Persistent Gradient Disagreement Reweighting: Detecting Shortcut Biases Without Training Group Labels

Nishant Sharma

Abstract

Deep learning models can learn rules that are predictive in the training distribution but misaligned with the intended concept. This report studies shortcut learning as an incorrect inductive bias, focusing on cases where models rely on spurious attributes such as image background or shallow lexical overlap. Waterbirds is used as the primary empirical setting: the target label is bird type, but land and water backgrounds are correlated with the label during training. I propose Persistent Gradient Disagreement Reweighting (PGDR), an approach that does not use training group labels to construct its selected set or training loss, and instead uses last-layer gradient geometry to identify examples that may conflict with the learned shortcut. The report evaluates gradient disagreement as a pseudo-group discovery signal, compares PGDR with ERM, JTT, and GroupDRO, and analyzes how warmup quality and refresh affect the reliability of the signal.

1 Introduction

Inductive bias is necessary for learning: a model must prefer some generalizations over others when many functions are consistent with the training data. The failure studied in this report occurs when that preference selects a rule that is predictive in the training distribution but misaligned with the intended concept. In deep learning, this often appears as shortcut learning: the model learns an easier correlated feature instead of the semantic feature the task is meant to measure [2, 13].

This problem appears across both vision and language benchmarks. In CelebA, hair-color classification can become entangled with gender presentation, so a model may rely on gender-associated visual cues instead of the hair-color attribute itself [7]. In Waterbirds, the target is landbird versus waterbird classification, but the training distribution correlates bird type with background: landbirds are often on land and waterbirds are often on water [5, 10]. In MNLI-style natural language inference, models can rely on shallow lexical or syntactic heuristics, such as assuming that high word overlap between premise and hypothesis implies entailment; HANS exposes these heuristics by constructing examples where they fail [8, 11]. These datasets differ in modality, but they share the same underlying structure: a spurious attribute is correlated with the label for many training examples, so ERM can achieve high average performance while failing on examples where the correlation breaks.

The Clever Hans Mirage survey frames this as a broad spurious-correlation problem: models can appear to solve a task while actually relying on non-essential cues, much like Clever Hans appeared to perform arithmetic while responding to subtle signals from the trainer [12]. The survey organizes existing responses into data-centric methods, representation-learning methods, post-hoc methods, optimization-based methods, and specialized methods. It also emphasizes several unresolved challenges: many approaches require group annotations or human-defined spurious attributes; mitigation often trades off worst-group robustness against average performance; and detecting why a model depends on a cue remains difficult.

This project is positioned within those gaps. PGDR is not a data augmentation method, because it does not create counterfactual images or text. It is not a fully group-aware method like GroupDRO, because it does not use training group labels to construct its selected set or loss. It is also not only a post-hoc diagnostic, because the discovered signal is used during training to reweight examples. Instead, PGDR is best viewed as a pseudo-group discovery and reweighting method without training group labels: it tries to infer a useful hidden group from the model's own training dynamics, then uses that inferred group to intervene.

The motivation is that shortcut-conflicting examples should not only be high-loss or misclassified; they may also request different parameter updates from the classifier. Easy shortcut-aligned examples of a class tend to reinforce the dominant learned rule. A hard example that conflicts with the shortcut may require a different update direction. PGDR operationalizes this idea by measuring the cosine alignment between a hard example's last-layer gradient and the average gradient of easy same-class examples. Low alignment is treated as evidence that the example may belong to a hidden shortcut-conflicting group.

The goal of this report is twofold. First, it evaluates whether gradient disagreement is an empirical pseudo-group discovery signal rather than only a plausible intuition. Second, it tests whether reweighting the examples

selected by this signal improves robustness while preserving average accuracy. In the taxonomy of the Clever Hans Mirage survey, PGDR sits at the intersection of concept or pseudo-label discovery and optimization-based mitigation, with a focus on detection without training group labels and the robustness–utility tradeoff.

