Interactive Model Correction with Natural Language

Yoonho Lee
Stanford University

Michelle Lam
Stanford University

Helena Vasconcelos
Stanford University

Michael S. Bernstein
Stanford University

Chelsea Finn
Stanford University

Abstract

In supervised learning, models are trained to extract correlations from a static dataset. This often leads to models that rely on spurious correlations that fail to generalize to new data distributions, such as a bird classifier that relies on the background of an image. Preventing models from latching on to spurious correlations necessarily requires additional information beyond labeled data. Existing methods incorporate forms of additional instance-level supervision, such as labels for spurious features or additional labeled data from a balanced distribution. Such strategies can become prohibitively costly for large-scale datasets since they require additional annotation at a scale close to the original training data. We hypothesize that far less supervision suffices if we provide targeted feedback about the misconceptions of models trained on a given dataset. We introduce CLARIFY, a novel natural language interface and method for interactively correcting model misconceptions. Through CLARIFY, users need only provide a short text description to describe a model's consistent failure patterns, such as "water background" for a bird classifier. Then, in an entirely automated way, we use such descriptions to improve the training process by reweighting the training data or gathering additional targeted data. Our empirical results show that non-expert users can successfully describe model misconceptions via CLARIFY, improving worst-group accuracy by an average of 7.3% in two datasets with spurious correlations. Finally, we use CLARIFY to find and rectify 31 novel spurious correlations in ImageNet, improving minority-split accuracy from 21.1% to 28.7%.

1 Introduction

Supervised learning fundamentally hinges on the premise of extracting correlations from labeled data to make predictions on new inputs. While effective in controlled environments, this paradigm often leads to models that are brittle in real-world conditions. This is because some correlations in the training data may be *spurious*, i.e. they may no longer hold in conditions we would like models to generalize to. To steer models away from the spurious correlations in a given dataset and towards reliable prediction rules, we must necessarily provide additional information beyond the original labeled data. However, this task has traditionally been labor-intensive due to the need to gather instance-level annotations, such as labels for spurious features (e.g., labeling each training datapoint by background category) or additional labeled data (e.g., gathering data where background and bird species is not correlated). These annotations are needed at a scale comparable to that of the original training data, making such strategies prohibitively costly for settings where the original training data is already close to the full annotation budget. We posit that far less supervision suffices if we provide *targeted concept-level feedback* about misconceptions of models trained on a given dataset.

We introduce Corrective Language Annotations for Robust InFerence (CLARIFY), a novel framework that allows humans to interactively correct failures of image classifiers with natural language

I Can't Believe it's Not Better: Failure Modes in the Age of Foundation Models @ NeurIPS 2023

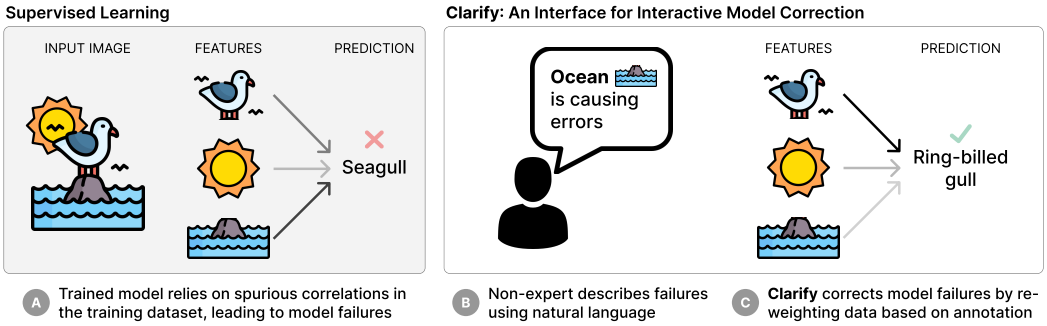


Figure 1: CLARIFY is an interface for interactively correcting model failures due to spurious correlations. (a) Given a model trained with supervised learning, (b) a human describes consistent failure modes of the model entirely in natural language. (c) We automatically incorporate these descriptions to improve the training process by reweighting the training data based on image-text similarity.

alone. CLARIFY consists of a web interface for collecting human feedback and a method for automatically incorporating this feedback into the training process. During interactions with the interface, users observe a trained model's predictions on a validation dataset, and write short text descriptions of consistent model misconceptions. For instance, for a bird classifier relying on a spurious correlation between bird species and their backgrounds, a human user can succinctly write that the model is mistakenly focusing on the "water background". We then use such textual feedback to improve the training process by reweighting the training data.

We highlight two advantageous ways in which the CLARIFY framework diverges substantially from standard supervised learning. First, in CLARIFY, annotations are collected *after* initial training, allowing the model's inductive biases to inform the annotation process. Specifically, CLARIFY focuses on eliciting *negative knowledge*, i.e., directing the model on what *not* to focus on; this is a core design decision since it is easier for humans to identify errors than to fully articulate complex rules. This negative knowledge serves as a complementary form of guidance to the positive knowledge in the original labeled dataset. Second, annotations from CLARIFY have a substantially higher density of information than conventional forms of annotations. Unlike instance-specific labels, textual feedback encapsulates concept-level "global" insights that are applicable across the entire dataset. Therefore, they more efficiently use the human effort required for annotation.

We instantiate CLARIFY in a web app implementation to carry out online experiments with non-expert users, and evaluate the gathered textual feedback in addition to the robustness of models fine-tuned based on them. We refer the reader to Figure 1 for an overview of CLARIFY in relation to traditional supervised learning, and Figure 2 for a visualization of key interface features. We find that non-expert feedback through CLARIFY (N=26) almost always helps in identifying a spurious correlation or difficult subpopulation. Models fine-tuned using these non-expert annotations consistently outperform zero-shot methods that use oracle text annotations of spurious features, achieving a 7.0-7.6 point improvement in worst-group accuracy on two datasets. Users achieve these performance gains with just a few minutes of interaction, averaging 2.7 minutes per dataset. A key advantage of the CLARIFY framework is its scalability, which we demonstrate by using the interface to identify 31 novel spurious correlations in the ImageNet dataset. We use these annotations to improve the average accuracy across the 31 minority splits from 21.1% to 28.7% with only a 0.21% drop in overall average accuracy, just by appropriately reweighting the ImageNet training set.

