SEBRA : Debiasing through Self-Guided Bias Ranking

ICLR
International Conference on Learning Representations

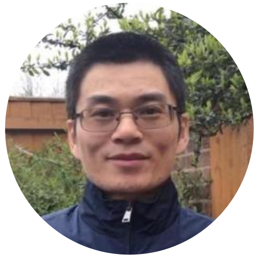

Adarsh Kappiyath [1]


Abhra Chaudhuri [2]


Ajay Kumar Jaiswal [3]


Ziquan Liu [4]


Yunpeng Li [1]


Xiatian Zhu [1]


Lu Yin [1]

[1] UNIVERSITY OF SURREY
[2] FUJITSU
[3] TEXAS The University of Texas at Austin
[4] Queen Mary University of London

Preliminaries

Empirical Risk Minimization (ERM) with CE Loss exhibit tendency to learn different attributes asynchronously during training.

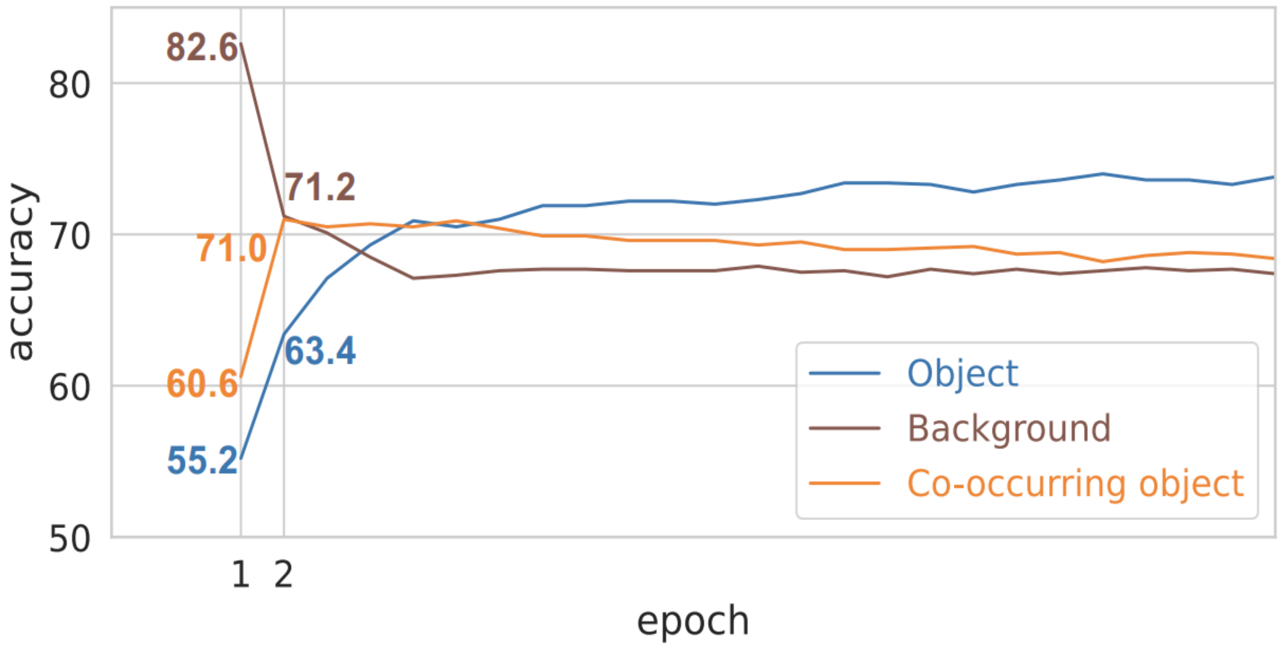


Fig 4 . Training Dynamics of Resnet50 with CE Loss on UrbanCars dataset.

Fig 4 taken from 'Whac-A-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others, Li et. CVPR 2023

Preliminaries

Empirical Risk Minimization (ERM) with CE Loss exhibit tendency to learn different attributes asynchronously during training.

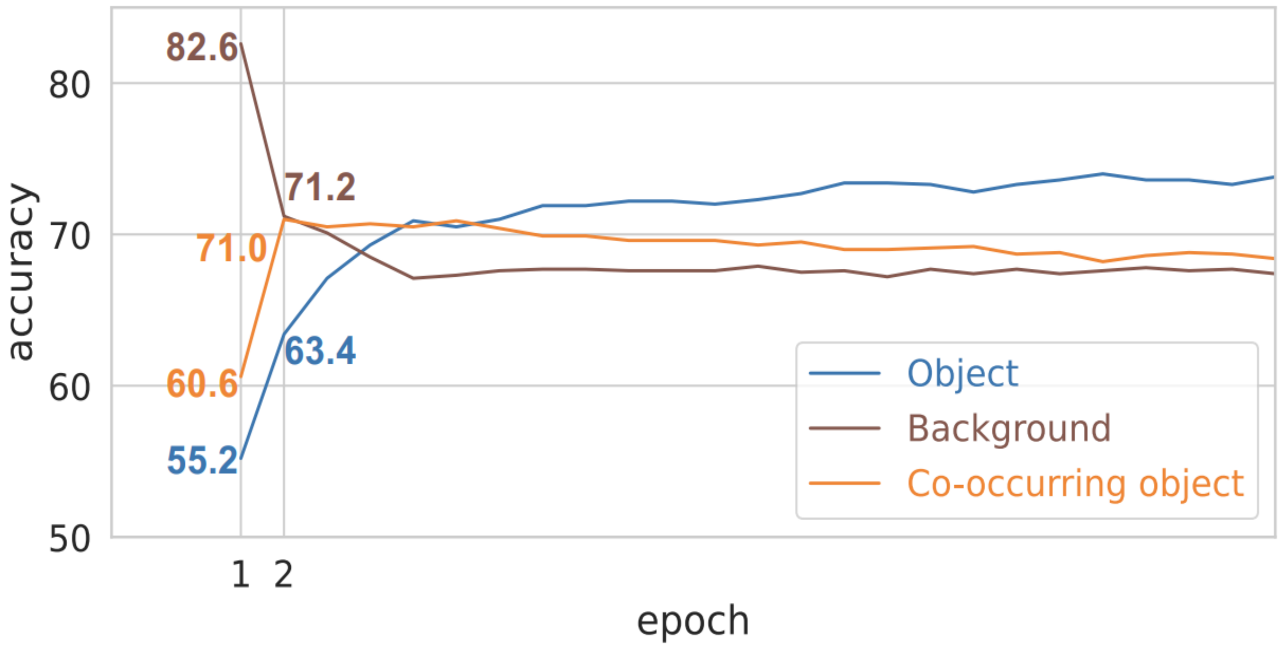


Fig 4 . Training Dynamics of Resnet50 with CE Loss on UrbanCars dataset.

