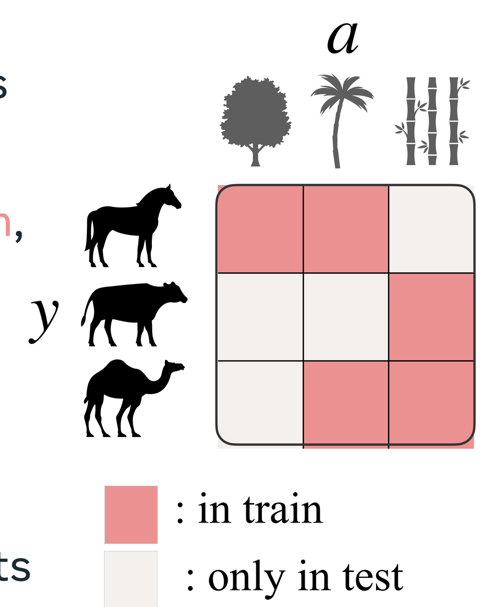
Compositional Risk Minimization

Divyat Mahajan, Mohammad Pezeshki, Charles Arnal, Ioannis Mitliagkas, Kartik Ahuja†, Pascal Vincent†

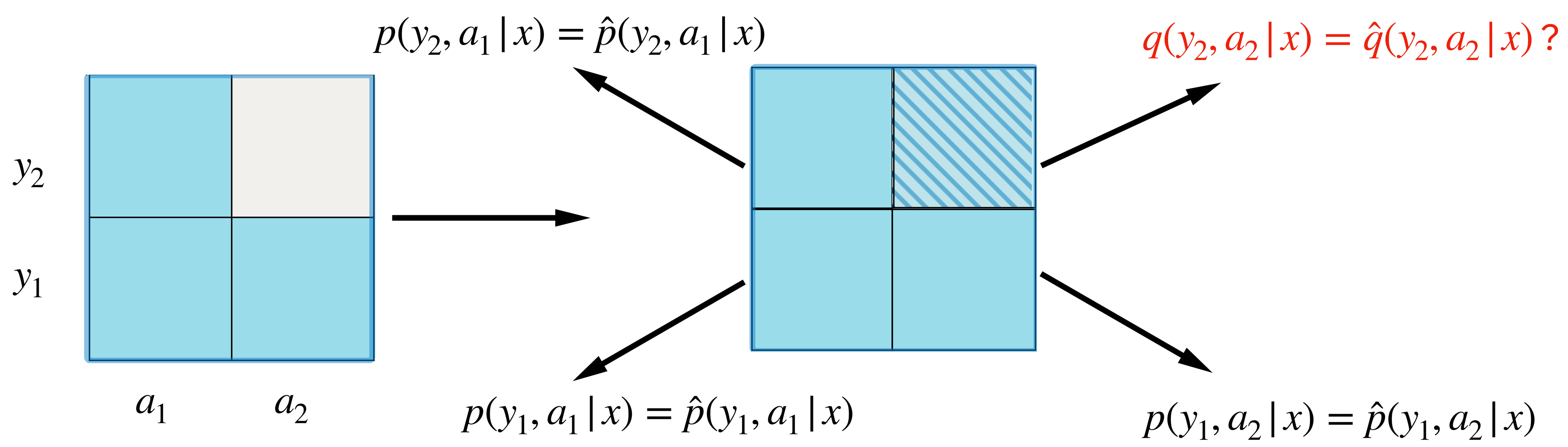
†:Joint Last Author

Contributions

- **Compositional Shifts:** Generalizing to novel combinations of attributes not in the training distribution
- **Theory of Compositional Shifts**. Assuming additive energy distribution, CRM provably generalizes compositionally to the discrete affine hull of training attributes.
- **A Practical Method.** Compositional Risk Minimization (CRM),
  - First train an additive energy classifier
  - Then extrapolates the learned bias for tackling compositional shifts