2 Related Work

Spurious correlations and incorrect inductive bias. This project studies shortcut learning as a form of incorrect inductive bias: the model learns a feature that is predictive in the training distribution but not part of the intended concept. Geirhos et al. [2] describe this behavior as shortcut learning, where neural networks exploit decision rules that work on standard data but fail under distribution shift. Ilyas et al. [3] make a related point through non-robust features: features can be genuinely predictive under the data distribution while still being brittle or misaligned with human perception. The Clever Hans Mirage survey places these failures under the broader category of spurious correlations, emphasizing that models may appear to solve a task while relying on non-essential cues such as background, texture, demographic attributes, or shallow linguistic patterns [12].

Datasets for studying shortcut learning. CelebA, Waterbirds, and MNLI/HANS expose the same failure structure in different domains. CelebA provides face-attribute labels and has been widely used to study cases where a target attribute, such as hair color, becomes correlated with demographic or visual attributes such as gender presentation [7]. Waterbirds constructs a spurious correlation between bird class and background: landbirds are often shown on land backgrounds and waterbirds on water backgrounds, so background can become an easier rule than bird recognition [5, 10]. MultiNLI provides a natural-language counterpart: NLI models can exploit annotation artifacts or shallow heuristics rather than sentence meaning [11]. HANS makes this failure explicit by constructing examples where lexical overlap, subsequence, and constituent heuristics give the wrong entailment prediction [8].

Group robustness methods. A direct way to address spurious correlations is to optimize for worst-group performance. GroupDRO assumes access to group labels and minimizes the worst group loss rather than the average loss [10]. It is an important oracle-style baseline because it directly targets the evaluation metric, but its supervision requirement is strong: the training procedure needs group annotations.

Reweighting without training group labels. JTT is the closest practical baseline to PGDR because it does not require training group labels [6]. JTT first trains an ERM model, identifies examples that the first model misclassifies, and then trains a second model that upweights those examples. Its key assumption is that early mistakes are enriched for minority or shortcut-conflicting examples. PGDR shares the two-stage spirit but uses a different signal. Instead of selecting examples only because they are misclassified or high-loss, PGDR asks whether a hard example's gradient direction disagrees with the easy same-class gradient direction.

Learning from biased models. Several prior methods exploit the idea that biased learning appears early in training. Learning from Failure trains a biased model that tends to capture easy spurious patterns, then uses the biased model's failures to guide a debiased classifier [9]. This is related to PGDR because both methods treat the model's own failure pattern as useful evidence. The difference is that PGDR tests whether an example is geometrically inconsistent with the dominant same-class update direction.

Invariant and representation-based solutions. Invariant Risk Minimization approaches the problem from a different angle: it tries to learn representations for which the optimal classifier is stable across environments [1]. This is conceptually aligned with avoiding shortcut features, but it assumes access to meaningful environments and can be difficult to apply when those environments or groups are unavailable. Last-layer retraining provides another relevant perspective. Kirichenko et al. [4] argue that even when a model relies on spurious correlations, its representation may still contain useful core features, and retraining the classifier head can improve robustness. PGDR is compatible with this view because it studies last-layer gradients: the method assumes that the classifier head contains useful information about how shortcut-aligned and shortcut-conflicting examples push the decision boundary differently.

Position of this work. PGDR is positioned between pseudo-group discovery and optimization-based mitigation. Unlike GroupDRO, it does not use training group labels. Unlike JTT, it does not rely only on misclassification. Unlike purely diagnostic shortcut analyses, it uses the discovered signal for reweighting. The central question is whether gradient disagreement is an empirically valid signal for hidden shortcut-conflicting examples, and whether that signal can support a useful robustness intervention.

3 Experimental Framing: Waterbirds and Baselines

The main empirical setting is Waterbirds, a binary image classification benchmark designed to expose spurious

Figure 1: Examples from the four Waterbirds groups [5, 10]. The shortcut-aligned majority groups preserve the training correlation between bird type and background; the shortcut-conflicting minority groups break it.

background reliance [5, 10]. Each example has an image x_i, a bird label $y_i \in \{0, 1\}$, and a background attribute $a_i \in \{0, 1\}$. The model is trained to predict bird type, but the training distribution correlates bird type with background. This creates four groups:

$$g_i = 2y_i + a_i. \tag{1}$$

The shortcut-aligned majority groups are landbird-on-land and waterbird-on-water. The shortcut-conflicting minority groups are landbird-on-water and waterbird-on-land.

