A. Appendix

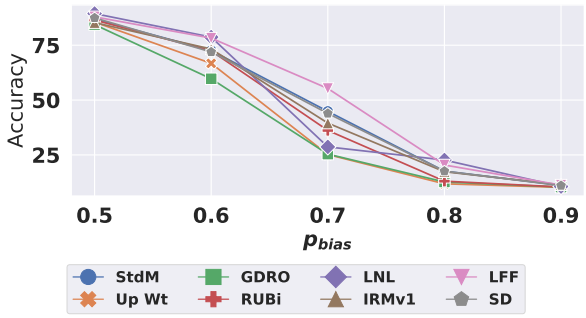


Figure A7: Unbiased accuracy on Biased MNIST as a function p_{bias}.

A.1. Additional Results

We now provide additional results for all datasets, that provide further details for the results presented in Sec. 6.

Biased MNIST. First, we show the unbiased accuracy as a function of p_{bias} in Fig. A7. None of the methods exhibit robustness across the range of p_{bias} values, with all methods showing drops in the unbiased accuracy as p_{bias} increases from 0.5 to 0.9. Second, in Fig. A8, we show the unbiased accuracies as functions of learning rate and weight decay, when p_{bias} is set to 0.7. The figure on the left shows unbiased accuracies as a function of learning rate with a fixed weight decay of 10^{-5} and the figure on the right shows unbiased accuracies as a function of weight decay, with learning rate set to 10^{-3}. Barring Up Wt, all methods show highest unbiased accuracies at learning rate = 10^{-3} and weight decay = 10^{-5}, and show large drops with other values, indicating high sensitivities to hyperparameters.

Third, in Table. A4, we compare the explicit methods with each other, when considering each of the seven variables as explicit bias variables ($b_{expl.}$) in separate experiments. For the explicit methods, variables apart from $b_{expl.}$ act as implicit biases, whereas implicit methods are not affected by the choice of the explicit variables. As previously discussed in Sec. 6.2, Up Wt, GDRO and RUBi show higher majority/minority difference for $b_{expl.}$ as compared to other variables, and all explicit methods are below StdM in terms of unbiased accuracies. Finally, in Table. A5, we show the per variable majority/minority accuracies when considering combinations of variables as explicit biases. As discussed in Sec. 6.3, explicit methods barring Up Wt fail when multiple variables are specified as explicit bias variables. The results have been averaged across three runs for both tables. We provide a ten samples for each digit in Fig. A10.

CelebA. We show accuracy for each group of CelebA in Table. A3. SD and GDRO obtain the highest accuracies. As discussed previously, we observe trade-offs between blond and non-blond classes with the improvements in the rare blond class incurring degradations in the non-blond class.

Table A3: Per group accuracies on CelebA.

Groups/ Methods	Male		Non-Male		Unbiased Accuracy
	Non-Blond	Blond	Non-Blond	Blond	
StdM	***99.3***	42.8	***95.9***	83.3	80.3
Up Wt	89.3	**77.8**	87.4	**95.0**	87.4
GDRO	91.2	***80.0***	90.9	92.1	**88.5**
RUBi	96.6	68.9	94.3	89.1	87.2
LNL	**99.1**	39.4	**95.4**	82.8	79.2
IRMv1	**99.1**	40.6	***95.9***	83.6	79.8
LFF	96.0	46.1	90.4	80.0	77.8
SD	96.4	73.3	88.0	***96.6***	***88.6***

GQA-OOD. GQA-OOD [36] defines the tail accuracy (Acc-tail) metric which is computed on the samples of the tail of the answer class distribution. Specifically, an answer class a_i is considered to be a tail class for a local group if:

$$|a_i| \leq (1+\beta)\mu(a),$$

where, $|a_i|$ is the number of instances for answer a_i in the given group, $\mu(a)$ is the mean number of answers in the group and β can be used to control the tail size. In Fig. A9, we plot the tail accuracies at different tail sizes, considering different explicit bias variables for the explicit methods. For implicit methods: StdM, LFF and SD, same tail accuracies are repeated on all four charts since they are not affected by the choice of explicit variables during training. Explicit methods fail when the explicit variables entail generalization to large number of groups, whereas implicit methods are close to or above StdM.