Cartesian Product Extrapolation



Additive Energy Distribution

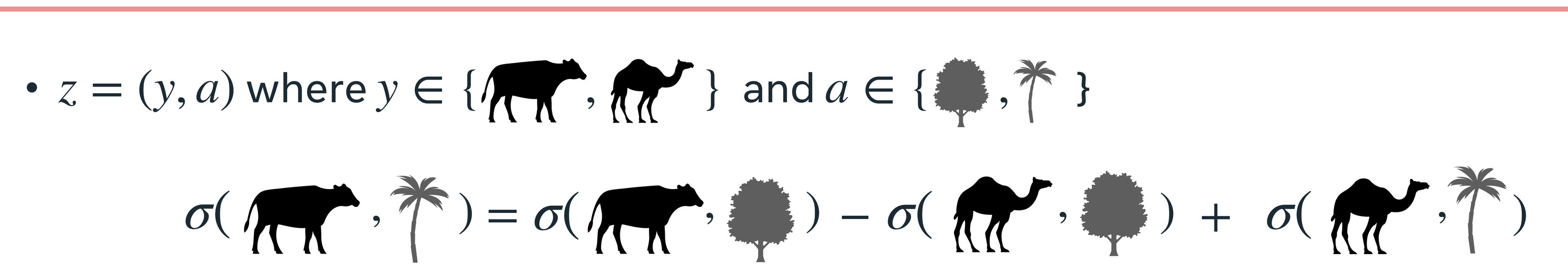
$$p(x|z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-1^T E(x,z) \Big) \quad \text{where} \quad 1^T E(x,z) = \sum_{i=1}^{m} E_i(x, z_i)$$

$$p(x|z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\langle \sigma(z), E(x) \rangle \Big)$$

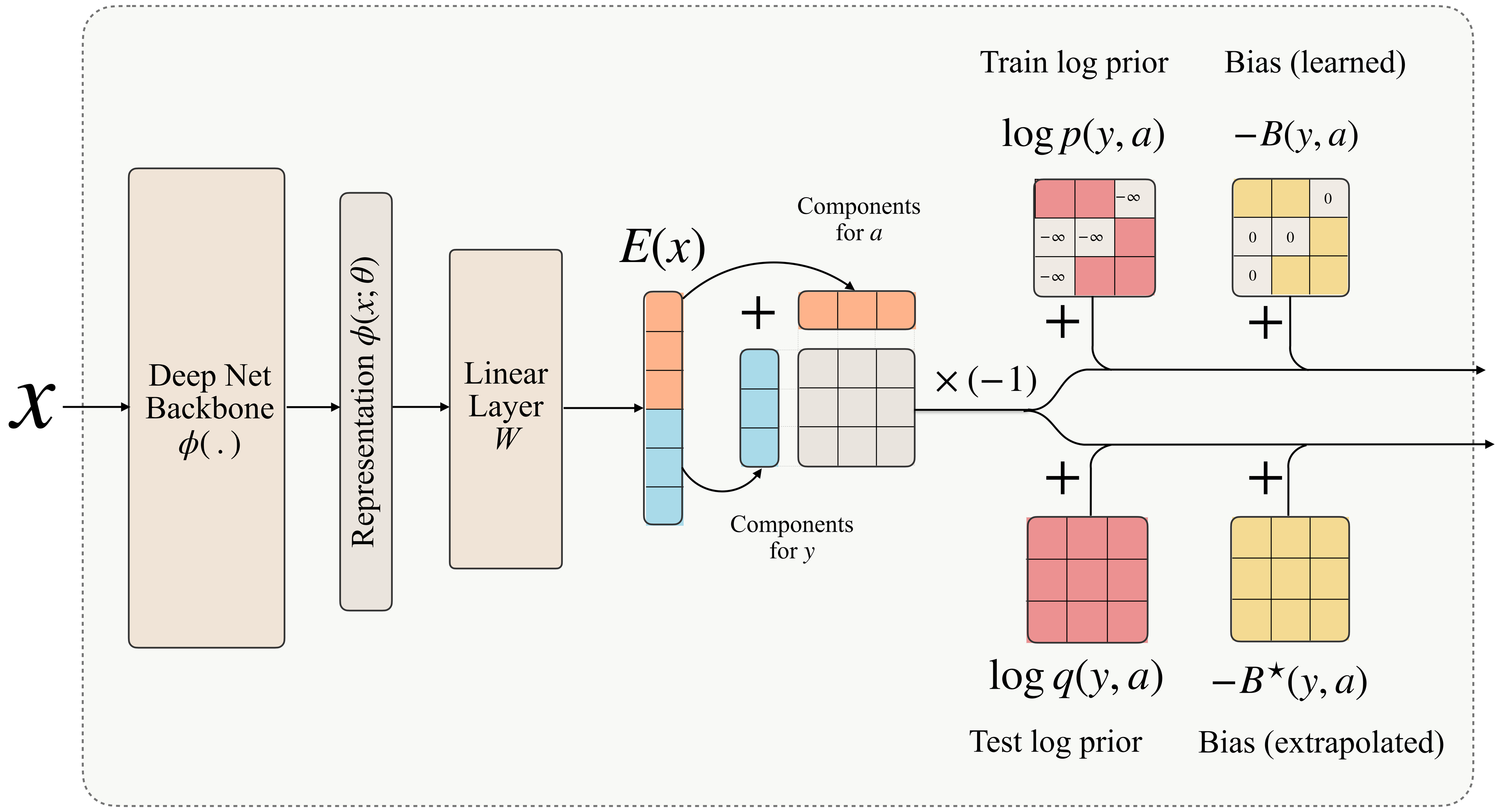
$$\sigma(z) = [\text{onehot}(z_1), \ldots, \text{onehot}(z_m)]^\top$$

