standard deviation over three independent runs

Method	Waterbirds		CelebA		MultiNLI		CiviComments	
	Average	WGA	Average	WGA	Average	WGA	Average	WGA
ERM	$87.4_{\pm 1.1}$	$73.2_{\pm 1.0}$	$93.5_{\pm 0.2}$	$71.5_{\pm 1.9}$	$81.8_{\pm 0.2}$	$62.6_{\pm 1.6}$	$90.1_{\pm 0.1}$	$70.5_{\pm 0.8}$
Retraining	$94.2_{\pm 0.1}$	$89.9_{\pm 1.0}$	$90.2_{\pm 0.7}$	$83.5_{\pm 4.0}$	$81.8_{\pm 0.1}$	$61.7_{\pm 2.7}$	$90.1_{\pm 0.1}$	$77.5_{\pm 0.9}$
Finetuning	$94.3_{\pm 0.2}$	$91.0_{\pm 0.6}$	$90.7_{\pm 0.6}$	$85.6_{\pm 3.8}$	$81.6_{\pm 0.4}$	$64.2_{\pm 4.3}$	$89.9_{\pm 0.2}$	$78.6_{\pm 1.3}$

Gradient distribution across groups. Figure 4 complements Figure 2 in the main paper and shows the average gradient norm within bias-aligned and bias-conflicting groups. We report the average and standard deviation of the majority and the minority groups over three independent runs. We observe that across the dataset, the worst-group data in the held-out dataset have higher gradient norms. Intuitively, as the ERM-trained model poorly generalizes on minority group data, gradient updates for these points in the held-out set will be much higher as they will provide more signal to the pretrained model for parameters' updates.

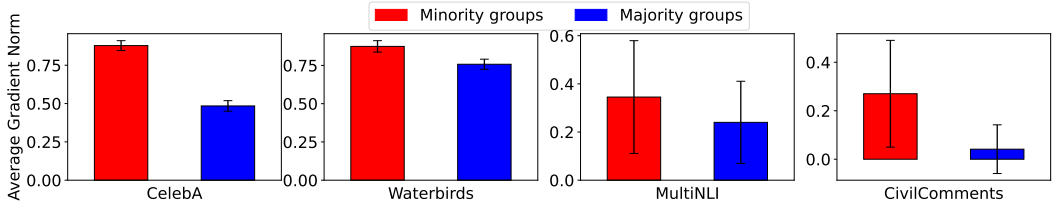

Figure 4: Average gradient norm of different groups across different datasets. We trained ERM models and computed the gradient norm of samples in the held-out validation dataset. The gradient norms are averaged over three independent runs.

Ablation on the impact of top-k gradient norms sampling. In the main paper, we fixed the fine-tuning size to 500 and used the top-400 gradient norm samples, and the remaining 100 points were sampled randomly with probability proportional to their gradient norm. In this experiment, we vary the value of k in Waterbirds and Celeba and analyze its impact on the WGA.

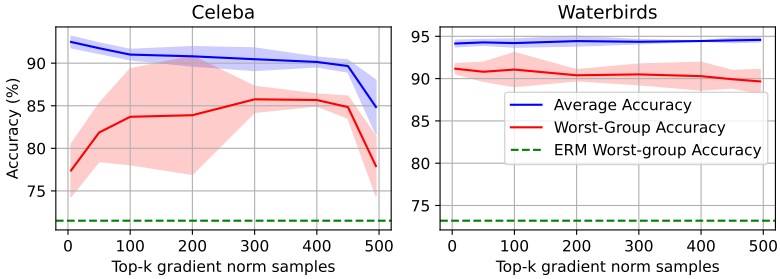

Figure 5: Impact of the top-k gradient norm samples

Figure 5 shows the overall accuracy and the worst-group accuracy for different values of $k \in \{5, 50, 100, 200, 300, 400, 450, 500\}$. When k is closer to zero, most data points in the fine-tuning set are mainly randomly sampled with a probability proportional to their gradient norm; for higher values of k, higher gradient norm samples are included in the fine-tuning set first. When k is close to zero, on the Celeba dataset, the fine-tuning set does not ensure enough samples from the minority group, and when the k is closer to 500, the fine-tuning set only contains samples from minority groups. This is why the worst-group accuracy decreases in both cases. Furthermore, as can be seen in Table 6, the group experiencing the worst performance changes depending on the value of k: for $k = 5$, the corresponding worst-group is `blond,male` with 77.4% accuracy and for $k = 495$ the worst-group shifts to `nonblond,female` with 77.9% accuracy. This justifies combining the top-k gradient norm samples and randomly sampled data points to avoid overfitting on a specific group. We also observe higher standard deviations for values of k below have 50% of M'; this is due to the fact that most data points are randomly drawn based on gradient norms.

