RELIABILITY-AWARE ENVIRONMENT DISCOVERY: LEVERAGING FEATURE ENTANGLEMENT FOR SUBPOPULATION ROBUSTNESS

Harim Lee and Dong-Kyu Chae
Department of Artificial Intelligence
Hanyang University
Seoul, South Korea
{hrimlee,dongkyu}@hanyang.ac.kr

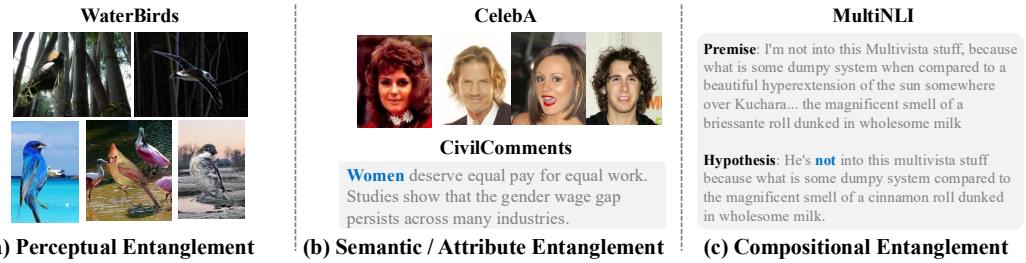
ABSTRACT

Machine learning models often fail under subpopulation shift, where latent subgroup distributions differ between training and test environments despite stable within-group relationships between features and labels. While empirical risk minimization achieves high average accuracy, it frequently performs poorly on minority subpopulations. 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 implicitly assume that failures are driven by dominant spurious correlations, leading to coherent error patterns. However, this assumption breaks down when spurious and invariant features are entangled, yielding diffuse and heterogeneous failures that are not well captured by error frequency alone. We propose Reliability-Aware Environment Discovery (RAE), an annotation-free framework that incorporates prediction reliability as an auxiliary signal for discovering vulnerable environments. RAE quantifies reliability using split conformal nonconformity scores, enabling the identification of regions with insufficient or contradictory predictive evidence. Then, define group label by combining prediction errors with reliability signals and apply group-robust optimization over the discovered partition. Experiments on vision and language benchmarks demonstrate that RAE improves both average and worst-group performance, with discovered environments closely approximating oracle group supervision.

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, a substantial body of work has 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).



(a) Perceptual Entanglement **(b) Semantic / Attribute Entanglement** **(c) Compositional Entanglement**

Figure 1: Examples of spurious-invariant feature entanglement across three categories. **(a) Perceptual entanglement:** A bird partially occluded by foliage shares color and texture with the background, making object and context features visually inseparable. **(b) Semantic entanglement:** In text, identity terms like "women" serve as grammatical subjects; their removal destroys sentence structure. In vision, attributes like hair color are integral to appearance; isolating them is semantically inconsistent. **(c) Compositional entanglement:** Negation words like "not" are essential for logical inference yet may be exploited as shallow lexical cues.

Since these approaches often require access to subgroup annotations, recent studies have proposed an alternative paradigm that consist of two phase, first **automatically discovers vulnerable subpopulations** from training data, typically based on error patterns or feature biases and then applies group-robust optimization over the inferred groups (Pezeshki et al., 2023; Han & Zou, 2024; Creager et al., 2021).

Despite algorithmic differences, these methods assume that suppressing spurious features during training improves worst-group generalization. However, this assumption does not always hold. Prior work has shown that aggressively removing spurious features, such as eliminating background information in the Waterbirds dataset, can actually degrade performance compared to training on the original data that includes these features (Kirichenko et al., 2022). This observation suggests that spurious features may, in some cases, encode information that is not easily disentangled from class-relevant signals. Building on this insight, we hypothesize that in real-world data, as illustrated in Figure 1, **spurious and invariant features are often entangled**, co-occurring and interacting in ways that make clean separation infeasible or even counterproductive. If spurious features are entangled with invariant features, then methods that aggressively suppress or ignore spurious cues may discard informative signals. Instead, we argue that spurious features should be leveraged to improve robustness. Specifically, regions where spurious and invariant features conflict or overlap tend to exhibit unreliable model predictions, high uncertainty, inconsistent outputs, or brittle decision boundaries even when error rates are moderate. Therefore, incorporating both unreliability and error frequency provides a more direct signal for identifying vulnerable subpopulations under feature entanglement.

Motivated by this insight, we propose **Reliability-Aware Environment Discovery (RAE)**, a framework that incorporates prediction reliability as an auxiliary signal for discovering vulnerable environments without subgroup annotations. RAE quantifies reliability through split conformal calibration (Angelopoulos et al., 2023; Yeh et al., 2024) and uses it for both training-time label interpolation and post-training environment discovery. During training, RAE employs a twin-network architecture to maintain split conformal calibration (Angelopoulos et al., 2023), ensuring that reliability estimates remain valid throughout learning. Samples identified as unreliable utilize adaptive label interpolation, preventing overfitting to spurious patterns while preserving supervision on confident samples. After training, RAE constructs pseudo-environments by flagging samples that are either misclassified or exhibit low reliability, then applies group-robust optimization (Sagawa et al., 2019) over the discovered partition.