2 Problem Setup

We consider a standard supervised learning setting, where we are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ of N labeled samples. Each label y_i belongs to one of C different classes: $y_i \in \{1, \ldots, C\}$. A model is trained to minimize the average loss across the training set, i.e. $\frac{1}{N}\sum_{i=1}^N \ell(f(x_i;\theta), y_i)$, where ℓ is a pointwise loss function such as cross-entropy, f is the model, and θ denotes model parameters. To formalize spurious correlations, we can consider an extended dataset that includes an unknown attribute s_i for each instance, resulting in $\{(x_i, y_i, s_i)\}_{i=1}^N$ where $s_i \in \{1, \ldots, S\}$. A model trained on $\mathcal{D}$ may learn to rely on s_i to make predictions, thereby failing on new distributions where the previous correlation between s_i and y_i no longer holds. In general, we do not have annotations for these spurious attributes s_i or even know in advance what they are. Our goal is to correct the model's reliance on these spurious attributes without knowing a priori what they are.

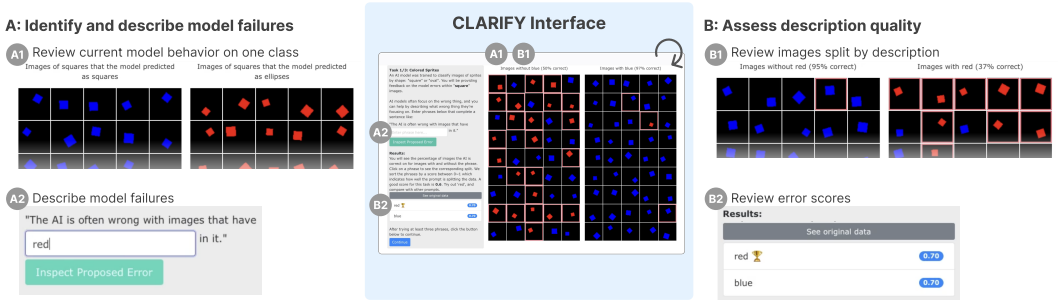


Figure 2: The CLARIFY interface enables users to iteratively (A) identify and describe model failures and (B) assess the quality of these descriptions. Users can review image examples of correct and incorrect predictions on one class, such as "square" (A1). Based on observed differences, they can input short, natural language descriptions of model failures, such as "red" squares (A2). The system surfaces feedback by splitting the data using the provided description (B1) and displaying an error score (B2). Users can repeat the process to generate improved descriptions.

To describe spurious attributes given only class-labeled image data, we leverage the capabilities of multimodal models such as CLIP [41], which encodes images and text into a shared embedding space. For a given image input I and text input T, CLIP outputs representations from seperate vision and language branches, $e_i = f_i(I)$ and $e_t = f_t(T)$ respectively. This model is trained to maximize the similarity between the image and text representations for corresponding image-text pairs and minimize it for non-corresponding pairs, through a contrastive loss function. We can estimate the similarity between a pair of image and text inputs by computing the cosine similarity of their respective representations:

$$\text{sim}(I, T) = \frac{e_i \cdot e_t}{\|e_i\| \|e_t\|}. \tag{1}$$

This black-box similarity function allows us to determine the relevance of a given image and text pair. In the next section, we describe how CLARIFY leverages this relevance function to mitigate spurious correlations based solely on natural language feedback on a labeled validation set.

3 CLARIFY: A Natural Language Interface for Model Correction

We now describe Corrective Language Annotations for Robust InFerence (CLARIFY), a novel framework for identifying and mitigating spurious correlations in models trained with supervised learning. The main idea behind CLARIFY is to allow humans to provide targeted natural language feedback to a model, helping the model focus on relevant features and ignore spurious ones. We employ a natural language interface to facilitate this process, which we describe in detail in this section. We will first describe a concrete example of an interaction with the interface in Section 3.1, and then describe two methods for incorporating this feedback into the training process in Section 3.2.

3.1 Interaction Workflow

User interaction. To demonstrate how CLARIFY enables non-expert users to correct model misconceptions, we will walk through a user's experience with the system, shown in Figure 2. CLARIFY takes as input a classification model trained with supervised learning. Here, we use an example of a model trained to classify images of sprites as squares or ovals.

Reviewing model behavior. First, the user is presented with a summary view of the model's current behavior. The goal of this interface is to scaffold the user in rapidly identifying reasons underlying model failures. Drawing from a validation dataset, we display one class at a time (i.e., images of squares) and divide the examples into those that the model correctly classified (i.e., images classified as squares) on the left versus those that it incorrectly classified (i.e., images classified as ovals) on the right (Figure 2, A1). By presenting the images in this way, CLARIFY streamlines the user's task to one of identifying differences between sets. In our example, all of the images on the page are indeed squares, but the model is only making accurate predictions for the examples on the left and not those on the right. Comparing the images on the two sides, the user notices that the correct cases contain blue squares while the incorrect cases contain red squares.

Describing model failures. Once the user has an initial idea of the model's misconception, they are tasked with describing this failure mode. Our system accepts short text descriptions of model failures

(Figure 2, A2). In particular, users are asked to complete the following fill-in-the-blank sentence: "The AI is often wrong on images that have ___ in it." We find that this question is effective since users may not be familiar with the general concept of spurious correlations. Continuing our running example, the user enters the phrase "red" here to describe what they observed.

Assessing descriptions. After the user submits their description, CLARIFY helps them to assess whether the description can successfully correct model errors. First, we provide a $0-1$ *Error Score* that indicates how well the description separates the error cases from the correct predictions (Figure 2, B2). Then, we present a summary visualization that partitions the validation dataset based on a threshold on image-text similarity from CLIP (Figure 2, B1). Images sufficiently similar to the description ("red") are on the right while others are on the left. For our example, the user sees an Error Score of 0.70, and they see a set of images *without red* on the left and a set of images *with red* on the right. This high Error Score indicates that they successfully achieved a strong level of separation, and they see in the image view that most of the red squares were captured by their description. We note that while the interface only shows *validation data* using the provided description, the user's natural language annotation will later be incorporated to partition the *training data* for model retraining.

Iterating on descriptions. However, users may not always be so successful on their first attempt, so CLARIFY aids users in iterating on their descriptions. Descriptions can fail for two reasons: (1) the description may not indeed differentiate the correct and incorrect cases, or (2) the description may be a valid differentiator, but may not be modeled accurately due to the user's word choice and CLIP-based similarity scoring. CLARIFY allows users to identify both of these failure modes. In our example, the user can see if the model is not accurately identifying images with the "red" keyword (case 2), and they can experiment with alternate rewordings to better isolate the difference (e.g., "red square," "crimson"). After iterating and isolating the red examples, the user can see if the Error Score is still low, indicating that this description is not sufficient to repair model errors (case 1). With this information, they can revisit the original view and brainstorm additional descriptions, like phrases related to the size and orientation of sprites. We give additional details about the interface, including the error score, similarity threshold, and modifications for ImageNet, in Appendix A

3.2 Automatic Fine-Tuning

After collecting textual feedback from users, we incorporate this feedback into the training process for fine-tuning a foundation model. In this paper, we consider fine-tuning only the last layer on top of a frozen backbone, though the strategy is more general. Given an error annotation (c, T, τ), we can partition the training data within class c into two subsets: $D_> = \{(x_i, y_i) \mid \text{sim}(x_i, T) > \tau\}$ and $D_< = \{(x_i, y_i) \mid \text{sim}(x_i, T) \leq \tau\}$. These two subsets correspond to images that are more and less similar to the provided text prompt, respectively, and serve as indicators of the spurious attribute identified by the annotator. Having identified these two subsets, we want to train a final model to achieve low training loss while *not* using the feature that separates the two subsets.

