RELIABILITY-AWARE ENVIRONMENT DISCOVERY FOR SUBPOPULATION ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models often fail under subpopulation shift, where subgroup proportions differ between training and test environments. In this setting, empirical risk minimization may rely on spurious correlations and underperform on minority subgroups. Existing group-robust methods address this issue by optimizing worst-group risk, but typically require subgroup annotations or rely on error-based environment discovery. These approaches identify vulnerable samples through observed prediction errors, overlooking correctly classified examples that may become misclassified when spurious cues change. To address this limitation, we propose Reliability-Aware Environment Discovery (RAE), an annotation-free framework that supplements prediction errors with margin-based uncertainty to discover candidate vulnerable environments. RAE employs twin discovery networks with reliability-guided label updates and auxiliary regularization, and combines prediction errors with uncertainty-based selection to construct environment partitions. Experiments on vision and language benchmarks demonstrate competitive worst-group performance. Further analyses show that prediction reliability complements error-based discovery by identifying currently correct but unreliable examples that would otherwise be excluded.

1 INTRODUCTION

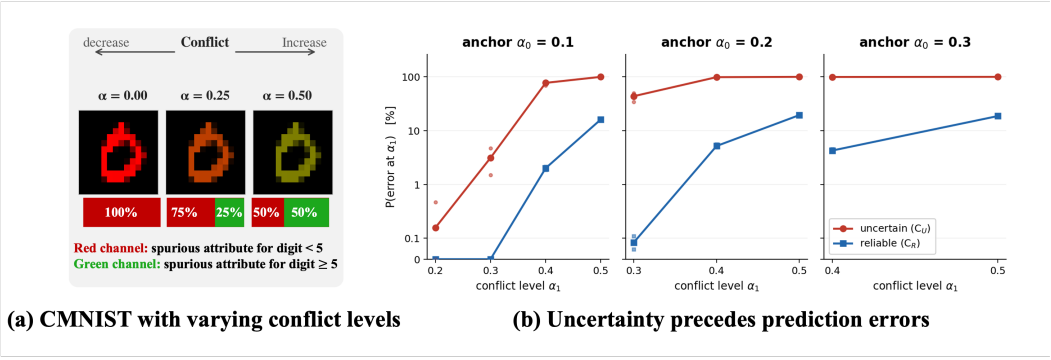
Machine learning models deployed in real-world applications aim to generalize reliably across multiple data sources. Ideally, a system should perform consistently across deployment environments. For example, in medical diagnosis, a model should remain robust when applied to patient data originating from multiple hospitals, each characterized by distinct patient demographics, imaging protocols, and clinical practices (Badgeley et al., 2019). However, *subpopulation shift* complicates this goal by changing the composition of subpopulations across environments, even when the conditional relationship between features and labels remains stable within each subgroup (Koh et al., 2021; Sagawa et al., 2019; Duchi et al., 2019; Beery et al., 2018).

To address this challenge, prior works have investigated the learning dynamics of spurious correlations and invariant (class-relevant) features (Arjovsky et al., 2019; Sagawa et al., 2019; Geirhos et al., 2020). The prevailing view is that subpopulation failures arise because models rely on spurious features associated with particular groups, which in turn interfere with the learning of invariant features that generalize across groups. Consequently, subgroup robustness methods aim to suppress the learning of spurious features during training to promote invariant representations, for example, by reweighting samples (Sagawa et al., 2019), enforcing invariance constraints (Arjovsky et al., 2019), or identifying and penalizing spurious patterns (Nam et al., 2020; Bao & Barzilay, 2022).

Since these approaches often require access to subgroup annotations, recent studies have proposed an alternative paradigm, **automatic environment discovery**, which consists of two phases: (i) inferring candidate vulnerable environments from training data and (ii) applying group-robust optimization over the inferred groups (Pezeshki et al., 2023; Han & Zou, 2024; Creager et al., 2021). Error-based approaches identify vulnerable samples using observed prediction failures, based on the assumption that *vulnerability can be identified through error-related signals*. While effective for capturing realized failures, such signals characterize vulnerability only after it has manifested as an error.

However, current correctness does not necessarily imply robustness to changes in spurious cues. A correctly classified example may become misclassified as the conflict between the target label and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) CMNIST with varying conflict levels **(b) Uncertainty precedes prediction errors**

Figure 1: **Prediction reliability under increasing spurious-attribute conflict.** (a) ColoredMNIST examples with increasing conflict between the digit label and the spurious color cue. As α increases, digit intensity is redistributed from the original label-associated color channel to the conflicting channel. (b) Error rates of initially correct examples as color conflict increases. At each anchor α_0, correctly classified examples are partitioned into reliable (C_R) and unreliable (C_U) subsets based on their prediction margins. The same examples are evaluated at $\alpha_1 > \alpha_0$ using a fixed classifier. Unreliable examples exhibit higher error rates despite both subsets being correctly classified at the anchor condition. Lines indicate mean error rates over three random seeds.

spurious cues increases, yet an error-only partition treats it identically to other correctly classified examples. Although their error indicators are identical, their prediction margins can vary substantially. We hypothesize that this variation in prediction reliability provides an additional signal for identifying vulnerable examples.

To examine this hypothesis, we conduct a controlled ColoredMNIST experiment (Figure 1). We progressively increase the conflict between the digit label and the spurious color cue while keeping the classifier, digit images, and ground-truth labels fixed. At each initial condition, we divide correctly classified examples into reliable and unreliable subsets based on their prediction margins. We then evaluate the same examples under increased color conflict. For example, when the conflict level increases from $\alpha = 0.2$ to $\alpha = 0.3$, 43.75% of initially correct but unreliable examples become misclassified, compared with only 0.08% of reliable examples. Notably, both subsets are correctly classified at the initial condition, although their prediction margins differ. The substantial difference in their error rates under increased color conflict indicates that this variation is informative about vulnerability to spurious-cue changes. These results show that prediction reliability distinguishes currently correct examples that an error-only partition treats identically.