2 Related Work

Most studies address subpopulation shift through group-robust optimization, which aims to optimize worst-case performance across predefined subgroups. To relax the requirement for subgroup annotations, recent methods propose automatically inferring vulnerable subpopulations or environments directly from training data. Early approaches such as Just Train Twice (JTT) (Liu et al., 2021)

and Learning from Failure (LfF) (Nam et al., 2020) 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 Learning to Split (LS) (Bao & Barzilay, 2022) searches for adversarial partitions that expose spurious correlations. Correct-n-Contrast (CnC) (Zhang et al., 2022) incorporates representation learning and contrastive objectives to improve the quality of discovered environments.

While these methods automate environment discovery, some still rely on annotated environments for validation or model selection. To fully eliminate such annotation requirements, (Han & Zou, 2024) improves group robustness through more precise group inference, training a spurious-attribute classifier that exploits two key properties where strong correlated with class labels and the variability across different data distributions. Complementarily, (Pezeshki et al., 2023) introduces Cross-Risk Minimization (XRM) to eliminate both validation annotations and early-stopping requirements. XRM employs 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 their success, most existing methods assume that vulnerable subpopulations exhibit coherent error patterns, dominant spurious correlations, or consistent prediction biases. Our work complements these approaches by addressing settings where spurious and invariant features are entangled, yielding diffuse failure signals not captured by error or bias patterns alone.

3 METHOD

In this section, we introduce Reliability-Aware Environment Discovery (RAE). We first formalize the problem of achieving group robustness under subpopulation shift without access to explicit subgroup annotations. We then conceptualize the limitations of existing environment discovery methods, particularly in settings where invariant and spurious features are entangled (Section 3.1). Motivated by this intuition, we propose RAE, which leverages calibrated prediction reliability as a proxy for feature entanglement to identify vulnerable environments (Section 3.2), and applies group-robust learning over the discovered pseudo-environments (Section 3.3).

3.1 PROBLEM DEFINITION

Subpopulation shift. The general setting of subpopulation shift assumes that the training and test distributions share the same set of latent groups, while the relative proportions of these groups differ between training and test (Koh et al., 2021). Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denote inputs and labels, and let G denote a set of latent subpopulations that partition the data according to underlying attributes or feature combinations. Under subpopulation shift, the marginal distribution over subpopulations differ between training and test:

$$\mathcal{D}^{\text{train}}(G = g) \neq \mathcal{D}^{\text{test}}(G = g), \tag{1}$$

for some $g \in G$, while the conditional distribution within each subpopulation remains invariant:

$$\mathcal{D}^{\text{train}}(y \mid x, G = g) = \mathcal{D}^{\text{test}}(y \mid x, G = g). \tag{2}$$

This shift often causes models trained to minimize average error under $\mathcal{D}^{\text{train}}$ to overfit majority subpopulations, leading to degraded performance on rare or underrepresented groups at test time, even when overall accuracy remains high (Koh et al., 2021; Sagawa et al., 2019).

Group robustness without subgroup annotations. In many real-world settings, explicit subgroup annotations are unavailable. We consider the problem of achieving group robustness in the absence of such annotations and assume the existence of an underlying latent subgroup variable G that partitions the data into subpopulations associated with different failure modes. Since G is unobserved, the training procedure first infers subgroup structure from the training data, yielding an estimated partition $\hat{G}$. Given the inferred subgroups, robustness is enforced by minimizing the worst-case risk across the estimated groups:

$$\min_{\theta} \max_{g \in \hat{G}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\cdot \mid \hat{G} = g)} \left[\ell(f_\theta(x), y) \right] \tag{3}$$

where f_θ denotes a generic predictor parameterized by θ and $\ell(\cdot, \cdot)$ a loss function. This objective emphasizes reliable performance on vulnerable inferred subpopulations while avoiding reliance on explicit subgroup supervision.