The core failure is that average risk does not distinguish between correct concept learning and shortcut learning. A classifier can achieve high average accuracy by using background when the training distribution makes background predictive. The intended concept, however, is bird type. Therefore, the evaluation must measure performance on the groups where the shortcut breaks. I use worst-group accuracy:

$$\mathrm{WGA} = \min_{g \in \mathcal{G}} \mathrm{Acc}(g). \tag{2}$$

Worst-group accuracy is the central robustness metric because it directly captures whether the model fails on the hidden shortcut-conflicting groups.

The comparison methods are chosen to separate three different levels of information. ERM uses only image-label pairs and minimizes average cross-entropy loss:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i). \tag{3}$$

Table 1: Matched Waterbirds comparison, reported as mean ± standard deviation over three seeds. GroupDRO uses train group labels and is therefore an oracle reference.

Method	Uses train group labels?	Test WGA (%)	Role
ERM	No	68.38 ± 0.78	Average-loss baseline
JTT	No	85.41 ± 1.04	Strong label-free WGA baseline
GroupDRO	Yes	83.84 ± 0.26	Supervised oracle reference
PGDR-FixedW	No	81.52 ± 0.91	Gradient-based pseudo-group method

ERM is the natural baseline because it exposes the incorrect inductive bias: when the shortcut is easier and frequent, average-loss training can prefer it.

JTT is the strongest label-free baseline in the matched Waterbirds experiments [6]. It first trains an ERM model, identifies the examples misclassified by that model,

$$E_{\mathrm{JTT}} = \{i : \hat{y}_i \neq y_i\}, \tag{4}$$

and then trains a second model that upweights those examples. JTT is relevant because it also tries to infer hidden difficult examples without training group labels. Its signal is error membership.

GroupDRO is the supervised oracle reference [10]. It assumes access to group labels during training and minimizes the worst group loss:

$$\min_{\theta} \max_{q \in \Delta^{G-1}} \sum_{g=1}^{G} q_g L_g(\theta). \tag{5}$$

Because GroupDRO uses training group labels, it is not directly comparable to label-free methods, but it gives useful context for the performance that can be achieved when the hidden groups are known.

All methods use the same Waterbirds train, validation, and test split, an ImageNet-pretrained ResNet-50 backbone, three seeds, a total training budget of 150 epochs, and model selection by validation worst-group accuracy. PGDR does not use training group labels to construct $W^\star$ or to compute the training loss. Group labels are used for validation-based checkpoint selection, test evaluation, and mechanism diagnostics.

Table 1 shows the main robustness result. PGDR-FixedW improves worst-group accuracy from 68.38 ± 0.78 for ERM to 81.52 ± 0.91, an approximately 13-point gain without using training group labels. It also preserves high average accuracy, achieving 92.56 ± 0.48.

However, JTT remains stronger on pure worst-group accuracy at 85.41 ± 1.04. This separates two evaluation goals that can otherwise be conflated. Worst-group accuracy rewards broad coverage of minority examples, while average accuracy also penalizes interventions that over-correct or add noise to majority groups. JTT upweights the broader error set from a first-stage model, which can cover more minority examples and help worst-group

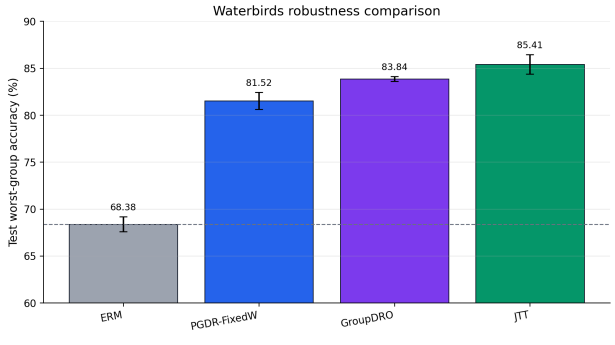

Figure 2: Matched Waterbirds worst-group accuracy comparison. GroupDRO uses training group labels and is included as an oracle reference.

accuracy. PGDR selects a narrower pseudo-group using gradient disagreement, which improves over ERM and preserves average accuracy, but can miss some minority examples. The appropriate interpretation is that PGDR is not optimized solely for maximizing WGA in this run; it is a mechanism-driven pseudo-group discovery method without training group labels and with a different accuracy–robustness tradeoff.