Building on this insight, we propose **Reliability-Aware Environment Discovery (RAE)**, which incorporates prediction reliability as an auxiliary signal to discover candidate vulnerable environments without requiring subgroup annotations. Our controlled experiments support this motivation: among currently correct examples, those identified as less reliable exhibit substantially higher subsequent error rates as predictive cues change. RAE therefore supplements observed prediction errors with margin-based uncertainty when constructing vulnerable environments.

RAE employs twin networks and a common margin-based uncertainty score for discovery training and environment construction. During training, RAE regularizes uncertainty and uses it to selectively update hard supervision labels through label interpolation. After training, RAE fixes an empirical uncertainty threshold and identifies candidate vulnerable samples as those that are either misclassified by either network or exceed this threshold. The discovered environments are then combined with class labels to define groups for robust optimization. We evaluate whether this additional reliability signal improves vulnerable-group recovery and worst-group performance on subpopulation shift benchmarks.

2 Related Work

Most studies address subpopulation shift through group-robust learning, which aims to optimize worst-case performance across predefined subgroups. To relax the requirement for subgroup annotations,

recent methods propose automatically inferring vulnerable subpopulations directly from training data. Early approaches such as Just Train Twice (JTT) (Liu et al., 2021) identify hard or minority samples based on the error patterns of a biased or undertrained model. Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021) formulates environment discovery as an optimization problem that maximally violates invariance constraints.

While these methods automate environment discovery, some still rely on annotated environments for hyperparameter tuning and model selection. To fully eliminate such annotation requirements, Group Inference via Data Comparison (GIC) (Han & Zou, 2024) leverages distributional differences between datasets to train a spurious-attribute predictor, enabling more accurate subgroup discovery. Bias Amplification (BAM) (Li et al., 2023) utilizes class differences for model selection to remove the need for group annotations. Complementarily, Cross Risk Minimization (XRM) (Pezeshki et al., 2023) removes the need for validation annotations and early stopping by employing twin networks trained on disjoint data splits, where each network identifies vulnerable samples by learning from confident mistakes made by its counterpart on held-out data.

Despite these advances, annotation-free approaches typically rely on the presence of coherent error patterns, identifying vulnerable subgroups based on misclassified samples. Our work complements these approaches by incorporating prediction reliability to identify currently correct but unreliable examples that error-based discovery may overlook.

3 METHOD

In this section, we introduce Reliability-Aware Environment Discovery (RAE). We first formulate the problem of group robustness under subpopulation shift without explicit subgroup annotations (Section 3.1) and examine the limitations of error-based environment discovery (Section 3.2). Building on these observations, we present RAE, which supplements observed prediction errors with margin-based uncertainty to identify candidate vulnerable environments (Section 3.3). Finally, we describe how the inferred environments are used for group-robust learning (Section 3.4).

3.1 PROBLEM DEFINITION

Group robustness without subgroup annotations. Following the standard group robustness setting (Liu et al., 2021), we consider a classification problem with inputs $X \in \mathcal{X}$, labels $Y \in \mathcal{Y}$, and spurious attributes $A \in \mathcal{A}$ (i.e., $\mathcal{G} = \mathcal{Y} \times \mathcal{A}$), where each example belongs to a group $G = (Y, A)$. These groups differ in their spurious attributes, such as background, gender, or negation cues, which may be spuriously correlated with the target label in the training distribution (Figure 4).

Under subpopulation shift, the marginal distribution over subpopulations differs between training and test for some $g \in \mathcal{G}$, while the conditional distribution within each subpopulation remains invariant:

$$P_{\mathrm{tr}}(G = g) \neq P_{\mathrm{te}}(G = g), \quad P_{\mathrm{tr}}(X \mid G = g) = P_{\mathrm{te}}(X \mid G = g). \tag{1}$$

When spurious attributes are correlated with target labels in the training distribution, average-risk minimization may favor patterns prevalent in majority groups, leading to degraded performance on rare or underrepresented groups at test time (Koh et al., 2021; Sagawa et al., 2019).

A general approach to handling subpopulation shift is to improve subgroup robustness using oracle group annotations $\mathcal{G}$, by explicitly optimizing the worst-group risk.

$$R_{\mathrm{wg}}(\theta) = \max_{g \in \mathcal{G}} \mathbb{E}\left[\ell(f_\theta(X), Y) \mid G = g\right]. \tag{2}$$

However, spurious-attribute annotations are often unavailable during training. We therefore consider environment discovery, which constructs surrogate groups from observable training signals. Specifically, let $\hat{e} : \mathcal{X} \times \mathcal{Y} \to 0, 1$ denote an inferred environment assignment, where $\hat{e}(x, y) = 1$ identifies a candidate vulnerable subset and $\hat{e}(x, y) = 0$ denotes its complement.

A common approach uses observed prediction errors to construct vulnerable subset. Given a discovery predictor $\hat{y}_{\mathrm{disc}}$, error-based discovery assigns $\hat{e}(x, y) = \mathbf{1}[\hat{y}_{\mathrm{disc}}(x) \neq y]$. The inferred environments are combined with class labels to form surrogate groups $\widehat{G} = (Y, \hat{E})$, where $\hat{E} = \hat{e}(X, Y)$. The model is then optimized over these inferred groups:

$$\min_{\theta} \max_{g \in \widehat{\mathcal{G}}} \mathbb{E}_{\mathcal{D}_{\mathrm{tr}}} \left[\ell(f_\theta(X), Y) \mid \widehat{G} = g \right], \qquad \widehat{\mathcal{G}} = \mathcal{Y} \times \{0, 1\}. \tag{3}$$

This enables group-robust optimization without spurious-attribute annotations during training. The oracle groups are used only to evaluate worst-group performance on the test set.