A.2. Hyperparameters and Other Details

We select hyperparameters based on the best unbiased validation set accuracy on each dataset, which is reflective of the unbiased test distribution. For all datasets and methods, we first perform a grid search over the learning rates $\in$ {1e-3, 1e-4, 1e-5} and weight decays $\in$ {0, 0.1, 1e-3, 1e-5}, and then tune the method-specific hyperparameters. For Biased MNIST, the hyperparameters were selected using single run, considering 'distractor shape' as the explicit bias variable for explicit methods. For CelebA, they were selected based on the unbiased accuracy/mean per group on the validation set and for GQA-OOD, they were selected based on the best mean head/tail accuracy when setting $\beta = 0.2$ (the default value in the original paper).

Next we specify the ranges considered for method-specific hyperparameters. For GDRO, we search the group weight step size between $\{0.001, 0.01, 0.1\}$. For LNL, we perform a grid search over gradient reversal weights

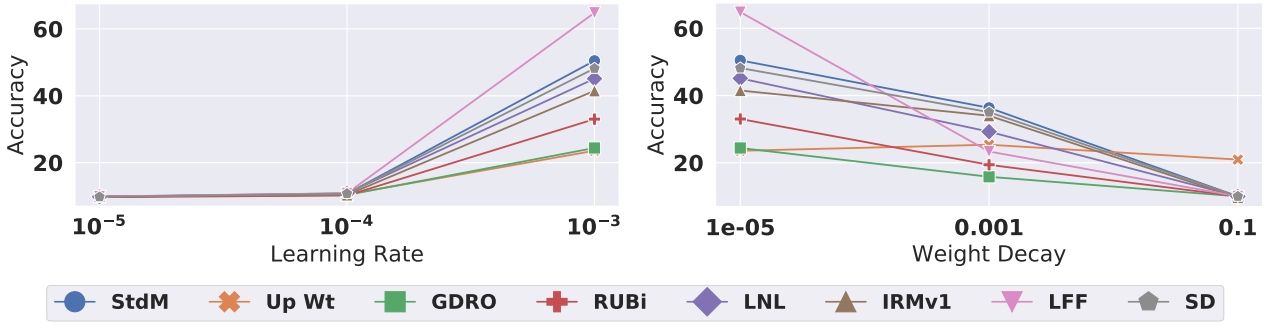


Figure A8: Unbiased test accuracy (mean per group) as a function of learning rate and weight decay (using single random seed).

$\lambda_{adv} \in \{-1.0, -0.1, -0.01\}$ and entropy loss weights $\lambda_{ent} \in \{1.0, 0.1, 0.01, 0\}$. For IRMv1, we search for λ_{grad} values from {1, 10, 100, 1000, 10000}. The original implementation of IRMv1 samples from all environments in a single mini-batch during training. While this is feasible for small scale problems with few environments e.g., four environments/explicit groups for CelebA, it is computationally infeasible for Biased MNIST and GQA, where the number of environments is larger than the batch size itself. So, for Biased MNIST and GQA, we sample from 16 randomly selected groups or environments within each mini-batch during training. This implies that our implementation of IRMv1 samples uniformly from all environments.