Feature entanglement under subpopulation shift. Standard approaches to subpopulation robustness often assume that invariant and spurious features can be cleanly separated. Following prior work on spurious correlations (Arjovsky et al., 2019; Geirhos et al., 2020), Conceptually, we can decompose each input x into invariant features $z(x)$ that are label-relevant and stable across subpopulations, and spurious features $s(x)$ that vary systematically across subpopulations:

$$z(x) \perp\!\!\!\perp G \mid y, \quad s(x) \not\perp\!\!\!\perp G \mid y. \tag{4}$$

A common assumption is that invariant and spurious features are conditionally independent given the label, i.e., $s(x) \perp\!\!\!\perp z(x) \mid y$. Under this separability assumption, vulnerable subpopulations can be identified by partitioning the input space based on spurious features, as models relying on spurious shortcuts tend to fail systematically on minority groups where the spurious–label correlation differs from that of the training majority. However, this separability assumption often fails in practice, as invariant and spurious features can be *entangled*, co-occurring or interacting in a label-dependent manner (Kirichenko et al., 2022; Rosenfeld et al., 2020). Formally, a region $\mathcal{R} \subseteq \mathcal{X}$ exhibits feature entanglement if

$$s(x) \not\perp\!\!\!\perp z(x) \mid y \quad \text{for } x \in \mathcal{R}. \tag{5}$$

Crucially, feature entanglement does not imply reliance on a single spurious attribute. Instead, spurious and invariant cues may provide conflicting, overlapping, or ambiguous evidence for the label. This yields diffuse and heterogeneous failure patterns not aligned with any predefined spurious factor, preventing error-based discovery methods from reliably identifying vulnerable subpopulations.

These entanglement phenomena arise at different levels, depending on how spurious and invariant features interact within the data. We categorize them into three types, illustrated in Figure 1. **(a) Perceptual entanglement** occurs when low-level visual features (texture, color, spatial layout) are inseparable across spurious and invariant attributes. **(b) Semantic entanglement** arises when group-correlated features are grammatically or semantically integral to the input, such that their removal destroys structure or meaning. **(c) Compositional entanglement** refers to cases where structural or logical patterns are essential for correct reasoning yet superficially resemble spurious shortcuts.

Across all three categories, a common theme emerges that suppressing or removing spurious features discards information necessary for robust prediction. Rather than eliminating spurious cues, we propose to leverage them as signals of predictive instability by quantifying this unreliability.

3.2 Reliability-Aware Environment Discovery

The key intuition behind RAE is that regions exhibiting feature entanglement tend to coincide with unreliable model behavior. When spurious and label-relevant cues interact, they may provide conflicting evidence for prediction, thereby increasing uncertainty even when empirical error remains low. This motivates using calibrated prediction reliability as a signal to identify vulnerable subpopulations, beyond error frequency alone. To this end, RAE consists of two key components: (1) **quantifying prediction reliability** as a proxy for feature entanglement using split-conformal calibration, and (2) leveraging these estimates to adaptively **interpolate training label**. After training, the calibrated nonconformity scores are used to identify vulnerable samples and construct pseudo-environments for group-robust optimization.

Reliability as a Proxy for Feature Entanglement Split conformal prediction (Angelopoulos et al., 2023) provides a principled framework for estimating prediction reliability through nonconformity scores computed on held-out data. Building on this idea, Yeh et al. (2024) show that conformal calibration can be incorporated into end-to-end training through differentiable coverage constraints, enabling well-calibrated uncertainty estimates during learning. Motivated by this, we incorporate split conformal calibration into the training process to obtain reliable uncertainty estimates. To mimic the held-out split required for conformal calibration without sacrificing training data, we employ twin networks f_θ^{in} and f_θ^{out} with a randomly initialized fixed sample-wise assignment $a_i \in \{0, 1\}$. For each sample x_i, the subset with $a_i = 1$ is used to train f_θ^{in}, while the subset with $a_i = 0$ is routed to f_θ^{out} for reliability estimation.

From the held out branch f_θ^{out}, we compute nonconformity scores to quantify prediction reliability. Let $p_\theta^{\text{out}}(y \mid x) = \text{softmax}(f_\theta^{\text{out}}(x))_y$ denote the predicted probability from the held-out branch. We define the nonconformity score $s_\theta^{\text{out}}(x, y) = 1 - p_\theta^{\text{out}}(y \mid x)$, and compute a batch-wise threshold via a tail quantile:

$$\hat{q} = \text{Quantile}_{1-\beta}\big(\{s_\theta^{\text{out}}(x_i, y_i)\}_{i \in \mathcal{B}}\big)\,. \tag{6}$$

To maintain calibration throughout training, we introduce two complementary regularization terms that prevent trivial solutions and enforce coverage. First, we regularize the quantile threshold directly to avoid collapse (i.e., $\hat{q} \to 0$):

$$\mathcal{L}_{\text{quant}} = \lambda_q\, \hat{q}^{\,2}. \tag{7}$$

Second, following the differentiable conformal calibration formulation of Yeh et al. (2024), we enforce coverage through a differentiable estimate. We define the prediction margin

$$\text{margin}(x) = p_\theta^{\text{out}}(\hat{y} \mid x) - \max_{y \neq \hat{y}} p_\theta^{\text{out}}(y \mid x), \qquad \hat{y} = \arg\max_y p_\theta^{\text{out}}(y \mid x). \tag{8}$$

Using learnable class-specific thresholds $\{\tau_y\}_{y=1}^C$, we construct a differentiable estimate of class-conditional coverage:

$$\text{Cov}_\theta^d = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sigma\big(k\big(\text{margin}(x) - \tau_y\big)\big)\,, \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function and $k > 0$ controls the sharpness of the approximation to the indicator function. We penalize deviation from the target coverage level $1 - \beta$:

$$\mathcal{L}_{\text{cov}} = \big(\text{Cov}_\theta^d - (1 - \beta)\big)^2. \tag{10}$$

Together, $\mathcal{L}_{\text{quant}}$ and $\mathcal{L}_{\text{cov}}$ maintain a calibrated reliability threshold $\hat{q}$ throughout training, allowing prediction reliability to serve as a stable signal for detecting feature entanglement.