We propose to adjust the loss weights for each subset so that their total weights are balanced:

$$w_i = \begin{cases} \frac{1}{C\|D_>\|} & \text{if } (x_i, y_i) \in D_> \\ \frac{1}{C\|D_<\|} & \text{if } (x_i, y_i) \in D_< \end{cases}. \tag{2}$$

This weight balancing discourages the model from exploiting the spurious attribute for prediction by reducing the statistical correlation between the spurious attribute and the class label in the training data. For classes without any error annotations, we use uniform weights during training as in standard supervised learning. Given such weights over the training dataset, we train the last layer with a weighted cross-entropy loss. In Section 4, we will measure the effectiveness of this fine-tuning approach based on language feedback. We note that this stage is fully automated, and there are no additional hyperparameters to tune beyond what was in the original training process.

4 Experiments

We first note that our framework diverges substantially from assumptions in traditional supervised learning. CLARIFY involves collecting annotations *after* an initial round of training, and these annotations consist of targeted concept-level feedback rather than model-agnostic instance-level feedback. We consider this deviation from the conventional setup as necessary for efficiently addressing

Table 1: Evaluation of methods for group robustness using the CLIP-ResNet50 backbone. Fine-tuning with annotations from CLARIFY consistently outperforms methods that use only text (zero-shot) or label information. We denote our implementation of other methods with (ours); all other results are from **(author?)** [58].

		Waterbirds			CelebA		
Data Assumptions	Method	WG	Avg	Gap	WG	Avg	Gap
Zero-Shot	Class Prompt	36.6	92.2	55.6	74.0	81.9	7.9
	Group Prompt	55.9	87.8	31.9	70.8	82.6	11.8
Labels	ERM	7.9	93.5	85.6	11.9	94.7	82.8
	ERM (ours)	63.4	96.0	32.6	31.1	95.4	64.3
	ERM (ours, class-balanced)	48.6	95.2	46.7	65.8	93.4	27.6
	ERM (ours, worst-class)	55.9	95.8	39.9	56.9	94.1	37.2
Labels, Text Feedback	CLARIFY (slice-balanced)	68.4	93.6	25.2	**89.3**	92.2	**2.8**
	CLARIFY (worst-slice)	**75.7**	83.8	**8.1**	89.1	92.1	3.0
Labels, Group Labels	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
	DFR (ours)	78.7	90.8	12.1	**90.6**	91.9	**1.3**
	Group DRO (ours)	**81.3**	88.1	**6.8**	89.2	91.8	2.7

the challenge of learning robust prediction rules from observational data. We seek to empirically answer the following questions about this framework for interactively correcting model errors:

1. How does re-training with annotations from CLARIFY compare to automated methods for addressing spurious correlations?
2. Can non-expert users use CLARIFY to identify and describe spurious correlations in models trained with supervised learning?
3. Can CLARIFY discover and rectify novel spurious correlations in ImageNet?

For detailed experimental setup including datasets, models, and human participants, see Appendix D.

4.1 Comparison With Automated Methods

We assess how re-training a model with expert annotations from CLARIFY compares to existing automated methods for addressing spurious correlations. We compare with representative prior methods which similarly fine-tune CLIP backbones and/or reweight training data. We conduct experiments on the Waterbirds and CelebA datasets. In addition to CLARIFY, we evaluate zero-shot CLIP [41] with class-based and group-based prompts, DFR [25], and Group DRO [44]. We desribe experimental details for each method in Appendix D. Our results, summarized in Table 1, demonstrate that CLARIFY consistently outperforms approaches that that use zero-shot prompts or class labels in terms of worst-group accuracy and robustness gaps. On this experiment, CLARIFY underperforms specialized methods on Waterbirds and is competitive on CelebA, while using considerably cheaper supervision. We also show comparisons with zero-shot classification methods in in Table 2, where CLARIFY shows substantially better worst-group accuracy and robustness gap on the Waterbirds and CelebA datasets. Among these points of comparison, RoboShot [1] is notable as it is an automated method that leverages state-of-the-art foundation models such as ALIGN [22], AltCLIP [6], and GPT-4 [36]. This result demonstrates that retraining with feedback from CLARIFY is a substantially more effective way of using natural language to guide an image classifier.

We emphasize that these experiments do not aim to conduct a head-to-head comparison with the best automated methods for addressing spurious correlations. The body of work on automated spurious correlations is large [44, 57, 35, 7, 31, 25, 40], and these methods are often designed for specific benchmarks including the Waterbirds and CelebA datasets. Instead, our primary goal is to show that CLARIFY, with minimal human supervision and no additional hyperparameter tuning, can yield results that yield benefits comparable with prior methods. We also note that prior methods often require a substantial amount of additional supervision, such as instance-level annotation for spurious attributes for either training or hyperparameter tuning, which CLARIFY does not require. Moreover, the key advantage of CLARIFY is in its scalability to large datasets, a feature that no prior automated method has demonstrated. Such scalability is crucial when applying these ideas to real-world problems, where the scale and diversity of data are ever-increasing. We will elaborate on and provide empirical evidence for the scalability of CLARIFY in Section 4.3.

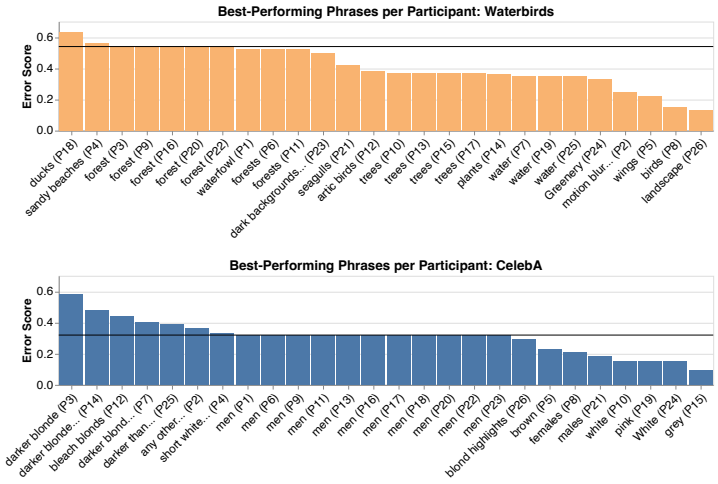


Figure 3: Non-experts used CLARIFY to identify high-quality descriptions with Error Scores that matched or exceeded the authors' expert annotations.

