Contrastive Adapters for Foundation Model Group Robustness

Michael Zhang [1] **Christopher Ré** [1]

Abstract

While large pretrained foundation models (FMs) have shown remarkable zero-shot classification robustness to dataset-level distribution shifts, their robustness to group shifts is relatively underexplored. We study this problem, and first find that popular FMs such as CLIP may not be robust to various group shifts. On prior robustness benchmarks, they achieve up to an 80.7 percentage point (pp) gap between average and worst-group accuracy. Unfortunately, current methods to improve robustness require retraining, which can be prohibitively expensive for large FMs. We find existing ways to efficiently improve large model inference, *e.g*, by training adapters (lightweight MLPs) on top of FM embeddings, can also *hurt* group robustness compared to zero-shot. We thus propose a first adapter training method designed to improve FM robustness to group shifts. While prior work only trains adapters with class labels, we add a contrastive objective to explicitly learn similar embeddings for initially dissimilar FM embeddings. Across the same benchmarks, contrastive adapting effectively and efficiently improves group robustness, raising worst-group accuracy by 16.0 to 56.0 pp over zero-shot without any FM finetuning. Beyond FM robustness, contrastive adapting achieves near-state-of-the-art robustness on Waterbirds and CelebA, while only training 1% of other methods' model parameters.

1. Introduction

Foundation models (FMs)—large pretrained models trained on massive datasets—offer an exciting new deep learning paradigm. Recent works show that with no finetuning, FMs can generalize to various tasks (Brown et al., 2020; Radford et al., 2021) with impressive robustness to certain distribution shifts (Kumar et al., 2022; Wortsman et al., 2021).

[1]Stanford University. Stanford, CA, USA. Correspondence to: Michael Zhang <mzhang@cs.stanford.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

However, an underexplored question is how robust this no finetuning, or *zero-shot*, paradigm is to "group shifts," distribution shifts between subpopulations or meaningful groups in data. Prior works have established that *group robustness*–*i.e* performing well on all groups—is a fundamental and real-world challenge for modern deep learning (Beery et al., 2018; Buolamwini & Gebru, 2018; Oakden-Rayner et al., 2020; Koh et al., 2021). Yet most prior foundation model evaluations focus on overall or average performance; few works consider their accuracy across groups.

In this work, we thus study FM group robustness. We first motivate this problem by showing that FMs can have poor zero-shot group robustness. Evaluating 9 FMs across 7 robustness benchmarks, we find up to an 80.7 percentage point (pp) gap between their average and worst group accuracy.

We therefore aim to improve FM group robustness. This poses several challenges and open questions. First, current robustness methods often retrain at least one model (Sagawa et al., 2019; Nam et al., 2020; Creager et al., 2021; Liu et al., 2021; Ahmed et al., 2021; Zhang et al., 2022). This can be prohibitively expensive for FMs due to their size and scale, and it is unclear if we can improve FM robustness without any retraining or finetuning. Second, many practitioners may only access FM outputs or embeddings (*e.g*, via APIs[1]), using zero-shot classification for their downstream tasks. Ideal robustness solutions would then also only require FM embeddings. However, if these same embeddings lead to poor zero-shot robustness, then it is unclear if they encode the information needed to classify all groups correctly.

Motivated by these challenges and questions, we study effective *and* efficient solutions for better FM group robustness. As a baseline, we first find that while efficient methods to improve FM inference–*e.g*, training linear probes (Radford et al., 2021; Kumar et al., 2022) and adapters (Houlsby et al., 2019; Gao et al., 2021) on top of FM embeddings–can improve robustness over zero-shot (*e.g*, reducing the accuracy gap by 20.4 pp on the Waterbirds dataset (Welinder et al., 2010)), they fail to do so consistently, and can *hurt* robustness (increasing the gap by up to 74.9 pp on the CelebA dataset (Liu et al., 2015)). To improve robustness, we note that while poor zero-shot robustness occurs if FMs embed

[1]https://beta.openai.com/docs/introduction., https://studio.ai21.com/docs/, https://docs.cohere.ai/

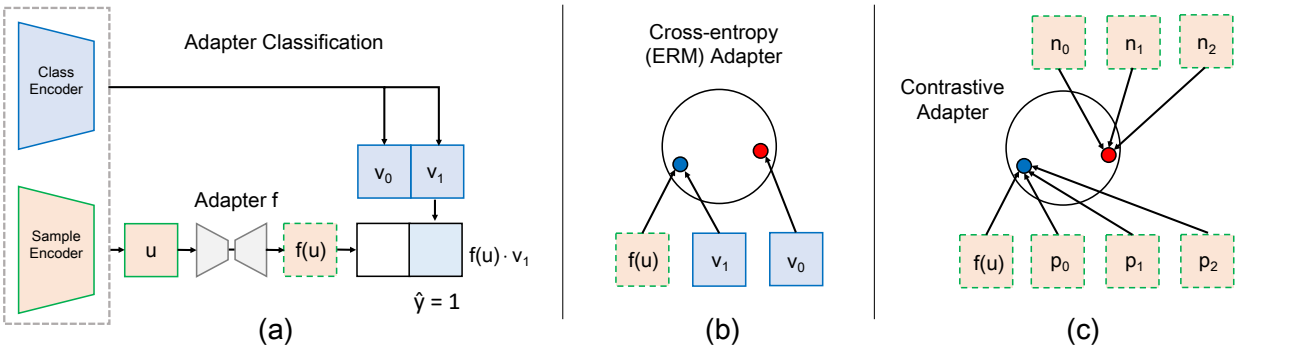

Figure 1. (a) Adapter classification with FM embeddings. Adapters learn transformations to align sample embeddings to class embeddings. (b) Standard cross-entropy loss only uses class embeddings. (c) Contrastive adapting adds other sample embeddings for greater alignment.

same-class samples in different groups "far apart", we find instances where standard adapters fail to close this distance.

To consistently improve group robustness over zero-shot, we propose *contrastive adapting*, a simple adapter training method that places greater emphasis on bringing these initially "far apart" points together. For each task, we first compute FM embeddings for each training sample and class. We then train adapters—small bottleneck MLPs—on these embeddings. Like prior work (Gao et al., 2021), these adapters take sample embeddings as inputs, and output transformed embeddings with greater cosine similarity to their ground-truth class embeddings. However, the key difference is that contrastive adapting also applies contrastive learning over other *sample* embeddings. We provide a way to "pull together" far apart sample embeddings in the same class, and "push apart" nearby sample embeddings in different classes.

In our experiments, we validate that contrastive adapting effectively and efficiently improves FM group robustness. First, across all 9 robustness benchmarks, we find contrastive adapting consistently improves worst-group accuracy over zero-shot (by 16.0 to 56.0 pp), using no training group labels and adapters with 0.1% to 0.3% of the original FM parameters. Then, on the popular Waterbirds and CelebA robustness datasets, we find contrastive adapting can substantially outperform other methods that only use FM embeddings. Finally we find contrastive adapting can enable effective and efficient group robustness in general. On CelebA, we achieve +0.2 pp worst-group accuracy over the prior state-of-the-art with only 1.0% of its parameters.

2. Problem

2.1. Preliminaries: group robustness and task setup

For setup, we follow prior work (Sagawa et al., 2019). For some task, we have N samples $\{(x_i, y_i, g_i)\}_{i=1}^N$, with sample inputs $x_i \in \mathcal{X}$, class labels $y_i \in \mathcal{Y}$, and group labels $g_i \in \mathcal{G}$. Let $C = |\mathcal{Y}|$ be the number of classes. g_i indicates each sample's group, but we do not see group labels during training. Distribution shifts may occur between same-class samples in different groups. Every sample (x_i, y_i, g_i) is drawn from some joint distribution P. Let P_g be the specific distribution conditioned on g for any $g \in \mathcal{G}$. For classification loss $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ and classifier $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, we want f_θ to be accurate, achieving low average error:

$$\mathcal{L}_{\text{avg}}(f_\theta) := \mathbb{E}_{(x,y,g)\sim P}[\ell(f_\theta(x), y)] \tag{1}$$

and *group robust*, achieving a small gap between average error and worst-group error (implying low worst-group error):

$$\mathcal{L}_{\text{wg}}(f_\theta) := \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y,g)\sim P_g}[\ell(f_\theta(x), y)] \tag{2}$$

Unlike out-of-distribution (OOD) evaluation, we observe each group in all data splits. However, standard training via empirical risk minimization (ERM) can lead to poor group robustness due to imbalanced groups (Shimodaira, 2000). Here, FMs are *not trained* on the training data, but we show their zero-shot classification may still not be group robust.

2.2. Empirical findings of poor FM group robustness

We now demonstrate the group robustness problem with foundation models. We first briefly describe zero-shot classification and baseline methods (linear probes, adapters) to improve inference over the former. We then show that popular FMs such as CLIP (Radford et al., 2021) may not be group robust on various robustness benchmarks from prior work. We finally use two datasets, Waterbirds (Welinder et al., 2010; Sagawa et al., 2019) and CelebA (Liu et al., 2015), to present two outcomes where baselines can help and hurt robustness. Additional benchmark details and results for all 9 FMs evaluated are deferred to Appendix A.

