Appendix

A. Additional Details

A.1. Group Composition in Datasets

In Sec. 5.1 we introduce the various datasets used in our experiments. We present the group-wise distribution of data samples in Table 3.

Dataset		Groupwise composition					
Waterbirds		(waterbird, water bg)	(waterbird, land bg)	**(landbird, water bg)**	(landbird, land bg)		
	Train	72.7%	3.9%	1.2%	22.2%		
	Val	38.9%	38.9%	11.1%	11.1%		
	Test	38.9%	38.9%	11.1%	11.1%		
CelebA		(blond, female)	(non-blond, male)	(non-blond, female)	**(blond, male)**		
	Train	44.0%	41.1%	14.0%	0.9%		
	Val	43.0%	41.7%	14.5%	0.9%		
	Test	48.9%	37.7%	12.4%	0.9%		
MultiNLI		(contradiction, no-negation)	(contradiction, negation)	(entailment, no-negation)	**(entailment, negation)**	(neutral, no-negation)	(neutral, negation)
	Train	27.9%	5.4%	32.7%	0.7%	32.2%	1.0%
	Val	27.7%	5.6%	32.7%	0.7%	32.3%	1.0%
	Test	28.0%	5.4%	32.7%	0.7%	32.3%	0.9%
CivilComments		(non-toxic, no-identity)	(non-toxic, identity)	**(toxic, no-identity)**	(toxic, identity)		
	Train	53.4%	35.3%	4.4%	6.9%		
	Val	53.9%	34.9%	4.4%	6.8%		
	Test	54.1%	34.5%	4.5%	6.8%		

Table 3. Different groups and their compositions in the training, validation and test splits of the four datasets. Boxed text highlights the minority group by frequency in training split–these groups by definition are "in violation" of the incidental / spurious feature correlation that is established by the majority group.

A.2. Training Details

In this section we follow up on the details provided in the Section 5.4. For all vision experiments, we consistently used ResNet-18 as the student model and ResNet-50 GroupDRO trained model as the teacher. For auxiliary layer we used 1 depth BasicBlock of ResNet. The ResNet network is composed of stages (each itself contains multiple BasicBlocks), we apply auxiliary layer only at the end of stages (except stage 4), this gives us three different choice for the hyperparameter aux position ($\mathcal{A}_P$). For text datasets, we used DistilBert as the student model and Bert GroupDRO trained model as the teacher. We used 2 layer neural network as an auxiliary layer that is applied at the end of encoder layers present in DistilBert. Table 4 shows the hyperparameter search space for all the hyperparameters that we tune on the basis of worst group accuracy of validation set. Table 5 shows the best hyperparameter configurations choosen for each dataset from the result of this grid search.

Hyperparameter	Range
α	[0.01, 0.05, 0.1, 0.2, 0.5]
β	[3, 3.5, 4., 4.5]
$\mathcal{A}_\mathcal{P}$	[1, 2, 3]
lr	[1e-5, 2e-5, 5e-4, 1e-3]
wd	[0.1, 0.01, 0.001]

Table 4. Range for hyperparameter search. Here α and β are the weighting parameters, $\mathcal{A}_\mathcal{P}$ determines the position of auxiliary layer, lr is the learning and wd denotes weight decay.

Dataset	α	β	$\mathcal{A}_{\mathcal{P}}$	lr	wd
Waterbirds	0.05	4	1	5e-4	0.1
CelebA	0.1	3.5	1	5e-4	0.01
CivilComments	0.05	3	1	2e-5	0.01
MultiNLI	0.2	3	2	2e-5	0.01

Table 5. Best hyperparameters for each dataset

B. Additional experiments

B.1. Sensitivity Analysis

In this section we show the sensitivity analysis of DEDIER on the Waterbirds dataset. Table 6 shows the average accuracy and worst group accuracy while distilling using DEDIER from GroupDRO teacher in a setting similar to Table 1 for different values of the weighting parameters α and β. We observe that tuning weighting parameters according to the dataset does help, but the performance is not too sensitive to these hyperparameters (as the standard deviation of the performance metric remains low across different hyperparameter values).

α	β	Avg. Acc.	Worst-group Acc.
0.05	**4**	**92.3**	**90.3**
0.05	3	91.5	89.7
0.05	3.5	92.1	90.2
0.05	4.5	91.1	90.3
0.01	4	92.2	90.3
0.1	4	91.1	88.9
0.2	4	91.8	90.8
0.5	4	91.1	89.5
	Mean	91.6	90.0
	Std dev.	0.52	0.60

Table 6. Sensitivity analysis of DEDIER's performance to the weighting hyperparameters α and β on the Waterbirds dataset.