Reliability-Guided Label Interpolation Using the calibrated reliability estimates, we identify uncertain samples and selectively smooth their supervision through label interpolation. The uncertainty estimates from the held-out branch f_θ^{out} are used to construct the sampling mask. Label interpolation is then applied during training of the held-in branch f_θ^{in}.

Given the mini-batch output of the held-out branch $p_\theta^{\text{out}}(\cdot \mid x)$, we compute a sample-wise uncertainty score $u(x)$ and derive a Bernoulli mask $\mu(x)$:

$$u(x) = 1 - \max_y p_\theta^{\text{out}}(y \mid x), \qquad \mu(x) \sim \text{Bernoulli}(u(x)). \tag{11}$$

In this way, samples with higher uncertainty are more likely to be selected for label interpolation.

For selected samples (i.e., $\mu(x) = 1$), we interpolate the one-hot label of the predicted class $\hat{y}(x) = \arg\max_y p_\theta^{\text{out}}(y \mid x)$ with that of a randomly permuted sample $\hat{y}_\pi$ in the mini-batch:

$$\tilde{y} = \lambda\, e_{\hat{y}} + (1 - \lambda)\, e_{\hat{y}_\pi}, \qquad \lambda \sim \text{Beta}(0.1, 0.1). \tag{12}$$

Then, the label is updated dynamically as:

$$y^{\text{dyn}} \leftarrow \begin{cases} \arg\max \tilde{y}, & \text{if } \mu = 1, \\ y^{\text{dyn}}, & \text{otherwise.} \end{cases} \tag{13}$$

Note that samples with $\mu(x) = 0$ retain their current labels, while selected samples are assigned updated hard labels derived from the interpolated targets. The updated labels are used as supervision in subsequent optimization steps.

Training Objective. Let $\ell(\cdot, \cdot)$ denote the cross-entropy loss and $f_\theta^{\text{in}}(x)$ the logits from held in branch. Given dynamic labels y^{dyn}, we jointly optimize:

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{(x, y^{\text{dyn}})}\big[\ell\big(f_\theta^{\text{in}}(x), y^{\text{dyn}}\big)\big]}_{\mathcal{L}_{\text{sup}}} + \underbrace{\lambda_q \hat{q}^{\,2}}_{\mathcal{L}_{\text{quant}}} + \underbrace{\lambda_c\big(\text{Cov}_\theta^d - \beta\big)^2}_{\mathcal{L}_{\text{cov}}}, \tag{14}$$

where $\mathcal{L}_{\text{sup}}$ trains the held-in branch using reliability-guided labels, while $\mathcal{L}_{\text{quant}}$ and $\mathcal{L}_{\text{cov}}$ calibrate uncertainty in the held-out branch to guide training of the held-in branch.

Environment Discovery. After training converges, RAE constructs pseudo-environments for robust optimization using the full training set. We compute a global nonconformity threshold

$$q = \text{Quantile}_{1-\beta}\big(\{s_\theta(x_i, y_i)\}_{i=1}^N\big)\,, \tag{15}$$

and flag each sample as either misclassified or unreliable:

$$\text{mistake}(x) = \mathbf{1}[\hat{y}(x) \neq y], \quad \text{unreliable}(x) = \mathbf{1}\big[s_\theta\big(x, \hat{y}(x)\big) \geq q\big]\,. \tag{16}$$

We define the vulnerable environment as samples satisfying either condition:

$$\hat{G}_{\text{vul}} = \{x : \text{mistake}(x) \vee \text{unreliable}(x)\}, \tag{17}$$

with the remaining samples forming the complementary environment $\hat{G}_{\text{other}}$. By combining prediction errors with reliability signals, this construction identifies vulnerable samples that are either consistently misclassified or predicted with low reliability. The full training procedure is described in Appendix A.4.