Baseline methods. To evaluate foundation model group robustness, we consider the following baseline methods:

- **Zero-shot classification** (Radford et al., 2021): Assuming C classes, we first use a FM to compute sample embeddings for each test sample, and C total class embeddings. We then use the nearest class embedding (via cosine similarity) to each sample embedding for class prediction. For class embeddings, we convert each class to a text prompt, *e.g*, "this is a `[class name]`", and use the FM text encoder output of the tokenized prompt (c.f. Fig. 1).

- **Linear Probe** (Radford et al., 2021; Wortsman et al., 2021): We train a linear classifier on top of training data sample embeddings to predict their class labels. We then use the linear probe to classify test sample embeddings.
- **Adapter** (Gao et al., 2021; Rebuffi et al., 2017): We train a single 2-layer bottleneck multilayer perception (MLP) to output transformed sample embeddings most similar to their ground-truth class embeddings (c.f. Fig. 1). Specifically, with adapter f_θ, sample embedding u, ground-truth class embedding v, temperature τ, and l2-norm operator $\hat{\cdot}$, we update adapter weights θ with a cross-entropy loss:

$$\ell(f_\theta(u), y) = -\log \frac{\exp(\hat{f}_\theta(u)^\top \hat{v}/\tau)}{\sum_{c=1}^{C} \exp(\hat{f}_\theta(u)^\top \hat{v}_c/\tau)} \quad (3)$$

  At test time, we compute a sample embedding for each test sample, and use these "adapted" embeddings instead of the FM embeddings in the zero-shot procedure above.

For class embeddings, we tried several templates, choosing by val. set worst-group acc. App C.6 lists the templates.

Zero-shot classification results. Fig. 2 charts the average and worst-group accuracies achieved via zero-shot classification with a CLIP ResNet-50 (RN-50) FM (Radford et al., 2021) on the 7 image benchmarks. We find evidence of poor group robustness (with accuracy gaps up to 80.7 pp). The larger gaps suggest that while FMs may learn correlations that apply to many unseen samples on *average* (*e.g*, classifying 80+% of samples correctly on Waterbirds and BREEDS datasets (Santurkar et al., 2020)), such correlations may not hold for all groups, leading to poor worst-group accuracy (only 49.8% and 6% for Waterbirds and BREEDS datasets).

Efficient baseline results. In Table 1, we report worst-group and average accuracy with all baseline methods on the popular Waterbirds and CelebA benchmarks. This presents two different outcomes. Perhaps surprisingly, ERM-trained adapters are sufficient to improve robustness significantly on Waterbirds (+24.4 pp worst-group accuracy, -20.4 pp accuracy gap). However, on CelebA, both baseline methods result in *poorer* group robustness than zero-shot.

2.3. Motivation for improving robustness over baselines

Towards improving robustness consistently, we expand on the FM group robustness problem with possible limitations of ERM adapter training. Note that if zero-shot classification for a class is accurate on average but not group robust, then in the FM embedding space there exist groups embedded "far apart" despite being in the same class. One view of ERM adapter training is that it trains adapters to learn an embedding space that brings these groups together. We can interpret Eq. 3 as a contrastive loss (Oord et al., 2018; Chen et al., 2020) where u is an anchor, v is a positive, and the other $C-1$ class embeddings are negatives. While this works for Waterbirds, only using class embeddings as positives and negatives may not always bring desired points together. We verify this in Fig. 3 by computing the mean pairwise cosine similarity of *intra-class* sample embeddings in the same class but different groups. On adapter-trained and pretrained embeddings from multiple CLIP models, poorer robustness tracks lower intra-class cosine similarity.

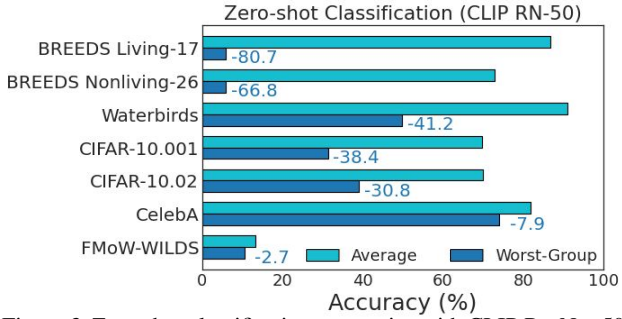

Figure 2. Zero-shot classification accuracies with CLIP ResNet-50 (RN-50). We consistently find large worst-group versus average accuracy gaps across many group robustness benchmarks.

Table 1. Worst-group (WG) and average (Avg) accuracies (in %) of linear probes and adapters compared to zero-shot. Best metrics **in bold**. Baselines can hurt robustness vs. zero-shot (in red).

	Waterbirds			CelebA		
Method	WG	Avg	Gap	WG	Avg	Gap
Zero shot	36.6	92.2	55.6	**74.0**	81.9	**7.9**
Linear Probe	7.9	93.5	85.6	11.9	94.7	82.8
Adapter	**60.8**	96.0	**35.2**	36.1	94.2	58.1

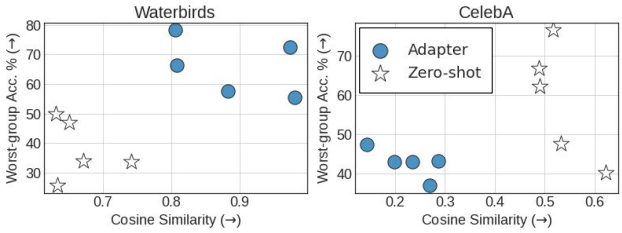

Figure 3. Despite task-specific training, adapter embeddings may not carry greater intra-class cosine similarity than initial FM embeddings and result in poorer worst-group accuracy (c.f. CelebA).

3. Method: Contrastive Adapting

To improve robustness, we thus aim to more explicitly bring far away samples together. For an anchor sample, instead of using only 1 class embedding positive and $C-1$ negatives, we add positives with *sample* embeddings in the same class but far away from the anchor (*e.g*, likely in different groups). We add negatives with sample embeddings nearest to the anchor but in different classes (as Ge et al. (2021); Zhang et al. (2022) show *hard* negatives benefit contrastive learning for robustness). As dataset size is often much larger than number of classes, contrastive adapting is also supported by Khosla et al. (2020); Robinson et al. (2021) that show more positives and negatives benefit contrastive learning. Our method is simple to implement with three components:

Table 2. On Waterbirds and CelebA, contrastive adapters achieve comparable to state-of-the-art worst-group acc. with 1% of the trainable parameters. ΔAcc. is pp gap with prior SoTA. **1st** / 2nd best metrics **bolded** / underlined. We report numbers from original works.

Model	# Trained Params	% Params	Method	Waterbirds WG Acc. (%)	Waterbirds ΔAcc.	CelebA WG Acc. (%)	CelebA ΔAcc.
ResNet-50	25557032	100	EIIL (Creager et al., 2021)	78.7	-10.3	83.3	-6.5
			CIM (Taghanaki et al., 2021)	83.6	-5.4	83.6	-6.2
			JTT (Liu et al., 2021)	86.7	-2.3	81.1	-8.7
			RWY (Idrissi et al., 2021)	86.1	-2.9	82.9	-6.9
			CNC (Zhang et al., 2022)	88.5	-0.5	88.8	-1.0
			SSA (Nam et al., 2022)	**89.0**	0.0	89.8	0.0
Adapter + CLIP RN-50	263424	1.03	Contrastive Adapting	83.7	-5.3	**90.0**	0.2

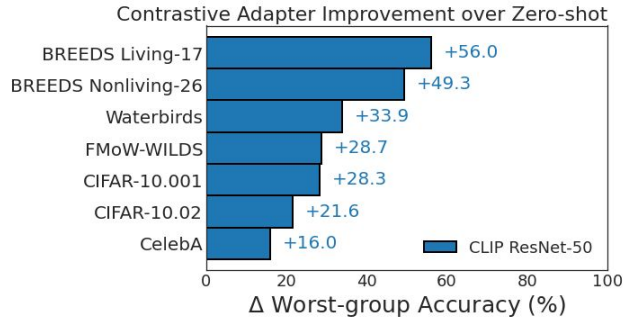

Figure 4. On group robustness datasets, contrastive adapting consistently improves worst-group acc. over zero-shot CLIP-RN50.

(1) Foundation model embedding and prediction: We compute FM embeddings over labeled training data. For (2), we also compute zero-shot predictions on this data.
(2) Contrastive sampling: For each class, we identify "anchor" sample embeddings that zero-shot predicts incorrectly. For each anchor, we sample P "positive" sample embeddings that zero-shot classifies correctly (as a heuristic for finding "far apart" samples), and M hard "negative" sample embeddings by computing the nearest neighbors to the anchors in different classes by cosine similarity.
(3) Training objective: We update adapters with a contrastive sample loss (Khosla et al., 2020) on the $1 + P + M$ sample embeddings. To ensure that the adapter learns to keep sample embeddings close to their ground-truth class embeddings, we also use a cross-entropy loss over random minibatches of sample embeddings (c.f. Eq. 3).