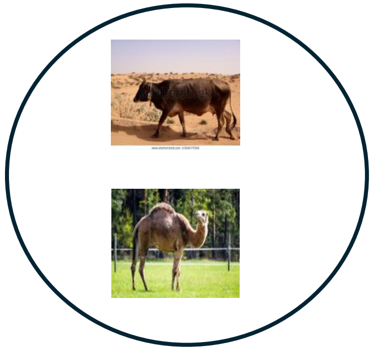
***Idea**: Modulate ERM dynamics to **<u>rank/order</u>** samples according to **<u>spuriosity</u>**. Mitigate biases based on ranking.*

Fig 4 taken from 'Whac-A-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others, Li et. CVPR 2023

Prior Works

Spurious Correlation Identification → Mitigating Impact of Spurious Correlations

- Identification of Spurious Correlations.
  - GCE Loss, Training with limited capacity models etc.

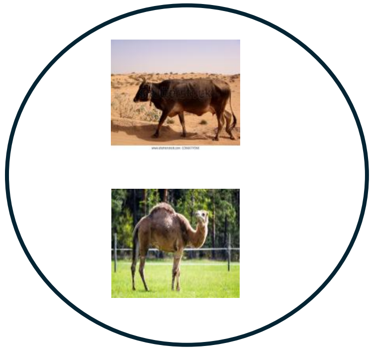


- Spuriosity of samples within a cluster
- Relative spuriosity across clusters

Prior Works

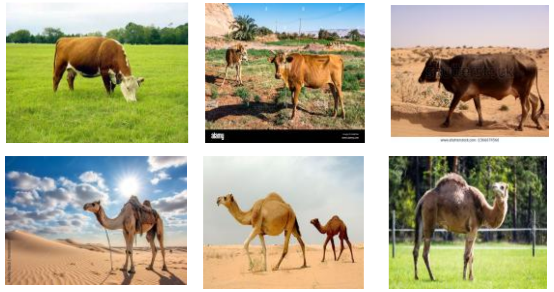
Spurious Correlation Identification → Mitigating Impact of Spurious Correlations

- Identification of Spurious Correlations.
  - GCE Loss, Training with limited capacity models etc.

- Spuriosity of samples within a cluster ✗
- Relative spuriosity across clusters ✗

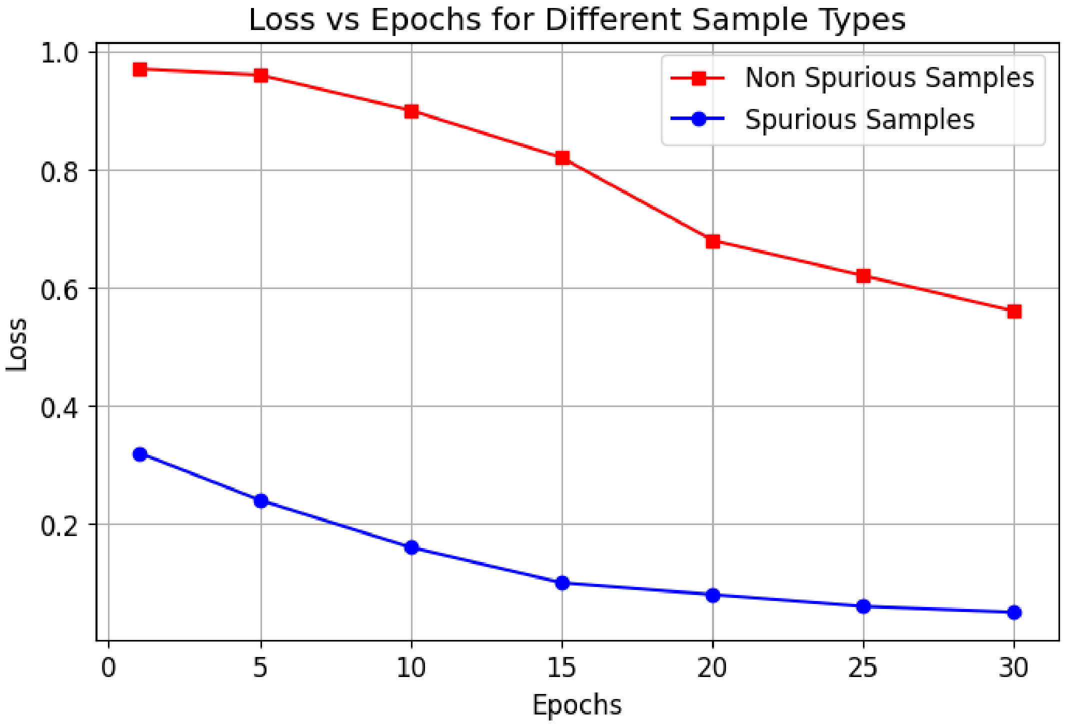
  - Class wise ranking of instances based on the *strength of spurious correlations*(*spuriosity*).