3.2 Limitations of Error-Based Discovery

Prediction Reliability under subpopulation shift A key limitation of error-based environment discovery is that it identifies vulnerable samples only through observed prediction failures. Since the error indicator $e(x, y) = \mathbf{1}[\hat{y}_{\mathrm{disc}}(x) \neq y]$ is binary, all correctly classified examples satisfy $e(x, y) = 0$, regardless of differences in their prediction reliability. Consequently, error-based discovery cannot distinguish among currently correct examples that may become misclassified under distribution shift.

To characterize this variation, we consider prediction reliability based on the separation between the two most probable classes. Let $p_{(1)}(x)$ and $p_{(2)}(x)$ denote the largest and second-largest predictive probabilities, respectively. We define the corresponding uncertainty score as

$$s(x) = 1 - \left(p_{(1)}(x) - p_{(2)}(x) \right), \tag{4}$$

where a larger $s(x)$ indicates lower prediction reliability. For an example with label y, we additionally define the signed prediction margin as

$$d(x, y) = p_y(x) - \max_{k \neq y} p_k(x). \tag{5}$$

A prediction is correct when $d(x, y) > 0$, assuming no ties. For a correctly classified example, $p_y(x) = p_{(1)}(x)$ and $\max_{k \neq y} p_k(x) = p_{(2)}(x)$. Therefore,

$$s(x) = 1 - d(x, y). \tag{6}$$

Thus, while the error indicator records only whether a prediction is correct, the uncertainty score preserves variation in the prediction margin among correctly classified examples.

To examine how prediction reliability relates to vulnerability under changes in spurious cues, we introduce a label-preserving intervention $x_\alpha = T_\alpha(x)$, where α controls the intervention strength and α_0 denotes the initial condition. We define the signed prediction margin as

$$d_\alpha(x, y) = p_y(x_\alpha) - \max_{k \neq y} p_k(x_\alpha). \tag{7}$$

An initially correct example becomes misclassified when its signed margin crosses zero. Suppose that the intervention locally decreases the margin around α_0:

$$d_\alpha(x, y) \approx d_{\alpha_0}(x, y) - \gamma(x, y)(\alpha - \alpha_0), \qquad \gamma(x, y) > 0. \tag{8}$$

Under this local linear approximation, the intervention strength at which the margin reaches zero satisfies

$$\alpha^\star(x, y) - \alpha_0 \approx \frac{d_{\alpha_0}(x, y)}{\gamma(x, y)}. \tag{9}$$

Hence, among examples with comparable sensitivity to the intervention, a smaller initial margin equivalently, higher uncertainty among correct predictions corresponds to a smaller change required to reach the decision boundary. This conditional relationship motivates examining whether reliability variation is informative about sensitivity to spurious attributes.

Controlled Synthetic Experiment We further examine this relationship using ColoredMNIST, where digit shape provides the label-relevant signal and digit color serves as the spurious attribute. As illustrated in Figure 1, we progressively increase color-cue conflict by redistributing digit intensity between the original label-associated and conflicting color channels in proportions $(1 - \alpha) : \alpha$.

At each anchor α_0, we retain correctly classified examples and partition them into reliable (C_R) and unreliable (C_U) subsets using a fixed empirical uncertainty threshold. We then evaluate the same examples at $\alpha_1 > \alpha_0$. Both subsets have the same initial error indicator, $e_{\alpha_0}(x, y) = 0$.

Table 1 reports subsequent error rates and the fraction of subsequent failures captured by C_U. Unreliable examples exhibit higher subsequent error rates than reliable examples and the expectation under uniform random selection of the same size. The results show that uncertainty distinguishes sensitivity to color-cue changes among examples that an error indicator treats identically. These observations motivate supplementing observed errors with reliability when constructing candidate vulnerable environments.

Table 1: **Failure under color-cue changes.** All examples are correct at α_0. 'Sel.' denotes fraction selected as C_U; 'Rand.' denotes expected error rate under size-matched random selection; 'Rec.' denotes fraction of subsequent failures captured by C_U. Values are percentages averaged over three seeds.

		Error at α_1			
$\alpha_0 \to \alpha_1$	Sel.	C_R	C_U	Rand.	Rec.
$0.1 \to 0.3$	2.58	0.00	3.13	0.08	100.00
$0.2 \to 0.3$	1.31	0.08	43.75	0.65	88.11
$0.3 \to 0.4$	1.65	4.27	99.49	5.84	28.01

3.3 Reliability-Aware Environment Discovery

We propose **Reliability-Aware Environment Discovery (RAE)**, which supplements observed errors with margin-based uncertainty when constructing candidate vulnerable environments. RAE learns this reliability signal through twin discovery networks with uncertainty-guided label updates and auxiliary regularization, and subsequently uses the trained networks to infer environments for group-robust learning.

Margin-based reliability. We employ two discovery networks, f_a and f_b, with a fixed random assignment of training examples. For each example, one network provides the supervised prediction, while the other provides the uncertainty estimate. We refer to them as the supervised and opposite branches, respectively, with their roles determined by the example's assignment.

Let p_i denote the predicted class probabilities from the opposite branch for example i. Following the uncertainty measure introduced in Sec. 3.2, we compute

$$m_i = p_{i,(1)} - p_{i,(2)}, \qquad s_i = 1 - m_i, \tag{10}$$

where $p_{i,(1)}$ and $p_{i,(2)}$ are the two largest predicted probabilities. A larger s_i indicates a smaller separation between the competing classes and hence lower prediction reliability. This score serves as a common signal for discovery training, label updates, and subsequent environment construction.

Reliability-guided discovery training. Instead of relying solely on fixed supervision, RAE selectively updates the labels of uncertain examples. This allows reliability estimates to guide discovery training, while auxiliary regularization controls the uncertainty scores used for label updates.