4. Experiments

We aim to validate that contrastive adapting improves and enables efficient and effective group robustness with FM embeddings. First, across image benchmarks, contrastive adapting consistently improves group robustness over zero-shot. Next, on Waterbirds and CelebA, contrastive adapting is more effective than other strategies to improve inference with only FM embeddings. Finally, on these datasets, contrastive adapting enables state-of-the-art robustness with greater parameter efficiency. We select models based on val. set worst-group accuracy as in prior work (Koh et al., 2021). App C.2 contains method details and hyperparameters.

Table 3. Evaluation of efficient methods for improving group robustness on CLIP RN-50. Contrastive adapters best improve group robustness. **1st** / 2nd best metrics **bolded** / underlined.

Method / Acc. (%)	Waterbirds WG	Waterbirds Avg	Waterbirds Gap	CelebA WG	CelebA Avg	CelebA Gap
Best Baseline (c.f. Table 1)	60.8	96.0	35.2	74.0	81.9	7.9
WiSE-FT	49.8	91.0	41.2	85.6	88.6	3.0
DFR (Subsample)	63.9	91.8	27.9	76.9	92.5	15.6
DFR (Upsample)	51.3	92.4	41.1	89.6	91.8	2.2
Group Prompt ZS	55.9	87.8	31.9	70.8	82.6	11.8
Contrastive Adapter	**83.7**	89.4	**5.7**	**90.0**	90.7	**0.7**

Consistent robustness improvements over zero-shot. In Fig 4, we find that unlike ERM adapters, contrastive adapting consistently improves group robustness on the 7 image robustness benchmarks and CLIP RN-50 over zero-shot, achieving 16.0 to 56.0 pp higher worst-group accuracy.

Robustness improvements over efficient FM methods. In Table 3, we find contrastive adapting improves group robustness over prior baselines and other strategies to improve FM inference with training sample FM embeddings: WiSE-FT (Wortsman et al., 2021), which ensembles linear probe and zero-shot weights; DFR (Kirichenko et al., 2022) which trains linear probes on resampled data; and group prompt zero-shot, where we use group-informed prompts for class embeddings. On average, contrastive adapting improves worst-group accuracy over the next best method by 16.4 pp.

Efficiency improvements over recent robust methods. In Table 2, we find that contrastive adapting with CLIP RN-50 embeddings achieves comparable worst-group accuracy to recent state-of-the-art robustness methods on Waterbirds and CelebA, despite only training 1% of their parameters. Notably, compared to the state-of-the-art Spread Spurious Attribute (SSA) (Nam et al., 2022), contrastive adapting achieves +0.2 pp worst-group accuracy on CelebA.

5. Conclusion

We find that FM zero-shot classification may not be group-robust, and present a simple first-step approach to significantly improve robustness without any finetuning.

References

Ahmed, F., Bengio, Y., van Seijen, H., and Courville, A. C. Systematic generalisation with group invariant predictions. In *ICLR*, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bahng, H., Jahanian, A., Sankaranarayanan, S., and Isola, P. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., and Bain, M. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL http://arxiv.org/abs/1903.04561.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11162–11173, 2021.

Fürst, A., Rumetshofer, E., Tran, V., Ramsauer, H., Tang, F., Lehner, J., Kreil, D., Kopp, M., Klambauer, G., Bitto-Nemling, A., et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Ge, S., Mishra, S., Li, C.-L., Wang, H., and Jacobs, D. Robust contrastive learning using negative samples with diminished semantics. *Advances in Neural Information Processing Systems*, 34, 2021.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine*

Learning, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. *arXiv preprint arXiv:2110.14503*, 2021.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Levine, Y., Dalmedigos, I., Ram, O., Zeldes, Y., Jannai, D., Muhlgay, D., Osin, Y., Lieber, O., Lenz, B., Shalev-Shwartz, S., et al. Standing on the shoulders of giant frozen language models. *arXiv preprint arXiv:2204.10019*, 2022.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46–54, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.7. URL https://aclanthology.org/2020.emnlp-demos.7.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online, April 2021. Association for Computational Linguis-

tics. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, 2017.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Robinson, J., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Singla, S., Moayeri, M., and Feizi, S. Core risk minimization using salient imagenet. *arXiv preprint arXiv:2203.15566*, 2022.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020.

Taghanaki, S. A., Choi, K., Khasahmadi, A., and Goyal, A. Robust representation learning via perceptual similarity metrics. *arXiv preprint arXiv:2106.06620*, 2021.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.-S., and Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022.

A. Expanded zero-shot evaluation for group robustness

In this section, we expand on the zero-shot evaluation of various foundation models on group robustness benchmarks discussed in Section 2. We first describe the datasets and models used in Appendix A.1. We then include results in Appendix A.2. We also describe and evaluate on two additional text group robustness datasets with two natural language foundation models. We find consistent trends of poor group robustness with zero-shot classification, marked by poor worst-group accuracy and large gaps between average and worst-group accuracy.

A.1. Additional details on robustness datasets and foundation models

Datasets. To benchmark zero-shot group robustness, we use a diverse set of datasets with group shifts from prior robustness literature. We describe them below and include details on size of groups and type of group shift in Table 4:

- **Waterbirds** (Welinder et al., 2010; Sagawa et al., 2019). We classify images by bird type. Each class $\in$ {`waterbird`, `landbird`} carries two groups: birds on water backgrounds, and birds on land backgrounds.
- **CelebA** (Liu et al., 2015; Sagawa et al., 2019). We classify images by celebrity hair color. Each class $\in$ {`not blond`, `blond`} carries two groups: celebrities labeled as male, and celebrities labeled as female.
- **BREEDS (`Living-17`, `Nonliving-26`)** (Santurkar et al., 2020). For the `Living-17` and `Nonliving-26` datasets in the BREEDS benchmark sourced from ImageNet (Santurkar et al., 2020), we classify images by one of several categories. Each class is a coarse category consisting of multiple fine-grained groups. Groups in the same class may be visually distinct (*e.g*, the `ape` class includes images of gibbons and gorillas). While the original benchmark evaluates how classifiers trained on seen `source` groups generalize to unseen `target` groups, we adapt the datasets for our group robustness setting by adding 5% of the images in each `target` group to the `source` groups, and evaluating worst-group accuracy over all `source` and `target` groups.
- **CIFAR-10.001**, **CIFAR-10.02** (Krizhevsky, 2009; Recht et al., 2018; Lu et al., 2020). We classify images by one of 10 categories. We combine CIFAR-10 (Krizhevsky, 2009) and either CIFAR-10.1 (Recht et al., 2018) or CIFAR-10.2 (Lu et al., 2020), which are collected from different sources. The new datasets' classes carry two groups determined by the source dataset.
- **FMoW-WILDS** (Christie et al., 2018; Koh et al., 2021). We classify satellite images into one of 62 building or land-use categories (*e.g*, `airport`, `zoo`). Each images belongs to one of five groups based on continental region. To test group robustness, we compare the accuracies over all samples in each group as in the WILDS benchmark (Koh et al., 2021). We also evaluate only over test images from the same time period as training images (the "IID" split in the original WILDS benchmark (Koh et al., 2021)).
- **CivilComments-WILDS** (Borkan et al., 2019; Koh et al., 2021). We classify if a text comment is toxic or not. Samples are organized into 8 groups based on mention of a demographic identity (*e.g*, "female", "LGBTQ").
- **Amazon-WILDS** (Ni et al., 2019; Koh et al., 2021). We classify if an online text review is positive or negative. Reviews are organized into different groups based on the product category (*e.g*, `books`, `electronics`). We adapt this dataset from the official Amazon-WILDS split by using the `category_subpopulation` split. We also map the original class labels, which are star-ratings from 1 to 5, to positive or negative reviews by discarding samples with a 3-star rating, and re-labeling 1- and 2-star ratings as `negative` and 4- and 5-start ratings as `positive`.

Foundation models. For image datasets, we evaluate pretrained CLIP (Radford et al., 2021) and CLOOB (Fürst et al., 2021) vision-language models using publicly available weights[2][3]. We evaluate 7 available CLIP models: 3 ResNet image encoder backbones (`RN-50`, `RN-101`, `RN-50x4`), and 5 Vision Transformer image encoder backbones: (`ViT-B/32`, `ViT-B/16`, `ViT-L/14`, `ViT-L/14@336px`) and 2 CLOOB models (all available: `RN-50`, `RN-50x4`). For text datasets, we evaluate 2 pretrained GPT-Neo (Black et al., 2021) text models trained on the Pile (Gao et al., 2020) (`GPT-Neo-125M`, `GPT-Neo-1.3B`) available on HuggingFace[4][5].

