PSEUDO GROUP-GUIDED AND RELIABILITY-AWARE DEBIASING FRAMEWORK FOR UNSUPERVISED BIAS MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks often exploit spurious correlations between target labels and task-irrelevant attributes, leading to substantial performance degradation on samples that deviate from dominant spurious patterns. To address this issue, we propose a selective and reliability-aware debiasing framework for settings where spurious attributes are unavailable. Our approach extends representation-based pseudo-group discovery through Gaussian Noise Intervention (GNI), which evaluates the stability of intra-class geometric evidence under bounded local perturbations and derives a conflict score for selective pseudo-group inference. Within each target class, samples are divided into pseudo-aligned, pseudo-conflicting, and uncertain groups. Retained pseudo-conflicting samples are further assigned reliability scores based on multi-view consistency and class-wise ranking. These pseudo-group assignments and reliability estimates are incorporated into a reliability-weighted multi-level debiasing objective, where they play distinct roles in emphasizing bias-conflicting samples and modulating their debiasing contributions. To further mitigate residual shortcut influence, we employ trained bias experts to perform Debiased Correction Of Residual Effects (D-CORE) at inference stage. Experiments on medical and natural-image benchmarks demonstrate improved worst-group robustness across most evaluated settings while maintaining competitive average accuracy.

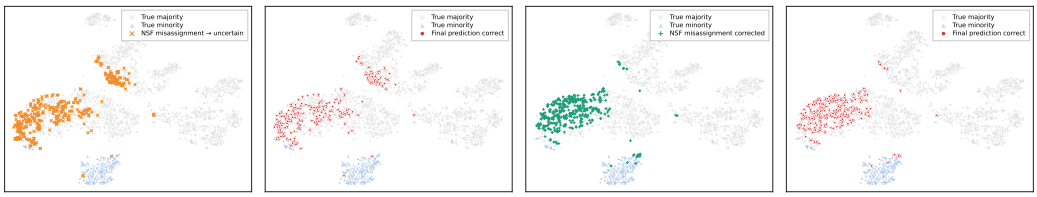
1 INTRODUCTION

Deep neural networks can exploit correlations between target labels and task-irrelevant attributes, achieving strong average performance while failing on samples that violate dominant spurious patterns. In medical imaging, for example, classifiers may rely on device-specific watermarks rather than pathology-related evidence, while models trained on Waterbirds may associate bird categories with background type. This problem becomes particularly challenging when spurious attributes are unavailable, since the latent target–attribute groups cannot be used directly during training or model selection.

Representation geometry provides a useful source of sample-level evidence for latent bias conflict. Neutralizing Spurious Features (NSF) (Li et al., 2025) characterizes samples through their relative geometry with respect to class-specific reference representations. However, its geometric evidence is evaluated from a single clean representation and converted into a deterministic binary pseudo-group assignment. Consequently, ambiguous samples are forced into a pseudo group, while assignments supported by different levels of evidence are treated uniformly. As illustrated in Figure 1, these limitations can introduce unreliable pseudo-group supervision even when the underlying samples remain informative for the target task.

To address this issue, we propose a selective and reliability-aware debiasing framework based on Gaussian Noise Intervention (GNI). Instead of relying on a single geometric observation, GNI evaluates the NSF relative-distance criterion over multiple bounded local perturbations and aggregates the resulting responses into intra-class evidence for pseudo-group discovery. This enables selective assignment into pseudo-aligned, pseudo-conflicting, and uncertain groups, while retained pseudo-conflicting samples are associated with sample-wise reliability estimates. The resulting signals are

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) **(b)** **(c)** **(d)**

Figure 1: **Selective pseudo-group inference and downstream prediction visualized with t-SNE.** Subfigures (a) and (c) highlight samples misassigned by the NSF binary rule in orange and green, respectively. Our procedure retains the samples in (a) as uncertain and reassigns those in (c) to the correct pseudo groups. Subfigures (b) and (d) show the correctly classified subsets of (a) and (c), respectively; each closely approximates its corresponding parent set.

incorporated into Reliability-Weighted Multi-Level Debiasing, where pseudo-group assignment and assignment reliability are used for distinct training roles. To further reduce residual shortcut influence after training, Debiased Correction Of Residual Effects (D-CORE) performs sample-adaptive inference-time correction using the trained bias experts.

Overall, the proposed framework combines selective pseudo-group discovery, reliability-aware debiasing, and inference-time correction without requiring spurious-attribute annotations during training or model selection. Experiments on medical and natural-image benchmarks demonstrate improved worst-group robustness while maintaining competitive average predictive performance.

2 Related Work

Group-robust learning seeks to improve performance on underrepresented groups that may be obscured by aggregate accuracy. When group annotations are available, Group DRO directly optimizes worst-group performance (Sagawa et al., 2020). In the absence of explicit bias annotations, prior methods instead derive debiasing signals from prediction errors, sample losses, model agreement, or salient features. Representative approaches include LfF and JTT (Nam et al., 2020; Liu et al., 2021), Biased Committee and Ada-ABC (Kim et al., 2022; Luo et al., 2025), and MaskTune (Asgari et al., 2022). Our framework differs in that its primary supervision is obtained from intra-class representation geometry rather than prediction behavior or feature masking.

A related line of work infers latent groups or environments before robust optimization. GEORGE clusters representations within each target class, EIIL infers environments from the behavior of a reference classifier, and NSF identifies samples that deviate from dominant intra-class geometry (Sohoni et al., 2020; Creager et al., 2021; Li et al., 2025). These approaches generally convert inferred structure into deterministic group or environment assignments. In contrast, our framework explicitly distinguishes ambiguous samples through selective assignment and further quantifies the reliability of retained pseudo-conflicting samples, allowing assignment uncertainty and reliability to affect subsequent optimization separately.

Representation-level debiasing provides another complementary direction. CNC aligns same-class representations across inferred bias groups, GERNE exploits gradient extrapolation across batches with different spurious-correlation strengths, and Partition-and-Debias introduces bias and debiasing experts with learned gating (Zhang et al., 2022; Asaad et al., 2025; Li et al., 2023). DFA and SelecMix further construct bias-conflicting training signals through feature recombination or bias-aware sample mixing (Lee et al., 2021; Hwang et al., 2022). Unlike these training-time interventions, the Gaussian perturbations in our method are used only to probe local representation geometry and do not augment the training distribution. Beyond training-stage debiasing, inference-time correction has also been explored for residual predictive bias (Cui et al., 2024); our D-CORE component uses the trained bias experts to construct sample-specific residual-shortcut signals for correcting the final decision margin.