We initialize the dynamic labels with the original class labels and train the supervised branch using the current dynamic labels. For a minibatch $\mathcal{B}$, we compute the upper uncertainty quantile

$$q_{\mathcal{B}} = Q_{1-\beta}\left(\{s_i : i \in \mathcal{B}\}\right) \tag{11}$$

and the following differentiable margin statistic:

$$r_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sigma(k(m_i - \tau_{y_i})), \tag{12}$$

where τ_{y_i} is a learnable threshold indexed by the original class label and k controls the sharpness of the sigmoid function. The discovery networks are then jointly optimized with

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_q q_{\mathcal{B}}^2 + \lambda_r \left(r_{\mathcal{B}} - (1-\beta)\right)^2, \tag{13}$$

where $\mathcal{L}_{\text{sup}}$ sums the mean cross-entropy within each represented dynamic-label class. The coefficients λ_q and λ_r control the contributions of the uncertainty quantile penalty and soft margin regularizer, respectively, while β determines the quantile level and the target of the soft margin statistic.

The uncertainty scores also determine which examples undergo label updates. After each optimization step, example i is selected with probability

$$\rho_i = \frac{s_i}{\max_{j \in \mathcal{B}} s_j + \varepsilon}, \tag{14}$$

where $\varepsilon > 0$ is a small constant for numerical stability. The normalization by the maximum batch uncertainty assigns higher selection probabilities to more uncertain examples.

For each selected example, we interpolate the one-hot opposite-branch predictions of the example and a randomly permuted minibatch partner:

$$\widetilde{\mathbf{y}}_i = \lambda \mathbf{e}_{\hat{y}_i^{\text{opp}}} + (1-\lambda)\mathbf{e}_{\hat{y}_{\pi(i)}^{\text{opp}}}, \qquad \lambda \sim \text{Beta}(1,1), \tag{15}$$

where $\mathbf{e}_{\hat{y}}$ denotes the one-hot vector of predicted class $\hat{y}$, π is a random permutation of the minibatch indices, and a single mixing coefficient λ is shared across the minibatch.

The updated dynamic label is obtained by taking $\arg\max_k \widetilde{\mathbf{y}}_{i,k}$; unselected examples retain their current dynamic labels.

These updates use detached predictions from the forward pass preceding the optimization step. The resulting hard labels provide supervision in subsequent steps rather than serving as soft training targets.

Environment construction. Once discovery training is complete, we freeze both networks and use their predictions to identify candidate vulnerable examples. Following the motivation in Sec. 3.2, we retain examples that are either already misclassified or exhibit high uncertainty, thereby extending error-based discovery to currently correct but unreliable examples.

We first compute opposite-branch uncertainty scores over the full training set and determine a shared empirical threshold:

$$q = Q_{1-\beta}\left(\{s_i : i \in \mathcal{D}_{\text{tr}}\}\right). \tag{16}$$

Let E_i indicate whether either discovery network misclassifies example i relative to its original label. The inferred environment assignment is

$$\hat{e}_i = \mathbf{1}\left[E_i = 1 \ \vee \ s_i \geq q\right]. \tag{17}$$

3.4 Group-Robust Learning with Discovered Environments

The inferred environments are then used to train a separate classifier with the original class labels. Specifically, we cross the environment assignments with class labels to form up to $2|\mathcal{Y}|$ surrogate groups and optimize the worst-group risk using GroupDRO, as defined in Eq. 3. The overall training procedure is described in Appendix A.3.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method on real-world subpopulation benchmarks. Table 2 summarizes the spurious attributes and class labels for each dataset. The Waterbirds (Sagawa et al., 2019) consists of waterbirds and landbirds, with background type (water vs. land) used as the oracle group annotation. CelebA (Liu et al., 2015) consists of face images, where hair color is the target label and gender serves as a spurious attribute.

CivilComments-WILDS (Borkan et al., 2019; Koh et al., 2021) aims to classify online comments as toxic or non-toxic, with identity mentions defining subgroup annotations. MultiNLI (Williams et al., 2018) aims to classify the textual relationship between a premise and a hypothesis, with the presence of negation treated as a spurious attribute.

Table 2: **Dataset overview.**

Dataset	Spurious attributes	Class Label
Waterbirds	Background type	Bird type
CelebA	Gender	Hair color
CivilComments	Identity mention	Toxicity
MultiNLI	Presence of negation	Textual relationship

Table 3: **Performance comparison between baselines and our method, averaged over three random seeds.** 'Avg' and 'Worst' denote the average and worst-group accuracy, respectively. 'tr/val' indicates whether group labels are used during training and validation. All experiments are reproduced based on the original implementations.

Group Annotation tr/val	Algorithm	Waterbirds		CelebA		MNLI		CivilComms		Average	
		Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Group annotations are required											
✓/✓	ERM	82.7	51.5	95.8	44.6	81.5	72.4	84.2	73.5	86.1	60.5
	GroupDRO	90.0	86.1	93.0	87.0	80.6	75.7	83.9	73.3	86.9	80.5
✗/✓	EIIL	86.5	69.3	95.2	43.5	82.3	67.5	-	-	-	-
	JTT	90.5	87.8	90.5	79.6	79.8	70.9	92.5	64.7	88.3	75.8
	BAM	91.0	89.3	87.0	77.7	81.3	71.2	90.1	72.9	87.4	77.8
Group annotations are not required											
✗/✗	EIIL	87.2	69.5	95.2	42.6	82.4	66.9	-	-	-	-
	JTT	91.8	87.4	94.3	48.7	80.7	66.8	92.9	54.9	89.9	64.5
	BAM	92.1	89.5	91.9	73.1	81.5	68.1	90.1	72.9	88.9	75.9
	GIC	89.0	74.1	93.5	67.0	82.0	67.2	90.8	73.1	88.8	70.4
	XRM	89.2	87.7	88.2	82.4	67.7	60.3	84.4	72.6	82.4	75.8
	RAE (Ours)	90.3	**88.0**	90.0	**87.9**	71.4	**68.7**	81.4	**73.3**	83.3	**79.5**

Table 4: **Comparison of environment discovery strategies on Waterbirds and CelebA. Oracle** uses ground-truth group annotations. **Error-based** and **Reliability-based** construct vulnerable environments using prediction errors and uncertainty scores, respectively. **Error-random** augments the error-based environment with randomly selected examples, while **Ours** combines prediction errors with uncertainty-based selection.'Avg' and 'Worst' denote the average and worst-group accuracy, respectively.