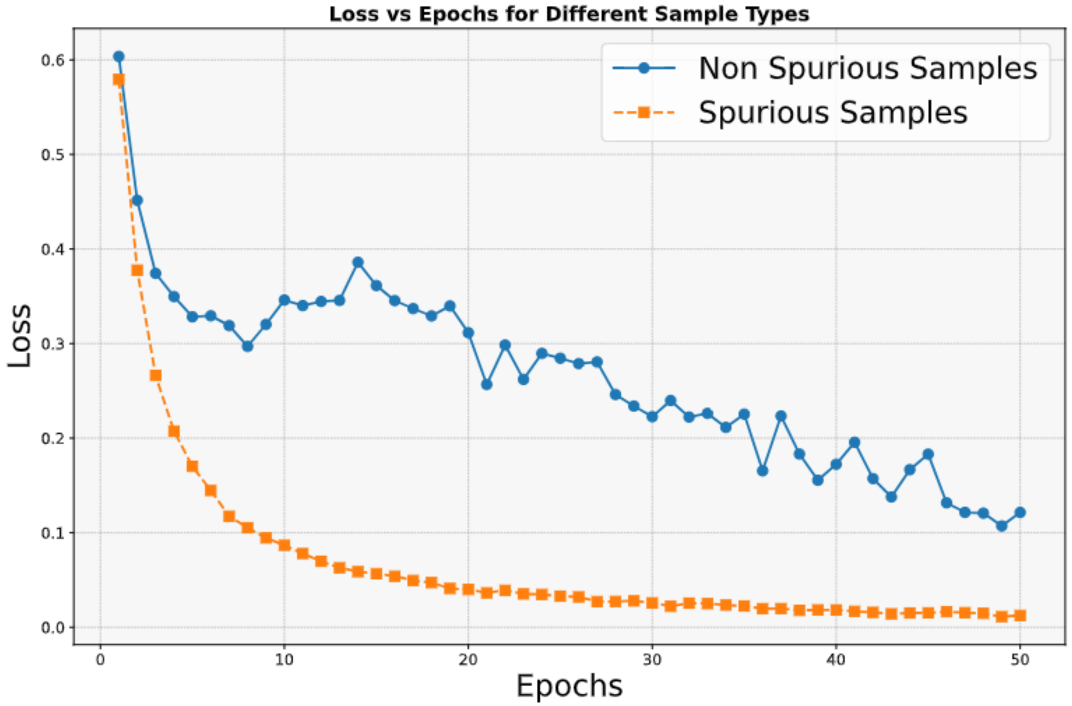


Human Supervision ☹

Assumption : Hardness Spuriosity Symmetry

The hardness of learning a sample, and its corresponding spuriosity measure, are symmetric to each other – the harder it is to learn a sample, the lower its spuriosity measure, and vice versa.





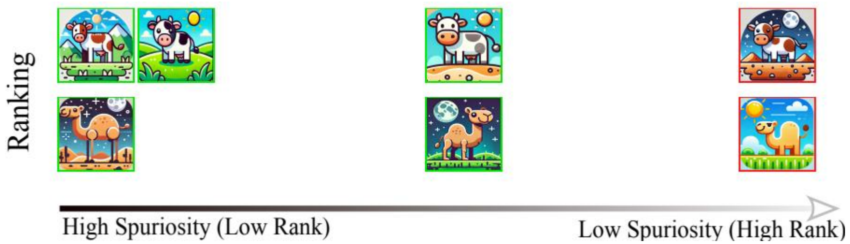
Deviation of ERM in the Multi-Bias Setting

Global trends of ERM deviate due to:

- **Reliance on spurious features**
- **Non-uniform gradient updates**.

Steering ERM in the Multi-Bias Setting

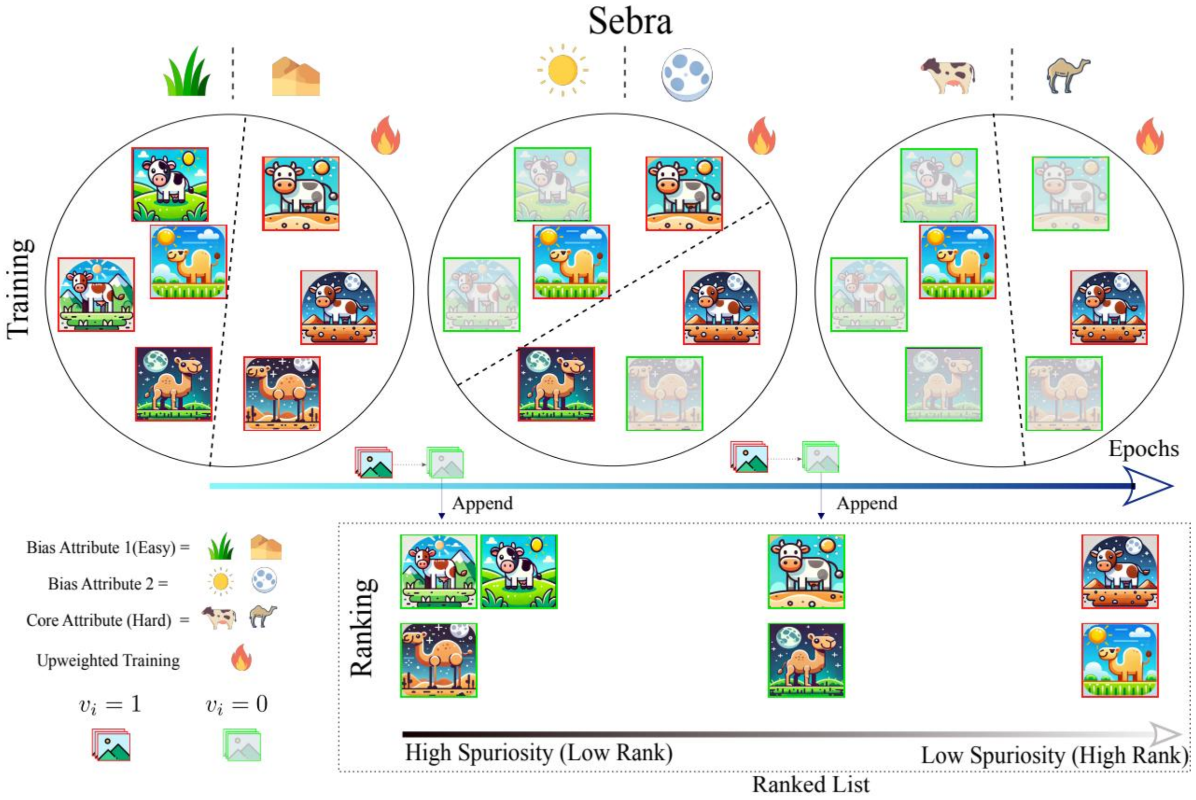
Correcting deviation requires explicit steering towards to maintain the Hardness-Spuriosity Symmetry.