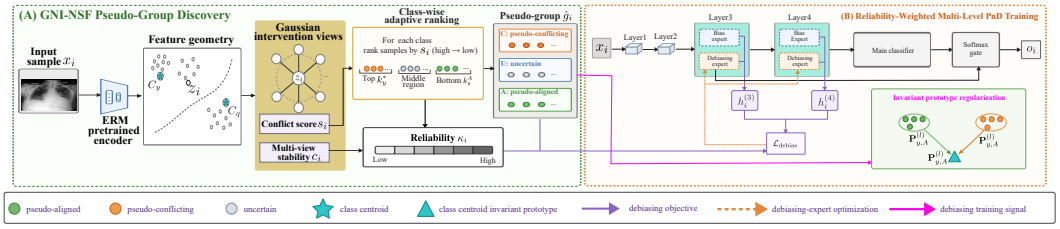


Figure 2: **Overview of our proposed two-stage training framework.** Stage 1 performs GNI-based selective pseudo-group discovery and reliability estimation, while Stage 2 conducts reliability-weighted multi-level debiasing with invariant-prototype regularization.

3 METHOD

3.1 PROBLEM FORMULATION AND METHOD OVERVIEW

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a supervised classification dataset with target label $y_i \in \{1, \dots, C\}$. Each sample may additionally be associated with an unobserved spurious attribute a_i, inducing the target–attribute group $g_i = (y_i, a_i)$. Neither a_i nor g_i is available during training, model selection, or inference. Within each target class, samples consistent with the dominant target–attribute association are termed *bias-aligned*, whereas those violating this association are termed *bias-conflicting*. Under target–attribute imbalance, ERM may preferentially encode the dominant spurious regularities represented by the more prevalent bias-aligned samples. We therefore use the ERM representation as the basis for extracting sample-level evidence of latent bias conflict rather than attempting to recover the true latent groups.

NSF infers latent bias structure from the relative geometry of sample representations with respect to class-specific references and converts the resulting geometric evidence into binary pseudo-group assignments. However, this evidence is evaluated from a single clean representation, without explicitly distinguishing geometrically ambiguous samples or quantifying the reliability of individual assignments. We retain the underlying geometric criterion and extend its evaluation through bounded GNI. Specifically, multiple locally perturbed views are evaluated against the fixed reference geometry and aggregated into a class-wise conflict score. This score supports a selective partition into pseudo-aligned, pseudo-conflicting, and uncertain groups, while retained pseudo-conflicting samples are further assigned sample-wise reliability estimates for subsequent debiasing.

Figure 2 summarizes the two-stage training framework. Stage 1 performs GNI-based pseudo-group discovery and reliability estimation from the ERM representation. Stage 2 incorporates the resulting pseudo-group assignments and reliability scores into Reliability-Weighted Multi-Level Debiasing, where pseudo-group assignment determines the additional weighting of pseudo-conflicting samples and reliability modulates their contribution to the debiasing objective. An invariant prototype further regularizes the debiased representation using the selected pseudo groups.

At inference, D-CORE uses the trained bias experts to compare factual intermediate representations with their zero-feature references. The resulting layer-wise response differences are aggregated into a sample-specific residual-shortcut estimate and used to correct the fused classification margin.

3.2 GNI-BASED PSEUDO-GROUP DISCOVERY

3.2.1 GAUSSIAN NOISE INTERVENTION

To reduce the sensitivity of single-view geometric evidence to local representation variation, we extend the NSF statistic through multi-view assessment within a bounded neighborhood of each clean representation. The resulting responses are subsequently used for selective pseudo-group assignment and reliability estimation.

After training the ERM model, let $z_i = \phi_\rho(x_i)$ denote the representation of sample i at level ρ. For target class k, define $\mathcal{I}_k = \{j : y_j = k\}$ and $C_k^\rho := |\mathcal{I}_k|^{-1} \sum_{j \in \mathcal{I}_k} \phi_\rho(x_j)$. Distances are measured using the regularized covariance-aware metric $D_\rho(u, v) := (u - v)^\top A_\rho (u - v)$, where

$A_\rho = (\widehat{\Sigma}_\rho + 10^{-4} I)^\dagger$, $\widehat{\Sigma}_\rho$ is the empirical covariance of all frozen ERM representations at level ρ, and † denotes the Moore–Penrose pseudoinverse.

For each sample, the geometric comparison requires a reference from the remaining target classes. We consider centroids of all classes other than y_i. The class with the centroid closest to z_i is then selected. Specifically, its class index q_i is determined by $q_i \in \arg\min_{k \neq y_i} D_\rho(z_i, C_k^\rho)$.

Following NSF, the geometric evidence is quantified by comparing the distance to the target-class centroid with that to the nearest centroid of another class. Consequently, the relative-distance statistic is defined as $d_i := D_\rho(z_i, C_{y_i}^\rho) - D_\rho(z_i, C_{q_i}^\rho)$. A positive value indicates that the sample lies relatively further away from its target-class centroid and closer to another class centroid, providing strong evidence that the underlying representation deviates from the dominant target-class geometry. The proposed method preserves the above relative-distance criterion, while extending its evaluation across multiple locally perturbed views.

To assess the local stability of the relative-distance criterion, we construct multiple bounded perturbation views around each clean representation. The representation for view v is defined as

$$\widetilde{z}_i^{(v)} = \mathrm{Clip}_{z_i}^{\epsilon}\left[z_i + \sigma S_{y_i} \varepsilon_i^{(v)}\right], \qquad \varepsilon_i^{(v)} \sim \mathcal{N}(0, I), \tag{1}$$

where $\sigma > 0$ controls the perturbation magnitude and $\epsilon > 0$ is an empirically selected clipping radius. For $\mathcal{I}_{y_i} = \{j : y_j = y_i\}$, component-wise standard deviations are computed from the frozen ERM representations, resulting in a feature-level standard-deviation vector. This vector then defines a diagonal scaling matrix as follows:

$$S_{y_i} = \mathrm{diag}\left(\left[\frac{1}{|\mathcal{I}_{y_i}|} \sum_{j \in \mathcal{I}_{y_i}} \left(z_j - C_{y_i}^\rho\right)^2\right]^{1/2}\right). \tag{2}$$

S_{y_i} is computed from the frozen ERM representations with respect to target label y_i. The perturbations are bounded around the clean representation and used only for multi-view evaluation. Each perturbation view is evaluated against the same clean-sample reference geometry. For the v-th perturbation view, the relative-distance statistic and the average over all V views are defined as

$$d_i^{(v)} = D_\rho\left(\widetilde{z}_i^{(v)}, C_{y_i}^\rho\right) - D_\rho\left(\widetilde{z}_i^{(v)}, C_{q_i}^\rho\right), \qquad s_i = \frac{1}{V}\sum_{v=1}^{V} d_i^{(v)}. \tag{3}$$