$$E(x) = [E_1(x,1), \ldots, E_1(x,d), \ldots, E_m(x,1), \ldots, E_m(x,d)]^\top$$

Discrete Affine Hull



Additive Energy Classifier



True Model:

$$p(z|x) = Softmax(\log p(x|z) + \log p(z)) \ \text{ where } \ p(x|z) = \frac{1}{\mathbb{Z}(z)} \exp\Big(-\sigma(z)^T E(x)\Big)$$

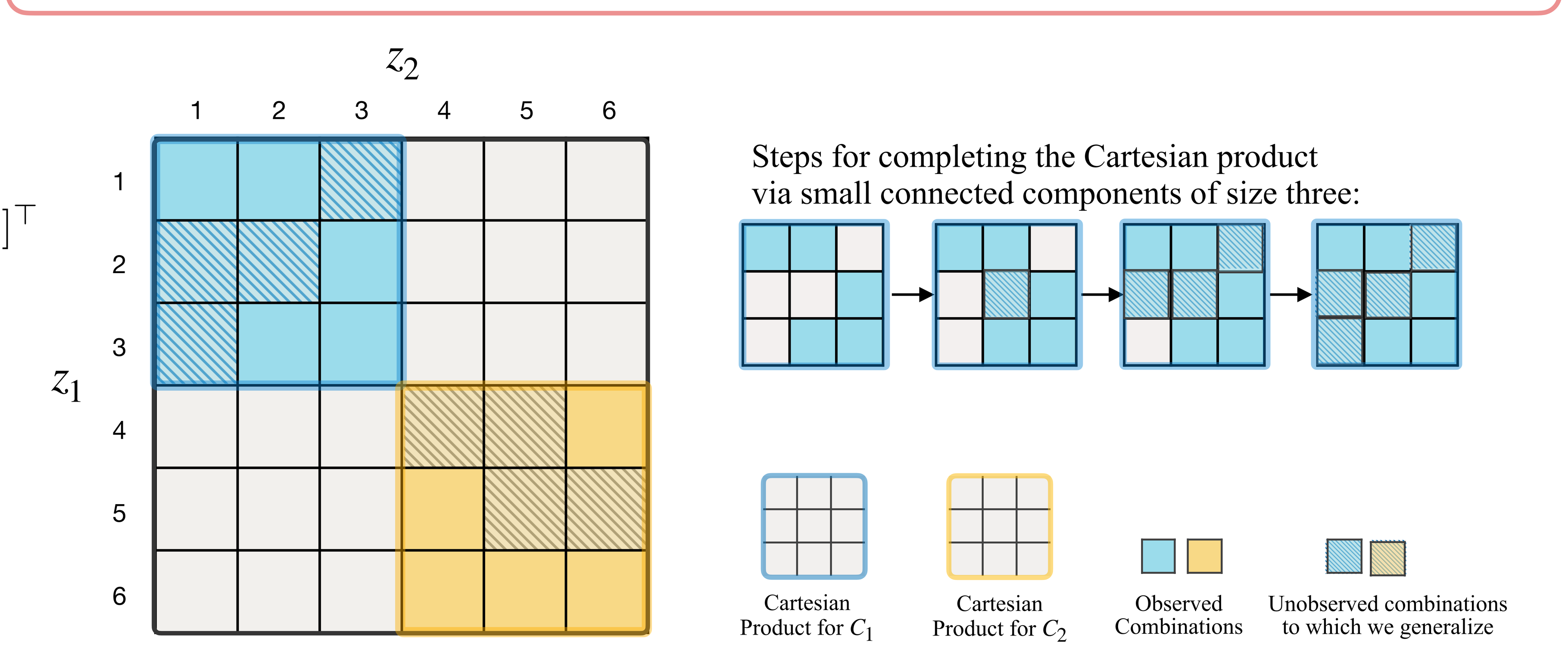
Learned Model (Train):