[2]CLIP: https://github.com/openai/CLIP/blob/main/clip/clip.py
[3]CLOOB: https://ml.jku.at/research/CLOOB/downloads/checkpoints/
[4]GPT-Neo 125M: https://huggingface.co/EleutherAI/gpt-neo-125M
[5]GPT-Neo 1.3B: https://huggingface.co/EleutherAI/gpt-neo-1.3B

Table 4. Group robustness datasets, source of group shift, and group sizes.

		(Class-wise) Group Size		
Dataset	Group Shift	Largest	Smallest	Class-Wise?
Waterbirds	Confounder	1057	56	Yes
CelebA	Confounder	22880	1387	Yes
BREEDS Living-17	Subclass	1076	1009	Yes
BREEDS Nonliving-26	Subclass	1043	712	Yes
CIFAR-10.001	Data source	1000	114	Yes
CIFAR-10.02	Data source	4039	431	Yes
FMoW-WILDS	Subclass	34816	1582	No
Amazon-WILDS	Subclass	496127	110	No
CivilComments-WILDS	Confounder	4962	1003	Yes

A.2. Group robustness results

In Figure 5, we chart worst-group and average accuracies achieved by various zero-shot foundation models across the group robustness datasets. Larger gaps between accuracies, *i.e* high average accuracy yet low worst-group accuracy, indicate poor group robustness. In aggregate, on all datasets except FMoW-WILDS and Amazon-WILDS, we observe a shared pattern of noticeable gaps between average and worst-group accuracy, suggesting that zero-shot classification with popular foundation models may not be group robust. We perform zero-shot classification as described in Section 2. As recommended by Radford et al. (2021), for each dataset we consider several prompt templates. We engineer prompts by using the single best template based on validation worst-group accuracy. In Appendix C.6 we include a full list of prompts used.

B. Contrastive adapter implementation details

We provide further details on the adapter architecture and training sampling.

B.1. Adapter architecture

Similar to prior works (Houlsby et al., 2019; Gao et al., 2021), the adapters we use are bottleneck 2-layer multilayer perceptrons (MLPs). We set the input dimension and output dimensions as the same as the pretrained foundation model embedding dimension, and pick a smaller dimension for the hidden layer (frequently 128, although this was chosen as a heuristic and not tuned). We also experimented with using a single residual connection (He et al., 2016) and batch normalization layer (Ioffe & Szegedy, 2015) between the input and output layers, but only found the latter to be helpful. Pytorch-like pseudocode is given below. The adapter is visualized in Figure 6.

```
import torch.nn as nn

class Adapter(nn.Module):
    def __init__(self, input_dim, hidden_dim):
        super().__init__()
        self.arch = nn.Sequential(
            nn.Linear(input_dim, hidden_dim),
            nn.BatchNorm1d(hidden_dim),
            nn.ReLU(),
            nn.Linear(hidden_dim, input_dim)
        )
    def __forward__(self, x):
        return self.arch(x)
```

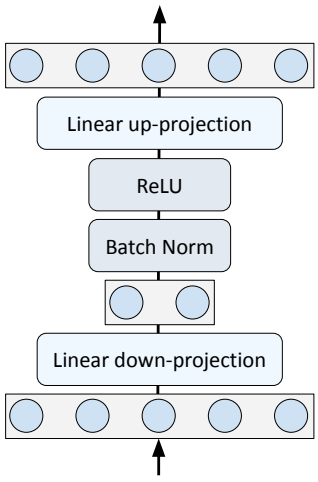

Figure 6. Adapter architecture

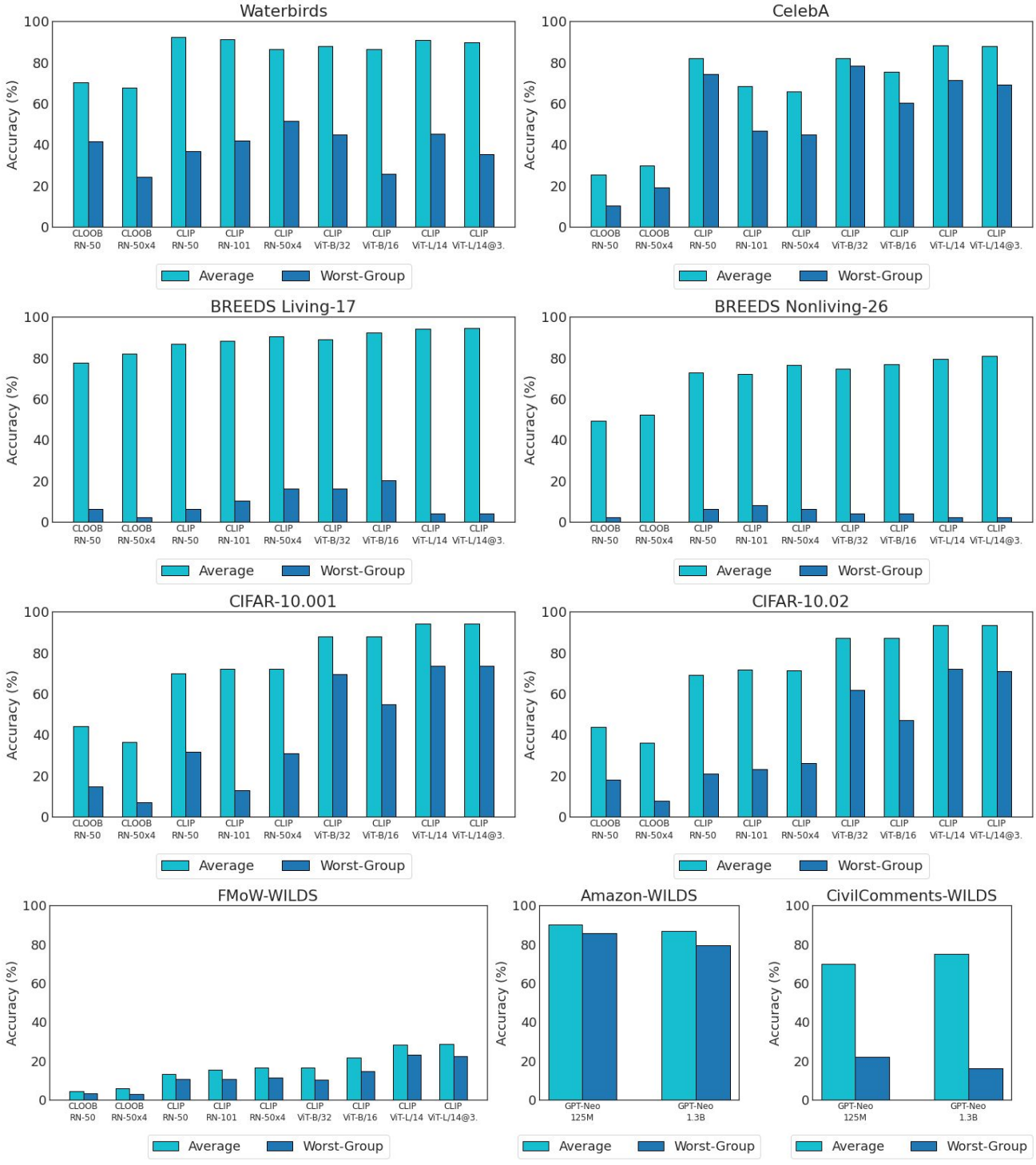

Figure 5. Foundation model zero-shot classification accuracies. We find poor zero-shot group robustness across datasets and models via large gaps between average and worst-group accuracies.

B.2. Adapter training sampling

Recall that we train the adapter with both a supervised contrastive loss over specifically sampled "contrastive batches", and a cross-entropy loss over resampled batches, both over the fixed pretrained foundation model embeddings. We use the

foundation model's zero-shot classification predictions to guide sampling for both.

We outline the algorithms for sampling training batches in Algorithm 1 and Algorithm 2. We then train an adapter by applying the sample-wise supervised contrastive loss (sample-wise SupCon) (Khosla et al., 2020) and the cross-entropy loss (Eq. 3) over these batches in Algorithm 3.