The conflict score s_i summarizes the local response of the above relative-distance criterion across all perturbed views and is subsequently used for target-class-wise ranking and selective pseudo-group assignment. The conflict score reflects the relative position of a sample within the intra-class score distribution. However, it does not directly measure the consistency of the local responses. The intra-class reference $\tau_{C,y}$ denotes the threshold for conflict region and is empirically determined. We define the fraction of perturbation views supporting the conflict region as $v_i^C := V^{-1} \sum_v \mathbf{1}[d_i^{(v)} > \tau_{C,y_i}]$. For retained pseudo-conflicting samples, the consistency term is defined as $c_i := v_i^C$, such that c_i gives the fraction of perturbation views supporting the pseudo-conflicting assignment. Consequently, s_i controls class-wise selection, whereas c_i measures cross-view consistency of the conflict-oriented geometric evidence.

Appendices A.3 and A.4 analyze the multi-view estimator under an idealized un-clipped Gaussian model with squared Euclidean distance. Under these assumptions, zero-mean probing preserves the relative-centroid statistic in expectation, while averaging V independent views reduces the intervention-induced variance at rate $1/V$. These results characterize the stability of repeated local evaluation and do not imply an improvement over an oracle noise-free statistic.

3.2.2 Pseudo-Group Assignment and Reliability Estimation

The scale and dispersion of the conflict scores may vary across target classes. Pseudo-group assignment is consequently performed within each target class rather than by a global threshold. For class y, let $n_y := |\mathcal{I}y|$ denote the number of samples in class y,and let $s_{y,(1)} \geq s_{y,(2)} \geq \cdots \geq s_{y,(n_y)}$ denote the descending score order. Within a prescribed candidate interval $\mathcal{K}_y$, the conflict-tail boundary is determined by the largest adjacent score separation. Accordingly, the boundary index is

defined by

$$k_y^\star \in \arg\max_{k \in \mathcal{K}_y} \left[s_{y,(k)} - s_{y,(k+1)} \right]. \tag{4}$$

The upper-ranked subset determined by $k_y^\star$ defines $\widehat{\mathcal{G}}_{C,y}$, whereas a lower-ranked subset of size k_y^A, determined by a prescribed class-wise lower-rank proportion proportion, defines $\widehat{\mathcal{G}}_{A,y}$. Samples lying between these two selected subsets constitute the uncertain group $\widehat{\mathcal{G}}_{U,y}$. Accordingly, the sample-wise pseudo-group label is defined as follows:

$$\widehat{g}_i = \begin{cases} C, & i \in \widehat{\mathcal{G}}_{C,y_i}, \\ A, & i \in \widehat{\mathcal{G}}_{A,y_i}, \\ U, & i \in \widehat{\mathcal{G}}_{U,y_i}, \end{cases} \tag{5}$$

where C, A, and U denote pseudo-conflicting, pseudo-aligned, and uncertain samples, respectively. The uncertain group is excluded from pseudo-group-specific supervision but remains available for ordinary target prediction. After Eq. 5 determines pseudo-group assignment, reliability is estimated for each retained pseudo-conflicting sample. Let π_i denote its descending rank within a conflict tail of size n_{C,y_i}. Define $r_i := 1 - (\pi_i - 1)/(n_{C,y_i} - 1)$ for $n_{C,y_i} > 1$, and $r_i := 1$ for a singleton tail. The rank term therefore measures relative position within the selected conflict tail.

The sample-wise reliability score is defined as $\kappa_i := c_i^\gamma r_i^\eta$, where $\gamma, \eta \geq 0$ are power parameters controlling the relative contributions of the two factors. The quantity κ_i quantifies the reliability of the retained pseudo-conflicting evidence and is applied only to the debiasing objective, where it regularizes the contribution of these samples during debiasing.

3.3 Reliability-Weighted Multi-Level Debiasing

3.3.1 Multi-Level Bias and Debiasing Experts

Representations from different backbone layers encode predictive structure at different levels of abstraction, and the associated spurious information may therefore vary with network depth. A bias-related component that remains entangled with target-relevant information at one layer may become more distinguishable at another, such that a single-layer expert may capture only a partial view of the spurious structure. Conversely, extending auxiliary experts to additional backbone layers can introduce redundant cross-layer information and increase optimization complexity without necessarily providing complementary bias-related evidence. We therefore place bias and debiasing experts at two selected high-level backbone layers, $\mathcal{S} = \{3, 4\}$, to balance cross-layer complementarity with architectural efficiency. For each $l \in \mathcal{S}$, let $z_i^{(l)}$ denote the corresponding backbone representation. The bias branch operates on the detached feature $o_{b,i}^{(l)} = f_b^{(l)}(\mathrm{sg}[z_i^{(l)}])$, whereas the debiasing branch uses $o_{d,i}^{(l)} = f_d^{(l)}(z_i^{(l)})$, where $\mathrm{sg}[\cdot]$ denotes stop-gradient. This separation allows the bias branch to capture dominant predictive regularities without propagating gradients to the shared backbone, while the debiasing branch remains coupled to representation learning.

The bias experts remain auxiliary and are excluded from the predictive mixture. Let o_i^m denote the main-branch logits and G be the learned gating network. A softmax gate $\boldsymbol{\psi}_i = \mathrm{softmax}(G[o_i^m, \{o_{d,i}^{(l)}\}_{l \in \mathcal{S}}])$ produces nonnegative weights satisfying $\psi_i^m + \sum_{l \in \mathcal{S}} \psi_i^{(l)} = 1$. The fused logit vector reads $o_i = \psi_i^m o_i^m + \sum_{l \in \mathcal{S}} \psi_i^{(l)} o_{d,i}^{(l)}$. This separation preserves the bias branches as auxiliary estimators of dominant predictive regularities and for subsequent D-CORE correction, while preventing their bias-expert logits from contributing directly to the final prediction.

3.3.2 Role-Decoupled Debiasing Optimization

Pseudo-group assignment and assignment reliability play distinct roles during optimization. Samples assigned to the pseudo-conflicting group receive additional weight in the prediction objective, whereas the reliability score κ_i modulates their contribution to the debiasing objective. The bias-expert objective is computed independently of pseudo-group assignment and assignment reliability.