4 Methodology: Persistent Gradient Disagreement Reweighting

4.1 Problem Setup

The goal of PGDR is to improve robustness without using training group labels. During training, the method observes only image-label pairs (x_i, y_i). The hidden background attribute a_i and group identity g_i are not used to construct the selected set. They are used only after training for validation-based checkpoint selection, test evaluation, and mechanism diagnostics.

The method starts from the hypothesis that shortcut-conflicting examples are not only high-loss examples. They can also ask for a different parameter update than easy shortcut-aligned examples of the same class. If easy landbird examples mostly share land backgrounds, their gradients may reinforce a landbird-with-land decision rule. A landbird on water has the same class label but violates that shortcut, so its classifier update can point in a different direction. PGDR tries to detect this difference using last-layer gradient geometry.

4.2 Comparison to Existing Methods and Novelty

PGDR is closest in spirit to JTT and Learning from Failure because all three use an imperfect early model to expose shortcut-related failures [6, 9]. The difference is the selection signal. JTT selects examples that are misclassified by the first model. Learning from Failure uses a biased model to guide a debiased model. PGDR instead selects examples whose gradients disagree with the easy same-class gradient direction.

This distinction matters because loss and error are scalar difficulty signals. They indicate that an example is hard, but they do not specify what kind of update the example demands. Gradient direction contains more structure: two examples can both be high-loss while pushing the classifier in different directions. PGDR uses this directional information to separate ordinary hard examples from shortcut-conflicting examples.

PGDR also differs from GroupDRO because it does not assume known groups. GroupDRO directly optimizes worst-group loss over provided group labels. PGDR attempts to infer a pseudo-group $W^\star$ from training dynamics and then reweights that pseudo-group. In the taxonomy of the Clever Hans Mirage survey, PGDR sits between pseudo-label or concept discovery and optimization-based mitigation [12].

4.3 PGDR Setup and Hyperparameters

The implementation uses an ImageNet-pretrained ResNet-50 with a two-class linear classifier head. Images are resized to 224×224, normalized with ImageNet statistics, and trained with random horizontal flips. The optimizer is SGD with momentum 0.9. The full training budget is 150 epochs. Warmup uses learning rate 10^{-5} and weight decay 1.0 for 60 epochs; the main phase uses learning rate 10^{-4} and weight decay 10^{-2}.

After warmup, define the class margin

$$m_i = z_{i,y_i} - \max_{c \neq y_i} z_{i,c}. \tag{6}$$

The hard set is the bottom-margin subset:

$$H = \{i : m_i \leq q_\rho(\{m_j\}_{j=1}^n)\}, \quad \rho = 0.15. \tag{7}$$

The easy set is

$$E = \{i : i \notin H\}. \tag{8}$$

For classifier head W and representation $\phi(x_i)$, the last-layer gradient is

$$g_i = \nabla_W \ell_i = (p_i - e_{y_i})\phi(x_i)^\top, \tag{9}$$

where $p_i = \mathrm{softmax}(z_i)$ and e_{y_i} is the one-hot label vector. For cosine computation, g_i is flattened into a vector. For each class c, define the confident easy same-class set

$$E_c^{\mathrm{conf}} = \{j \in E : y_j = c, \ \max_r p_{j,r} > 0.85\}. \tag{10}$$

If this set has fewer than 10 examples, the implementation falls back to all easy same-class examples. Let E_c^{ref} denote whichever set is used after this fallback. The reference gradient is

$$\bar{g}_c = \frac{1}{|E_c^{\mathrm{ref}}|} \sum_{j \in E_c^{\mathrm{ref}}} g_j, \tag{11}$$

The cosine alignment score is

$$s_i^{(t)} = \cos(g_i^{(t)}, \bar{g}_{y_i}^{(t)}). \tag{12}$$

Table 2: PGDR implementation settings.