- One bias at a time

Steering ERM in the Multi-Bias Setting

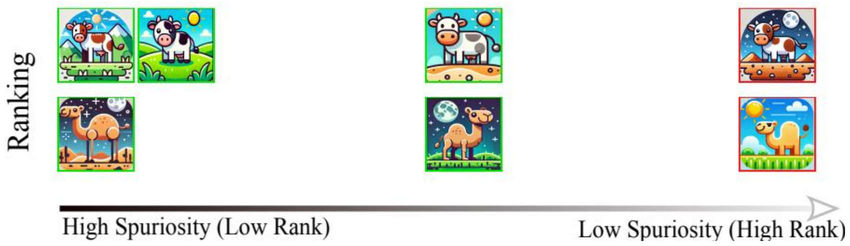
Correcting deviation requires explicit steering towards to maintain the Hardness-Spuriosity Symmetry.



- Spuriosity-Based Sequential Learning
- One bias at a time

Sequential Learning Based on Levels of Spuriosity

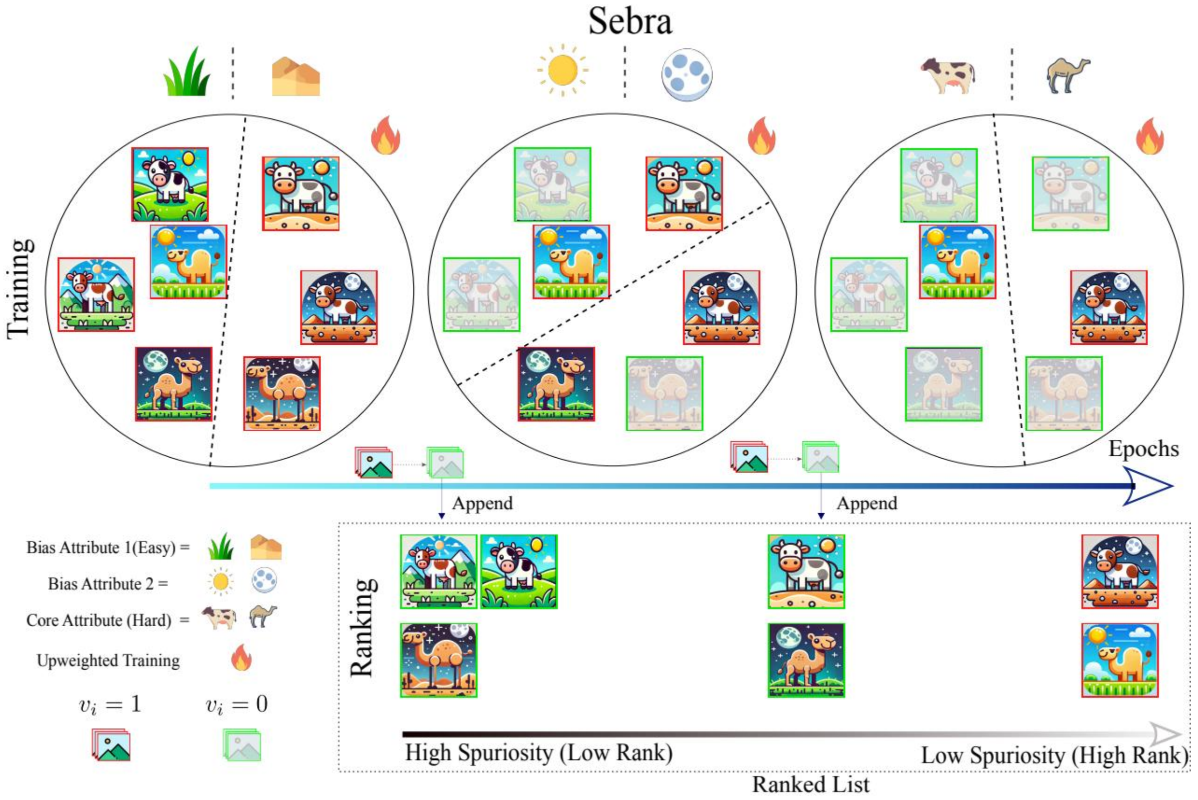
Selection variable ensures isolation among subgroups.



$$\max_{v} \sum_{i=1}^{N} \{ v_i^t \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) - \lambda v_i^t \}$$

Sequential Learning Based on Levels of Spuriosity

Reweight CE-Loss by some measure of spuriosity. Vulnerable to shortcut $u_i = 0$, for all u_is.

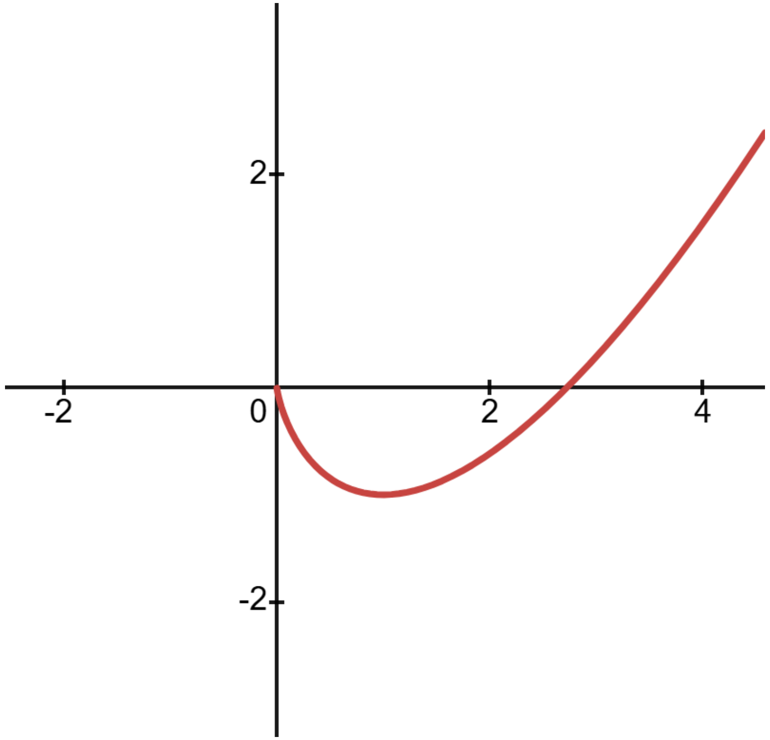


$$\min_{\theta} \min_{u} \sum_{i=1}^{N} \{\underline{u_i} \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i)\}$$

$$\max_{v} \sum_{i=1}^{N} \{v_i^t \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) - \lambda v_i^t\}$$

Hardness-Spuriosity Conservation Law

Remedy for collapse: u_is must belong to a specific manifold satisfying a certain conservation law.