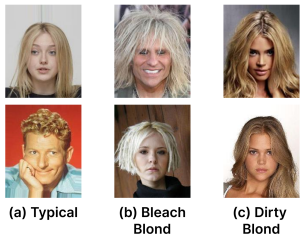


Figure 4: (a) Typical images from the "blond" class of CelebA. Non-experts provided textual feedback corresponding to hard subpopulations of (b) lighter and (c) darker hair colors.

Figure 5: An example of a spurious correlation found on ImageNet. Within the "sliding door" class, the model successfully classifies (a) images of sliding doors inside buildings. However, it is wrong on all instances of (b) sliding doors on cars. This is one of the 31 spurious correlations we found; please refer to Figure 9 in the appendix for more visualizations.

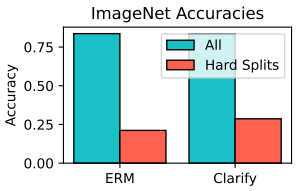


Figure 6: Average accuracies on ImageNet data. Fine-tuning with Clarify substantially improves accuracy on hard splits, while keeping overall accuracy intact.

4.2 Non-Expert Annotators Can Describe Model Errors

Identifying and annotating spurious correlations is a more nuanced task than conventional forms of annotation such as class labeling. This raises the question of whether non-expert annotators can perform this task. To answer this question, we conduct a user study (N=26) to assess the ability of non-expert users to identify and describe spurious correlations in models trained with supervised learning (see Appendix D for study details). We ask each participant to annotate the Waterbirds and CelebA datasets using the CLARIFY interface, and we summarize our results in Figures 3 and 7.

Taking the best-performing annotation from each user, the average worst-group accuracy was 63.5 (SD=4.7, max=69.5) for the Waterbirds dataset and 78.8 (SD=2.9, max=82.8) for the CelebA dataset. These results all exceed Class Prompt, Group Prompt, and ERM (best) baselines (Figure 7b). Promisingly, users were able to achieve these performance improvements with minimal annotation effort, averaging 2.7 minutes (SD=2.5) per dataset (Figure 7a). Overall, non-experts appeared proficient at this annotation task. For the Waterbirds dataset, the authors' expert annotation of "forest" achieved a 0.54 Error Score. In comparison, the best-performing Error Score for non-expert users was 0.41 on average (SD=0.13), and one participant achieved as high as 0.63. For the CelebA dataset, the expert annotation of "man" achieved a 0.32 Error Score. Across non-expert users, the best-performing Error Score averaged 0.31 (SD=0.11), and the highest Error Score was 0.58.

We also find that non-expert annotators propose novel model failures that had not been previously surfaced by experts. While experts had surfaced spurious correlations with gender in the CelebA dataset, participants also surfaced "dirty blonde" and "bleach blond" subpopulations, which achieved higher Error Scores than the "man" subpopulation (Figure 4). Our findings suggest that CLARIFY can enable non-expert annotators to identify and describe spurious correlations in models trained with supervised learning. This opens up the possibility of leveraging a broader workforce for annotating and mitigating spurious correlations in web-scale datasets such as ImageNet or LAION [8, 46].

4.3 Discovering and Mitigating Spurious Correlations in ImageNet

We now evaluate whether CLARIFY can be used to discover novel spurious correlations in models trained on the ImageNet training set. For such widely used large-scale datasets, it is important to develop both tools to find spurious correlations and methods to mitigate their effect. For this evaluation, the authors of this paper use the CLARIFY interface to identify spurious correlations in ImageNet, and additionally evaluate whether the resulting annotations can improve model robustness.

Discovered spurious correlations in ImageNet. Using CLARIFY, we identified 31 spurious correlations in ImageNet; we show a full list in Table 5. To our best knowledge, no prior works have identified these spurious correlations, despite ImageNet being a widely studied dataset. As an example, we visualize a spurious correlation in the "sliding door" class in Figure 5. Here, sliding doors are negatively correlated with cars in the training set, causing standard models to misclassify cars that have sliding doors. We visualize more such spurious correlations in Figure 9. We evaluate the performance of a standard ERM model trained on the ImageNet training set on each identified minority and majority split. Results in Figure 10 show that the ERM model consistently underperforms on the minority split for each class, indicating that the trained model is relying on each of these spurious correlations. We also note that this trend continues to hold on ImageNet-V2, which follows a different distribution from the validation set we use during interactions.

Fine-tuning while avoiding spurious correlations. We use the collected annotations to fine-tune a model on ImageNet, and evaluate this fine-tuned model on various splits of the ImageNet validation set. Results in Figure 8 show that the retrained model achieves higher minority split performance on many classes. Aggregate metrics in Figure 6 show that fine-tuning with CLARIFY annotations reduces the average minority-split accuracy from 21.1% to 28.7%, with only a 0.2% drop in overall average accuracy. We emphasize that no additional data was used during fine-tuning—the annotations from CLARIFY were only used to find a better reweighting of the exact same training data used to train the original ERM model.

5 Discussion

A limitation of CLARIFY is that it depends heavily on human knowledge and a model of text-image similarity, and is not straightforward to apply when either of these is limited, such as medical imaging or scientific domains. Future work could broaden the applicability of the CLARIFY framework by incorporating domain-specific knowledge or extending to data modalities beyond images. CLARIFY contributes to the democratization of machine learning by allowing laypeople to correct concept-level errors stemming from spurious correlations in data. This feature can potentially foster greater public trust, especially when users witness measurable improvements in the model after their interventions.