The Waterbirds dataset is less sensitive to k, and the worst-performing subgroups remain almost the same for all values of k but sharply decrease as the value of k gets closer to 100% of M', similarly in the CelebA dataset. The consistent worst-group performance in the Waterbirds dataset is due to the relatively smaller size of the validation set and the smaller discrepancy between the gradient norms of the minority and the majority group (Cf. Figure 4). Furthermore, Table 7 shows the group-wise accuracy comparison between the ERM model before and after applying GradTune. These results show the performance of the worst-performing subgroups significantly improves after fine-tuning the last layer with the fine-tuning set. We also observe that in most cases, the worst-performing subgroup becomes the majority group; we argue this is due to catastrophic forgetting after continual training of the last layer of the model. A better fine-tuning strategy can be derived to ensure the model maintains its performance on the majority group. A validation set with group labels can also be used to select a model that better compromises the performance across subgroups.

Table 6: Average and worst-group accuracy (%) for different Top-k gradient norm samples. We report the average and standard deviation over three independent runs

Top-k	CelebA			Waterbirds		
	Avg	WGA	Worst Group	Avg	WGA	Worst Group
5	$92.5_{\pm 0.7}$	$77.4_{\pm 3.1}$	`blond,male`	$94.1_{\pm 0.4}$	$91.2_{\pm 0.6}$	`waterbirds,water`
50	$91.8_{\pm 0.6}$	$81.9_{\pm 3.4}$	`blond,male`	$94.3_{\pm 0.3}$	$90.8_{\pm 1.1}$	`waterbirds,water`
100	$91.0_{\pm 0.6}$	$83.7_{\pm 5.6}$	`blond,male`	$94.2_{\pm 0.5}$	$91.1_{\pm 2.0}$	`waterbirds,water`
200	$90.8_{\pm 1.1}$	$83.9_{\pm 6.9}$	`blond,male`	$94.4_{\pm 0.6}$	$90.4_{\pm 0.6}$	`waterbirds,water`
300	$90.5_{\pm 1.3}$	$85.8_{\pm 1.5}$	`nonblond,female`	$94.4_{\pm 0.2}$	$90.5_{\pm 1.2}$	`waterbirds,water`
400	$90.1_{\pm 0.6}$	$85.7_{\pm 0.7}$	`nonblond,female`	$94.4_{\pm 0.0}$	$90.3_{\pm 1.6}$	`waterbirds,water`
450	$89.7_{\pm 0.7}$	$84.8_{\pm 1.3}$	`nonblond,female`	$94.5_{\pm 0.3}$	$89.9_{\pm 1.0}$	`waterbirds,water`
495	$84.9_{\pm 3.1}$	$77.9_{\pm 3.5}$	`nonblond,female`	$94.6_{\pm 0.2}$	$89.7_{\pm 1.4}$	`waterbirds,water`

Table 7: Group-wise accuracy comparison between ERM before and after applying GradTune.

Dataset	Groups	ERM	ERM after GradTune
Waterbirds	landbirds,land	$99.5_{\pm 0.2}$	$95.6_{\pm 0.9}$
	landbirds,water	$76.9_{\pm 3.2}$	$94.4_{\pm 0.3}$
	waterbirds,land	$73.2_{\pm 1.1}$	$92.7_{\pm 0.5}$
	waterbirds,water	$96.6_{\pm 0.3}$	$91.0_{\pm 0.7}$
Celeba	nonblond,female	$90.1_{\pm 0.3}$	$85.6_{\pm 3.85}$
	nonblond,male	$97.5_{\pm 0.2}$	$93.3_{\pm 1.8}$
	blond,female	$96.2_{\pm 0.2}$	$97.8_{\pm 0.4}$
	blond,male	$71.5_{\pm 2.0}$	$86.7_{\pm 6.0}$
Civilcomments	neutral,no-identity	$95.0_{\pm 0.1}$	$94.8_{\pm 0.2}$
	neutral,identity	$85.9_{\pm 0.6}$	$85.3_{\pm 0.8}$
	toxic,no-identity	$79.4_{\pm 0.3}$	$79.6_{\pm 0.8}$
	toxic,identity	$70.5_{\pm 0.9}$	$78.6_{\pm 1.4}$
Multinli	contradiction,no-negation	$81.2_{\pm 0.5}$	$80.5_{\pm 1.6}$
	contradiction,negation	$95.2_{\pm 0.2}$	$95.0_{\pm 0.6}$
	entailment,no-negation	$83.5_{\pm 0.5}$	$83.3_{\pm 1.1}$
	entailment,negation	$77.3_{\pm 0.8}$	$78.3_{\pm 1.0}$
	neither,no-negation	$78.9_{\pm 1.1}$	$79.2_{\pm 1.1}$
	neither,negation	$62.6_{\pm 1.7}$	$64.2_{\pm 4.4}$