Component	Value
Backbone	ImageNet-pretrained ResNet-50
Seeds	0, 1, 2
Batch size	64
Total epochs	150
Warmup epochs	60
Warmup optimizer	SGD, lr 10^{-5}, weight decay 1.0, momentum 0.9
Main optimizer	SGD, lr 10^{-4}, weight decay 10^{-2}, momentum 0.9
Hard-set fraction ρ	0.15
Persistence snapshots T	3
Required votes k	3
Low-cosine quantile	0.25
Confidence threshold for easy reference	0.85
Minimum $W^\star$ size	20
Loss mixing α	0.5
Refresh	Disabled in final FixedW method

Low $s_i^{(t)}$ means weaker alignment with the easy same-class update direction. The cosine does not need to be negative; the signal is relative within the hard set.

At snapshot t, example i receives a vote if it falls in the bottom cosine quantile:

$$v_i^{(t)} = \mathbf{1}\left[s_i^{(t)} \le q_{0.25}^{(t)}\right]. \tag{13}$$

The selected pseudo-group is

$$W^\star = \left\{ i \in H : \sum_{t=1}^{T} v_i^{(t)} \ge k \right\}, \quad T = 3, \quad k = 3. \tag{14}$$

A minimum selected-set size of 20 is enforced. If enough persistent examples exist, the method selects up to the target budget $\max(20, 0.25|H|)$ ranked by mean cosine.

The final training objective is

$$\mathcal{L}_{\mathrm{PGDR}} = \alpha \frac{1}{|E|} \sum_{i \in E} \ell_i + (1-\alpha) \frac{1}{|W^\star|} \sum_{i \in W^\star} \ell_i, \quad \alpha = 0.5. \tag{15}$$

The final FixedW method keeps $W^\star$ fixed after selection because naive refresh was unstable in ablations.

5 Mechanism Evidence and Failure Analysis

A central requirement for this method is to show that gradient disagreement is an empirically meaningful signal, rather than only a plausible intuition. I test this directly before interpreting the final accuracy results. Among hard examples H, the question is whether low cosine-to-easy-gradient predicts hidden minority membership better than chance. The positive class is whether an example belongs to a shortcut-conflicting Waterbirds group:

$$\mathbb{1}_{\mathrm{minority}}(i) = \mathbb{1}[g_i \in \{1, 2\}]. \tag{16}$$

The method selects low cosine scores s_i. For AUC and AP, I use

$$d_i = -s_i \tag{17}$$

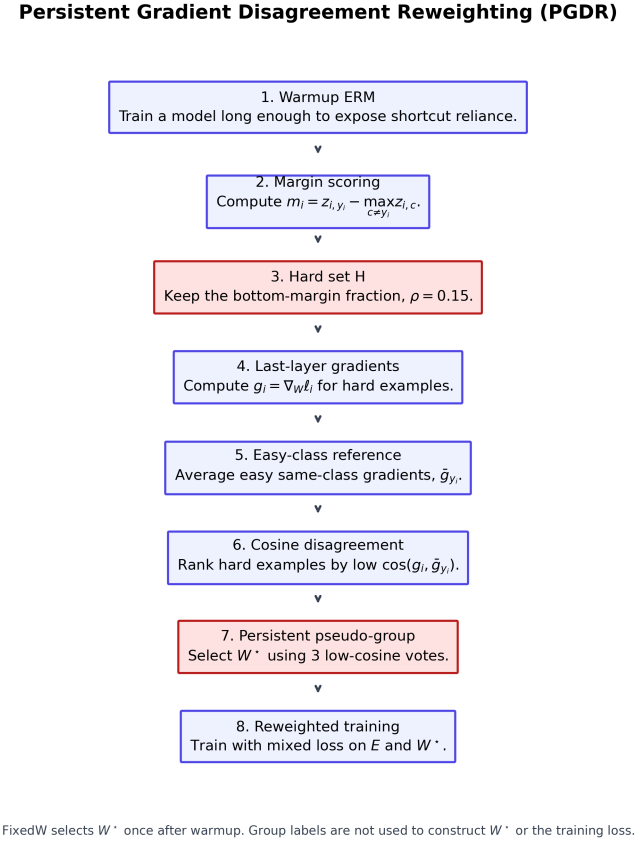

Figure 3: PGDR pipeline: warmup, margin-based hard-set construction, gradient-disagreement scoring, persistent pseudo-group selection, and reweighted training.