	Oracle		Error-based		Reliabitliy based		Error random		Ours	
	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Waterbirds	90.0	86.1	89.9	86.8	80.5	59.9	88.7	83.1	**90.18**	**87.54**
CelebA	93.0	87.0	87.9	83.3	91.9	78.9	91.2	83.3	**89.77**	**87.78**

Detailed dataset statistics for each subgroup annotation are provided in Appendix A.1.

Models. Across all the datasets in our experiment, we use the same backbone architectures, following prior works Sagawa et al. (2019); Pezeshki et al. (2023). For image datasets (Waterbirds and CelebA), we use a ResNet-50 (He et al., 2016) pretrained on the ImageNet ILSVRC dataset as the image encoder. For text datasets (CivilComments and MNLI), we use a pretrained BERT-base (Devlin et al., 2019) model as the encoder. Details of the training configuration are provided in Appendix A.2.

Evaluation Strategy. We evaluate our method against existing group discovery approaches under subpopulation shift, including (i) methods that require access to validation group annotations, such as EIIL (Creager et al., 2021), JTT (Liu et al., 2021), and BAM (Li et al., 2023), and (ii) methods that do not require any group supervision for group discovery, such as GIC (Han & Zou, 2024) and XRM (Pezeshki et al., 2023). For methods that rely on validation group annotations, we additionally report results using model selection based on average validation performance. We then further analyze the environments discovered by our method by comparing them with oracle group annotations and our variants. For evaluation metrics, we report average accuracy across groups (Avg) and worst-group accuracy (Worst), following (Han & Zou, 2024; Koh et al., 2021).

4.2 Experimental Results

Main Result. Table 3 compares our method with existing baselines. All results are averaged over three random seeds, with standard deviations reported in Appendix **??**. We include GroupDRO as an upper bound, which has access to oracle group annotations during both training and validation.

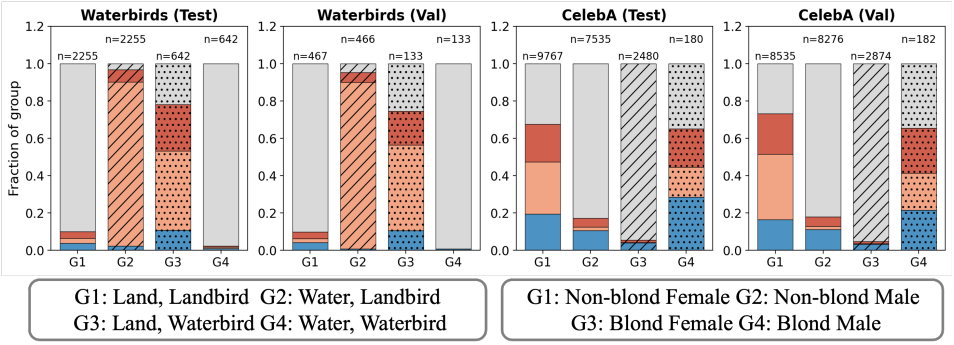


Figure 2: **Decomposition of Phase-1 group discovery across groups for Waterbirds and CelebA.** Each bar shows the fraction of samples assigned to the discovered vulnerable environment $\hat{\mathcal{G}}_{\text{vul}}$ (colored area) and its complement $\hat{\mathcal{G}}_{\text{oth}}$ (grey area). $\hat{\mathcal{G}}_{\text{vul}}$ is further decomposed based on the criterion used for discovery: prediction mistakes only (■), unreliability only (■), both criteria (■). Bars with diagonal hatching indicate groups underrepresented in the training distribution, while dotted bars indicate the most underrepresented group, where the largest distribution shift occurs.

As shown in the table, our method achieves performance comparable to oracle GroupDRO while requiring no group supervision. Notably, RAE consistently improves worst-group accuracy over existing annotation-free methods across multiple datasets, while remaining competitive in average accuracy. Moreover, it outperforms baselines that require validation group annotations, demonstrating the effectiveness of reliability-based group discovery.

Experiments on Different Group Discovery Table 4 compares different strategies for constructing candidate vulnerable environments, including an oracle setting using ground-truth group annotations. Our method achieves the highest worst-group accuracy on both Waterbirds and CelebA, reaching 87.54% and 87.78%, respectively. While error-based discovery performs competitively on Waterbirds, its worst-group accuracy decreases on CelebA. Reliability-based discovery alone yields lower worst-group accuracy on both datasets, indicating that uncertainty cannot fully replace observed prediction errors. Furthermore, augmenting the error-based environment with randomly selected examples does not achieve the same improvements as our reliability-guided selection. These results support combining prediction errors with reliability to construct candidate vulnerable environments for robust learning.

4.3 Analysis on Group Discovery

Group Discovery Decomposition. Figure 2 decomposes the discovered environments according to whether samples are identified through prediction mistakes, reliability, both signals, or neither. We observe that the two signals capture distinct subsets of vulnerable samples. Mistake-based discovery primarily identifies underrepresented groups, where error rates are high. However, a substantial portion of vulnerable samples is captured only by the reliability-based signal, particularly in groups where predictions remain correct but uncertain. This complementary behavior indicates that error-based methods alone miss a significant fraction of vulnerable samples, while incorporating reliability signals enables RAE to identify these otherwise overlooked cases.