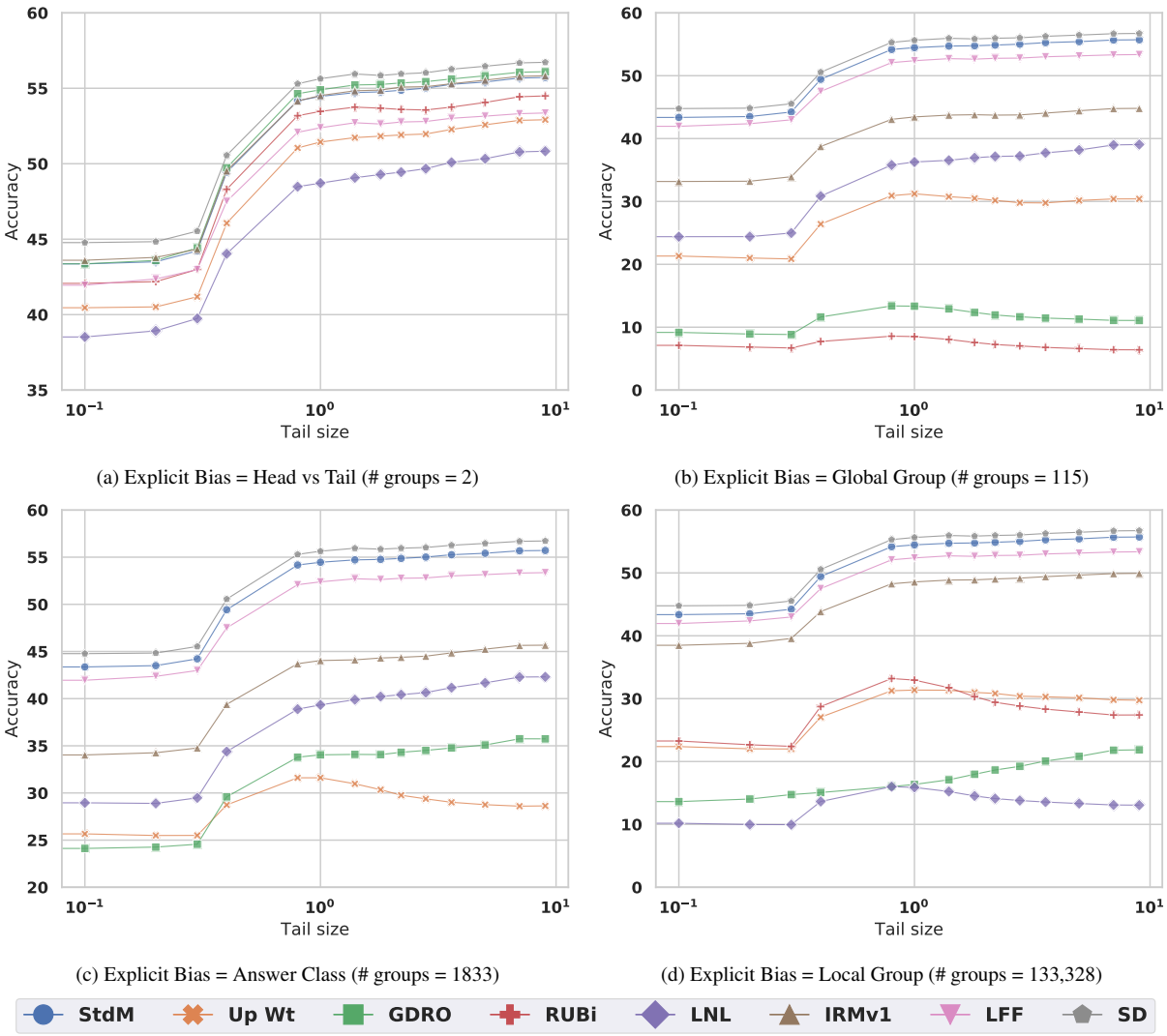
Even though the original paper proposing LFF uses a fixed $\gamma = 0.7$ for all datasets, we search over $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and find that LFF is indeed sensitive to γ. The default value: $\gamma = 0.7$ is not the optimal value for all cases. For SD, we consider $\gamma, \lambda \in \{10^{-3}, 10^{-2}, 0.1, 1.0, 10.0, 100.0\}$. For CelebA, we use the class-specific γ values specified in the paper and only tune λ. The procedure for obtaining class-specific hyperparameters when there are large number of classes e.g., in GQA however remains unclear for SD. The complete specification of hyperparameters is provided in Table. A6.

Table A4: Majority (Maj.) and Minority (Min.) group accuracies for each variable in Biased MNIST when using one of the seven variables as explicit bias for the explicit methods.