References

[1] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models with foundation models. *arXiv preprint arXiv:2309.04344*, 2023. [page 5, 14]

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017. [page 13]

[3] Yujia Bao and Regina Barzilay. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*, 2022. [page 13]

[4] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022. [page 12]

[5] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *International conference on machine learning*, pages 1617–1629. PMLR, 2021. [page 13]

[6] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022. [page 5]

[7] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021. [page 5, 13]

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [page 6, 13]

[9] Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022. [page 13]

[10] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602, 2008. [page 12]

[11] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018. [page 12]

[12] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2022. [page 13]

[13] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022. [page 13]

[14] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003. [page 12]

[15] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019. [page 13]

[16] Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. *arXiv preprint arXiv:2212.02774*, 2022. [page 12]

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. [page 13]

[18] Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, COLT '91, page 303314, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. [page 12]

[19] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. [page 12]

[20] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2018. [page 13]

[21] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. [page 13]

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. [page 5]

[23] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019. [page 13]

[24] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023. [page 12, 13]

[25] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. [page 5, 13, 14]

[26] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. Model sketching: Centering concepts in early-stage machine learning model design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2023. [page 13]

[27] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022. [page 13]

[28] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995. [page 12]

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. [page 12]

[30] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. 2022. [page 13]

[31] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. [page 5, 13]

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. [page 13]

[33] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009. [page 12]

[34] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021. [page 13]

[35] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. [page 5, 13]

[36] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. [page 5]

[37] Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*, 2022. [page 13]

[38] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. [page 13]

[39] Michael Polanyi. The tacit dimension. In *Knowledge in organisations*, pages 135–146. Routledge, 2009. [page 13]

[40] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. *arXiv preprint arXiv:2306.11074*, 2023. [page 5, 13]

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [page 3, 5, 13]

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. [page 12]

[43] Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*, 2021. [page 13]

[44] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. [page 5, 13]

[45] Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. *Advances in Neural Information Processing Systems*, 34:23359–23373, 2021. [page 13]

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. [page 6]

[47] Burr Settles. Active learning literature survey. 2009. [page 12]

[48] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Conference on Neural Information Processing Systems*, 2020. [page 13]

[49] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742*, 2017. [page 12]

[50] Otilia Stretcu, Edward Vendrow, Kenji Hata, Krishnamurthy Viswanathan, Vittorio Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Enming Luo, Neil Gordon Alldrin, et al. Agile modeling: Image classification with domain experts in the loop. *arXiv preprint arXiv:2302.12948*, 2023. [page 13]

[51] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022. [page 13]

[52] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021. [page 13]

[53] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023. [page 13]

[54] Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. *arXiv preprint arXiv:2208.08831*, 2022. [page 13]

[55] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021. [page 13]

[56] Yilun Xu, Hao He, Tianxiao Shen, and Tommi Jaakkola. Controlling directions orthogonal to a classifier. *arXiv preprint arXiv:2201.11259*, 2022. [page 13]

[57] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*, 2019. [page 5, 13]

[58] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022. [page 5, 14]

[59] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023. [page 13]

[60] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. [page 12]

[61] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018. [page 12]

A Additional Interface Details

Error score. We now describe how we calculate the Error Score shown in the CLARIFY interface to guide description generation. Consider input text prompt T, and let D_{correct} and D_{error} be subsets of the validation dataset for a given class that the model made correct and incorrect predictions on, respectively. We denote the similarities between the T and the images in each subset as $S_{\text{correct}} = \{\text{sim}(I, T) \mid I \in D_{\text{correct}}\}$ and $S_{\text{error}} = \{\text{sim}(I, T) \mid I \in D_{\text{error}}\}$. To quantify how well image similarity with T predicts model errors, we compute the "Error Score" as the highest achievable binary classification accuracy from a threshold τ on the similarity:

$$\text{Error Score} = \max\left(\max_{\tau}\left(\frac{|\{s \mid s > \tau, s \in S^c_{\text{correct}}\}| + |\{s \mid s \leq \tau, s \in S^c_{\text{error}}\}|}{|S^c_{\text{correct}}| + |S^c_{\text{error}}|}\right),\right.$$
$$\left.\max_{\tau}\left(\frac{|\{s \mid s \leq \tau, s \in S^c_{\text{correct}}\}| + |\{s \mid s > \tau, s \in S^c_{\text{error}}\}|}{|S^c_{\text{correct}}| + |S^c_{\text{error}}|}\right)\right). \quad (3)$$

As mentioned earlier, showing this score during interactions allows users to quickly assess the quality of their descriptions.

Similarity threshold. Finally, for each natural language threshold, we determine a similarity threshold τ, which can be chosen by the user after inspecting the similarity scores for a representative sample of images, or can be automatically chosen as the threshold that maximizes the Error Score. For each class, only the textual feedback with the highest Error Score is used for retraining. Together with this threshold, we can specify a spurious correlation using a tuple of the form (class label, text prompt, similarity threshold), which corresponds to a binary classifier that is predictive of model errors on that class.

Additional features for large datasets. We found that a few more optional features are helpful for annotating spurious correlations ImageNet, and expect that these features will similarly be helpful for other datasets. We begin by narrowing down the 1000 classes to 100 classes that are most likely to have identifiable spurious correlations. To do so, we first prune out classes with too low or too high accuracy (i.e. accuracy below 0.2 or above 0.8), to ensure a sufficient number of correct and incorrect predictions for each class. For the remaining classes, we caption each image with an image captioning model [29, BLIP] and use a keyword extraction model [19, KeyBERT] to suggest a pool of up to 50 keywords for each class, a procedure inspired by **(author?)** [24]. Through CLARIFY, we perform interactions with the top 100 classes according to maximum Error Score (3) across the candidate keywords. During interactions, the user is shown the top 10 candidate keywords as a helpful starting point.

B Related Work

Teaching ML models. As machine learning models require more and more resources to train, it becomes increasingly important to optimize the training process. The machine teaching problem setting aims to formalize what an optimal training set for a given task is and characterize the so-called training complexity. While this setting has been well-studied [18, 10, 33, 60, 49, 61], its application to large-scale models has been limited. Supervised learning, the dominant paradigm for training task-specific models, requires explicit labels for each instance in the dataset, which is often large and expensive to collect. Although active learning methods aim to reduce this annotation burden by selecting the most informative datapoints for labeling [28, 47], they still require humans to label individual datapoints. Our work proposes a form of supervision which can be used to find and rectify spurious correlations in labeled datasets: natural language descriptions of model errors. Such textual feedback is immediately useful since it describes failure modes that the model would otherwise fall into. Compared to labels, these descriptions hold substantially more information per annotation, as they hold global information about the model's behavior on the entire dataset, rather than just a single datapoint.

Human-computer interaction for ML. There is also a rich literature on the interaction between humans and machine learning models. Improving the interface between humans and models has benefits in many points of the machine learning pipeline, including interactive feature selection [14, 11], interpretability [42, 4], and human-in-the-loop data collection [16]. Perhaps the most closely related

works are those that elicit high-level concepts from humans [50, 26]. However, a key difference between these works and ours is that we focus on *negative knowledge*—teaching the model what not to learn—as opposed to these works which specify what features the model should use. Especially for intuitive tasks like image classification, human knowledge is often *tacit* rather than explicit, making it hard to define and put into words [39]; thus, it is easier for annotators to describe the failures of an existing model rather than define its behavior upfront. Restricting the feedback to negative knowledge is also important for scalability, as it is much easier to identify a few failure modes in an otherwise well-performing model, than to specify the full set of useful concepts. This scalability is crucial for our goal of correcting spurious correlations in large-scale datasets such as ImageNet.