Algorithm 1 Contrastive batch sampling

input Training dataset sample embeddings $U = \{u_n\}_{n=1}^N$. Ground-truth class labels $Y = \{y_n\}_{n=1}^N$. Foundation model zero-shot predictions $\hat{Y} = \{\hat{y}_n\}_{n=1}^N$.
Require: Number of positives P per anchor. Number of negatives M per anchor. Number of nearest neighbors M^* per anchor to sample negatives from.
Initialize set of contrastive batches $B = \{\}$
for anchor $u_a \in \{u_i \in U : \hat{y}_i \neq y_i\}$ **do**
    (Positive sampling)
    Sample P positives $\{u_p\}_{p=1}^P$ uniform-randomly from U where $\hat{y}_p = y_p$ (and $\hat{y}_p \neq \hat{y}_a$)
    (Negative sampling)
    Sample M negatives $\{u_m\}_{m=1}^M$ by computing the M^* sample embeddings with the highest cosine similarity to u_a where $y_m \neq y_a$, then randomly sampling M of these embeddings
    Update contrastive batch sets $B \leftarrow B \cup \left(u_a, \{u_p\}_{p=1}^P, \{u_m\}_{m=1}^M\right)$
end for=0

Algorithm 2 Resampled training set sampling

input Training dataset sample embeddings $U = \{u_n\}_{n=1}^N$. Ground-truth class labels $Y = \{y_n\}_{n=1}^N$. Foundation model zero-shot predictions $\hat{Y} = \{\hat{y}_n\}_{n=1}^N$. All unique classes C.
Initialize resampled training samples $U^* = \{\}$
for class $c \in C$ **do**
    Identify incorrect samples $U^- = \{u_i\}$ where $\hat{y}_i \neq c$
    Identify correct samples $U^+ = \{u_i\}$ where $\hat{y}_i = c$
    Obtain upsampled samples $\tilde{U}^-$ by uniform-randomly sampling from U^- s.t. $|\tilde{U}^-| = |U^+|$
    Update resampled samples $U^* \leftarrow U^* \cup \left(\tilde{U}^- \cup U^+\right)$
end for=0

Algorithm 3 Contrastive adapting

input Set of contrastive batches B, resampled training samples U^*, number of epochs K.
Randomly initialize adapter f_θ
for epoch $1, \ldots, K$ **do**
    Sample contrastive batch $\{b\}$ from B
    Sample randomly-shuffled minibatch of samples $\{u\}$ from U^*
    Update f_θ with sample-wise SupCon loss over $\{b\}$
    Update f_θ with Equation 3 over $\{u\}$
end for=0

C. Additional experimental details

C.1. Main results comparison methods

In Table 3, we compared contrastive adapting with several other approaches designed to improve downstream transfer in related settings, while similarly only requiring pretrained model embeddings. We describe them in more detail below:

- **Weight space ensembling (WiSE-FT)** (Wortsman et al., 2021), which first trains a linear classifier with standard ERM, and then ensembles the classifier outputs with the initial zero-shot predictions. While proposed for both training linear classifiers and finetuning the original weights of a foundation model, we focus on the linear classifier version for fair comparison in our setting.

- **Deep feature reweighting (DFR)** (Kirichenko et al., 2022), which first trains a linear probe on embeddings computed from a pretrained model over group-balanced data. As we do not assume training group labels, we first infer groups using zero-shot classification with foundation model embeddings. As in prior work (Liu et al., 2021; Zhang et al., 2022), we treat the incorrect and correctly classified samples as proxies for different groups.

Finally, assuming we have validation group labels, we know what groups could plausibly be in our test data. We thus also compare against **group-informed prompting**, which performs zero-shot classification using prompts with group information (*e.g*, "a waterbird on a land background").

C.2. Model selection and hyperparameters

For each dataset and method, we use the following hyperparameters. As in prior group robustness work (Koh et al., 2021), we select the best model and hyperparameters based on early stopping that achieves highest worst-group validation accuracy. For all methods and datasets, we train both linear probes and adapters with SGD, and sweep over learning rate $\in$ {1e-3, 1e-4, 1e-5} and weight decay $\in$ {5e-5, 1e-5, 5e-4}. For adapter classification, we used the default temperature used for zero-shot classification in CLIP (Radford et al., 2021). We did not tune the contrastive temperature. Unless noted, we ran all numbers.

We list hyperparameters for linear probes (Table 5), adapters (Table 6, both ERM and contrastive), and contrastive-specific hyperparameters (Table 7). We discuss method-specific hyperparameters:

- **Contrastive adapting** requires selecting three additional hyperparameters: the number of positives and negatives, and the number of nearest neighbors to sample negatives from. For these we swept over the following combinations of (number positives, number negatives, number neighbors): (2048, 2048, 2146), (2048, 2048, 4096), (512, 512, 1024).

- **Weight-space ensembling** (WiSE-FT): WiSE-FT requires picking a value $\alpha \in [0, 1]$ to compute a weighted combination of the zero-shot classifier parameters and the trained linear probe parameters. We sweep over intervals of size 0.1, *i.e* $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

Table 5. Linear probe hyperparameters

Dataset	Max Epochs	Learning Rate	Weight Decay	Momentum	Batch Size
Waterbirds	100	1e-3	5e-5	0.9	128
CelebA	50	1e-3	5e-5	0.9	128
BREEDS Living-17	100	1e-3	5e-5	0.9	128
BREEDS Nonliving-26	100	1e-3	5e-5	0.9	128
CIFAR-10.001	100	1e-3	5e-5	0.9	128
CIFAR-10.02	100	1e-3	5e-5	0.9	128
FMoW-WILDS	100	1e-3	5e-5	0.9	128
Amazon-WILDS	100	1e-3	5e-5	0.9	16
CivilComments-WILDS	100	1e-3	5e-5	0.9	16

C.3. Data splits

We use the same train, validation, and test splits for Waterbirds, CelebA, FMoW-WILDS, Amazon-WILDS, and CivilComments-WILDS as in prior work. For BREEDS and CIFAR datasets that we adapt for our problem setting, we construct test splits by combining official test splits from the original benchmarks. We then create training and validation splits by combining the rest of the data from these benchmarks, and randomly splitting this into 80% training data and 20% validation data. No original test data is seen during training on our splits.

Table 6. Adapter hyperparameters. For contrastive adapters, batch size refers to the size of each minibatch sampled for updating with cross-entropy loss.

Dataset	Max Epochs	Learning Rate	Weight Decay	Momentum	Batch Size	Hidden Dimension	Temperature
Waterbirds	100	1e-3	5e-5	0.9	128	128	0.01
CelebA	50	1e-3	5e-5	0.9	128	128	0.01
BREEDS Living-17	100	1e-3	5e-5	0.9	128	128	0.01
BREEDS Nonliving-26	100	1e-3	5e-5	0.9	128	128	0.01
CIFAR-10.001	100	1e-3	5e-5	0.9	128	128	0.01
CIFAR-10.02	100	1e-3	5e-5	0.9	128	128	0.01
FMoW-WILDS	100	1e-3	5e-5	0.9	128	512	0.01
Amazon-WILDS	100	1e-3	5e-5	0.9	16	512	0.01
CivilComments-WILDS	100	1e-3	5e-5	0.9	16	512	0.01

Table 7. Specific contrastive adapter hyperparameters.

Dataset	Number Positives	Number Negatives	Number Nearest Neighbors	Contrastive Temperature
Waterbirds	2048	2048	4096	0.1
CelebA	2048	2048	4096	0.1
BREEDS Living-17	2048	2048	4096	0.1
BREEDS Nonliving-26	512	512	1024	0.1
CIFAR-10.001	512	512	1024	0.1
CIFAR-10.02	512	512	1024	0.1
FMoW-WILDS	2048	2048	2146	0.1
Amazon-WILDS	2048	2048	2146	0.1
CivilComments-WILDS	2048	2048	2146	0.1

C.4. Additional dataset assets details and discussion

Dataset licenses. To curate CIFAR-10.0001 we use the CIFAR-10.1 dataset, which is distributed under the MIT License. The FMoW-WILDS dataset is distributed under the FMoW Challenge Public License[6]. The CivilComments-WILDS dataset is distributed under CC0 1.0. The Amazon-WILDS dataset does not have a license, but is requested to be used for research purposes only (Koh et al., 2021). We were not able to find explicit license information for CIFAR-10.2, Waterbirds, CelebA, or the BREEDS datasets. We note that the BREEDS datasets are sourced from ImageNet, which is distributed under the BSD 3-Clause License, and set up with code from the MadryLab robustness GitHub repository[7], which is distributed under a MIT license. The authors of the CelebA dataset provide a list of agreements[8], including that the dataset is used only for non-commercial research purposes.

Existing assets personally identifiable information and offensive content. The CelebA dataset consists of images of celebrity faces, which are personally identifiable. The dataset is also categorized by male and female identification at the time of curation, which may be outdated. The CivilComments-WILDS contains text samples flagged as toxic by toxicity classifiers (Koh et al., 2021), which contain potentially offensive content. Both datasets are existing assets, and both personal identifiability for CelebA and offensive content for CivilComments-WILDS can be checked by inspecting the original data inputs (images and text comments).

C.5. Compute and resources

All experiments were run on a machine with 14 CPU cores and a single NVIDIA Tesla P100 GPU. For training a contrastive adapter on top of CLIP ResNet-50 Waterbirds embeddings, this took approximately 30 minutes to run 100 epochs. Other than the numbers reported from their original publications in Table 2, we report all numbers from running experiments on the same machine.