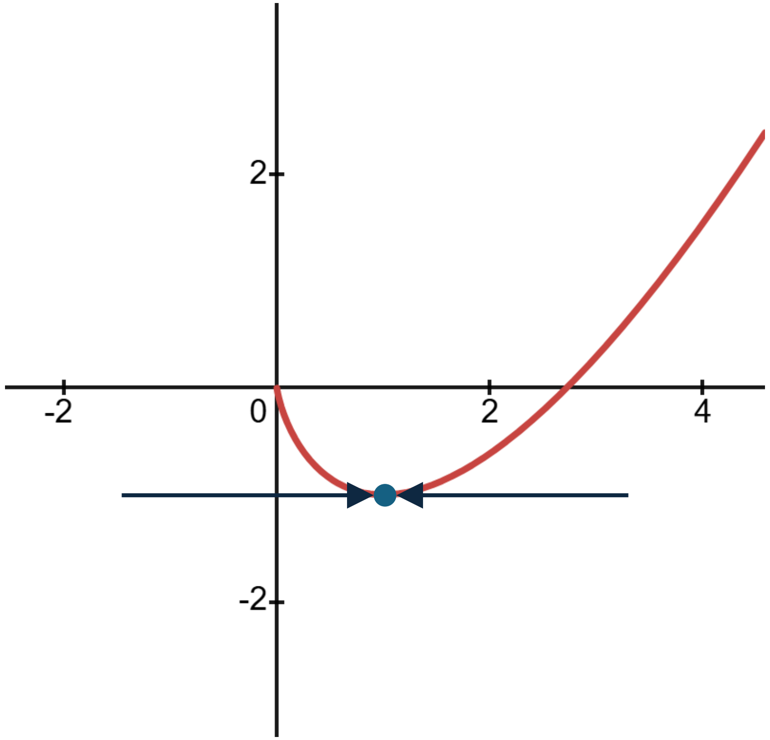


$$\min_{\theta} \min_{u} \sum_{i=1}^{N} \{\underline{u_i} \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i)\}$$

$$\boxed{u_i \mathcal{L}_{CE}(f_\theta(x_i), y_i) + \beta(u_i \ln u_i - u_i) = c}$$

Hardness-Spuriosity Conservation Law

Remedy for collapse: u_is must belong to a specific manifold satisfying a certain conservation law. Ensure adherence to manifold through β-weighted regularization.



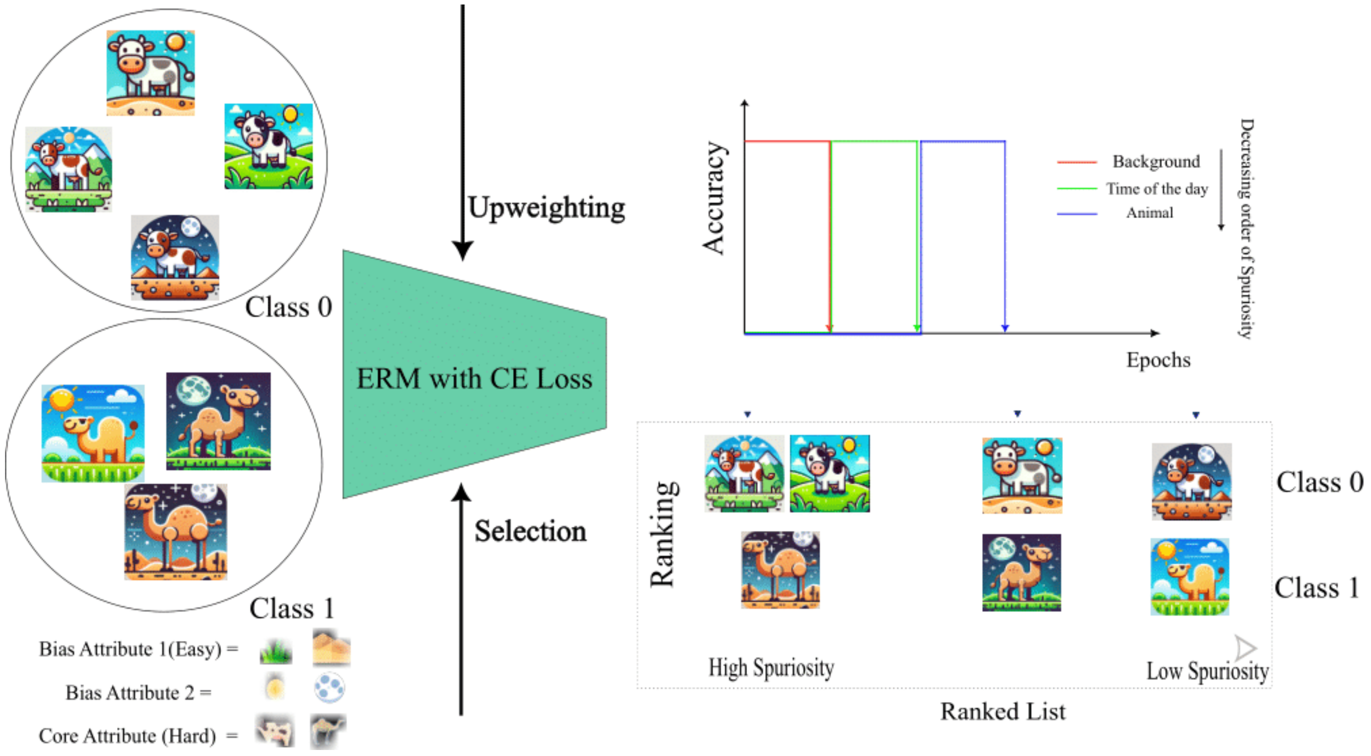
$$\boxed{u_i \mathcal{L}_{CE}(f_\theta(x_i), y_i) + \beta(u_i \ln u_i - u_i) = c}$$

$$\min_\theta \min_u \sum_{i=1}^{N} \{u_i \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i)\}$$

$$g(u_i) \;=\; (u_i \ln u_i - u_i)$$

$$\min_{\theta,u} \sum_{i=1}^{N} \{u_i \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) + \beta g(u_i)\}$$