Coverage across subgroups. To further investigate the role of conformal prediction, we analyze nonconformity score distributions across subgroups on Waterbirds and CelebA (Figure 3). Across both datasets, the most underrepresented subgroups (land + waterbird, 3.8% in Waterbirds, and blond male, 0.9% in CelebA) are consistently identified as vulnerable. In Waterbirds, the most underrepresented subgroup shows a large fraction of samples exceeding the conformal threshold, resulting in substantially lower empirical coverage compared to other groups. In CelebA, samples exceeding the conformal threshold are distributed more broadly across subgroups, likely due to high intra-group diversity (e.g., short-haired females or long-haired males). This diversity introduces ambiguity beyond subgroup labels, making the flagged set less concentrated in minority groups.

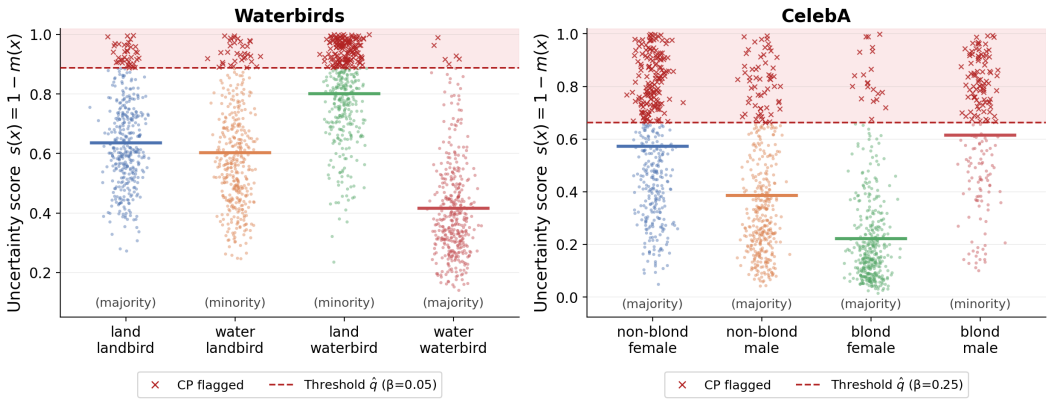


Figure 3: **Group-wise uncertainty scores on Waterbirds (left) and CelebA (right).** Each point represents a training example, and the dashed line indicates the empirical uncertainty threshold q used to identify unreliable examples. Red crosses denote examples with $s(x) \geq q$. The plots illustrate how uncertainty scores vary across ground-truth subgroups, although subgroup annotations are not used for environment discovery.

Despite these differences, nonconformity scores consistently assign higher uncertainty to minority samples under distribution shift.

Ablation Study. We investigate the contributions of the three components of RAE: the uncertainty quantile regularizer (λ_q), the soft margin regularizer (λ_r), and reliability-guided dynamic label updates (Dynamic label). Table 5 reports the results obtained by removing each component individually and in combination. Removing either regularizer reduces worst-group accuracy on Waterbirds and CelebA, indicating that both objectives contribute to reliability-aware discovery. Disabling dynamic label updates leads to a further reduction in worst-group accuracy, particularly on CelebA, highlighting the importance of incorporating uncertainty into the discovery training process. When both regularizers are removed, worst-group performance deteriorates further on Waterbirds. Removing all three components also substantially reduces performance, demonstrating the contribution of the combined training strategy.

Table 5: Ablation study of RAE components on Waterbirds and CelebA.

λ_q	λ_r	Dynamic label	Waterbirds		CelebA	
			Avg	Worst	Avg	Worst
✓	✓	✓	**90.18**	**87.54**	**89.77**	**87.78**
✗	✓	✓	88.13 (↓2.05)	82.75 (↓4.79)	90.79 (↑1.02)	83.89 (↓3.89)
✓	✗	✓	88.70 (↓1.48)	82.62 (↓4.92)	90.39 (↑0.63)	83.89 (↓3.89)
✓	✓	✗	81.12 (↓9.06)	79.29 (↓8.25)	93.52 (↑3.76)	71.67 (↓16.11)
✗	✗	✓	84.43 (↓5.75)	73.17 (↓14.37)	91.16 (↑1.40)	83.89 (↓3.89)
✗	✗	✗	77.55 (↓12.63)	75.39 (↓12.15)	93.42 (↑3.66)	70.00 (↓17.78)

5 Conclusion

In this work, we introduced Reliability-Aware Environment Discovery (RAE), a framework for achieving group robustness under subpopulation shift without requiring explicit subgroup annotations. RAE supplements observed prediction errors with margin-based uncertainty to identify currently correct but unreliable examples that error-based discovery may overlook. By incorporating reliability-guided label updates and auxiliary regularization into twin-network training, RAE learns uncertainty scores that are subsequently combined with prediction errors to construct candidate vulnerable environments. Empirical results on vision and language benchmarks demonstrate competitive worst-group performance while maintaining average accuracy. Further analyses show that combining prediction errors with uncertainty-based selection improves environment discovery over using either signal alone. These findings highlight the importance of prediction reliability as a complementary signal for identifying vulnerable examples that may become misclassified when spurious cues change.