[6] https://github.com/fMoW/dataset/blob/master/LICENSE
[7] https://github.com/MadryLab/robustness
[8] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Table 8. Class prompt templates or example prompts

Dataset	Foundation Model	Prompt template / example of prompt
Waterbirds	CLIP	"This is a picture of a `[class_name]`."
	CLOOB	"a `[class_name]`"
CelebA	CLIP	"A photo of a celebrity with blond hair."
	CLOOB	"A photo of a celebrity with blond hair."
BREEDS Living-17	CLIP	"This is a picture of a `[class_name]`."
	CLOOB	"This is a picture of a `[class_name]`."
BREEDS Nonliving-26	CLIP	"A photo of a `[class_name]`."
	CLOOB	"a `[class_name]`"
CIFAR-10.001	CLIP	"a `[class_name]`"
	CLOOB	"a `[class_name]`"
CIFAR-10.02	CLIP	"a `[class_name]`"
	CLOOB	"a `[class_name]`"
FMoW-WILDS	CLIP	"satellite view of the `[class_name]`"
	CLOOB	"aerial view of an `[class_name]`"
Amazon-WILDS	GPT-Neo	"Negative"
CivilComments-WILDS	GPT-Neo	"Not toxic"

C.6. Class prompt templates

In Table 8, we list the templates used to generate class prompts for each dataset. As a reminder, for each provided class name in a dataset, we create a prompt by inserting the class name into the prompt template. We then encode this prompt with a foundation model text encoder to get class embeddings.

D. Additional related work discussion

We provide additional discussion of related work and connections to our work below.

Zero-shot classification with foundation models. Our work builds on a growing literature on applying foundation models, large pretrained models that can be applied to various downstream tasks. These models demonstrate exciting promise in their ability to achieve accurate downstream transfer *without* any additional finetuning (Brown et al., 2020; Radford et al., 2021; Bommasani et al., 2021). In particular we consider the zero-shot capabilities of pretrained vision-language foundation models. These models, such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and CLOOB (Fürst et al., 2021) are trained on massive amounts of naturally paired image-text data, *e.g*, Internet images and their corresponding captions. Consisting of an image encoder (usually a ResNet or Vision Transformer) and a text encoder (usually a Transformer), such foundation models are commonly trained to learn a shared image-text embedding space where embeddings of images are most similar to embeddings of their corresponding caption text. While these objectives have been shown to lead to powerful representations (Zhang et al., 2020; Desai & Johnson, 2021), a crucial element for successful zero-shot classification is training data scale (Radford et al., 2021). However, added scale can also be a double-edged sword; when zero-shot classification still makes undesirable mistakes, standard ways to correct for these mistakes via retraining can become prohibitively expensive. We study one such motivating instance via group robustness, and provide a first-step solution towards improving group robustness efficiently.

Improving foundation model inference efficiently. Prior works aim to improve foundation model downstream performance, without having to finetune or update original model weights. *Prompt tuning* optimizes the inputs of a FM while keeping the original model weights frozen. Optimizing either text (Li & Liang, 2021; Zhou et al., 2021; 2022; Levine et al., 2022) or image (Bahng et al., 2022; Yao et al., 2021) inputs can improve a frozen foundation model's downstream task accuracy. However, doing so can require multiple passes through the foundation model, which may become expensive in certain situations (*e.g*, interacting with a foundation model via a commercial API). Another paradigm adds small trainable

parameters to the original model, either within its layers or on top of its embeddings. These include linear probes (linear classifiers) (Radford et al., 2021) and adapters (small bottleneck MLPs) (Houlsby et al., 2019; Rebuffi et al., 2017; Pfeiffer et al., 2021; 2020). Recently, Kumar et al. (2022); Wortsman et al. (2021) propose methods that use linear probes to improve robustness to dataset-level out-of-distribution shifts. Gao et al. (2021) propose to train single adapters on top of FM embeddings to improve average downstream task accuracy. We focus on *group shifts* that occur within a dataset. We also show that the latter can hurt group robustness, and propose alternatives to consistently improve group robustness.

Robustness of foundation models. Prior works have studied the robustness of foundation model inference to natural distribution shifts. Radford et al. (2021) show that zero-shot CLIP models can be more robust to out-of-distribution (OOD) shifts than prior ImageNet-trained models, measured via better generalization to various dataset-level distribution shifts on ImageNet classes. However, they also show that finetuning, or updating the original weights, of CLIP models on ImageNet can reduce this OOD robustness. Kumar et al. (2022); Wortsman et al. (2021) thus propose finetuning methods that improve downstream in-distribution accuracy while maintaining out-of-distribution robustness. Kumar et al. (2022) specifically study the trade-off between linear probing and finetuning, finding that finetuning on downstream data can improve generalization on in-distribution data over linear probing but more substantially hurt performance OOD data than linear probing. They show theoretically and empirically that a two-step strategy of first linear probing then full fine-tuning can combine the performance boosts of both. Wortsman et al. (2021) focus on the OOD trade-off presented by Radford et al. (2021) between a finetuned foundation model and its pretrained zero-shot weights. They propose weight-space ensembling (WiSE-FT), which computes a weighted average of the finetuned and pretraiend foundation model parameters, and show that the resulting averaged parameters can in some instances achieve higher performance on both data distributions that the model was finetuned on and unseen OOD data than the initial finetuned and zero-shot or pretrained models. They show this effect with both full finetuning and training a linear probe. Unlike these works, we focus on foundation model robustness to group shifts that occur within a dataset. We also compare against the linear probe version of WiSE-FT, and find that training adapters can be advantageous for achieving higher group robustness on various datasets.

Recently, other works also study foundation model learned spurious correlations and biases. Singla et al. (2022) show how various models (including CLIP models) may rely on spurious artifacts to classify ImageNet images. Berg et al. (2022) aim to debias CLIP image embeddings of human faces using extra metadata (textual concepts or attributes) that the embeddings should ignore. Our evaluation is complementary, noting poor group robustness across multiple types of data sources (objects, animals, human faces, text). We also provide a method that works without additional training metadata.

Improving group robustness of deep learning models. Improving the group robustness of deep learning models is a common deep learning challenge, where models may learn biases during training that lead to poor performance on certain groups. This is a widespread issue presented in contexts ranging from algorithmic fairness to healthcare diagnosis (Blodgett et al., 2016; Hashimoto et al., 2018; Buolamwini & Gebru, 2018). Several methods exist to improve group robustness. We compare against several recent approaches in Section 4. While one effective strategy to improve group robustness is to upweight the error of worst-performing group during training (Sagawa et al., 2019; 2020), training group labels may be impractical to obtain in practice (Sohoni et al., 2020; Oakden-Rayner et al., 2020). We thus consider robustness approaches which aim to work without training group labels. Several approaches involve training two models; one model is first trained with standard ERM to help infer groups, and another trained with a robust objective using these inferred groups. Just Train Twice (JTT) (Liu et al., 2021) treats samples that the first model misclassifies as inferred minority group samples to upweight. JTT then upweights these samples by a hyperparameter factor, and trains a second model with ERM on this upsampled data. Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021) infers groups by assigning samples to group under which the ERM model maximally violates an Invariant Risk Minimization (Arjovsky et al., 2019) principle. It then trains a robust model with Group DRO (Sagawa et al., 2019) using the inferred groups, which dynamically upweights the worst-performing groups during training. Correct-N-Contrast (CNC) (Zhang et al., 2022) instead identifies samples with the same class labels but different ERM model predictions, and trains a robust model by using a contrastive loss to learn similar representations between these samples. Spread Spurious Attribute (SSA) (Nam et al., 2022) specifically trains the first model to predict groups using a small set of group labels, before using Group DRO to train a robust model. Contrastive Input Morphing (CIM) (Taghanaki et al., 2021) trains a network to transform the input features of an image to better present class-specific information shared across groups. Idrissi et al. (2021) suggest that simply changing the training data by subsampling large classes (SUBY) or balancing the class sampling probabilities (RWY), then training a model with ERM, can also improve group robustness.

E. Additional experimental results

E.1. Extended main results

In Table 11 and Table 12 we report group robustness results evaluating all methods discussed in Section 4 on all group robustness benchmarks. Table 11 contains results for image datasets, using CLIP-RN50 embeddings. Table 12 contains results for text datasets, using GPT-Neo 1.3B embeddings. As in Table 3, we report the worst-group and average accuracies, along with their gap. Higher worst-group accuracy and smaller accuracy gap are indicative of better group robustness. All results are computed over three random seeds, with mean and one standard deviation included (error bars deferred to here from the main paper). Compared to alternative methods, contrastive adapting consistently improves group robustness over zero-shot classification, and obtains highest worst-group accuracy and smallest accuracy gap on datasets where training adapters with ERM fails. On datasets where ERM-trained adapters achieve best group robustness, contrastive adapters are also competitive or closest to ERM-trained adapters among other robustness methods.

E.2. Contrastive adapter ablations

In this section, we ablate different training components of contrastive adapting. We show that the presented combination leads to best worst-group accuracy on the Waterbirds dataset. We also study how the number of positives and negatives used in contrastive sampling affects performance, and find that models do seem to benefit from a greater number of samples.

E.2.1. Training component ablations

We study the importance of the contrastive and cross-entropy components in contrastive adapting. For evaluation, we use the Waterbirds dataset, and run ablations comparing adapters trained on top of CLIP embeddings with (i) no contrastive component, (ii) no cross-entropy component, or the default proposed approach. We evaluate across five different CLIP models. All other training procedures are kept consistent.