3.3 Group-Robust Learning with Discovered Environments

Finally, we perform group-robust optimization using the discovered environments. Let $\hat{e}(i) \in \{0, 1\}$ denote the inferred environment indicator obtained from Algorithm 1, which partitions the training data into two environments: the vulnerable environment $\hat{G}_{\text{vul}} = \{i : \hat{e}(i) = 1\}$ and the remaining environment $\hat{G}_{\text{oth}} = \{i : \hat{e}(i) = 0\}$. We treat these environments as groups and optimize the standard worst-group objective:

$$\min_\theta \max_{g \in \{\hat{G}_{\text{vul}}, \hat{G}_{\text{oth}}\}} \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_g}\big[\ell(f_\theta(x), y)\big], \tag{18}$$

where $\hat{\mathcal{D}}_g$ denotes the empirical distribution restricted to group g. By defining environments using both prediction mistakes and unreliability, RAE targets failure regions induced by feature entanglement without requiring explicit subgroup annotations or assuming separability between invariant and spurious features.

4 Experiments

4.1 Experimental Settings

We evaluate our method on real-world subpopulation benchmarks. Table 4.1 summarizes the oracle group annotations 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.

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

Table 1: Dataset overview.

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. We first evaluate our method against existing subgroup-robustness approaches under subpopulation shift, including methods that require access to group annotations and those that automatically infer groups without supervision. 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 and worst-group accuracy, following (Han & Zou, 2024; Koh et al., 2021). Details of the training configuration are provided in Appendix A.1.

4.2 Experimental Results

Experiments on Subpopulation Shift Table 2 presents a performance comparison between our proposed method and existing baselines. As shown in the table, our approach is competitive with

Group Annotation tr/val	Algorithm	Image Data				Text Data				Average	
		Waterbirds		CelebA		MNLI		CivilComms			
		Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Group annotation are required											
✓/✓	ERM†	83.8	66.4	95.5	55.1	81.6	72.0	84.3	74.0	86.3	66.9
	GroupDRO†	90.2	86.5	93.1	88.3	80.6	73.4	84.2	73.8	87.0	80.5
✗/✓	ERM†	97.3	72.6	95.6	47.2	82.4	67.9	83.1	69.5	89.6	64.3
	LfF†	91.2	78	85.1	77.2	80.8	70.2	68.2	50.3	81.3	68.9
	EIIL†	96.9	78.7	89.5	77.8	79.4	70.0	90.5	67.0	89.1	73.4
	JTT†	93.3	86.7	88.0	81.1	78.6	72.6	83.3	64.3	85.8	76.2
	CnC†	90.9	88.5	89.9	88.8	-	-	-	-	-	-
	AFR†	94.4	90.4	91.3	82.0	81.4	73.4	89.8	68.7	89.2	78.6
Group annotation are not required											
✗/✗	ERM†	83.5	66.4	95.4	54.3	82.1	67.9	81.3	67.2	85.6	63.9
	LfF†	86.6	75.0	81.1	53.0	71.4	57.3	69.1	42.2	77.1	56.9
	EIIL†	90.8	64.5	95.7	41.7	80.3	64.7	-	-	-	-
	JTT†	88.9	71.2	95.9	48.3	81.4	65.1	79.0	51.0	86.3	58.9
	LS†	91.2	86.1	87.2	83.3	78.7	72.1	-	-	-	-
	BAM†	91.4	89.1	88.4	80.1	80.3	70.8	88.3	79.3	87.1	79.8
	GICC†	89.3	85.4	92.1	89.5	-	-	89.7	72.3	-	-
	XRM†	89.3	88.1	91.4	89.1	75.8	72.1	84.0	72.2	85.1	80.4
	RAE (Ours)	90.2	87.0	90.3	88.4	75.3	72.6	84.5	72.1	85.1	80.0

Table 2: Performance comparison between baselines and our method. 'Avg' and 'Worst' denote average and worst-group accuracy, respectively. '†' indicates results from original papers, and '–' denotes not reported values. 'tr/val' indicates the use of group labels during training/validation.

	Oracle		Balanced		Class-ratio		Mistake		Ours	
	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Waterbirds	90.2	86.5	77.9	54.9	76.7	52.3	87.8	79.7	**90.2**	**87.0**
CelebA	93.1	88.3	94.1	66.1	94.2	64.4	92.0	80.0	**90.3**	**88.4**

Table 3: Performance comparison under different minority group discovery strategies. Oracle uses ground-truth group annotations. Balanced and Class-ratio define minority groups via random partitioning with equal sizes and class-proportional sizes, respectively. Mistake uses the same discovery procedure as our method but treats only misclassified samples as the minority group.

oracle GroupDRO, which assumes access to true group annotations, while requiring no group supervision. Notably, our method outperforms all validation-time group-annotated baselines on the validation dataset. Among approaches that do not rely on group annotations, our method achieves strong performance in terms of both average accuracy and worst-group accuracy.

Experiments on Different Group Discovery Table 3 presents a performance comparison across different minority group discovery strategies. As an upper bound, we report results using oracle group annotations. To assess the effectiveness of automatic group discovery, we further consider three alternative strategies for constructing minority groups. Our approach achieves performance close to the oracle setting, whereas the other variants exhibit substantially weaker worst-group accuracy. In particular, defining the minority group solely based on misclassified samples improves average performance but yields limited gains in worst-group accuracy. This suggests that mistake-based grouping captures only a subset of vulnerable samples and fails to model the entangled structure of spurious and invariant features underlying subpopulation failures.