Variables/ Methods	Unbiased Accuracy	Background Color		Digit Color		Digit Position		Distractor Shape		Distractor Color		Texture		Texture Color	
		Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.
StdM	42.0	70.3	39.1	74.6	38.7	48.3	41.6	69.2	39.2	68.9	39.4	69.4	39.1	65.2	39.7
$b_{expl.}$ = Background Color															
Up Wt	39.0	42.0	38.9	72.7	35.4	43.3	38.7	71.2	35.6	70.3	35.8	67.6	35.9	65.4	36.2
GDRO	32.8	52.2	30.8	66.8	29.2	37.0	32.5	62.4	29.6	64.1	29.6	64.5	29.3	62.6	29.6
RUBi	47.9	51.9	47.7	78.8	44.6	52.4	47.6	77.4	44.8	77.8	44.9	76.5	44.8	72.9	45.2
LNL	44.0	55.8	43.5	59.7	43.0	46.8	44.5	57.5	43.3	57.6	43.3	56.6	43.3	53.2	43.7
IRMv1	38.0	66.0	35.0	70.5	34.5	41.2	37.8	68.1	34.8	68.3	34.9	64.5	35.1	65.6	35.0
$b_{expl.}$ = Digit Color															
Up Wt	19.4	48.6	16.1	22.8	19.0	21.5	19.2	52.0	15.7	47.7	16.3	50.5	15.7	44.9	16.5
GDRO	22.6	50.1	19.6	45.7	20.1	25.7	22.3	52.1	19.3	52.3	19.4	52.0	19.2	51.5	19.4
RUBi	24.0	54.5	20.6	27.9	23.6	25.0	23.9	57.3	20.2	55.4	20.6	55.4	20.3	52.6	20.7
LNL	40.2	53.7	39.5	57.9	39.0	46.2	40.3	56.1	39.2	55.1	39.4	56.1	39.2	50.9	39.8
IRMv1	40.9	66.8	38.4	72.3	37.8	43.8	41.0	71.1	37.9	68.9	38.3	67.8	38.2	66.5	38.4
$b_{expl.}$ = Digit Position															
Up Wt	21.4	46.7	18.6	50.4	18.2	21.9	21.4	53.4	17.8	50.4	18.3	49.0	18.2	45.8	18.6
GDRO	26.8	54.9	23.9	55.4	23.9	27.6	27.0	58.2	23.5	54.9	24.0	54.5	23.9	51.6	24.3
RUBi	26.6	54.9	23.5	59.1	23.0	28.3	26.5	56.8	23.2	56.9	23.4	56.0	23.2	53.5	23.6
LNL	41.4	51.8	40.9	56.6	40.4	46.4	41.5	54.4	40.6	53.7	40.7	52.7	40.7	49.4	41.1
IRMv1	38.1	67.8	35.0	70.8	34.7	42.9	37.8	69.3	34.8	68.0	35.1	67.0	34.9	66.0	35.1
$b_{expl.}$ = Distractor Shape															
Up Wt	21.5	52.0	18.2	56.0	17.8	24.2	21.4	25.6	21.2	53.0	18.2	54.0	17.9	51.9	18.2
GDRO	25.2	54.1	22.2	56.9	21.9	27.9	25.2	48.9	22.8	56.0	22.1	56.0	21.9	52.3	22.4
RUBi	34.5	66.9	31.0	70.9	30.5	35.2	34.5	40.4	33.9	67.7	31.0	67.5	30.7	64.6	31.1
LNL	44.1	55.8	43.1	59.6	42.6	50.5	43.7	57.9	42.8	57.4	42.9	58.8	42.7	55.0	43.1
IRMv1	38.3	67.0	35.5	70.5	35.1	41.5	38.3	69.3	35.2	65.9	35.7	67.9	35.2	64.6	35.7
$b_{expl.}$ = Distractor Color															
Up Wt	40.9	70.1	38.0	70.8	38.0	46.6	40.7	73.0	37.7	44.8	40.9	73.0	37.6	68.5	38.2
GDRO	28.1	58.6	24.8	57.4	24.9	30.4	28.0	62.1	24.4	49.1	25.9	59.1	24.6	55.8	25.1
RUBi	49.5	77.3	46.6	80.1	46.3	54.5	49.1	78.3	46.4	53.8	49.2	77.6	46.4	76.4	46.6
LNL	38.9	52.4	38.2	55.9	37.8	46.0	38.9	56.6	37.7	52.8	38.2	56.0	37.7	51.5	38.3
IRMv1	39.7	66.1	36.8	72.3	36.1	45.1	39.2	68.8	36.5	69.1	36.6	66.4	36.6	64.8	36.9
$b_{expl.}$ = Texture															
Up Wt	29.3	60.8	25.9	66.1	25.4	32.9	29.1	64.6	25.5	62.3	25.9	32.9	29.1	60.7	25.9
GDRO	24.5	54.9	21.3	58.0	21.0	27.3	24.4	56.8	21.1	54.8	21.4	42.7	22.6	51.6	21.6
RUBi	43.1	73.7	39.8	78.5	39.3	44.9	43.1	74.2	39.7	74.8	39.8	48.1	42.7	71.7	40.0
LNL	41.7	55.9	40.8	59.1	40.5	48.4	41.7	56.3	40.8	56.5	40.8	57.1	40.6	53.7	41.0
IRMv1	37.5	64.4	34.6	67.1	34.3	39.5	37.4	68.0	34.2	65.8	34.5	66.2	34.3	62.5	34.8
$b_{expl.}$ = Texture Color															
Up Wt	39.5	71.6	36.1	71.4	36.2	43.7	39.3	70.3	36.3	67.6	36.7	69.4	36.2	42.4	39.4
GDRO	30.4	59.5	27.4	60.7	27.2	33.5	30.3	61.6	27.1	60.9	27.3	61.9	26.9	45.1	28.9
RUBi	46.8	74.0	43.8	76.6	43.5	48.9	46.6	73.8	43.8	74.0	43.9	75.7	43.5	49.2	46.6
LNL	34.3	49.2	33.4	47.9	33.5	41.8	34.2	49.9	33.3	47.8	33.6	48.6	33.4	43.3	34.0
IRMv1	38.5	66.1	35.7	71.0	35.2	43.7	38.3	69.2	35.4	67.6	35.7	68.2	35.4	63.6	36.0
Implicit Methods															
LFF	56.6	71.2	55.4	82.6	54.2	63.8	56.3	77.0	54.8	74.6	55.1	76.2	54.8	75.4	54.9
SD	41.3	69.5	38.3	71.2	38.1	46.3	40.9	72.2	38.0	72.1	38.1	71.1	38.0	70.5	38.2