For the prediction branch, let $\ell_i := \mathrm{CE}(o_i, y_i)$. The pseudo-group assignment determines the sample weight w_i^m, where pseudo-conflicting samples receive $w_i^m = w_C > 1$ and all remaining samples

receive $w_i^m = 1$. The unweighted prediction loss is $\mathcal{L}_{\text{plain}} := N^{-1}\sum_i \ell_i$, while its weighted counterpart is $\mathcal{L}_{\text{weighted}} := (\sum_i w_i^m \ell_i)/(\sum_i w_i^m)$. The prediction objective is then defined as $\mathcal{L}_{\text{pred}} := \beta\mathcal{L}_{\text{plain}} + (1-\beta)\mathcal{L}_{\text{weighted}}$, where $\beta \in [0,1]$.

The bias experts are trained with generalized cross entropy (GCE) (Zhang & Sabuncu, 2018). For layer l, let $p_{b,i,y_i}^{(l)}$ denote the bias-expert probability assigned to the ground-truth class y_i; the per-sample loss is $\mathcal{L}_{\text{GCE}}(o_{b,i}^{(l)}, y_i) := [1 - (p_{b,i,y_i}^{(l)})^q]/q$, where $q > 0$ is the GCE shape parameter, and $\mathcal{L}_b^{(l)} := N^{-1}\sum_i \mathcal{L}_{\text{GCE}}(o_{b,i}^{(l)}, y_i)$. This unweighted objective allows the bias branch to capture dominant predictive regularities without pseudo-group supervision.

For each layer l, let $e_{b,i}^{(l)} := \text{CE}(o_{b,i}^{(l)}, y_i)$ and $e_{d,i}^{(l)} := \text{CE}(o_{d,i}^{(l)}, y_i)$ denote the bias- and debiasing-expert errors. We define $h_i^{(l)} := \text{sg}[e_{b,i}^{(l)}/(e_{b,i}^{(l)} + e_{d,i}^{(l)} + \delta_h)]$, where $\delta_h > 0$ is a small constant for numerical stability. $h_i^{(l)}$ determines each sample's contribution to the debiasing objective according to the relative prediction errors of the bias and debiasing experts.

For retained pseudo-conflicting samples, their contribution to the debiasing objective is further weighted by $w_i^d := 1 + (w_C - 1)\kappa_i$. Let $\mathcal{B}_A$, $\mathcal{B}_U$, and $\mathcal{B}_C$ denote the pseudo-aligned, uncertain, and pseudo-conflicting groups in the current batch, respectively. The layer-specific debiasing objective is then given by

$$\mathcal{L}_{\text{debias}}^{(l)} = \frac{\sum_{i\in\mathcal{B}_A} h_i^{(l)} e_{d,i}^{(l)} + \sum_{i\in\mathcal{B}_U} h_i^{(l)} e_{d,i}^{(l)} + \sum_{i\in\mathcal{B}_C} w_i^d h_i^{(l)} e_{d,i}^{(l)}}{|\mathcal{B}_A| + |\mathcal{B}_U| + \sum_{i\in\mathcal{B}_C} w_i^d}. \tag{6}$$

The debiasing and bias-expert objectives are averaged across the selected layers. Specifically, $\mathcal{L}_{\text{debias}} := |\mathcal{S}|^{-1}\sum_{l\in\mathcal{S}} \mathcal{L}_{\text{debias}}^{(l)}$ and $\mathcal{L}_b := |\mathcal{S}|^{-1}\sum_{l\in\mathcal{S}} \mathcal{L}_b^{(l)}$ denote the aggregated debiasing and bias-expert objectives, respectively.

3.3.3 Invariant Prototype Regularization and Overall Objective

A class centroid estimated from all samples may be dominated by the more prevalent pseudo-aligned group. To reduce the influence of group imbalance, the regularizer constructs separate pseudo-group centroids before forming the invariant prototype. For class y and layer l, let $\mathbf{p}_{y,A}^{(l)}$ and $\mathbf{p}_{y,C}^{(l)}$ denote the centroids of the corresponding pseudo-aligned and pseudo-conflicting debiasing representations. Their normalized combination, $\mathbf{p}_{y,\text{inv}}^{(l)} := \text{Norm}\Big[(1-\alpha_p)\,\text{Norm}\Big(\mathbf{p}_{y,A}^{(l)}\Big) + \alpha_p\,\text{Norm}\Big(\mathbf{p}_{y,C}^{(l)}\Big)\Big]$, defines the class-specific invariant prototype. Here, $\text{Norm}(\cdot)$ denotes L_2 normalization, and $\alpha_p \in [0,1]$ controls the relative contribution of the two pseudo groups.

To reduce minibatch variability, each pseudo-group class centroid is updated by the moving average $\mathbf{p}_{y,g}^{(l)} \leftarrow m_p\mathbf{p}_{y,g}^{(l)} + (1-m_p)\widehat{\mathbf{p}}_{y,g}^{(l)}$ for $g \in \{A, C\}$ and $0 \le m_p < 1$. Let $u_{d,i}^{(l)}$ denote the debiasing representation, the invariant-prototype alignment is defined over the selected pseudo-aligned and pseudo-conflicting samples as $\mathcal{L}_{\text{proto}}^{(l)} := (|\mathcal{B}_A| + |\mathcal{B}_C|)^{-1}\sum_{i\in\mathcal{B}_A\cup\mathcal{B}_C}\Big[1 - \cos\Big(u_{d,i}^{(l)}, \mathbf{p}_{y_i,\text{inv}}^{(l)}\Big)\Big]$. Specifically, $\widehat{\mathbf{p}}_{y,g}^{(l)}$ denotes the current-minibatch centroid for class y and pseudo group g. This objective regularizes the selected pseudo-aligned and pseudo-conflicting representations toward the corresponding class-specific invariant prototype, while uncertain samples are excluded from prototype-specific supervision.

Prototype alignment encourages the debiasing representation to approach the invariant prototype, but excessive deviation from the shared backbone representation may alter the original predictive structure. We therefore introduce a representation-preservation term to constrain this deviation. For each layer l, let $u_{d,i}^{(l)}$ denote the representation produced by the debiasing expert before its classification head. A fixed matching operator $R^{(l)}$ maps the shared backbone feature $z_i^{(l)}$ to the same representation space, yielding $u_{r,i}^{(l)} := R^{(l)}(z_i^{(l)})$. The corresponding preservation penalty is defined as $\mathcal{L}_{\text{pres}}^{(l)} := N^{-1}\sum_i \|u_{d,i}^{(l)} - \text{sg}[u_{r,i}^{(l)}]\|_2^2$. The stop-gradient operation keeps the reference representation fixed during this alignment, thereby constraining the debiasing representation without updating the shared reference branch through this term.