4.3 Analysis

Effectiveness of label interpolation To evaluate the effectiveness of label interpolation, we compare our reliability-aware strategy with three variants: **None**, which applies no label interpolation; **Random**, which assigns randomly interpolated labels; and **Error-based**, which flips labels only for misclassified samples. All variants share the same training objectives and differ only in their label

interpolation strategies. We additionally report standard ERM without group annotations as a lower bound.

Without label interpolation, our method already outperforms ERM on CelebA, indicating that the proposed training objectives, particularly split conformal calibration for identifying vulnerable samples, effectively leverage uncertainty during training. While error-based interpolation improves worst-group accuracy on CelebA, it performs poorly on Waterbirds, even underperforming random interpolation. In contrast, our calibrated reliability-based label interpolation consistently achieves the best performance on both Waterbirds and CelebA, demonstrating its robustness across datasets with different entanglement characteristics.

	Waterbirds		CelebA	
	Avg	Worst	Avg	Worst
ERM	83.5	66.4	95.4	54.3
None	76.1	50.3	92.9	75.6
Random	84.3	62.9	90.6	75.6
Error-based	84.7	59.6	89.7	84.4
Ours	**89.4**	**87.0**	**90.3**	**88.4**

Table 4: Performance comparison under different label interpolation strategies

When did RAE fail and succeed? Although RAE achieves near-oracle performance on average, it exhibits degraded worst-group accuracy on certain datasets such as Waterbirds and CelebA. To better understand these limitations, we analyze both the group discovery stage (Phase 1) and the downstream robust optimization stage (Phase 2).

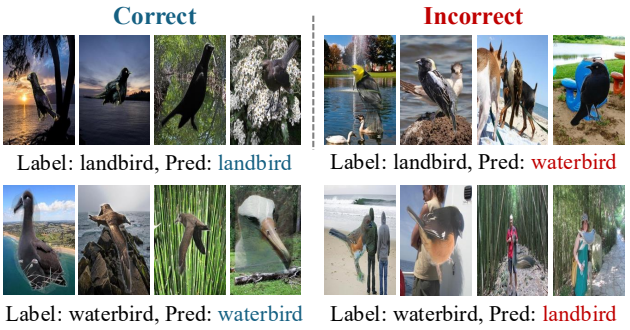


Figure 2: Error analysis.

As illustrated in Figure 4, some minority groups exhibit high discovery recall but still remain challenging during Phase-2 training. For example, the *water+landbird* group achieves 97.8% recall during Phase-1 discovery, yet the Phase-2 classifier only reaches 83.2% accuracy on this group (379 test errors). This indicates that the failure does not stem from incorrect environment discovery.

To further investigate this discrepancy, we qualitatively inspect the Phase-2 misclassified samples. As illustrated in Figure 2, many failure cases occur in visually complex scenes containing overlapping objects (e.g., birds appearing together with humans, dogs, or other birds). In such scenarios, the visual signal becomes ambiguous at the object level, as multiple objects interact or partially occlude each other. These cases are inherently more challenging than typical background-spurious examples.

Taken together, these observations suggest two bottlenecks. First, data scarcity limits what GroupDRO can learn even when minority environments are correctly identified during Phase 1. Second, visually complex scenes involving multi-object interactions introduce an additional challenge beyond what reliability-based environment discovery can address. These findings indicate that improving representation robustness for visually complex samples, for example through representation-level regularization such as manifold mixup (Verma et al., 2019), may further complement reliability-aware environment discovery. We leave the integration of such extensions to future work.

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 leverages calibrated prediction reliability as an auxiliary signal to identify vulnerable subgroups in the context of subpopulation shifts. By integrating split conformal calibration and label interpolation into the training network, RAE discovers an effective subgroup with low reliability. Empirical results on both vision and language benchmarks demonstrate that RAE consistently improves worst-group accuracy while maintaining competitive average performance. Moreover, the discovered environments closely approximate oracle group supervision, outperforming existing annotation-free discovery strategies that rely solely on errors or spurious-attribute classifiers. These findings highlight the importance of reliability as a signal for subgroup vulnerability in the presence of feature entanglement.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

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.

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.

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.

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.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

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.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pp. 6438–6447. PMLR, 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.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *arXiv preprint arXiv:2409.20534*, 2024.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

A Appendix

A.1 Training Configuration

Backbone Architectures. For fair comparison, we follow the backbone architectures commonly used in prior subpopulation shift benchmarks. 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. The encoder is frozen during training, and only the linear classification layer on top of the encoder is updated. For text datasets (CivilComments and MNLI), we use a pretrained BERT-base Devlin et al. (2019) model as the encoder. All baseline methods use the same backbone architectures for each dataset.