$$\hat{p}(z|x) = Softmax(\log \hat{p}(x|z) + \log p(z)) \ \text{ where } \ \hat{p}(x|z) = \frac{1}{\hat{B}(z)} \exp\Big(-\sigma(z)^T \hat{E}(x)\Big)$$

Learned Model (Eval):

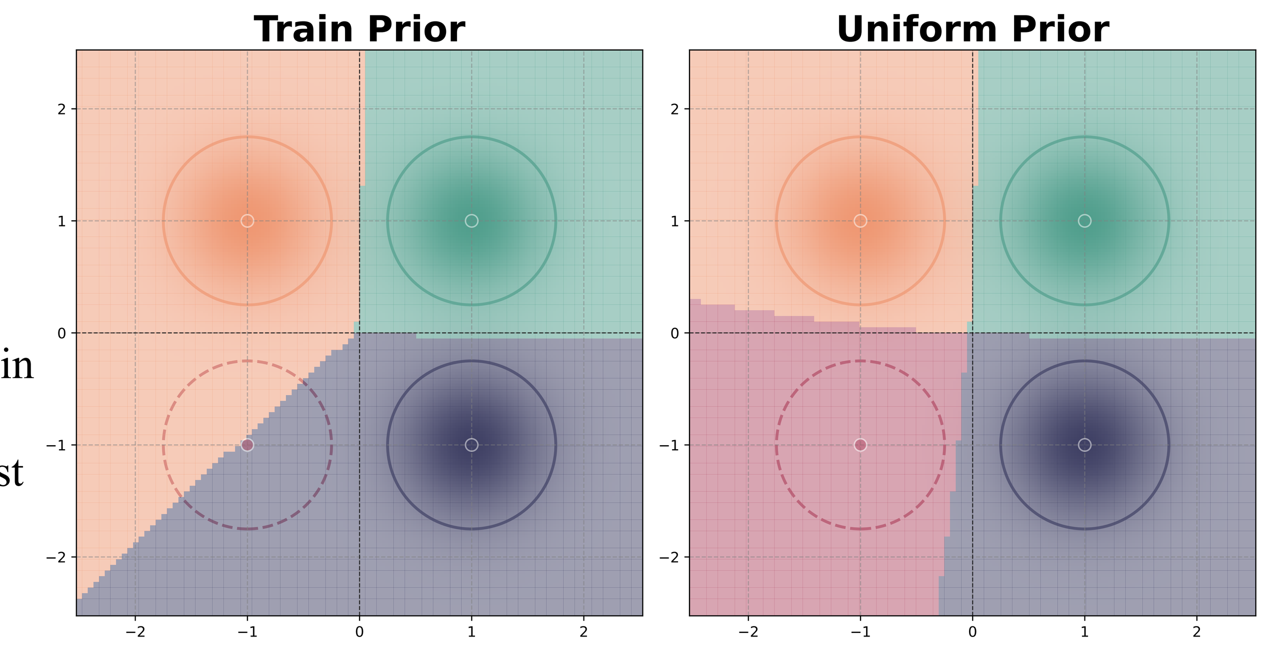
$$\hat{q}(z|x) = Softmax(\log \hat{q}(x|z) + \log \hat{q}(z)) \ \text{ where } \ \hat{q}(x|z) = \frac{1}{B^{\star}(z)} \exp\Big(-\sigma(z)^T \hat{E}(x)\Big)$$