In Table 9, we report worst-group accuracies. We find that both contrastive and cross-entropy components are necessary for best worst-group accuracy. The contrastive objective leads to a substantial improvement over just the resampled cross-entropy loss (+17.9 pp on average). However, we also note that without the cross-entropy objective to learn sample embeddings close to their ground-truth class embeddings, we observe high variance in classification accuracy. We improve +26.9 pp on average using both objectives compared to contrastive alone.

Table 9. Contrastive adapter training component ablation. For five CLIP models, we report the worst-group accuracy (%) on Waterbirds. Both contrastive and cross-entropy components are necessary for best worst-group accuracy. Without the cross-entropy objective to learn sample embeddings close to class embeddings (No cross-entropy), we observe high variance in classification accuracy.

Training Component Ablation	RN-50	RN-101	ViT-B/32	ViT-B/16	ViT-L/14
No contrastive	56.3 ± 1.5	68.8 ± 2.2	56.7 ± 2.4	70.2 ± 1.4	75.1 ± 1.0
No cross-entropy	60.7 ± 8.3	37.8 ± 12.0	23.1 ± 10.5	77.7 ± 2.9	82.4 ± 2.0
Default	$\mathbf{83.7 \pm 0.7}$	$\mathbf{82.0 \pm 1.3}$	$\mathbf{80.7 \pm 1.4}$	$\mathbf{83.1 \pm 2.1}$	$\mathbf{86.9 \pm 1.6}$

E.2.2. Effect of contrastive batch size

While one advantage of training adapters is that because we train on embeddings, the memory size of our data inputs during training is much smaller than the traditional alternative (*e.g*, storing an tensorized image). We can thus train with larger batch sizes. Here we study how contrastive batch size, *i.e* how many positives and negatives we sample per anchor, affects worst-group accuracy. On the Waterbirds and CelebA datasets and with CLIP RN-50 embeddings, we train a contrastive adapter with varying levels of positives and negatives. For both datasets, the default is 2048 positives and 2048 negatives per batch. We ablate these numbers with the following (positive, negative) combinations: (1, 1), (2, 2), (256, 256), (256, 512), (512, 256), (512, 512), (512, 1024), (1024, 512), (1024, 1024), (1024, 2048), (2048, 1024).

In Figure 7, we plot the effect of smaller batch sizes on worst-group accuracy. We find that larger batch sizes weakly correspond to higher worst-group accuracy on both Waterbirds and CelebA. However, perhaps surprisingly, we still maintain a substantial improvement over zero-shot classification with just a single positive and negative per anchor.

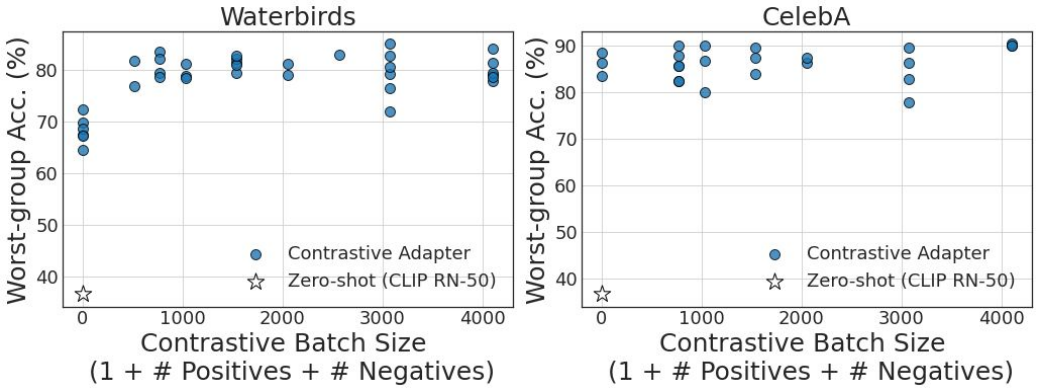

Figure 7. Effect of contrastive batch size on worst-group accuracy. With CLIP RN-50 embeddings, training contrastive adapters with larger batch sizes (greater number of positives and negatives) tends to help worst-group accuracy. However, even training with one positive and negative per batch leads to substantially greater worst-group accuracy than zero-shot classification.

E.3. Comparison to TIP-adapter and training sample nearest-neighbors lookup

On representative benchmarks, we perform further comparison to the nearest-neighbor look-up approach employed by TIP Adapter (Zhang et al., 2021). Instead of learning transformed representations of pretrained embeddings, another approach to better classify a test sample is to use the class of its nearest training sample. Under the assumption that the training and test data are sampled from the same broader distribution and share the same groups, then test samples in a given group should embed closest to training samples in the same group. The training sample ground-truth class should then apply to the test sample. TIP adapter operates accordingly, keeping a cache of training sample embeddings available at test-time. One advantage is this allows for potentially more accurate classification *without any* training. To test how well this idea fares for group robust classification, for each test sample we perform a look-up with *all* training samples, using cosine similarity to identify nearest neighbors.

In Table 10, we compare TIP-adapter with zero-shot classification and contrastive adapting on the Waterbirds, CelebA, BREEDS Living-17, and CIFAR-10.02 group robustness benchmarks. For all methods, we use CLIP RN-50 pretrained embeddings. We find that TIP adapter improves worst-group accuracy over zero-shot classification on 3 out of 4 datasets, and notably achieves best worst-group accuracy on BREEDS Living-17 without training any additional parameters. However, the improvements are more marginal on Waterbirds and CIFAR-10.02. Contrastive adapting still achieves 30.4 pp higher worst-group accuracy over TIP adapter on average. This may suggest that learning a nonlinear transformation of the pretrained embeddings can still be helpful for better "presenting" class-specific information to classify samples by.

Table 10. Group robustness comparison to nearest training sample look-up / TIP Adapter (Zhang et al., 2021). Across representative benchmarks, on average contrastive adapting achieves 30.4 pp higher worst-group accuracy than the nearest training sample look-up employed by TIP-adapter. This supports learning non-linear transformations of pretrained embeddings to better classify samples.

	Waterbirds			CelebA			BREEDS Living-17			CIFAR-10.02		
Acc (%)	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
Zero-shot (ZS)	36.6	92.2	55.6	74.0	81.9	7.9	6.0	86.7	80.7	39.1	69.9	30.8
TIP Adapter	39.9	93.9	54.0	19.4	91.1	71.7	**64.0**	90.7	**26.7**	51.5	75.4	23.9
Contrastive Adapter	**83.7**	89.4	**5.7**	**90.0**	90.7	**0.7**	62.0	90.9	28.9	**60.7**	80.9	**20.2**

E.4. Evaluation with respect to weight-space ensembling trade-off

To provide additional perspective on how different embedding-only methods trade-off worst-group and average accuracy, we compare how these methods perform with respect to the accuracy trade-off traced out by weight-space ensembles. Wortsman et al. (2021) show an interesting phenomenon where simply taking a weighted average of a trained linear probe and the original foundation model (either over the weights, or the outputs) can result in a "pareto frontier" of accuracy metrics. They specifically show that a weight-space ensemble can often achieve better OOD performance without sacrificing too much IID

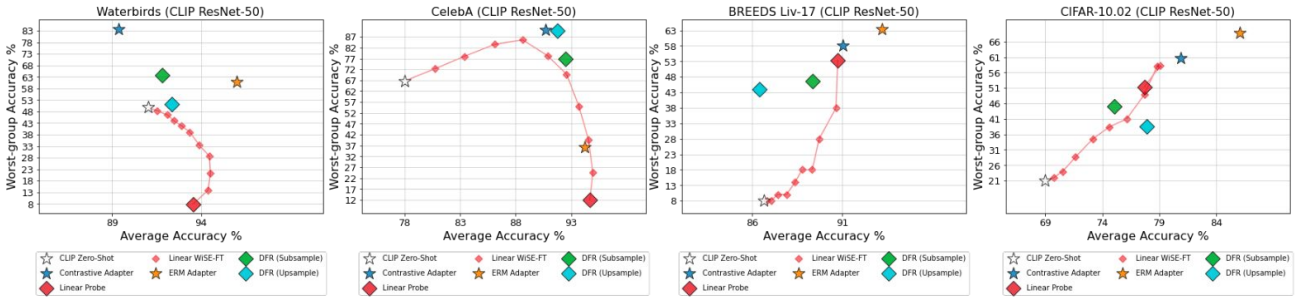

Figure 8. Plotting worst-group versus average accuracy trade-off against WiSE-FT ensemble (traced out) on representative datasets (Table 3). Contrastive adapters (dark blue stars) consistently achieve higher worst-group accuracy than weight-space ensembles.

performance. compared to a single linear probe. In this context, we see how this trade-off occurs over average accuracy and worst-group accuracy across our representative set of group-robustness datasets. We also evaluate how other approaches (ERM adapters, DFR (Kirichenko et al., 2022), contrastive adapters) fare along this trade-off.