Hyperparameter Selection We follow the training protocol and most hyperparameter settings of prior work (Pezeshki et al., 2023). We additionally introduce β for conformal calibration training objectives.

Algorithms	Params	Model	
		ResNet	BERT
RAE	learning rate	$10^{\text{Uniform}(-5,-3)}$	$10^{\text{Uniform}(-6,-4)}$
	weight decay	$10^{\text{Uniform}(-6,-3)}$	$10^{\text{Uniform}(-6,-3)}$
	batch size	$2^{\text{Uniform}(5,3)}$	$2^{\text{Uniform}(4,6)}$
	dropout	-	Random
	β	$[0.01, 0.3]$	
GroupDRO	η	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-3,-1)}$

Training Time All experiments are conducted on a single NVIDIA RTX 3090 GPU. Training RAE on Waterbirds takes approximately 60 seconds, comparable to the training time of other automatic environment discovery baselines Pezeshki et al. (2023); Han & Zou (2024), which also take 60 seconds on Waterbirds.

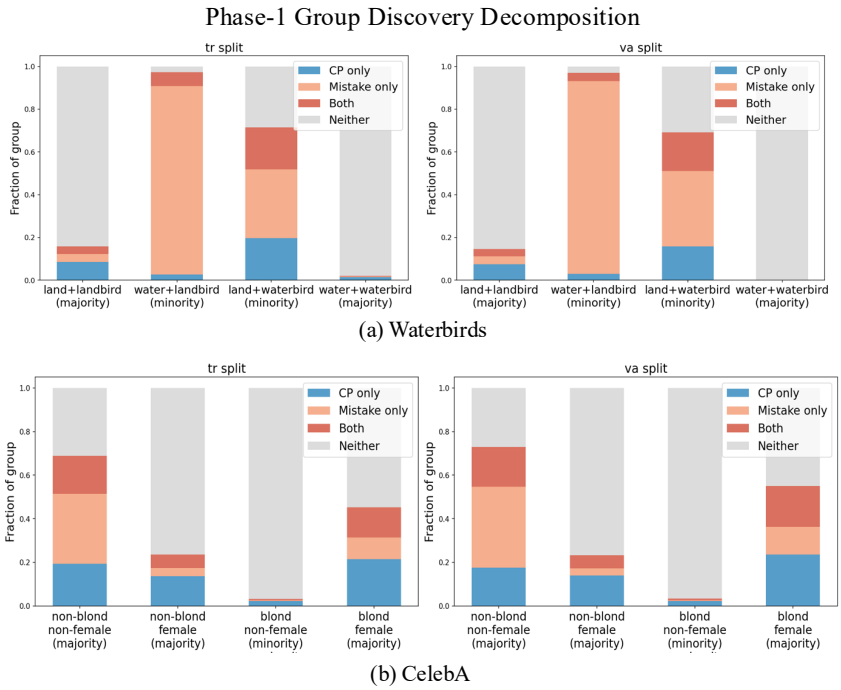
A.2 Dataset Details

Dataset	Num. of Groups	Num. of Classes	Total Samples
Waterbirds	2 {land, water}	2 {landbird, waterbird}	11,788
CelebA	2 {male, female}	2 {blond, non-blond}	202,599
CivilComments	8 {identity attributes: gender, age, ethnicity}	2 {toxic, non-toxic}	242,436
MultiNLI	2 {negation (present/absent)}	3 {entailment, neutral, contradiction}	412,349

A.3 Analysis

Group Discovery Decomposition To better understand how environments are discovered during Phase 1, we analyze which criterion is responsible for assigning samples to the vulnerable group. Figure 3 decomposes the discovered environments according to whether samples are identified through prediction mistakes, conformal unreliability, both signals, or neither.

The results show that the two criteria capture complementary failure modes. Mistake-based discovery accounts for most minority samples, while conformal unreliability becomes more important when the minority group is extremely small.



(a) Waterbirds

(b) CelebA

Figure 3: 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 based on prediction mistakes (Mistake only), unreliability (CP only), both criteria (Both), or neither.

Relationship Between Phase-1 Group Discovery and Phase-2 Accuracy We analyze the relationship between minority-group recall in Phase-1 and Phase-2 accuracy. We observe that some minority groups are well discovered in Phase-1 but still remain difficult in Phase-2, suggesting that downstream robustness is influenced not only by group discovery quality but also by data scarcity and visual complexity.

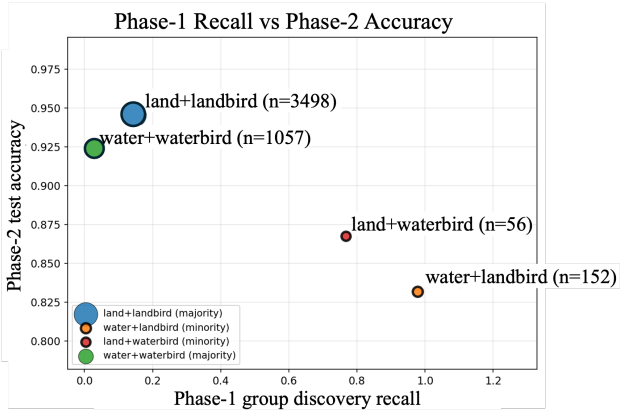


Figure 4: Relationship between Phase-1 group discovery recall and Phase-2 test accuracy across groups. Bubble size denotes the number of training samples in each group.