Table 3: Predictive validity of low cosine-to-easy-gradient inside the hard set H. AUC, AP, base rate, and precision@budget are means across runs; permutation p is the median across runs. AP should be compared with the minority base rate in H.

Variant	AUC	AP	Base Rate in H	Precision@Budget	Median perm. p
FixedW	0.800	0.576	0.296	0.595	0.0002
NoPersistence	0.808	0.586	0.296	0.596	0.0002
RefreshNoOracle	0.465	0.280	0.294	0.258	0.9188
WeakWarmup	0.575	0.265	0.245	0.268	0.0006

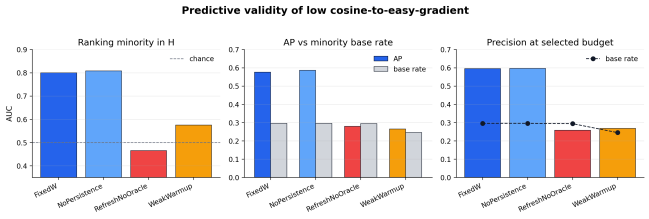

Figure 4: Predictive validity of low cosine-to-easy-gradient inside the hard set. Average precision and precision@budget should be compared with the minority base rate in H.

so that larger values mean stronger disagreement. This uses group labels only for evaluation, not for training or selection.

Table 3 supports the gradient-disagreement signal. In the FixedW setting, AUC 0.800 means that a random shortcut-conflicting example inside H receives a stronger disagreement score than a random shortcut-aligned hard example about 80% of the time. AP is 0.576 while the minority base rate in H is only 0.296, so the ranking nearly doubles the enrichment expected from a random ranking. Precision@Budget is 0.595, meaning that the examples selected at the PGDR budget are about 59.5% minority. The permutation p-value of 0.0002 indicates that this ranking is unlikely under shuffled hidden-group labels.

The ablations show where the method works and where it breaks. NoPersistence has similar predictive validity and similar WGA to FixedW, so the raw low-cosine score already contains strong signal in these runs. Persistence is still a reasonable conservative design, but the current evidence does not prove it is the decisive ingredient.

WeakWarmup is the clearest failure mode. Its WGA drops to 66.51 ± 0.47, close to ERM. Its AUC is only 0.575 and AP is 0.265 against a 0.245 base rate. This means the signal is statistically detectable but practically weak. The interpretation is that warmup must expose shortcut reliance before gradient disagreement can reveal hidden shortcut-conflicting structure.

RefreshNoOracle is primarily a selection-signal failure. Its AUC is 0.465, AP is below the hard-set base rate, and the permutation p-value is 0.9188. This suggests that once training begins correcting the selected pseudo-group, the gradient geometry changes. Reapplying the same low-cosine rule later can select noisy examples. This is why the final method uses fixed $W^\star$.

Table 4: PGDR robustness ablations. WGA values are mean $\pm$ standard deviation where multiple seeds were run; RefreshNoOracle is reported as a diagnostic selection-signal ablation.

Variant	Test WGA (%)	Interpretation
FixedW	81.52 ± 0.91	Main method
NoPersistence	80.32 ± 0.33	Similar WGA; persistence not proven necessary
WeakWarmup	66.51 ± 0.47	Warmup quality is critical
RefreshNoOracle	Diagnostic	Refresh damages selection quality

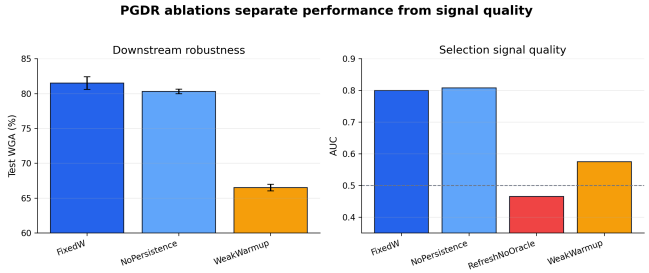

Figure 5: PGDR ablations show that downstream robustness and selection-signal quality can diverge. Weak warmup hurts both robustness and predictive validity, while refresh primarily damages the selection signal.