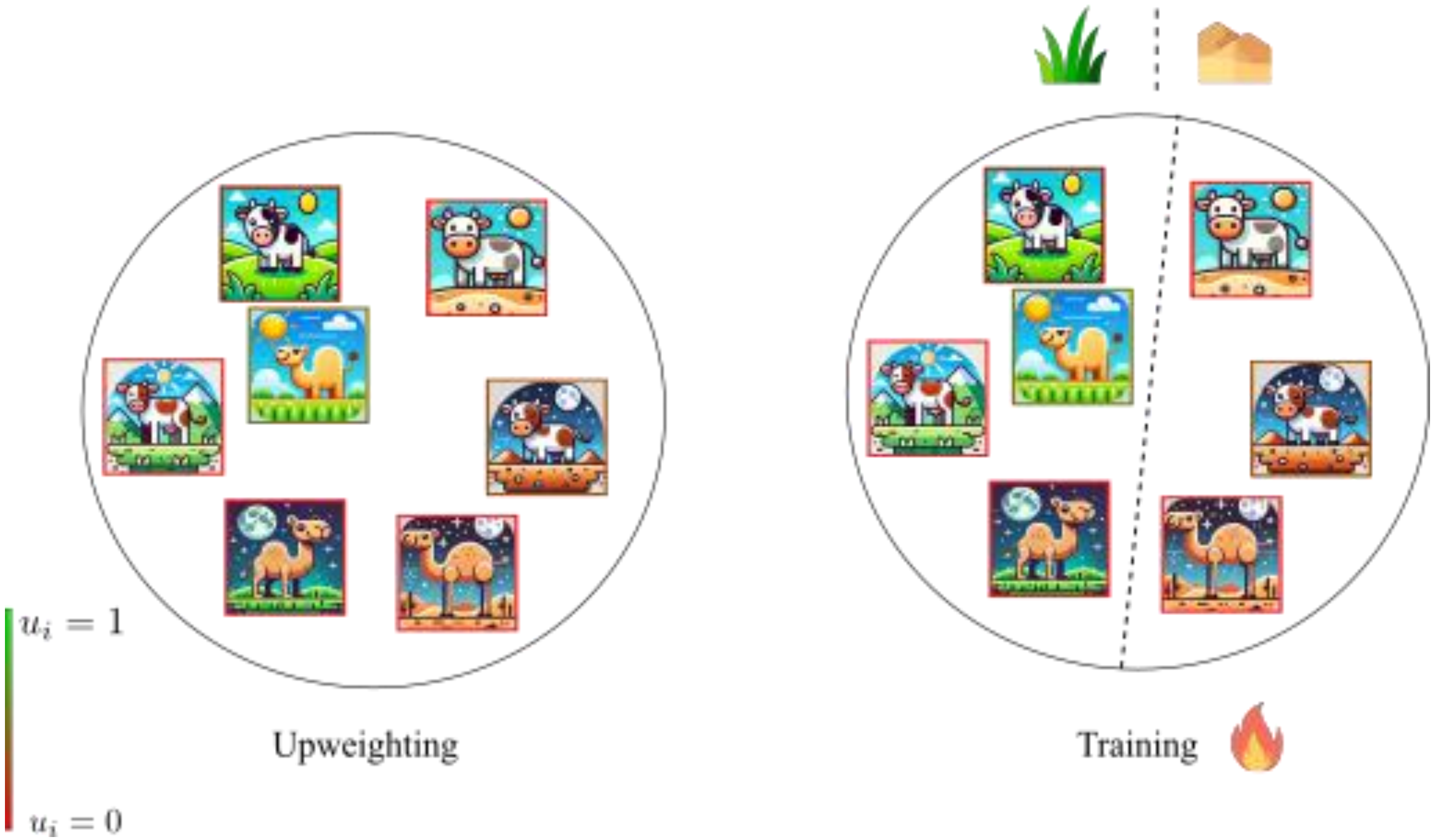
Sebra : Self-Guided Bias Ranking



$$\mathcal{L}_{\text{ranking}}(\theta, u, v) = \sum_{i=1}^{N} v_i^{t-1} \left\{ v_i^t u_i \mathcal{L}_{\text{CE}}(f_\theta(x_i), y_i) - \lambda v_i^t - \beta u_i + \beta u_i \ln u_i \right\}$$