AI USE STATEMENT

We used generative AI tools to assist with grammatical corrections, language polishing, and drafting and revising portions of the manuscript. We did not use generative AI tools for research ideation. Additionally, generative AI tools assisted in writing code for analysis figures and utility functions. All AI-assisted content was reviewed by the authors, and AI-generated code was verified and tested for correctness by two authors. We take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work aims to improve robustness under subpopulation shift, which may have positive societal impacts by reducing performance disparities for underrepresented or vulnerable groups in applications such as healthcare, content moderation, and other decision-support systems. However, the method also has potential negative impacts if deployed without sufficient validation: incorrectly inferred vulnerable environments may lead to misplaced trust, degraded performance for some groups, or unfair decisions in high-stakes settings. Since RAE does not require explicit subgroup annotations, it may also obscure which real-world demographic or semantic groups are being affected. We therefore recommend using the method as a robustness tool together with careful subgroup evaluation, domain-specific validation, and monitoring after deployment.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):31, 2019.

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

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1), 2019.

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.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. *arXiv preprint arXiv:2404.13815*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Gaotang Li, Jiarui Liu, and Wei Hu. Bias amplification enhances minority group performance. *arXiv preprint arXiv:2309.06717*, 2023.

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*, pp. 6781–6792. PMLR, 2021.

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

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.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm. *arXiv preprint arXiv:2309.16748*, 2023.

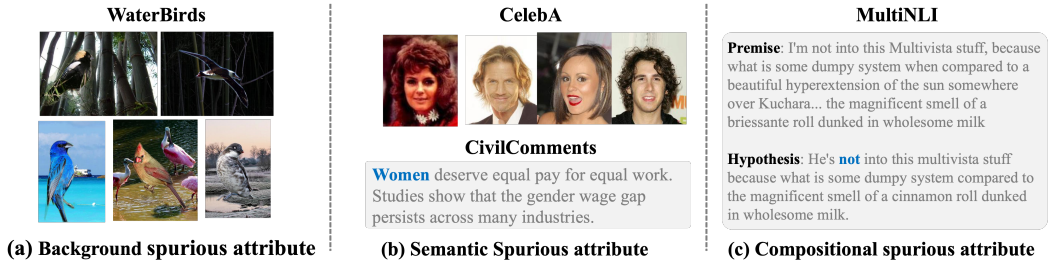
Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

A SUBPOPULATION SHIFT EXPERIMENTS

A.1 DATASET

Figure 4 illustrates examples of spurious attributes in the benchmark datasets, while Figure 5 summarizes the composition of each dataset, including subgroup annotations, class labels, and total sample sizes..



(a) Background spurious attribute **(b) Semantic Spurious attribute** **(c) Compositional spurious attribute**

Figure 4: **Examples of spurious attributes across benchmark datasets.** Waterbirds, CelebA, and MultiNLI exhibit spurious correlations involving background, gender, and negation cues, respectively. Subgroups are defined by the combination of target labels and spurious attributes, and their relative proportions may change under subpopulation shift.

A.2 TRAINING CONFIGURATION

In this subsection, we detail the training configurations and hyperparameter settings for all approaches used in the experiments. All experiments are conducted using the same three random seeds as in the original implementation. The details of each approach are as follows:

- **JTT, XRM, BAM**: As they provide all codes and scripts, we rerun their script on our local setting for fair comparison.
- **GIC**: Implemented with their original source, as the original paper conducts experiments on binary class classification datasets (e.g., CelebA, Waterbirds, CivilComments), we replace the binary classification layer with a multi-classification layer for experiments on MultiNLI.
- **RAE (Ours)**: We follow the training protocol and adopt most hyperparameter settings of prior work (Pezeshki et al., 2023). We additionally introduce β to control the uncertainty quantile used in reliability-aware environment discovery. Dataset-specific training configurations are provided in Table 6.

Table 6: Training configuration for each dataset.

Dataset	Optimizer	Arch	lr	Weight Decay	Batch Size	β (Phase 1)	η (Phase 2)
Waterbirds	SGD	ResNet	1.25e-4	6.43e-5	86	0.05	1.23e-02
CelebA	SGD	ResNet	1.25e-4	6.43e-5	86	0.25	1.23e-02
MultiNLI	SGD	BERT	1.25e-5	6.43e-5	43	0.05	1.77e-02
CivilComments	SGD	BERT	1.25e-5	6.43e-5	43	0.25	1.77e-02

Training Time All experiments on Waterbirds and CelebA are conducted on an NVIDIA RTX 3090 GPU, while MultiNLI and CivilComments are run on an NVIDIA A100 GPU. Training RAE on Waterbirds takes approximately 60 seconds and approximately 2 hours on CelebA, comparable to the training time of other automatic environment discovery baselines Pezeshki et al. (2023); Han & Zou (2024).

A.3 TRAINING PROCEDURE

Algorithm 1 summarizes the training and environment discovery procedure in Phase I. We employ two networks, f_a and f_b, with identical architectures and a fixed random sample-wise assignment.