Robustness to spurious correlations. Models trained with standard supervised learning often exhibit a bias towards shortcut features, i.e. simple features that perform well on the training distribution yet fail to capture the underlying causal structure [2, 20, 48, 17, 38, 30]. Many recent works have proposed methods to mitigate this issue, such as learning multiple functions consistent with the training data [15, 56, 51, 37, 27], and reweighting training instances to render shortcut features non-predictive [44, 57, 35, 7, 31, 25, 40]. However, these approaches often entail significant overhead for additional supervision, such as group labels indicating spurious features [44, 25], or labeled data from the target distribution [35, 7, 31]. In contrast, our method requires only a few natural language descriptions of model errors, which are substantially easier to collect. This lower annotation burden renders CLARIFY especially practical for addressing spurious correlations in large datasets.

Discovering failure modes. Our work builds upon a growing body of literature aimed at identifying and correcting failure models of machine learning models. Previous works in this area aim to discover data subsets on which models perform poorly [5, 3, 9] and devise methods to rectify such specific failures [45, 34, 21]. Some works perform counterfactual data augmentation to directly highlight model reliance on spurious features [23, 55, 43, 52, 53]. More closely related to our work are methods that leverage vision-language models to describe failure modes with natural language [13, 54, 12, 59, 24]. Natural language descriptions of error slices have the advantage of being interpretable and naturally grounded in human understanding. However, many of the descriptions generated by these fully automated methods do not correspond to true model failures. For example, **(author?)** [59] reports that DOMINO [13] can make nonsensical descriptions such as "mammoth" for a bird classification task. By incorporating humans in the loop, our approach avoids such errors, making it possible to discover spurious correlations in large datasets such as ImageNet.

C Discussion

A limitation of CLARIFY is that it depends heavily on human knowledge and a model of text-image similarity, and is not straightforward to apply when either of these is limited, such as medical imaging or scientific domains. Future work could broaden the applicability of the CLARIFY framework by incorporating domain-specific knowledge or extending to data modalities beyond images. CLARIFY contributes to the democratization of machine learning by allowing laypeople to correct concept-level errors stemming from spurious correlations in data. This feature can potentially foster greater public trust, especially when users witness measurable improvements in the model after their interventions.

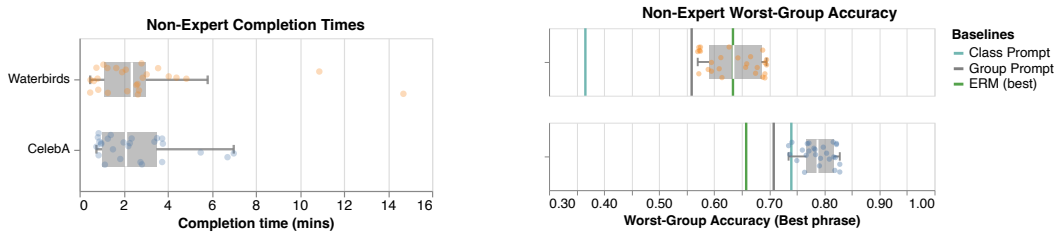
D Experimental Details

Datasets. We run experiments on three datasets: Waterbirds [44], CelebA [32], and ImageNet [8]. Waterbirds and CelebA have a known spurious correlation between the class label and a spurious attribute; for these datasets, we have access to ground truth spurious attribute labels. We use these datasets to evaluate whether CLARIFY can correct model failures due to spurious correlations. To our knowledge, ImageNet does not have any previously known spurious correlations.

Backbone models. All experiments use pre-trained CLIP models [41] as the feature extractor. The CLARIFY interface uses the CLIP ViT-L/14 vision and language backbones for calculating image-text similarity. We use the CLIP ResNet-50 and ViT-L/14 models for Waterbirds and CelebA, and only the CLIP ViT-L/14 model for ImageNet. We use frozen backbone models without any fine-tuning, and only train a final linear layer for classification, following related works for addressing

Table 2: Comparison with different zero-shot CLIP prompting strategies for group robustness. Fine-tuning with CLARIFY substantially outperforms RoboShot, a method that leverages state-of-the-art foundation models to automatically generate text prompts. All results besides ours are from **(author?)** [1].

Model	Method	Waterbirds			CelebA		
		Avg	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
ALIGN	Class Prompt	72.0	50.3	21.7	81.8	77.2	4.6
	Group Prompt	72.5	5.8	66.7	78.3	67.4	10.9
	RoboShot [1]	50.9	41.0	**9.9**	86.3	83.4	2.9
AltCLIP	Class Prompt	90.1	35.8	54.3	82.3	79.7	2.6
	Group Prompt	82.4	29.4	53.0	82.3	79.0	3.3
	RoboShot [1]	78.5	54.8	23.7	86.0	77.2	8.8
CLIP (ViT-L/14)	Class Prompt	88.7	27.3	61.4	80.6	74.3	6.3
	Group Prompt	70.7	10.4	60.3	77.9	68.9	9.0
	RoboShot [1]	79.9	45.2	34.7	85.5	82.6	2.9
	CLARIFY	96.8	**81.8**	14.9	90.9	**88.8**	**2.1**



(a) Non-experts completed annotation tasks for each dataset within a few minutes.

(b) Models retrained with non-expert annotations exceeded existing baselines.

Figure 7: CLARIFY achieves low annotation effort and improved model performance with non-experts.

spurious correlations [25, 58]. We use no data augmentation, and normalize all embeddings before computing similarity or training.

Human annotators. We recruit 26 non-expert users through Prolific (https://www.prolific.co/). These participants had no qualifications beyond being native English speakers and having some programming experience, and did not necessarily have any prior knowledge about machine learning. We provide a brief tutorial on how to use the interface, and ask each participant to annotate the class with highest error rate for each dataset. After the completion of the user study, we then retrained the models for both datasets using each user-provided annotation. Authors of this paper collected another set of annotations for Waterbirds and CelebA, which we use as a baseline for comparison. Additionally, annotations for the ImageNet dataset were collected by paper authors.

Table 3: The full set of model failure description phrases provided by non-expert annotators in our user study. The "Best WGA" and "Best Error Score" phrases were selected by identifying the phrase that achieved the highest Worst-Group Accuracy or Error Score, respectively, for each participant.