$$\min_{\theta, u} \max_{v} \mathcal{L}_{\text{ranking}}(\theta, u, v),$$

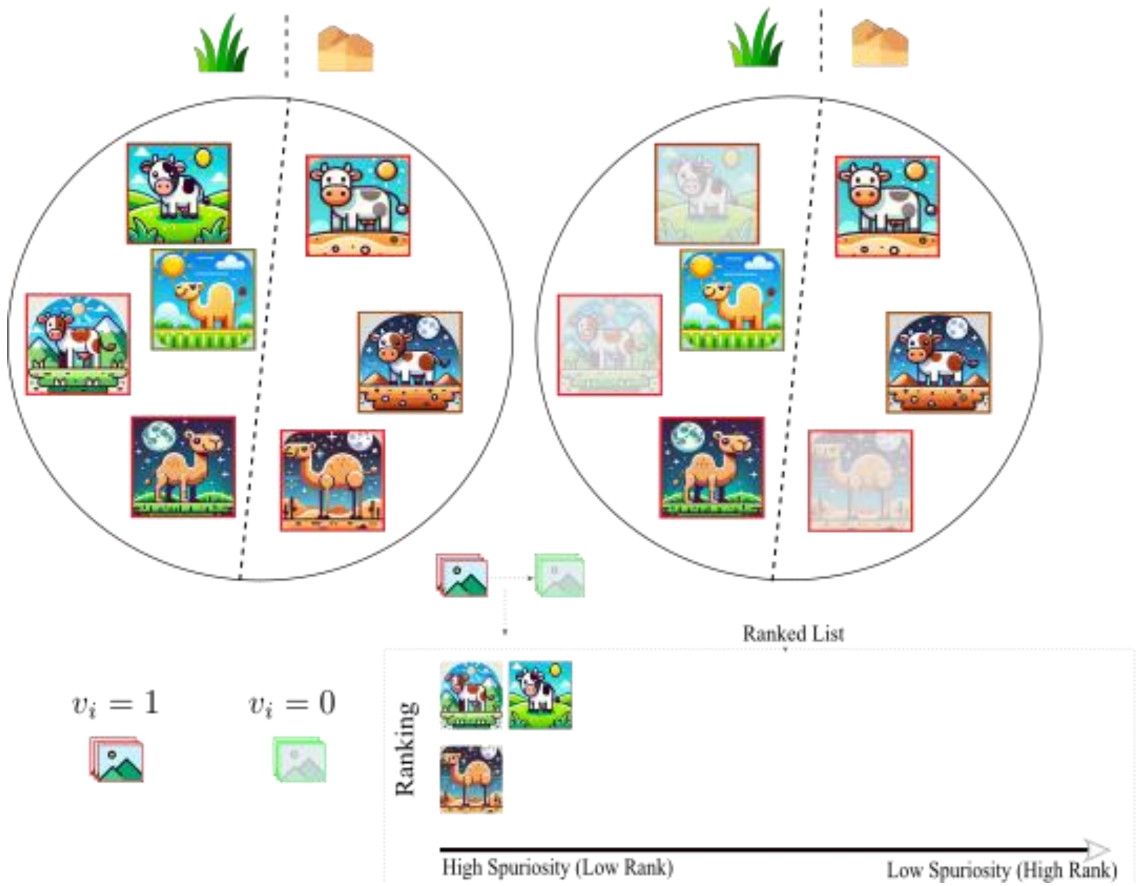
Sebra : Upweighted Training



$$\min_{\theta,u} \sum_{i=1}^{N} \{\underline{u_i}\mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) + \beta g(\underline{u_i})\}$$

$$\boxed{u_i^* = p_y^{\frac{1}{\beta}}}$$

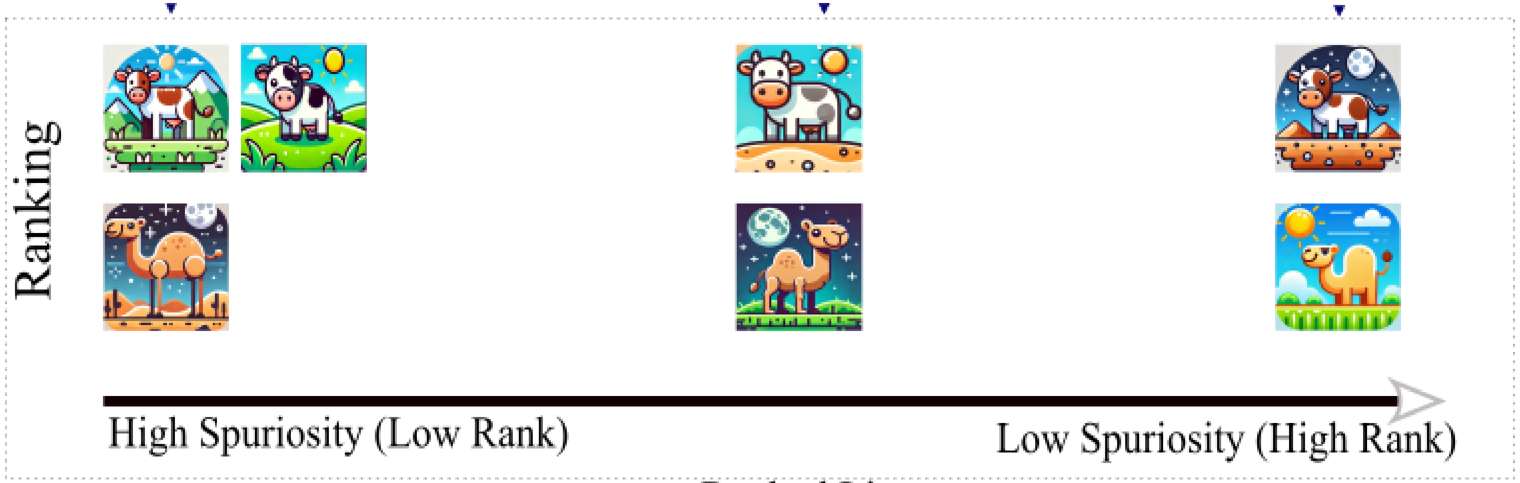
Sebra : Selection & Ranking



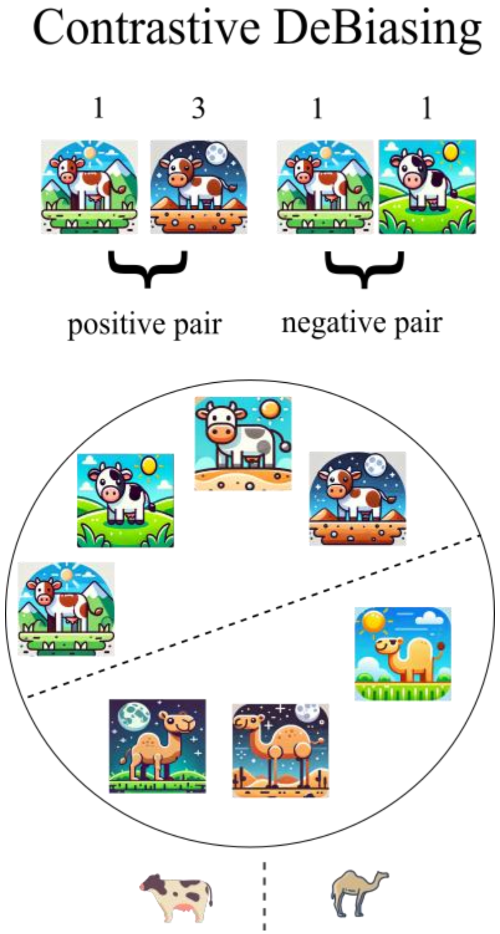
$$\max_{v} \sum_{i=1}^{N} \{v_i^t \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) - \lambda v_i^t\}$$

$$\boxed{v_i^{t*} = \begin{cases} 0, & \text{if} p_y > p_{critical}, \\ 1, & \text{otherwise.} \end{cases}}$$