In Figure 8, we plot the accuracies of these methods run on CLIP RN-50 embeddings. We note several observations. Weight-space ensembles (WiSE-FT) achieve the desired effect on certain datasets but not others. On CelebA and CIFAR-10.02, we find that an ensemble can obtain a better worst-group accuracy versus average accuracy trade-off than either zero-shot classification or linear probes. However, the single linear probe does at least as well as any ensemble in BREEDS Living-17, while the zero-shot classification does at least as well as any ensemble in Waterbirds.

We also find that among other methods, contrastive adapting is the only evaluated approach that consistently achieves higher worst-group accuracy than any weight-space ensemble. While contrastive adapting places "above" the trade-off curve traced out by WiSE-FT on 3 out of 4 datasets (CelebA, BREEDS Living-17, and CIFAR-10.02), it tends to degrade average performance in favor of higher worst-group performance compared to other approaches. Further work can improve how to raise worst-group performance without sacrificing any average performance.

F. Limitations and societal impact

Method limitations. While in this work, we demonstrated that we can substantially improve the group robustness of foundation model classification without any finetuning of the original model, several limitations still exist. First, this does not imply that we can get desirable performance in general without additional retraining. To obtain high worst-group performance in general, we are upper-bounded by whether the pretrained embeddings do contain the information needed to classify all groups. While our study suggests that in many cases they do carry this information—which can surprising given that the zero-shot classification with the same embeddings results in poor group robustness—in other situations the pretrained embeddings may lack this information. For example, if downstream task data is very different in distribution from the pretraining data, then the pretrained foundation model embeddings may not be sufficient to work with. While more efficient ways to improve robustness can democratize foundation model use, further finetuning may still be needed.

We also emphasize that our approach is a simple first-step method to improving the group robustness over existing baseline approaches. This is motivated by our observation that foundation model zero-shot classification may not be group robust, and that we would like to both (i) improve performance of these models when we realize they fail in certain aspects, and (ii) do so efficiently, such that fixing their failures is not bound by who can conduct costly retraining procedures, and when they can do so. We are excited for future work and expect further improvements as this thread of how to efficiently improve FM performance with limited access (*e.g*, only pretrained embeddings) is further explored.

Societal impact and related limitations. Finally, we note that it is important to carefully study the learned biases of foundation models, and to devise appropriate solutions evaluated outside of just computational metrics. Due to their promise of widespread and effective downstream transfer, foundation models may have a particularly strong impact on various parts of society. Individuals may get the sense that they can successfully apply these pretrained models to their desired downstream tasks "out-of-the-box". However, doing so also risks applying any learned biases of the model. Our work raises this issue with respect to group robustness as motivation for our problem setting, and also notes that additional evaluation beyond average accuracy can shed light on the negative qualities of existing models (where zero-shot FM classification may perform very well on average, but very poorly on certain groups, c.f. Figure 5). However we recognize the limitations of purely computational solutions to addressing group performance disparities in society. We also recognize the need to better understand foundation models and their potential uses in broader socio-technical systems (Bommasani et al., 2021).

Table 11. Worst-group (WG) and average (Avg) accuracies (in %) for zero-shot and efficient methods to improve CLIP-RN50 inference. **1st** / 2nd highest WG acc. and **1st** / 2nd smallest accuracy gap **bolded** / underlined respectively.

Waterbirds								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	49.8 ± 0.0	55.9 ± 0.0	7.9 ± 1.0	60.8 ± 0.9	49.8 ± 0.0	51.3 ± 1.4	63.9 ± 1.5	$\mathbf{83.7 \pm 0.7}$
Avg.	91.0 ± 0.0	87.8 ± 0.0	93.5 ± 0.1	96.0 ± 0.1	91.0 ± 0.0	92.4 ± 0.1	91.8 ± 3.1	89.4 ± 0.9
Gap	41.2	31.9	85.6	35.2	41.2	41.1	27.9	**5.7**
CelebA								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	74.0 ± 0.0	70.8	11.9 ± 0.3	36.1 ± 1.4	85.6 ± 0.0	76.9 ± 1.4	89.6 ± 0.3	$\mathbf{90.0 \pm 0.4}$
Avg.	81.9 ± 0.0	82.6	94.7 ± 0.0	94.2 ± 0.2	88.6 ± 0.0	92.5 ± 0.2	91.8 ± 0.1	90.7 ± 0.0
Gap	7.9	11.8	82.8	58.1	3.0	15.6	2.2	**0.7**
BREEDS Living-17								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	6.0 ± 0.0	30.0 ± 0.0	53.3 ± 0.9	$\mathbf{70.7 \pm 0.9}$	53.3 ± 0.9	46.7 ± 3.4	44.0 ± 0.0	62.0 ± 1.6
Avg	86.7 ± 0.0	90.6 ± 0.0	90.8 ± 0.0	93.9 ± 0.1	90.8 ± 0.0	89.3 ± 0.3	86.4 ± 0.0	90.9 ± 0.3
Gap	80.7	60.6	37.5	**23.2**	37.5	42.6	42.4	28.9
BREEDS Nonliving-26								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	6.0 ± 0.0	56.0 ± 0.0	32.0 ± 0.0	$\mathbf{61.3 \pm 1.9}$	36.7 ± 0.9	29.3 ± 1.9	30.0 ± 4.1	55.3 ± 4.2
Avg	72.3 ± 0.0	87.1 ± 0.0	82.3 ± 0.1	92.1 ± 0.2	83.6 ± 0.1	80.6 ± 0.1	83.6 ± 0.0	88.1 ± 0.6
Gap	66.3	31.1	50.3	**30.8**	46.9	51.3	53.6	32.8
CIFAR-10.001								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	31.4 ± 0.0	N/A	44.0 ± 1.4	$\mathbf{68.2 \pm 3.5}$	53.3 ± 0.0	18.1 ± 4.3	45.0 ± 1.6	59.7 ± 4.1
Avg	69.8 ± 0.0	N/A	75.2 ± 0.2	87.3 ± 0.3	81.1 ± 0.0	58.7 ± 1.7	78.3 ± 0.1	82.0 ± 0.1
Gap	38.4	N/A	31.2	**19.1**	27.8	40.6	33.3	22.3
CIFAR-10.02								
Acc.	Zero-Shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	39.1 ± 0.0	N/A	51.3 ± 0.2	$\mathbf{68.8 \pm 0.5}$	58.2 ± 0.2	45.0 ± 0.8	38.5 ± 2.1	60.7 ± 1.7
Avg	69.9 ± 0.0	N/A	77.7 ± 0.1	86.0 ± 0.5	79.1 ± 0.0	75.0 ± 0.3	77.9 ± 0.5	80.9 ± 0.2
Gap	48	N/A	26.4	**17.2**	20.9	30.0	39.4	20.2
FMoW-WILDS								
Acc.	Zero-shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	10.5 ± 0.0	-	21.6 ± 0.1	$\mathbf{41.3 \pm 0.5}$	21.6 ± 0.1	6.8 ± 0.6	27.0 ± 0.2	39.2 ± 0.7
Avg	13.2 ± 0.0	-	24.1 ± 0.1	43.6 ± 0.5	24.1 ± 0.1	10.2 ± 0.5	28.7 ± 0.2	41.9 ± 0.1
Gap	2.7	-	2.5	2.3	2.5	3.4	**1.7**	2.7

Table 12. Worst-group (WG) and average (Avg) accuracies (in %) for zero-shot and efficient methods to improve GPT-Neo 1.3B inference. **1st** / 2nd highest WG acc. and **1st** / 2nd smallest accuracy gap **bolded** / underlined respectively.

Amazon-WILDS								
Acc.	Zero-shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	79.4 ± 0.0	N/A	87.2 ± 0.3	87.2 ± 0.3	87.2 ± 0.3	87.2 ± 0.3	85.4 ± 0.8	$\mathbf{87.9 \pm 1.1}$
Avg	86.7 ± 0.0	N/A	93.3 ± 0.2	93.6 ± 0.1	93.3 ± 0.2	93.2 ± 0.3	92.7 ± 0.7	92.6 ± 0.8
Gap	7.3	N/A	6.1	6.4	6.1	6.0	7.3	**4.7**
CivilComments-WILDS								
Acc.	Zero-shot	Group Prompt	ERM LP	ERM Adapter	WiSE-FT	DFR (Sub)	DFR (Up)	**Contrastive Adapter**
WG	16.0 ± 0.0	N/A	46.7 ± 2.0	32.1 ± 1.5	46.7 ± 2.0	47.4 ± 0.9	48.2 ± 1.3	$\mathbf{50.1 \pm 1.5}$
Avg	74.8 ± 0.0	N/A	51.2 ± 0.26	37.7 ± 0.7	51.2 ± 0.26	51.9 ± 0.8	52.1 ± 1.3	54.2 ± 0.5
Gap	58.8	N/A	4.5	5.6	4.5	4.5	**3.9**	4.1