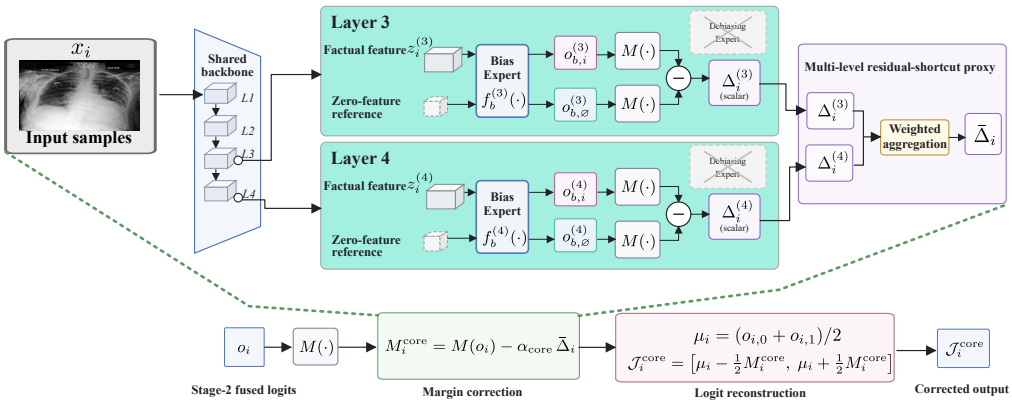


Figure 3: **D-CORE inference.** Residual-shortcut proxies are derived from the bias-expert responses under factual and zero-reference conditions, aggregated across the selected layers, and used to correct the Stage 2 decision margin. Here, $M(\cdot)$ maps the output logits to the corresponding binary classification margin.

The prototype-alignment and representation-preservation objectives are averaged over the selected layers as $\mathcal{L}_{\text{proto}} := |\mathcal{S}|^{-1} \sum_{l \in \mathcal{S}} \mathcal{L}_{\text{proto}}^{(l)}$ and $\mathcal{L}_{\text{pres}} := |\mathcal{S}|^{-1} \sum_{l \in \mathcal{S}} \mathcal{L}_{\text{pres}}^{(l)}$. To prevent the learned gate from collapsing onto a single predictive branch, we additionally use the gate entropy $\mathcal{H}_{\text{gate}} := -\frac{1}{N} \sum_{i=1}^{N} \sum_j \psi_{i,j} \log \psi_{i,j}$, where j denotes the predictive branch index. The complete Stage 2 training objective is then given by

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{debias}} + \mathcal{L}_b + \lambda_p \mathcal{L}_{\text{proto}} + \lambda_r \mathcal{L}_{\text{pres}} - \lambda_h \mathcal{H}_{\text{gate}}, \tag{7}$$

where λ_p, λ_r and λ_h control the contributions of prototype alignment, representation preservation, and gate-entropy regularization, respectively.

3.4 D-CORE: Inference-Time Residual Shortcut Correction

Residual shortcut influence may remain after debiasing training and affect the final prediction. To characterize this influence at inference, the trained bias experts are evaluated on both the factual intermediate representations and their corresponding zero-feature references. Figure 3 summarizes the resulting D-CORE inference pathway, in which bias-expert responses are used to estimate residual shortcut influence and correct the Stage 2 prediction margin.For binary classification, this correction is formulated in terms of the decision margin $M(o) := o_1 - o_0$. At layer l, the trained bias expert is evaluated on the factual feature, $o_{b,i}^{(l)} = f_b^{(l)}(z_i^{(l)})$, and on the corresponding zero-feature reference, $o_{b,\varnothing}^{(l)} = f_b^{(l)}(\mathbf{0})$. Their difference in decision margin, *i.e.*, $\Delta_i^{(l)} := M(o_{b,i}^{(l)}) - M(o_{b,\varnothing}^{(l)})$, defines the layer-specific residual-shortcut . The layer-wise proxies are then aggregated as $\bar{\Delta}_i := \sum_{l \in \mathcal{S}} \xi_{i,l} \Delta_i^{(l)}$, where $\xi_{i,l}$ denotes the empirically specified aggregation weight for layer l, with $\xi_{i,l} \geq 0$ and $\sum_{l \in \mathcal{S}} \xi_{i,l} = 1$.

The aggregated residual-shortcut estimate $\bar{\Delta}_i$ is used to adjust the fused decision margin. The corrected margin is defined as

$$M_i^{\text{core}} = M(o_i) - \alpha_{\text{core}} \bar{\Delta}_i, \qquad \alpha_{\text{core}} \geq 0, \tag{8}$$

where α_{core} controls the magnitude of the correction. The resulting margin M_i^{core} is subsequently used to reconstruct the corrected prediction logits.

The correction is applied to the decision margin, while the common logit offset is preserved during reconstruction. Let $\mu_i := (o_{i,0} + o_{i,1})/2$ denote the common offset. The corrected logits are then reconstructed as $\mathcal{J}_i^{\text{core}} = [o_{i,0}^{\text{core}}, o_{i,1}^{\text{core}}]$, where $o_{i,0}^{\text{core}} := \mu_i - \frac{1}{2} M_i^{\text{core}}$ and $o_{i,1}^{\text{core}} := \mu_i + \frac{1}{2} M_i^{\text{core}}$. We perform an expected invariant-margin variance analysis for D-CORE in Appendix B, where the theoretically optimal value of coefficient α_{core} can be proved to reduce the expected variance.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on four benchmarks spanning source, demographic, and background spurious correlations. MIMIC-CXR+NIH follows the source-bias construction of Li et al. (2023), with *No Finding* versus *Pneumonia* as the target task, dataset source as the spurious attribute, and training bias ratios of 0.85, 0.90, and 0.95. GbP follows Luo et al. (2022) and considers two gender-bias configurations for *No Finding* versus *Pneumothorax*. For CelebA (Liu et al., 2015), we use `Blond_Hair` as the target attribute and `Male` as the spurious attribute. For Waterbirds (Sagawa et al., 2020), bird category is the target and background type is the spurious attribute, with a training correlation of 0.95. Dataset construction and split details are provided in Appendix C.

Implementation and evaluation. We use XRV DenseNet-121 for GbP, ResNet-18 for MIMIC-CXR+NIH and CelebA, and ResNet-50 for Waterbirds. Model checkpoints, decision thresholds, and D-CORE parameters are selected using validation worst-class accuracy (WCA), defined over target classes and requiring no spurious-attribute information. Test performance is reported using mean accuracy and worst-group accuracy (WGA); true group annotations are used only for evaluation and post-hoc analysis. Additional implementation details are provided in Appendix C.