Table A5: Method accuracies with increasing number of explicit bias variables.

Variables/ Methods	Unbiased Accuracy	Background Color		Digit Color		Digit Position		Distractor Shape		Distractor Color		Texture Pattern		Texture Color	
		Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.
StdM	42.0	70.3	39.1	74.6	38.7	48.3	41.6	69.2	39.2	68.9	39.4	69.4	39.1	65.2	39.7
$b_{expl.}$ = Distractor Shape															
Up Wt	21.5	52.0	18.2	56.0	17.8	24.2	21.4	25.6	21.2	53.0	18.2	54.0	17.9	51.9	18.2
GDRO	25.2	54.1	22.2	56.9	21.9	27.9	25.2	48.9	22.8	56.0	22.1	56.0	21.9	52.3	22.4
RUBi	34.5	66.9	31.0	70.9	30.5	35.2	34.5	40.4	33.9	67.7	31.0	67.5	30.7	64.6	31.1
LNL	44.1	55.8	43.1	59.6	42.6	50.5	43.7	57.9	42.8	57.4	42.9	58.8	42.7	55.0	43.1
IRMv1	38.3	67.0	35.5	70.5	35.1	41.5	38.3	69.3	35.2	65.9	35.7	67.9	35.2	64.6	35.7
$b_{expl.}$ = Distractor Shape+Texture Pattern															
Up Wt	16.1	49.0	12.6	51.6	12.3	18.3	16.0	20.6	15.8	50.7	12.5	18.5	16.0	45.3	12.9
GDRO	17.4	43.7	14.6	44.3	14.6	20.1	17.3	42.5	14.8	46.4	14.4	42.6	14.6	41.8	14.8
RUBi	27.9	64.3	24.0	70.9	23.3	32.1	27.6	36.5	27.1	65.9	24.0	34.1	27.4	60.6	24.3
LNL	10.0	14.7	10.5	11.0	10.9	11.3	10.9	12.3	10.8	11.8	10.9	12.9	10.7	11.7	10.9
IRMv1	34.3	64.4	31.2	68.5	30.8	38.5	34.2	63.9	31.3	63.7	31.4	61.3	31.4	62.3	31.4
$b_{expl.}$ = Distractor Shape+Texture Pattern+Digit Color															
Up Wt	12.6	47.7	8.7	20.1	11.8	13.5	12.5	22.9	11.5	45.8	9.0	22.7	11.4	42.3	9.2
GDRO	18.4	44.1	15.6	42.6	15.8	20.0	18.3	44.8	15.5	45.7	15.6	44.2	15.5	43.6	15.7
RUBi	11.0	48.9	6.8	12.9	10.9	11.8	11.0	11.1	11.1	46.8	7.2	11.0	11.1	41.0	7.6
LNL	10.2	15.2	10.6	11.5	11.1	11.7	11.0	11.4	11.1	12.2	11.0	18.4	10.2	15.3	10.6
IRMv1	24.9	57.0	21.4	53.8	21.8	27.6	24.8	52.4	21.9	56.0	21.7	50.7	22.0	55.3	21.6
$b_{expl.}$ = Distractor Shape+Texture Pattern+Digit Color+Background Color															
Up Wt	30.1	54.5	27.6	59.3	27.1	35.0	29.8	56.0	27.4	64.0	26.7	56.3	27.3	63.4	26.5
GDRO	19.9	48.2	17.0	45.3	17.4	22.8	19.9	50.9	16.7	47.4	17.2	46.7	17.1	43.8	17.5
