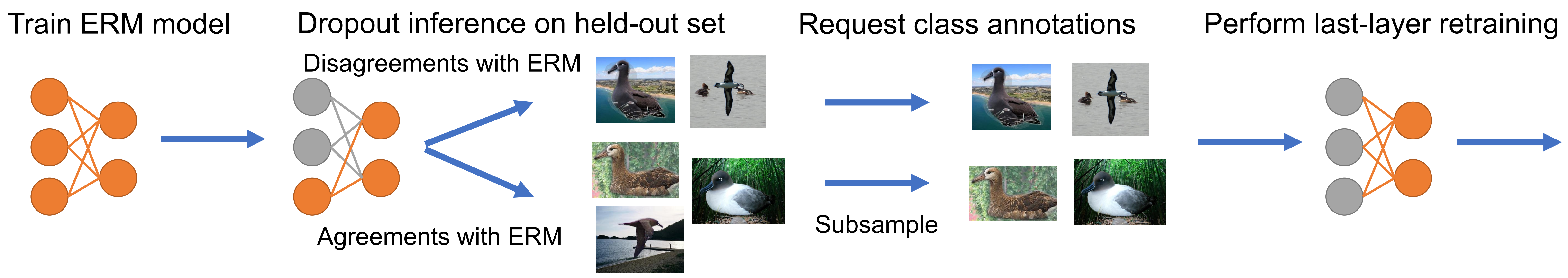
Dropout Disagreement: A Recipe for Group Robustness with Fewer Annotations

Tyler LaBonte[1], Vidya Muthukumar[1], Abhishek Kumar[2]
[1]Georgia Tech, [2]Google Research

Abstract: We perform last-layer retraining with dropout disagreements to improve worst-group accuracy with no group annotations and 20x fewer class annotations.



Train ERM model

Dropout inference on held-out set

Disagreements with ERM

Agreements with ERM

Request class annotations

Subsample

Perform last-layer retraining

Problem: Empirical risk minimization gives poor minority group performance

- Datasets often suffer from *spurious correlations* which are irrelevant for the true label
- Spurious features create minority groups which are underrepresented during training
- Maximize worst-group test accuracy instead of mean over the training distribution (ERM)

Landbird on land (73%)

Landbird on water (4%)

Waterbird on water (22%)

Waterbird on land (1%)

Prior Work: With group annotations, last-layer retraining boosts worst-group accuracy

- Models learn core features, but spurious features are overweighted in last layer [1]
- Last-layer retraining (DFR) on held-out group-balanced dataset is efficient and effective
- However, groups are often unknown ahead of time or are difficult to annotate

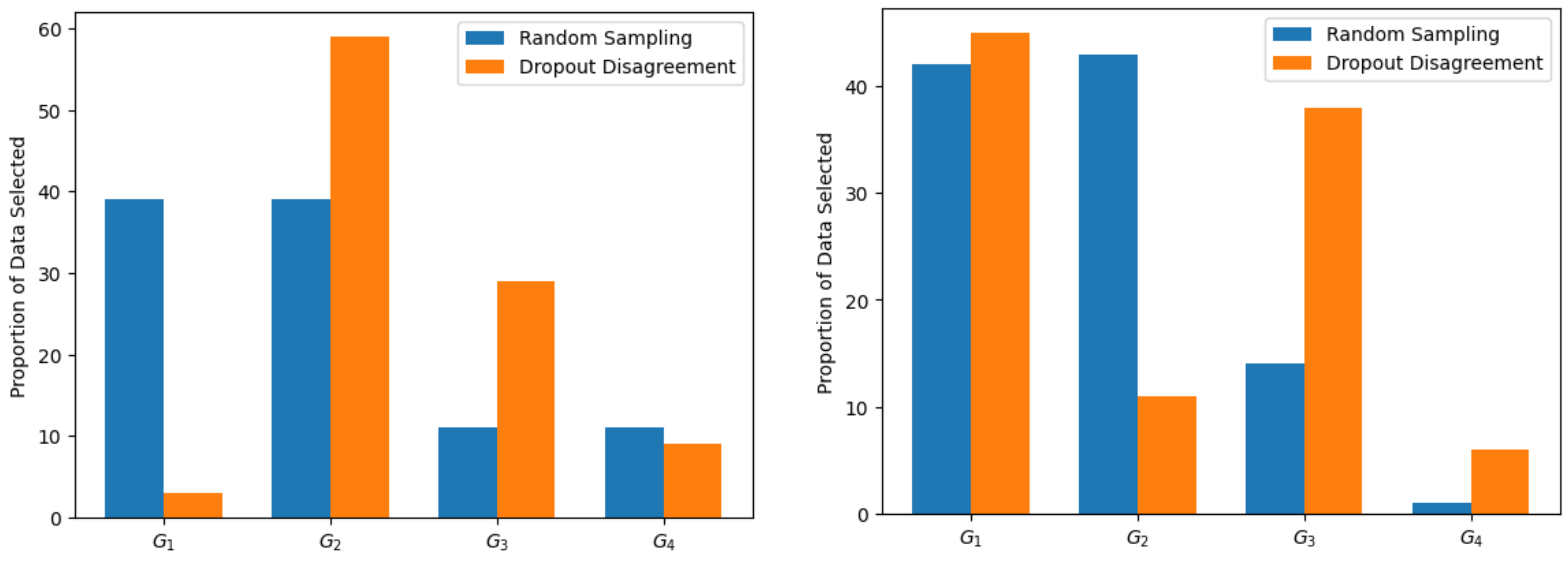
Dropout disagreement results on the Waterbirds dataset [6], averaged over 5 random seeds.

Method	Extra Annotations		Test Accuracy		
	Group	Class	Worst-Group	Train Dist. Mean	Test Dist. Mean
ERM	0	0	71.3	**97.8**	89.5
SSA [5]	0	599	89.0	92.2	-
DFR [1]	599	599	91.8	95.0	94.4
M-DFR (baseline)	0	599	89.7	92.6	93.7
DD-DFR (ours)	0	48	**91.6**	94.5	**93.8**

Our Work: Dropout disagreement matches DFR accuracy without group annotations

- Original and resource-constrained models disagree disproportionately on minority group
- Intuitive: early-stopping has simplicity bias [2, 3], dropout approximates uncertainty metric [4]
- Enables constructing nearly-group-balanced dataset without even knowing the groups
- Only need to request class annotations for disagreements – up to 20x fewer datapoints

Dropout disagreement proportions on the Waterbirds and CelebA datasets [6, 7].



References

[1] Kirichenko et al. “Last layer re-training is sufficient for robustness to spurious correlations.” NeurIPS 2022. **[2]** Arpit et al. “A closer look at memorization in deep networks.” ICML 2017. **[3]** Liu et al. “Just train twice: Improving group robustness without training group information.” ICML 2021. **[4]** Gal and Ghahramani. “Dropout as a Bayesian approximation: representing model uncertainty in deep learning.” ICML 2016. **[5]** Nam et al. “Spread spurious attribute: improving worst-group accuracy with spurious attribute estimation.” ICLR 2022. **[6]** Sagawa et al. ”Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization.” ICLR 2020. **[7]** Liu et al. “Deep learning face attributes in the wild.” ICCV 2015.