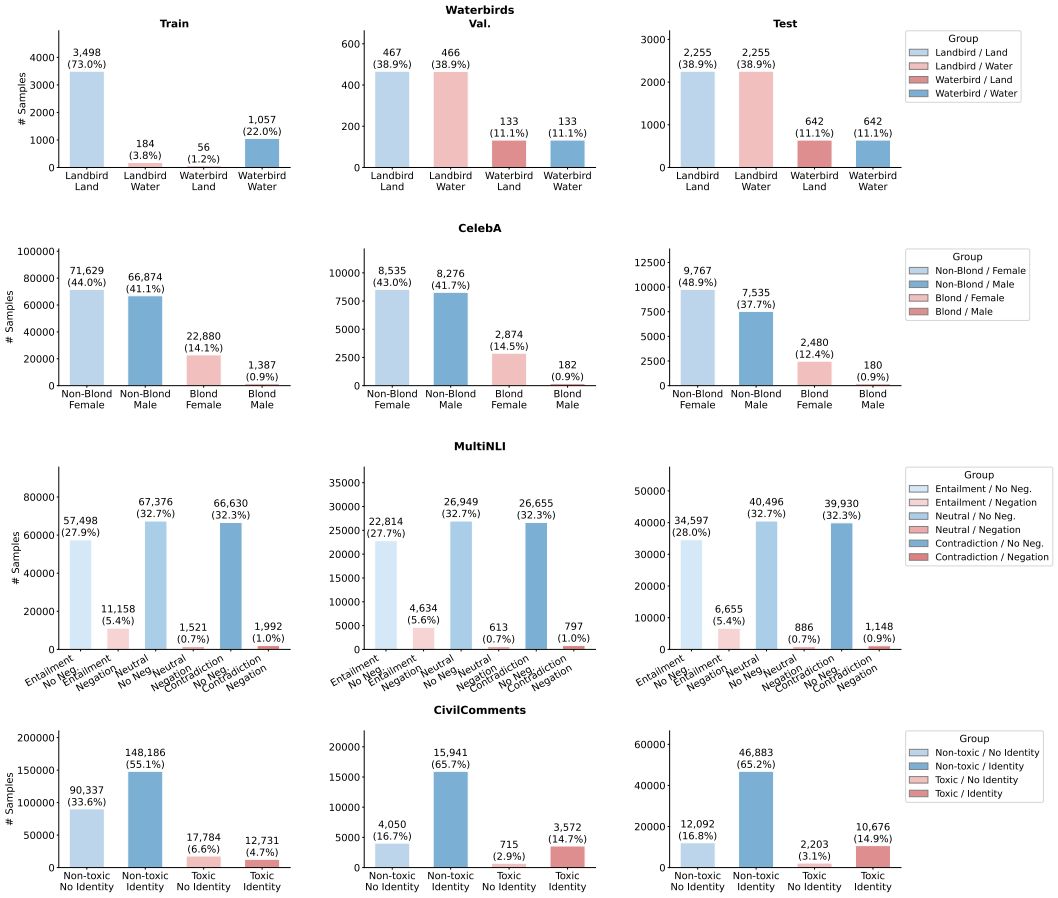


Figure 5: **Dataset statistics across train/val/test splits.** Each dataset contains spurious attributes correlated with class labels: background type (land/water) for **Waterbirds**, hair color (blond/non-blond) paired with gender for **CelebA**, negation words (present/absent) for **MultiNLI**, and identity attributes (gender, age, ethnicity) for **CivilComments**.

For each example, one network serves as the supervised branch, while the other serves as the opposite branch for uncertainty estimation. Their roles are determined by the example's assignment.

During training, uncertainty scores from the opposite branch guide selective label updates, while both networks are optimized using the supervised and auxiliary regularization objectives. After training, we freeze the discovery networks and combine their prediction errors and uncertainty scores to construct the inferred environments. Phase II then trains a separate classifier using GroupDRO over the fixed inferred groups.

B Controlled Synthetic Experiments: Dataset and Experimental Setup

Dataset Construction. We construct a modified version of ColoredMNIST to study feature entanglement. The base dataset follows the standard ColoredMNIST (Arjovsky et al., 2019) setup: MNIST digits are downsampled to 14×14 and colored red or green, where color is spuriously correlated with the binary label (digit < 5). Label noise and color noise are applied with probabilities drawn from $\{0.1, 0.2\}$ across two environments.

To introduce feature entanglement, we apply α-blending to majority samples (i.e., samples where the color matches the label). For a sample with label y and corresponding color channel $c = y \in \{0, 1\}$, let x_c and x_{1-c} denote the pixel intensities of the aligned and conflicting color channels, respectively.

Algorithm 1 Reliability-Aware Environment Discovery (RAE)

Input: Training set $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$; discovery networks f_a, f_b; parameter β
Output: Inferred environment assignments $\{\hat{e}_i\}_{i=1}^{N}$
// Phase I: Reliability-aware discovery
Fix random sample-wise assignments $a_i \in \{0, 1\}$
Initialize dynamic labels $y_i^{\mathrm{dyn}} \leftarrow y_i$
while not converged **do**
    Sample minibatch $\mathcal{B}$
    Route each example to its supervised and opposite branches according to a_i
    Compute supervised predictions and opposite-branch uncertainty scores s_i via Eq. 10
    Compute batch quantile $q_{\mathcal{B}}$ and soft margin statistic $r_{\mathcal{B}}$
    Update both networks and class thresholds by minimizing Eq. 13
    Select examples for label updates using ρ_i from Eq. 14
    Update selected dynamic labels using detached opposite-branch predictions; retain the remaining labels
end while
Freeze the discovery networks
Compute opposite-branch uncertainty scores over $\mathcal{D}_{\mathrm{tr}}$
Determine the empirical threshold q via Eq. 16
for $i = 1$ to N **do**
    Set $E_i \leftarrow \mathbf{1}[\hat{y}_i^a \neq y_i \ \vee \ \hat{y}_i^b \neq y_i]$
    Set $\hat{e}_i \leftarrow \mathbf{1}[E_i = 1 \ \vee \ s_i \geq q]$ (Eq. 17)
end for
// Phase II: Group-robust learning
Form surrogate groups $\widehat{G}_i = (y_i, \hat{e}_i)$
Train a separate classifier using the original labels and GroupDRO over the inferred groups
return $\{\hat{e}_i\}_{i=1}^{N}$

We modify the input as

$$x_c \leftarrow (1 - \alpha)\, x_c, \quad x_{1-c} \leftarrow \alpha\, x_c,$$

where α controls the degree of conflict between the spurious (color) and invariant (digit) features. Examples are shown in Figure 1(a).

Training and validation sets assign $\alpha \in \{0.00, 0.25, 0.50\}$ deterministically in equal thirds by sample index to the majority samples (color = label); minority samples (color $\neq$ label) are left unmodified. For evaluation we use a test split whose label and color noise match the average of the two training environments. The test set is replicated once per α level, with noise redrawn independently for each replicate, so it contains $3\times$ the number of underlying MNIST test images.