RUBi	10.0	10.3	9.9	11.9	9.7	10.0	10.0	9.2	10.0	59.4	4.6	10.4	9.9	48.0	5.6
LNL	10.0	10.0	11.5	10.6	11.4	11.2	11.4	11.9	11.3	13.0	11.2	11.3	11.4	10.6	11.4
IRMv1	25.6	52.9	22.7	58.1	22.1	30.1	25.2	52.3	22.7	57.5	22.3	52.4	22.6	56.0	22.3
$b_{expl.}$ = Distractor Shape+Texture Pattern+Digit Color+Background Color+Texture Color															
Up Wt	38.8	65.3	36.0	71.4	35.3	42.3	38.5	68.2	35.6	71.0	35.4	67.1	35.6	61.5	36.3
GDRO	21.9	49.9	19.1	47.2	19.4	25.1	21.8	53.2	18.7	50.8	19.1	51.6	18.7	45.6	19.5
RUBi	9.8	10.2	9.7	9.6	9.8	9.8	9.7	8.8	9.8	88.2	1.2	9.7	9.7	9.3	9.8
LNL	10.0	10.0	11.5	10.6	11.4	11.2	11.4	11.9	11.3	13.0	11.2	11.3	11.4	10.6	11.4
IRMv1	28.4	58.7	25.0	61.5	24.7	32.2	28.0	56.6	25.2	59.3	25.1	57.8	25.0	53.5	25.6
$b_{expl.}$ = Distractor Shape+Texture Pattern+Digit Color+Background Color+Texture Color+Distractor Color															
Up Wt	43.2	71.0	40.4	75.3	39.9	47.3	43.1	72.1	40.2	69.7	40.6	72.8	40.0	68.5	40.6
GDRO	19.7	47.3	16.7	42.2	17.3	22.2	19.5	50.7	16.3	48.5	16.7	48.1	16.5	44.4	17.0
RUBi	9.9	9.0	10.1	9.8	10.0	11.4	9.8	10.2	10.0	12.3	9.7	10.9	9.9	9.2	10.1
LNL	10.4	22.4	9.6	12.1	10.8	13.5	10.6	11.3	10.9	11.5	10.8	30.8	8.6	24.8	9.3
IRMv1	24.8	52.6	21.8	58.3	21.2	26.0	24.8	57.4	21.3	53.2	21.9	53.3	21.6	49.5	22.1
$b_{expl.}$ = Distractor Shape+Texture Pattern+Digit Color+Background Color+Texture Color+Distractor Color +Digit Position															
Up Wt	44.0	70.5	41.4	73.9	41.0	48.8	43.9	71.8	41.2	71.8	41.4	71.0	41.2	68.2	41.6
GDRO	16.6	42.6	13.7	34.6	14.7	19.0	16.4	46.8	13.3	46.5	13.4	44.1	13.4	38.9	14.1
RUBi	10.0	11.9	10.6	10.2	10.8	10.7	10.7	10.2	10.8	12.6	10.5	11.2	10.7	9.7	10.8
LNL	10.0	10.1	11.5	10.5	11.4	11.1	11.4	12.2	11.2	12.8	11.2	11.4	11.3	10.7	11.4
IRMv1	21.8	47.7	19.0	51.6	18.6	24.6	21.6	49.8	18.7	52.2	18.6	49.3	18.7	46.5	19.1
Implicit Methods															
LFF	56.6	71.2	55.4	82.6	54.2	63.8	56.3	77.0	54.8	74.6	55.1	76.2	54.8	75.4	54.9
SD	41.3	69.5	38.3	71.2	38.1	46.3	40.9	72.2	38.0	72.1	38.1	71.1	38.0	70.5	38.2