4.2 MAIN RESULTS

We compare our method with ERM, LfF (Nam et al., 2020), DFA (Lee et al., 2021), JTT (Liu et al., 2021), Ada-ABC (Luo et al., 2025), Debiasify (Bayasi et al., 2025), GERNE (Asaad et al., 2025), and NSF (Li et al., 2025). All debiasing methods are evaluated without training-set bias annotations, and GERNE is evaluated under its unknown-attribute setting.

Quantitative results. As shown in Table 1, ERM is consistently vulnerable to spurious correlations, with particularly severe degradation in worst-group performance as the training bias becomes stronger. This behavior is most evident on MIMIC-CXR+NIH, where ERM WGA decreases from 29.76% at a bias ratio of 0.85 to 5.76% at 0.95. WGA evaluates performance on the worst-performing target–attribute group and thus reflects worst-group robustness against group-dependent failures under spurious correlations. Existing debiasing methods substantially alleviate this degradation, although their improvements in worst-group robustness are not always accompanied by comparable aggregate performance. In contrast, our method achieves the highest WGA in all seven evaluated settings while exceeding the strongest competing method by an average of 2.37 percentage points. The advantage remains stable across different bias types and correlation strengths. On MIMIC-CXR+NIH, our WGA remains between 65.28% and 67.36% as the bias ratio increases from 0.85 to 0.95, indicating substantially reduced sensitivity to increasingly dominant source correlations. Consistent improvements are also observed under both gender-bias configurations of GbP and on CelebA, while on Waterbirds our method attains the strongest WGA together with the second-highest mean accuracy. Across all settings, it achieves the best mean accuracy in four cases and the second-best result in two additional cases. These results indicate that the proposed framework improves worst-group robustness across heterogeneous spurious correlations while generally preserving competitive aggregate predictive performance.

4.3 ABLATION STUDY

Component ablation. As shown in Table 2(a), WGA increases from 41.20% with selective GNI to 58.72% after incorporating reliability-aware training, and further reaches 66.24% with D-CORE. The progressive improvement indicates that selective pseudo-group discovery, reliability-aware optimization, and inference-time residual-shortcut correction provide complementary contributions to worst-group robustness.

Expert placement. Table 2(b) shows that placing the bias and debiasing experts at Layers $3{+}4$ yields the highest WGA of 66.24%. The lower performance obtained with earlier layers or all backbone stages suggests that the selected high-level representations provide more useful complementary bias-related information while avoiding unnecessary cross-layer redundancy. Additional fine-grained ablations on pseudo-group discovery, reliability-aware optimization, prototype regularization, and post-hoc pseudo-group quality are provided in Appendix D.

Table 1: Mean accuracy (Mean) and worst-group accuracy (WGA) of debiasing methods without training-set bias annotations. Results are reported as mean ± standard deviation over five random seeds; higher is better. Best and second-best values are shown in **bold** and underlined, respectively.

Dataset	Metric	ERM	LfF	DFA	JTT	Ada-ABC	Debiasify	GERNE	NSF	Ours
MIMIC-CXR+NIH (bias ratio: 0.85)	WGA ↑	$29.76_{\pm 3.22}$	$40.00_{\pm 5.51}$	$52.16_{\pm 3.17}$	$55.04_{\pm 5.70}$	$56.32_{\pm 5.08}$	$39.04_{\pm 1.28}$	$61.87_{\pm 3.02}$	$48.96_{\pm 0.32}$	$\mathbf{67.36_{\pm 1.91}}$
	Mean ↑	$61.08_{\pm 0.73}$	$53.76_{\pm 3.18}$	$61.60_{\pm 2.18}$	$65.72_{\pm 2.48}$	$69.24_{\pm 1.16}$	$64.00_{\pm 1.46}$	$68.47_{\pm 0.81}$	$62.16_{\pm 0.23}$	$\mathbf{69.64_{\pm 1.13}}$
MIMIC-CXR+NIH (bias ratio: 0.90)	WGA ↑	$20.16_{\pm 1.63}$	$24.48_{\pm 4.02}$	$41.60_{\pm 3.10}$	$51.20_{\pm 3.39}$	$65.76_{\pm 1.71}$	$26.56_{\pm 2.39}$	$61.48_{\pm 2.98}$	$47.84_{\pm 0.32}$	$\mathbf{66.24_{\pm 0.93}}$
	Mean ↑	$57.36_{\pm 0.83}$	$52.04_{\pm 3.73}$	$57.52_{\pm 2.27}$	$64.47_{\pm 0.38}$	$\mathbf{72.28_{\pm 0.91}}$	$60.16_{\pm 0.74}$	$59.96_{\pm 3.33}$	$65.44_{\pm 0.32}$	$70.56_{\pm 0.81}$
MIMIC-CXR+NIH (bias ratio: 0.95)	WGA ↑	$5.76_{\pm 1.71}$	$18.56_{\pm 5.10}$	$47.84_{\pm 3.81}$	$46.93_{\pm 1.64}$	$64.16_{\pm 2.15}$	$10.72_{\pm 2.06}$	$60.64_{\pm 2.91}$	$47.04_{\pm 0.78}$	$\mathbf{65.28_{\pm 3.47}}$
	Mean ↑	$51.68_{\pm 0.94}$	$54.44_{\pm 2.74}$	$60.56_{\pm 1.44}$	$64.13_{\pm 1.75}$	$70.00_{\pm 1.71}$	$53.88_{\pm 0.56}$	$66.84_{\pm 0.99}$	$63.12_{\pm 0.2}$	$\mathbf{71.20_{\pm 0.96}}$
GbP (Case 1)	WGA ↑	$42.40_{\pm 3.16}$	$55.04_{\pm 4.56}$	$51.28_{\pm 2.90}$	$41.28_{\pm 3.33}$	$66.16_{\pm 5.52}$	$46.72_{\pm 1.09}$	$48.16_{\pm 1.43}$	$53.12_{\pm 0.49}$	$\mathbf{69.36_{\pm 0.74}}$
	Mean ↑	$61.94_{\pm 1.86}$	$64.24_{\pm 3.50}$	$69.66_{\pm 1.90}$	$63.04_{\pm 4.74}$	$76.38_{\pm 1.39}$	$69.34_{\pm 0.78}$	$66.18_{\pm 0.82}$	$71.90_{\pm 0.14}$	$\mathbf{77.36_{\pm 0.42}}$
GbP (Case 2)	WGA ↑	$39.36_{\pm 4.54}$	$50.32_{\pm 7.69}$	$48.08_{\pm 4.68}$	$38.40_{\pm 3.41}$	$61.60_{\pm 3.14}$	$39.20_{\pm 4.79}$	$48.48_{\pm 2.18}$	$43.60_{\pm 2.23}$	$\mathbf{64.56_{\pm 1.73}}$
	Mean ↑	$61.82_{\pm 2.04}$	$63.18_{\pm 2.08}$	$59.50_{\pm 3.05}$	$51.20_{\pm 5.19}$	$74.18_{\pm 0.97}$	$68.32_{\pm 1.31}$	$65.70_{\pm 2.21}$	$69.82_{\pm 0.56}$	$\mathbf{74.74_{\pm 0.79}}$
CelebA	WGA ↑	$52.11_{\pm 2.54}$	$66.00_{\pm 3.05}$	$49.63_{\pm 3.26}$	$68.33_{\pm 4.32}$	$53.22_{\pm 2.97}$	$54.67_{\pm 5.89}$	$69.33_{\pm 1.44}$	$58.06_{\pm 1.18}$	$\mathbf{72.11_{\pm 0.46}}$
	Mean ↑	$\mathbf{95.80_{\pm 0.09}}$	$93.99_{\pm 0.57}$	$94.81_{\pm 0.37}$	$92.48_{\pm 0.96}$	$95.46_{\pm 0.22}$	$94.52_{\pm 0.78}$	$94.07_{\pm 0.12}$	$95.25_{\pm 0.04}$	$92.96_{\pm 0.16}$
Waterbirds	WGA ↑	$21.99_{\pm 5.27}$	$75.49_{\pm 4.39}$	$77.62_{\pm 4.18}$	$87.64_{\pm 1.11}$	$74.34_{\pm 1.88}$	$72.45_{\pm 1.15}$	$71.26_{\pm 2.30}$	$84.70_{\pm 0.27}$	$\mathbf{88.19_{\pm 0.83}}$
	Mean ↑	$67.88_{\pm 4.48}$	$87.44_{\pm 1.96}$	$88.88_{\pm 2.58}$	$90.13_{\pm 1.14}$	$86.92_{\pm 1.17}$	$87.04_{\pm 2.36}$	$85.29_{\pm 1.46}$	$\mathbf{95.80_{\pm 0.03}}$	$92.56_{\pm 0.23}$