Phrase Category	**Waterbirds**	**CelebA**
Best WGA (per-user)	a bird with no head or as landbirds and a red outline, a blurry vision and they don't look like real birds, artic birds, beak, bird swims water, dark backgrounds and tall trees, forest, forest, forest, forest, forest, forests, forests, grass, greenery, landscape, landscapes, leaves, no water, plants, red, sandy beaches, seagulls, seagulls, trees, water, water	any other hair color than blonde or light hair color, backgrounds, bleach blonds, brown hair, buns, curls, curly hair, dyed hair, females, glasses, light background, light colors, men, men, men or short hair, older women, orange hair, pink, red, red, short hair, short hair, short haired men, smiles, white backgrounds, women
Best Error Score (per-user)	artic birds, birds, dark backgrounds and tall trees, ducks, forest, forest, forest, forest, forest, forests, forests, greenery, landscape, motion blur or can't make out a real bird, plants, sandy beaches, seagulls, trees, trees, trees, trees, water, water, water, waterfowl, wings	any other hair color than blonde or light hair color, bleach blonds, blond highlights, brown, darker blond hair, darker blonde, darker blonde hair, darker than blond, females, grey, males, men, men, men, men, men, men, men, men, men, men, men, pink, short white hair, very short hair, white, white, white
All Others	a lot of dark colors and no blue water, a lot of tree trunks, aqua blue water, been generated by ai, bird, bird wading in water, birds, birds floating, birds floating in water, birds standing in water, birds water, black, blue, blue, branches, branches, dark backgrounds, dark backgrounds and small birds, dark colors, darker backgrounds and a lot of trees, extended wings, eyes, flightless birds, flowers, game birds, grass, green, green, green, green, green plants, humans, land, landscapes, length of leg, lots of tree trunks, more dark colors than light colors and a lot of trees, mountains, no water, no water, no water and dark backgrounds, ocean coasts, people, people, people, plants, reeds, seagulls, shadows, sticks, tree trunk, trees, trees, trees, trees, trees, trees, very dark backgrounds and a lot of trees, water plants, wings, woods	bad lighting, bangs, beards, black hair, blue, blue, blue background, blue or black, brown, brown or dark hair, dark hair, darker hair, dim lighting, fair hair, flaxen, gold, golden hair, hair, hair, hats, hats, hats or bows, hazy, letters, light hair, little visible hair, long hair, males, males, males, men, more dark colors than light colors, non-blond hair. dark hair color. not blond, nondarkened hair, not blond, orange hair, people not facing the camera, red hair, red hair, redheads, short, short hair, short or curly hair, short or pulled back hair, shoulders, signs, skin color that is similar to their hair color, smiles, smiling faces, sunglasses, tan skin, teenagers, teeth, very tan skin, women

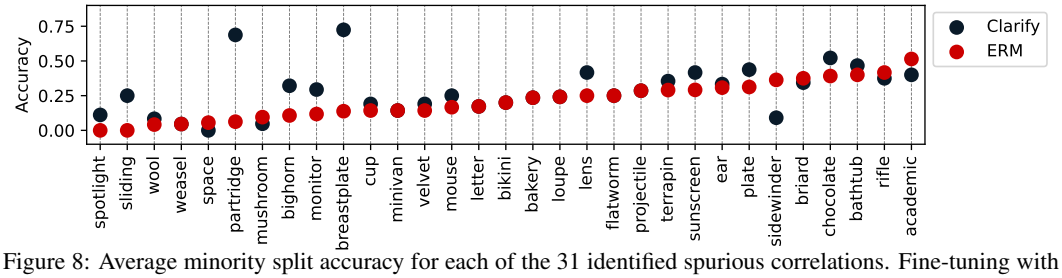


Figure 8: Average minority split accuracy for each of the 31 identified spurious correlations. Fine-tuning with textual feedback from CLARIFY improves minority split accuracy in many classes.

Table 4: Evaluation of methods for improving group robustness of CLIP models. Grouped by data and expressivity, with best worst-group (WG) and robustness gaps **bolded**. All metrics are averaged over three seeds.

			Waterbirds			CelebA		
	Assumptions	Method	WG	Avg	Gap	WG	Avg	Gap
CLIP ResNet-50	Zero-Shot	Class Prompt	36.6	92.2	55.6	74.0	81.9	7.9
		Group Prompt	55.9	87.8	31.9	70.8	82.6	11.8
	LP, Labels	ERM	7.9	93.5	85.6	11.9	94.7	82.8
		ERM (ours)	63.4	96.0	32.6	31.1	95.4	64.3
		ERM (ours, class-balanced)	48.6	95.2	46.7	65.8	93.4	27.6
		ERM (ours, worst-class)	55.9	95.8	39.9	56.9	94.1	37.2
	LP, Labels+Interaction	CLARIFY (group-balanced)	68.4	93.6	25.2	**89.3**	92.2	**2.8**
		CLARIFY (worst-group)	**75.7**	83.8	**8.1**	89.1	92.1	3.0
	LP, Labels+Groups	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
		DFR	78.7	90.8	12.1	**90.6**	91.9	**1.3**
		Group DRO	**81.3**	88.1	**6.8**	89.2	91.8	2.7
	Adapter, Labels	ERM Adapter	60.8	96.0	35.2	36.1	94.2	58.1
		WiSE-FT	49.8	91.0	41.2	85.6	88.6	3.0
		Contrastive Adapter	**83.7**	89.4	**5.7**	**90.0**	90.7	**0.7**
CLIP ViT-L/14	Zero-Shot	Class Prompt	25.7	87.3	61.6	62.1	71.9	9.8
		Group Prompt	27.4	85.5	58.1	72.4	81.8	9.4
	LP, Labels	ERM	65.9	97.6	31.7	28.3	94.7	66.4
		ERM (ours)	79.5	97.4	17.9	25.7	94.6	68.9
		ERM (ours, class-balanced)	71.1	97.2	26.1	63.7	92.6	28.9
		ERM (ours, worst-class)	74.3	97.1	22.8	56.9	93.3	36.4
	LP, Labels+Interaction	CLARIFY (worst-group)	**81.8**	96.8	**14.9**	**88.8**	90.9	**2.1**
	LP, Labels+Groups	DFR (subsample)	51.9	95.7	43.8	76.3	92.1	15.8
		DFR (upsample)	65.9	96.1	30.2	83.7	91.2	7.5
		DFR	85.9	93.5	7.6	**89.0**	90.9	**1.9**
		Group DRO	**88.5**	92.7	**4.1**	88.1	91.1	2.9
	Adapter, Labels	ERM Adapter	78.4	97.8	19.4	36.7	94.2	57.5
		WiSE-FT	65.9	97.6	31.7	80.0	87.4	7.4
		Contrastive Adapter	**86.9**	96.2	**9.3**	**84.6**	90.4	**5.8**

Table 5: The 31 identified spurious features in the ImageNet dataset. All annotation was performed on the validation split.