Rate of Growth of Discrete Affine Hull



Lower bound on train groups to generalize to d^m groups: $2c(md + d \log d)$

Results



CRM adapts to the Bayes optimal classifier of test distribution

Dataset	Method	Average Acc	WGA	WGA (No Groups Dropped)
Waterbirds	ERM	77.9 (0.1)	43.0 (0.2)	62.3 (1.2)
	G-DRO	77.9 (0.9)	42.3 (2.6)	87.3 (0.3)
	LC	88.3 (0.9)	75.5 (1.8)	88.7 (0.3)
	sLA	89.3 (0.4)	77.3 (1.4)	89.7 (0.3)
	CRM	87.1 (0.7)	78.7 (1.0)	86.0 (0.6)
CelebA	ERM	85.8 (0.3)	39.0 (0.3)	52.0 (1.0)
	G-DRO	89.2 (0.5)	67.8 (0.8)	91.0 (0.6)
	LC	91.1 (0.2)	57.4 (0.5)	90.0 (0.6)
	sLA	90.9 (0.2)	57.4 (1.3)	86.7 (1.9)
	CRM	91.1 (0.2)	81.8 (0.5)	89.0 (0.6)
MetaShift	ERM	85.7 (0.4)	60.5 (0.5)	63.0 (0.0)
	G-DRO	86.0 (0.3)	63.8 (1.1)	80.7 (1.3)
	LC	88.5 (0.0)	68.2 (0.5)	80.0 (1.2)
	sLA	88.4 (0.1)	63.0 (0.5)	80.0 (1.2)
	CRM	87.6 (0.3)	73.4 (0.4)	74.7 (1.5)
MultiNLI	ERM	68.4 (2.1)	7.5 (1.3)	68.0 (1.7)
	G-DRO	70.4 (0.2)	34.3 (0.2)	57.0 (2.3)
	LC	75.9 (0.1)	54.3 (1.0)	74.3 (1.2)
	sLA	76.4 (0.3)	55.0 (1.5)	71.7 (0.3)
	CRM	74.3 (0.3)	58.7 (1.4)	74.7 (1.3)
CivilComments	ERM	80.4 (0.2)	55.9 (0.2)	61.0 (2.5)
	G-DRO	80.1 (0.1)	61.6 (0.5)	64.7 (1.5)
	LC	80.7 (0.1)	65.7 (0.5)	67.3 (0.3)
	sLA	80.6 (0.1)	65.6 (0.2)	66.3 (0.9)
	CRM	83.7 (0.1)	67.9 (0.5)	70.0 (0.6)
NICO++	ERM	85.0 (0.0)	35.3 (2.3)	35.3 (2.3)
	G-DRO	84.0 (0.0)	36.7 (0.7)	33.7 (1.2)
	LC	85.0 (0.0)	35.3 (2.3)	35.3 (2.3)
	sLA	85.0 (0.0)	33.0 (0.0)	35.3 (2.3)
	CRM	84.7 (0.3)	40.3 (4.3)	39.0 (3.2)

CRM outperforms the baselines w.r.t worst group accuracy