Coverage Analysis To examine whether the samples identified as unreliable by our criterion in Eq. (16) correspond to vulnerable subpopulations, we analyze conformal coverage across subgroups on Waterbirds. Figure 5 shows that while marginal coverage remains close to the nominal target, subgroup coverage varies substantially. In particular, the land+waterbird minority group exhibits substantially lower coverage (0.590 vs. 0.90 nominal). Since lower coverage implies higher non-

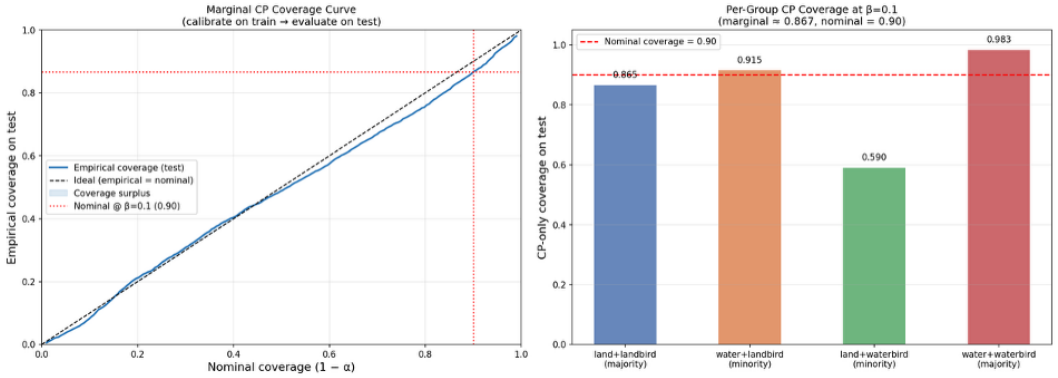


Figure 5: Conformal coverage across subgroups.

conformity scores relative to the threshold q, this indicates that nonconformity scores are higher in distribution-shifted subpopulations. Consequently, our criterion in Eq. (16) tends to flag samples from these vulnerable minority groups, supporting the use of nonconformity scores as a proxy for identifying vulnerable subgroups.

A.4 Training Procedure

Algorithm 1 summarizes the full procedure. Our method trains two models with identical architectures: a held-in branch f_θ^{in} and a held-out branch f_θ^{out}. Each training sample is assigned to one of the branches through a random assignment.

Algorithm 1 Reliability-Aware Learning and Environment Discovery (RAE)

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$; twin networks $f_\theta^{in}, f_\theta^{out}$; miscoverage β
Output: Environment partition $\hat{G}_{\text{vul}}, \hat{G}_{\text{oth}}$
// Phase I: Reliability-guided learning
Fix random assignment $a_i \in \{0, 1\}$; initialize $y_i^{\text{dyn}} \leftarrow y_i$
while not converged **do**
  For each (x_i, y_i) in minibatch $\mathcal{B}$:
    compute in-prediction from the assigned branch f_θ^{in} and held-out prediction from f_θ^{out}
    compute calibration losses (quantile/coverage) via Eq. (6)–(9)
    update dynamic labels for unreliable samples via Eq. (11)–(12)
  Update θ by minimizing Eq. (13)
end while
// Phase II: Environment discovery
for $i = 1$ to N **do**
  Compute predictions and reliability scores
  Determine mistake/unreliable indicators via Eq. (14)
  Set $\hat{e}(i)$ and assign i to $\hat{G}_{\text{vul}}$ (if $\hat{e}(i) = 1$) else $\hat{G}_{\text{oth}}$ via Eq. (15)
end for
return $\hat{G}_{\text{vul}}, \hat{G}_{\text{oth}}$

In **Phase I**, the two branches are jointly optimized. The held-out branch estimates predictive uncertainty, which serves as a reliability proxy. These reliability estimates are used to construct reliability guided labels for the held-in branch via label interpolation. Meanwhile, the held-out branch is optimized to produce calibrated uncertainty through the quantile and coverage objectives. After Phase I, the calibrated held-out branch is used to compute quantile thresholds and identify vulnerable samples in the training set. Based on these reliability signals, the dataset is partitioned into environments $\hat{G}_{\text{vul}}$ and $\hat{G}_{\text{oth}}$.

In **Phase II**, we train the model is optimized through GroupDRO algorithms on the discovered environments $\hat{G}_{\text{vul}}$ and $\hat{G}_{\text{oth}}$.