Class Name	Spurious Feature
cup	tea cup
weasel	snow weasel
wool	yarn ball
space bar	computer mouse
letter opener	silver
loupe	person holding a magnifying glass
mouse	desk and laptop
bakery	store front
sunscreen	person with sunburns
minivan	black minivan
plate rack	machine
briard	shaggy dog
lens cap	camera equipment
bighorn	rocky hillside
mushroom	red
rifle	wooden barrel
spotlight	shining
chocolate sauce	pastries with chocolate
terrapin	pond
sidewinder	sand
bikini	group of people
flatworm	coral reef
monitor	monitor on a desk
breastplate	museum display
projectile	rocket in a building
academic gown	many people in robes
velvet	pink velvet
bathtub	person
sliding door	car
partridge	tall grass
ear	green

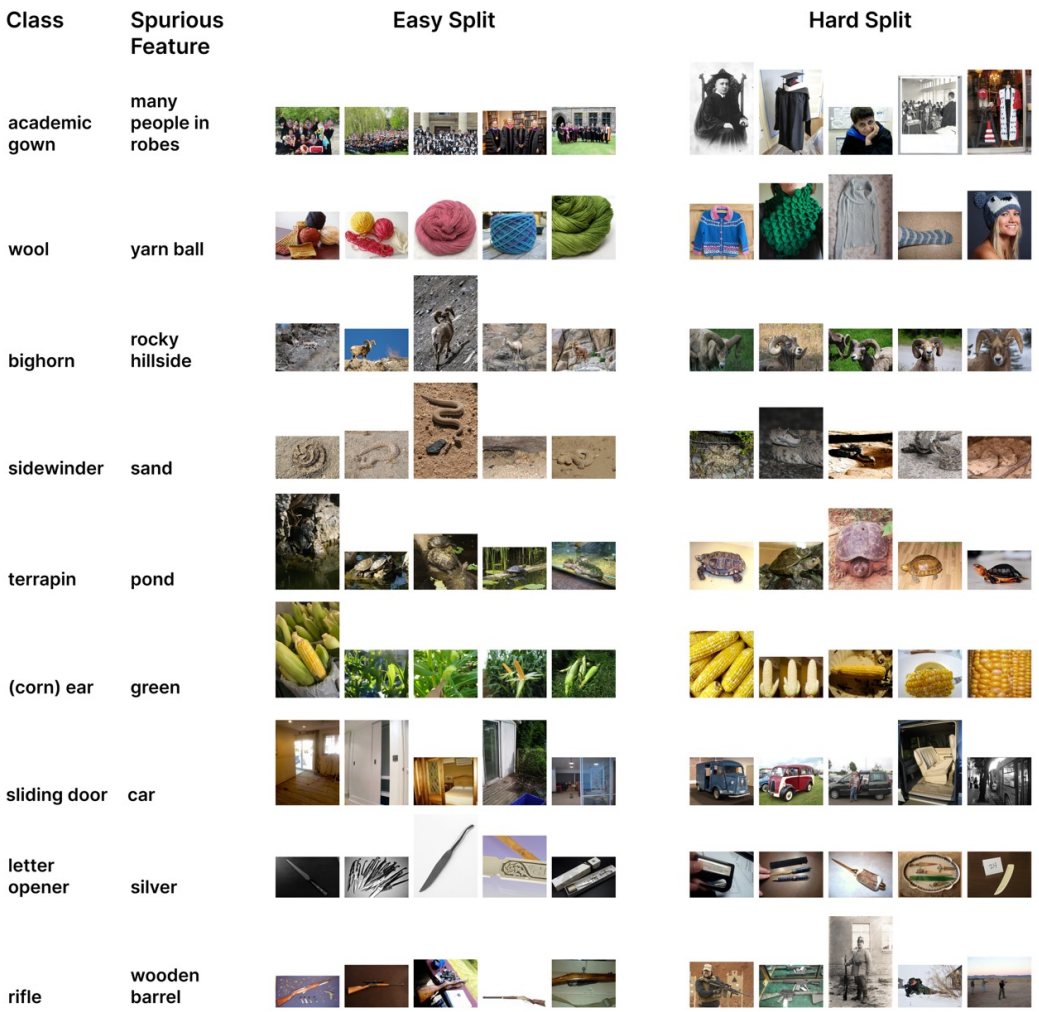


Figure 9: Representative samples corresponding to nine identified spurious correlations in ImageNet. All images shown are in the ImageNet validation set, and belong to the class shown in the first column. Similarity to the specified text annotation splits separates the "easy" and "hard" examples.

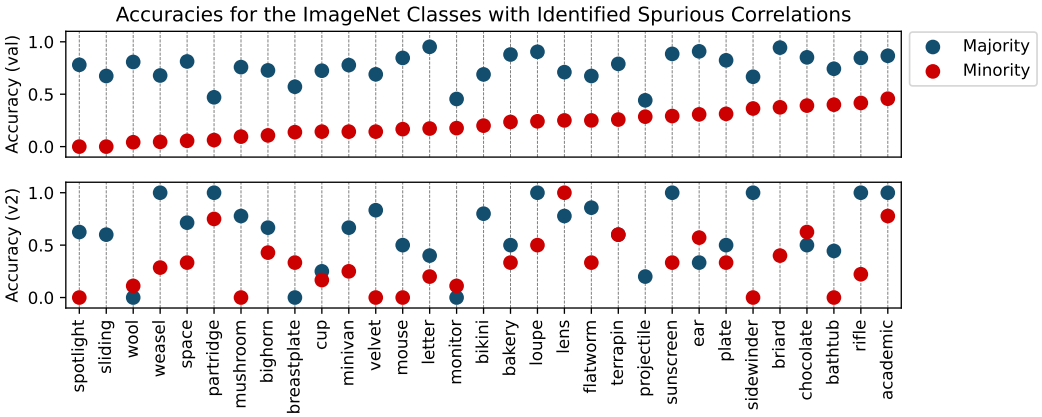


Figure 10: Accuracy of a model trained on the ImageNet train set, on the ImageNet validation set (top) and ImageNet-V2 (bottom). For each class with an identified spurious correlation, we show average, majority split, and minority split accuracy. The model achieves lower accuracy on the minority split for all classes in the validation set and all but 6 classes in ImageNet-V2, indicating that the model is relying on each identified spurious feature.