(a) Explicit Bias = Head vs Tail (# groups = 2)

(b) Explicit Bias = Global Group (# groups = 115)

(c) Explicit Bias = Answer Class (# groups = 1833)

(d) Explicit Bias = Local Group (# groups = 133,328)

Figure A9: Tail accuracies on GQA-OOD, when considering 4 different explicit biases.

Table A6: Hyperparameters for all datasets for all the comparison methods selected using unbiased accuracies of the validation sets after performing grid search over hyperparameters.

Methods/Datasets	**Parameters**	**CelebA**	**Biased MNIST**	**GQA-OOD**
Common to all methods unless specified otherwise	Optimizer	SGD	Adam	Adam
	Batch Size	128	128	128
	Epochs	50	30	30
Standard Model (StdM)	Learning Rate	10^{-3}	10^{-3}	10^{-4}
	Weight Decay	0	10^{-5}	0
Group Upweighting (Up Wt) [56]	Learning Rate	10^{-5}	10^{-3}	10^{-3}
	Weight Decay	0.1	10^{-5}	0
Group DRO (GDRO) [55]	Learning Rate	10^{-5}	10^{-3}	10^{-4}
	Weight Decay	0.1	10^{-5}	0
	Step Size	0.01	10^{-3}	0.01
Reduction of Unimodal Biases (RUBi) [12]	Learning Rate	10^{-4}	10^{-3}	10^{-4}
	Weight Decay	10^{-5}	10^{-5}	0
Adversarial Regularization (LNL) [37]	Learning Rate	10^{-4}	10^{-3}	10^{-3}
	Weight Decay	10^{-4}	10^{-5}	0
	Gradient Reversal Weight ($\lambda_{grad.}$)	-0.1	-0.1	-0.1
	Entropy Loss Weight ($\lambda_{ent.}$)	0	0.01	0.01
Invariant Risk Minimization (IRMv1) [5]	Learning Rate	10^{-4}	10^{-3}	10^{-4}
	Weight Decay	0	10^{-5}	0
	Gradient Regularization Weight ($\lambda_{grad.}$)	1.0	0.01	0.01
	Number of environments per mini-batch	4	16	16
Learning From Failure (LFF) [46]	Optimizer	Adam	Adam	Adam
	Learning Rate	10^{-4}	10^{-3}	10^{-4}
	Weight Decay	0	10^{-5}	0
	Amplification Factor (γ)	0.1	0.5	0.7
Spectral Decoupling (SD) [50]	Learning Rate	10^{-4}	10^{-3}	10^{-4}
	Weight Decay	10^{-5}	10^{-5}	0
	Output Decay (λ)	λ_0 = 10 λ_1 = 10	10^{-3}	10^{-3}
	Output Shift (γ)	γ_0 = 0.44 γ_1 = 2.5	10^{-3}	10^{-3}