This analysis also explains why JTT can be better on WGA while PGDR preserves higher average accuracy. JTT upweights a broader first-stage error set, while PGDR selects a smaller pseudo-group whose selected-budget precision is 0.595 in the FixedW predictive-validity test. Broader coverage can help worst-group accuracy. A cleaner but smaller intervention can preserve average accuracy, but it may leave some worst-group examples underweighted.

6 Current Limitations and Future Work

The first limitation is warmup dependence. PGDR works when warmup exposes shortcut reliance clearly enough for shortcut-conflicting examples to become geometrically unusual. If the warmup model is too weak, or if it has not learned the shortcut in a stable way, the pseudo-group becomes noisy.

The second limitation is the precision-recall tradeoff. FixedW selects a cleaner pseudo-group, but it does not cover all minority examples. This helps explain why PGDR improves over ERM but does not beat JTT on WGA in the matched Waterbirds run. Future work should test adaptive budgets or confidence-calibrated selection rules that increase recall without losing too much precision.

The third limitation is refresh instability. Naive refresh assumes that the same gradient-disagreement rule remains valid after reweighting changes the classifier. The RefreshNoOracle result suggests this assumption is false. A better refresh method would need a stability criterion,

a validation-based acceptance rule, or a way to compare new pseudo-groups against earlier selections.

The fourth limitation is model selection. PGDR does not use training group labels to construct $W^\star$ or compute the loss, but validation worst-group accuracy still requires group labels for checkpoint selection. A more deployment-ready version would need a group-label-free model selection criterion.

The fifth limitation is scope. Waterbirds is the main benchmark in this report. The MNLI/HANS experiment is only exploratory: it suggests that gradient disagreement can produce a signal beyond vision, but it does not establish competitive NLI robustness. Future work should evaluate PGDR more systematically on CelebA, MultiNLI group-shift, CivilComments, and other spurious-correlation benchmarks.

7 Conclusion

This report studies shortcut learning as an incorrect inductive bias: the model learns a rule that is predictive in the training distribution but misaligned with the intended concept. On Waterbirds, that rule is background reliance. PGDR addresses this setting by discovering a pseudo-group from persistent gradient disagreement, without using training group labels to construct the pseudo-group or training loss.

The empirical evidence supports the use of gradient disagreement as more than an intuition. Inside the hard set, low cosine-to-easy-gradient predicts hidden minority membership with AUC 0.800, AP 0.576 against a 0.296 base rate, precision@budget 0.595, and a permutation p-value of 0.0002. Reweighting the selected pseudo-group improves Waterbirds WGA substantially over ERM while preserving high average accuracy. At the same time, JTT remains stronger on pure WGA in this matched run. The strongest positioning of PGDR is therefore as a mechanism-driven pseudo-group discovery method, without training group labels, that exposes a gradient-space signature of incorrect shortcut learning.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

[2] 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*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

[3] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019. URL https://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.

[4] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022.

[5] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

[6] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

[7] 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*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

[8] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019. doi: 10.18653/v1/P19-1334. URL https://arxiv.org/abs/1902.01007.

[9] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier, 2020.

[10] 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. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

[11] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

[12] Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Megan Coffee, Zhe Zeng, Sheng Li, Ziran Wang, Ting-Hao Kenneth Huang, James M. Rehg, Henry Kautz, and Aidong Zhang. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=kIuqPmS1b1.

[13] Eliezer Yudkowsky. Inductive bias. https://www.lesswrong.com/posts/H59YqogX94z5jb8xx/inductive-bias, 2007. Accessed 2026-04-28.