Table 2: **Ablation study on MIMIC-CXR+NIH at a bias ratio of** 0.90**.** Panel (a) evaluates the contributions of selective GNI, reliability-aware training, and D-CORE, while Panel (b) examines the placement of the bias and debiasing experts across backbone stages. Results are reported as mean ± standard deviation over five random seeds; higher is better. The best result in each panel is shown in **bold**.

(a) Component ablation

Selective GNI	Reliability Training	D-CORE	WGA ↑
✓			$41.20_{\pm 2.00}$
✓	✓		$58.72_{\pm 3.87}$
✓	✓	✓	$\mathbf{66.24_{\pm 0.93}}$

(b) Expert placement

Expert Placement	WGA ↑
Layers 1+2	$50.80_{\pm 2.80}$
Layers 3+4	$\mathbf{66.24_{\pm 0.93}}$
All Layers	$54.40_{\pm 2.40}$

4.4 Qualitative Visualization of Model Attention

Figure 4 compares Grad-CAM (Selvaraju et al., 2017) attention patterns of the proposed method and ERM across source, gender, and background spurious correlations. Across all four benchmarks, our method concentrates on target-relevant regions, whereas ERM concentrates on regions associated with the respective spurious attributes. These results indicate that the proposed method reduces dependence on spurious features.

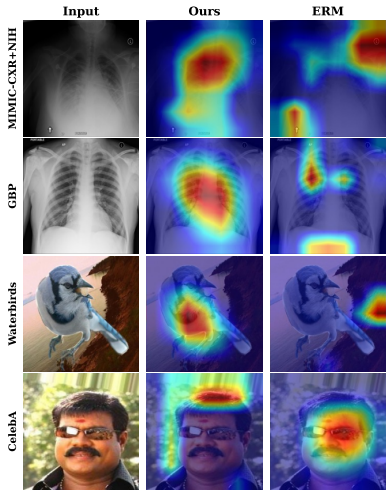


Figure 4: **Grad-CAM attention comparison between our proposed method and ERM.** Columns show the input, our method, and ERM, respectively; rows denote four benchmarks.

5 Conclusion

We developed a debiasing framework for settings without spurious-attribute annotations. It combines GNI-based pseudo-group discovery, reliability-aware multi-level debiasing, and D-CORE correction. Experiments show improved worst-group robustness without group annotations for training or model selection. Numerical analysis supports the proposed design. However, pseudo-group discovery relies on group-dependent geometry in ERM representations and may weaken when such separation is limited. Moreover, D-CORE currently targets binary decision margins; multiclass extension remains future work.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing of this manuscript, including language editing and improvements to grammar, clarity, and readability. They were not used to develop the research methodology, design experiments, formulate scientific claims, analyze or interpret experimental results, or draw scientific conclusions. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Ihab Asaad, Maha Shadaydeh, and Joachim Denzler. Gradient extrapolation for debiased representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3819–3829, 2025.

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. MaskTune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23284–23296, 2022.

Nourhan Bayasi, Jamil Fayyad, Ghassan Hamarneh, Rafeef Garbi, and Homayoun Najjaran. Debiasify: Self-distillation for unsupervised bias mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3227–3236, 2025.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 2021. URL `https://proceedings.mlr.press/v139/creager21a.html`.

Chaoran Cui, Hebo Ma, Xiaolin Dong, Chen Zhang, Chunyun Zhang, Yumo Yao, Meng Chen, and Yuling Ma. Model-agnostic counterfactual reasoning for identifying and mitigating answer bias in knowledge tracing. *Neural Networks*, 178:106495, 2024. doi: 10.1016/j.neunet.2024.106495.

Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. SelecMix: Debiased learning by contradicting-pair sampling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14345–14357, 2022.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0.

Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18403–18415, 2022.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25123–25133, 2021.

Jiaxuan Li, Duc Minh Vo, and Hideki Nakayama. Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4924–4934, 2023.

Weiwei Li, Junzhuo Liu, Yuanyuan Ren, Yuchen Zheng, Yahao Liu, and Wen Li. Let samples speak: Mitigating spurious correlation by exploiting the clusterness of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15486–15496, 2025.

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*, volume 139 of *Proceedings of Machine Learning Research*, 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.

Luyang Luo, Dunyuan Xu, Hao Chen, Tien-Tsin Wong, and Pheng-Ann Heng. Pseudo bias-balanced learning for debiased chest x-ray classification. In *Medical Image Computing and Computer Assisted Intervention*, pp. 621–631. Springer, 2022. doi: 10.1007/978-3-031-16452-1_59.