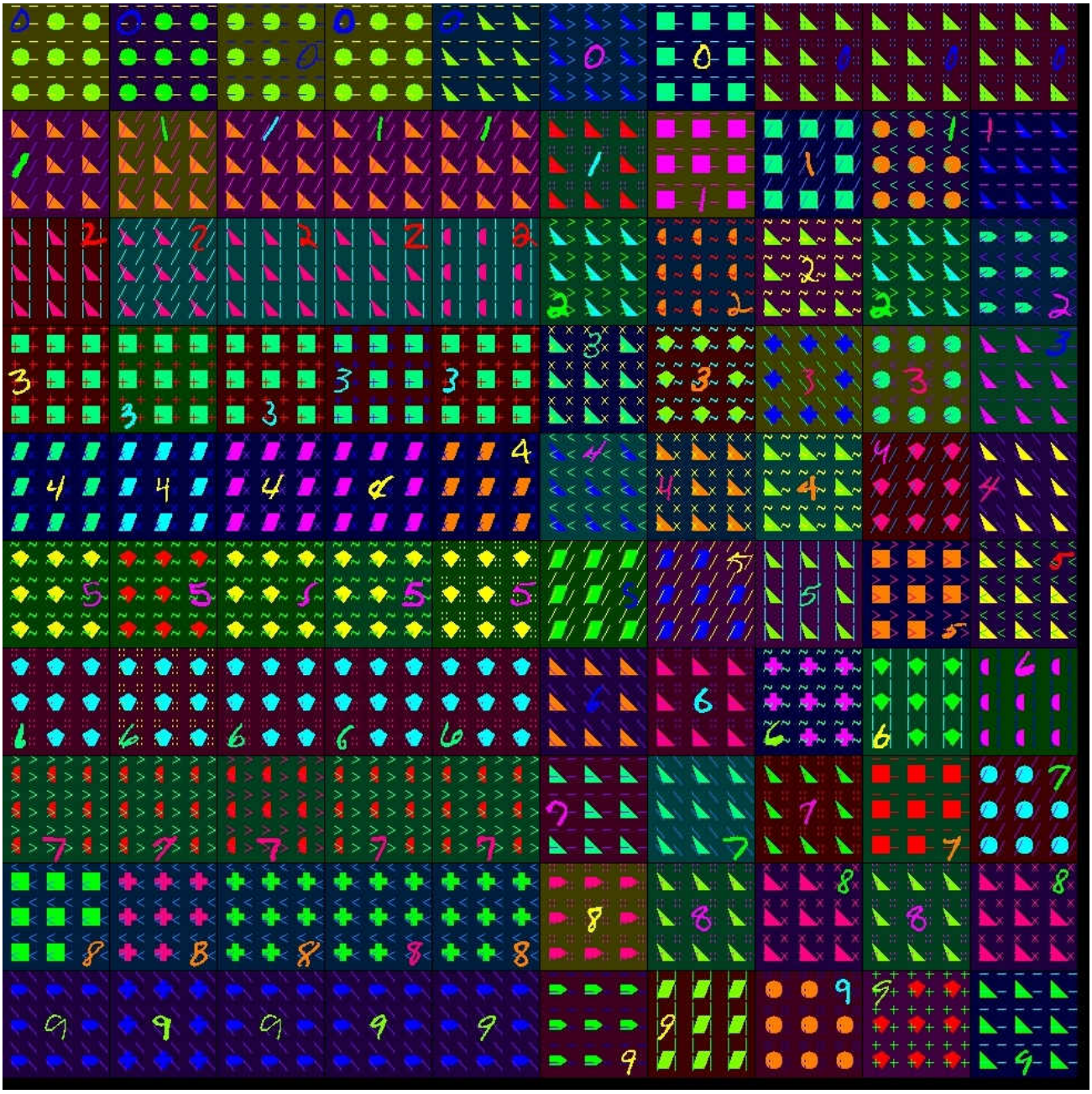


Figure A10: Biased MNIST requires the methods to classify the target digit while remaining invariant to biases.