Sebra : Debiasing



Ranked List

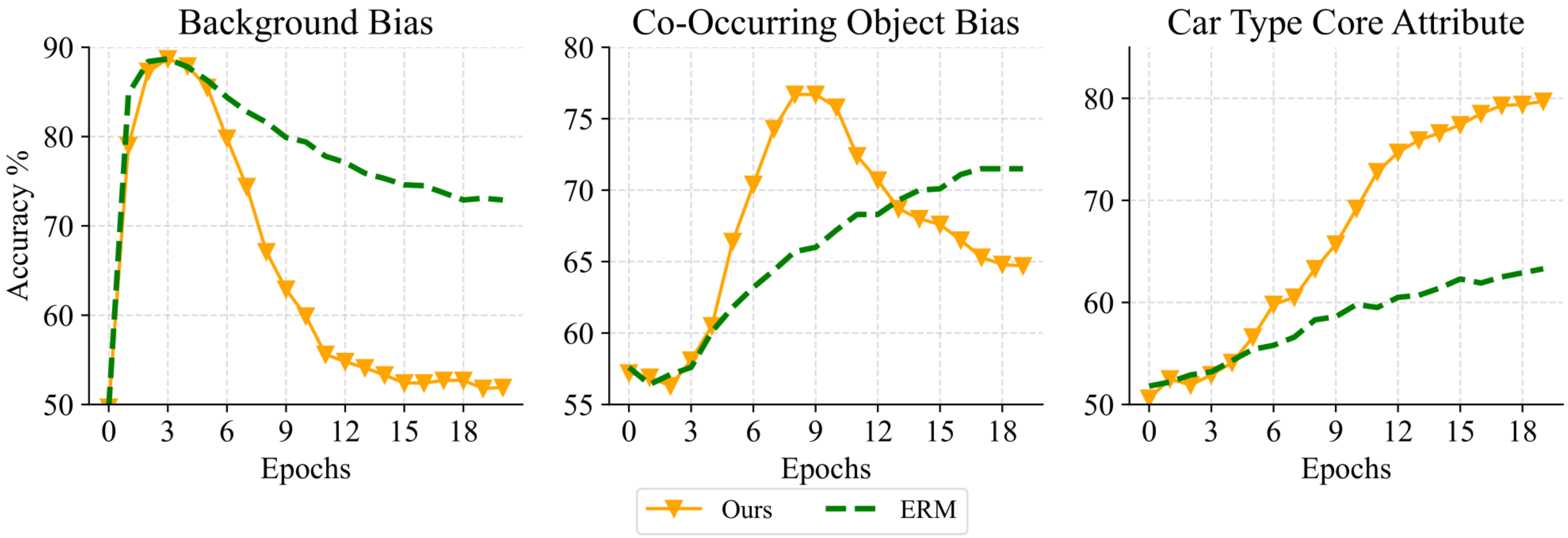


$$\boxed{\mathcal{L}_{\text{con}}^{\text{sup}}(x; f_{\text{enc}}) = \mathbb{E}\left[-\log \frac{\exp(z^\top z_m^+/\tau)}{\sum_{m=1}^{M}\exp(z^\top z_m^+/\tau) + \sum_{n=1}^{N}\exp(z^\top z_n^-/\tau)}\right]},$$

where τ is the temperature coefficent, z_m^+, z_n^- and $z^\top$ are the embeddings of positive, negative, and reference samples respectively.

Modulation of ERM Dynamics

Sebra successfully mitigates the Whac-a-Mole Dilemma.



Sebra : Ranking Results

Diving

Top Ranked

Bottom Ranked

Pole Vaulting

Top Ranked

Bottom Ranked

Table 1: Quantitative comparison of Sebra with various baselines. The results are shown in terms of Kendall's τ for Urban Cars and CelebA, and Performance Disparity (PD) for BAR.

Method	**Urban Cars**	**CelebA**	**BAR**
Metric	*Kendall's* τ ($\uparrow$)	*Kendall's* τ ($\uparrow$)	*PD* ($\uparrow$)
Random Ordering	0.02	-0.01	0.25
ERM-based Ranking	0.12	0.14	4.55
Spuriosity Ranking	0.40	0.38	28.88
Sebra (Ours)	**0.85**	**0.69**	**32.32**

Sebra : Debiasing Results

Methods	Sup.	UrbanCars			CelebA			BAR
		I.D. Acc. (↑)	WG Acc. (↑)	Avg GAP (↑)	I.D. Acc. (↑)	WG Acc. (↑)	Avg GAP (↑)	Test Acc. (↑)
Group DRO	✓	91.60 (1.23)	75.70 (1.79)	-10.30 (1.35)	90.08 (0.70)	37.9 (1.6)	-5.79 (1.63)	-
ERM	✗	97.60 (0.86)	33.20 (0.86)	-31.90 (3.92)	96.43 (0.13)	36.0 (1.7)	-22.83 (0.84)	68.00 (0.43)
LfF	✗	97.20 (2.40)	35.60 (2.40)	-31.06 (3.56)	95.12 (0.35)	35.5 (2.0)	-22.57 (1.26)	68.30 (0.97)
JTT	✗	95.80 (1.45)	33.30 (6.90)	-20.50 (2.61)	91.86 (1.48)	38.7 (2.4)	-26.81 (2.53)	68.14 (0.28)
Debian	✗	98.00 (0.89)	30.10 (0.89)	-31.40 (1.44)	96.28 (0.37)	41.1 (4.3)	-22.56 (0.54)	69.88 (2.92)
DFR	✗	89.70 (1.21)	-	-20.93 (2.61)	60.12 (1.28)	-	-19.16 (3.27)	69.22 (1.25)
Sebra (Ours)	✗	92.54 (2.10)	**73.8 (3.28)**	**-10.57 (1.72)**	88.61 (3.36)	**65.3 (4.1)**	**-9.82 (3.06)**	**75.36 (2.23)**

Method	Sup.	ImageNet-1K					MultiNLI
		I.D. Acc. (↑)	IN-W Gap (↑)	IN-9 Gap (↑)	IN-R Gap (↑)	Carton Gap (↑)	WG. Acc (↑)
LLE	✓	76.25	-6.18	-3.82	-54.89	+10	-
ERM	✗	76.13	-26.64	-5.53	-55.96	+40	66.8
LfF	✗	70.26	-17.57	-8.10	-56.54	+40	63.6
JTT	✗	75.64	-15.74	-6.75	-55.70	+32	69.1
Debian	✗	74.05	-20.00	-7.29	-56.70	+30	-
Sebra (Ours)	✗	74.89	**-14.77**	**-3.15**	**-54.81**	**+25**	**72.3**

Thank You

https://kadarsh22.github.io/sebra_iclr25/