B.2. Additional Results

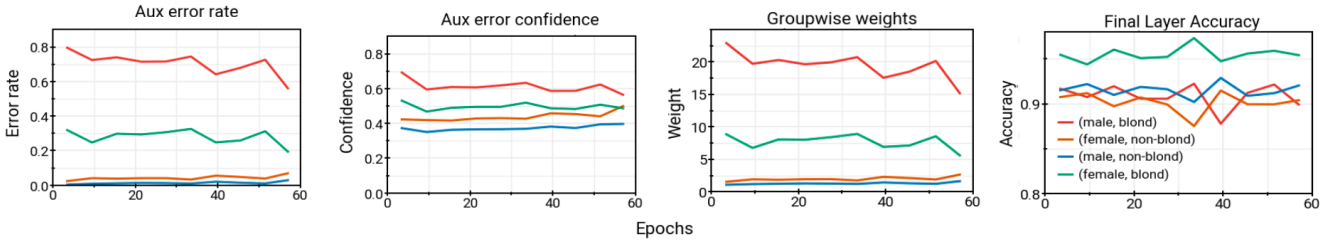

Figure 6. Evolution of reweighting during distillation (CelebA dataset). As expected, minority groups have high error rate; and through the distillation process, the overconfidence reduces

We replicate Figure 5 in main paper for the training dynamics on the CelebA dataset with similar findings, thus showing that the dynamics of our method play a crucial role in training of the final classifier.

B.3. Analyzing the importance of early readouts in DEDIER

To demonstrate the effect of *early* readouts on the final performance, we show the results on Waterbirds dataset of performance on varying the position of auxiliary layer. As shown in figure 7 the worst group accuracy shows a downward trend

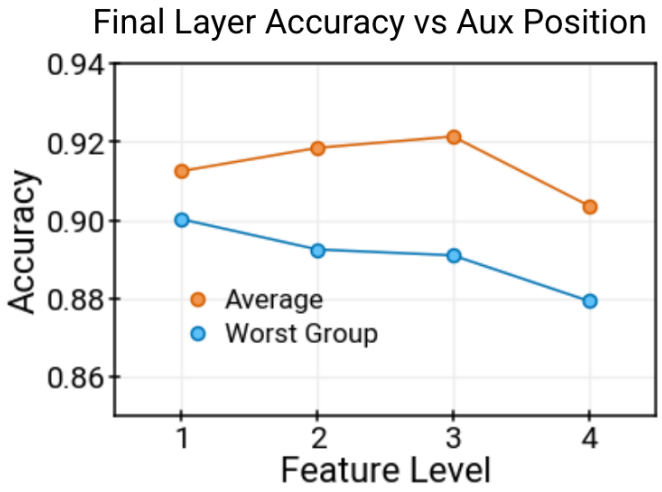

Figure 7. Final performance on the Waterbirds dataset by DEDIER wrt feature layer chosen for readout. As expected, early readout leads to better debiased performance than using final layer.

wrt to depth from which readout is taken. Overall, early-readouts (ie. feature level 1-3) are better than using the final layer (feature level 4), thus reaffirming the hypothesis mentioned in the main paper (Section 4.1). We have the same result on other datasets as well: ref. Table 5 where the optimum $\mathcal{A}_{\mathcal{P}}$ is on the earliest layers.

B.4. Analyzing weight assignments by DEDIER

For the datasets WaterBird and CelebA we analysed the weights assigned (at the end of training) and report the findings in Table 7. In this table, we provide the mean weight for each group along with its standard deviation. Notably, we observe a discernible correlation between the assigned weights and the number of data points within each group. It is noteworthy that our method, DEDIER, tends to allocate larger weights to groups with fewer data points, effectively addressing the dataset's inherent imbalance. Remarkably, DEDIER accomplishes this without using label information to explicitly balance the dataset.

Furthermore, the weights detailed in Table 7 underscore the effectiveness of our novel method for identifying the worst-performing group, even in the absence of explicit group information. It's apparent that the highest weights are consistently assigned to the group where the spurious correlations are broken, in both datasets. This finding further validates the improvements highlighted in Table 1.

Waterbirds' groups	#Samples	Mean weight
(waterbird, water background)	3498	3.2 ($\pm$ 3)
(waterbird, land background)	184	39.07 ($\pm$ 6.73)
(landbird, water background)	56	49.03 ($\pm$ 5.17)
(landbird, land background)	1057	10.6 ($\pm$ 4.17)
CelebA's groups		
(blond, male)	138	19.52 ($\pm$ 3.37)
(blond, female)	22880	7.75 ($\pm$ 2.17)
(non-blond, male)	66874	1.28 ($\pm$ 0.22)
(non-blond, female)	71629	1.97 ($\pm$ 0.5)

Table 7. Table presents the group distribution in training and the mean weightage given by DEDIER across the training. Note that minority groups are upweighted the most.