Luyang Luo, Xin Huang, Minghao Wang, Zhuoyue Wan, Wanteng Ma, and Hao Chen. Mitigating medical dataset bias by learning adaptive agreement from a biased council. *Medical Image Analysis*, 105:103629, 2025. doi: 10.1016/j.media.2025.103629.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 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. In *International Conference on Learning Representations*, 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e0688d13958a19e087e123148555e4b4-Abstract.html.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

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. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22z.html.

A Analytical Properties of Selective Multi-View GNI

This appendix provides an idealized analysis of the geometric and multi-view properties underlying Section 3.2. The practical method uses a regularized covariance-aware distance and bounded Gaussian perturbations. For analytical tractability, we consider a binary representation model, squared Euclidean distance, and the unclipped zero-mean Gaussian component. The analysis characterizes population ordering and the statistical stability of repeated local evaluation; it does not establish exact pseudo-group recovery for arbitrary learned representations.

A.1 Idealized Core–Bias Representation Model

Let $Y, A \in \{-1, +1\}$ denote the target and spurious variables, respectively, and consider the representation model $Z = Y\mu + A\nu$, where $\|\mu\|_2 = 1$, $\|\nu\|_2 = B$, and $\mu^\top \nu = 0$. For each target value $y \in \{-1, +1\}$, assume $\Pr(A = y \mid Y = y) = \theta$ with $\theta \in [1/2, 1]$, and define $\chi := 2\theta - 1$. It follows that $\mathbb{E}[A \mid Y = y] = \chi y$, and hence the class-conditional population centroid is $C_y = \mathbb{E}[Z \mid Y = y] = y(\mu + \chi\nu)$.

Following the sign convention of the main text, we define the population relative-distance statistic as $d(Z, Y) := \|Z - C_Y\|_2^2 - \|Z - C_{-Y}\|_2^2$.

A.2 Population Separation of Relative-Distance Evidence

Theorem A.1. *Under the model above, the population statistics for bias-aligned and bias-conflicting samples satisfy*

$$d_{\text{align}} = -4(1+\chi B^2), \qquad d_{\text{conflict}} = -4(1-\chi B^2), \tag{9}$$

and therefore

$$\Delta := d_{\text{conflict}} - d_{\text{align}} = 8\chi B^2. \tag{10}$$

Consequently, $d_{\text{conflict}} > d_{\text{align}}$ *whenever* $\chi B^2 > 0$.

Proof. For a bias-aligned sample, $A = Y$ and $Z_{\text{align}} = Y(\mu+\nu)$. Using $C_Y = Y(\mu+\chi\nu)$ and $C_{-Y} = -Y(\mu+\chi\nu)$ gives $Z_{\text{align}} - C_Y = Y(1-\chi)\nu$ and $Z_{\text{align}} - C_{-Y} = Y[2\mu + (1+\chi)\nu]$. The orthogonality of μ and ν therefore yields

$$d_{\text{align}} = (1-\chi)^2 B^2 - \left[4 + (1+\chi)^2 B^2\right] = -4(1+\chi B^2).$$

For a bias-conflicting sample, $A = -Y$ and $Z_{\text{conflict}} = Y(\mu-\nu)$. In this case, $Z_{\text{conflict}} - C_Y = -Y(1+\chi)\nu$ and $Z_{\text{conflict}} - C_{-Y} = Y[2\mu - (1-\chi)\nu]$, which gives

$$d_{\text{conflict}} = (1+\chi)^2 B^2 - \left[4 + (1-\chi)^2 B^2\right] = -4(1-\chi B^2).$$

Subtracting the two expressions gives $\Delta = 8\chi B^2$. □

The theorem establishes only a population ordering under the stated model. When $\chi = 0$, the two statistics coincide and the assumed representation contains no group-dependent geometric separation.

A.3 Expectation Preservation and Multi-View Variance Reduction

We next consider the unclipped Gaussian component of GNI. Let $\widetilde{Z}^{(v)} = Z + E^{(v)}$, where the perturbations are independent across views and satisfy $E^{(v)} \sim \mathcal{N}(0, \sigma^2 S_Y^2)$ for a class-conditional diagonal scaling matrix S_Y. Let $\widetilde{d}^{(v)}$ denote the relative-distance statistic evaluated at $\widetilde{Z}^{(v)}$, and define $\overline{d}_V := V^{-1}\sum_{v=1}^{V} \widetilde{d}^{(v)}$.

Proposition A.2. *Under squared Euclidean distance,*

$$\begin{aligned} \widetilde{d}^{(v)} &= d(Z,Y) + 2(C_{-Y} - C_Y)^\top E^{(v)}, \\ \mathbb{E}[\widetilde{d}^{(v)} \mid Z, Y] &= d(Z,Y), \\ \mathrm{Var}[\overline{d}_V \mid Z, Y] &= \frac{\sigma_d^2}{V}, \end{aligned} \tag{11}$$

where $\sigma_d^2 := 4\sigma^2 (C_Y - C_{-Y})^\top S_Y^2 (C_Y - C_{-Y})$.

Proof. Expanding the two squared distances in $\widetilde{d}^{(v)}$ causes the quadratic perturbation terms $\|E^{(v)}\|_2^2$ to cancel, leaving Equation 11. Zero-mean perturbations therefore preserve the clean statistic in conditional expectation. The linear perturbation term has conditional variance σ_d^2, and independence across the V views reduces the variance of their average by a factor of V. □

Thus, under the idealized unclipped model, multi-view averaging does not shift the expected relative-distance statistic; it reduces only the perturbation-induced estimation variance.

A.4 Ranking Stability of Class-Wise GNI

The practical pseudo-group procedure uses within-class ranking rather than a global score threshold. Consider a pseudo-conflicting sample and a pseudo-aligned sample with V-view scores s_C and s_A, respectively. Under the Gaussian model above, assume their perturbation terms are independent and have per-view variances σ_C^2 and σ_A^2. By Theorem A.1 and the expectation-preservation result, $s_C - s_A$ is Gaussian with mean $\Delta > 0$ and variance $(\sigma_C^2 + \sigma_A^2)/V$.

A standard Gaussian tail bound therefore gives

$$\Pr(s_C \le s_A) \le \exp\left[-\frac{V\Delta^2}{2(\sigma_C^2+\sigma_A^2)}\right] = \exp\left[-\frac{32V\chi^2B^4}{\sigma_C^2+\sigma_A^2}\right]. \tag{12}$$

Equation 12 shows that, under the stated assumptions, the probability of reversing the population ordering decreases exponentially with the number of independent views when $\Delta > 0$. This result concerns pairwise ranking stability rather than exact recovery of the complete adaptive pseudo-group partition. When $\chi = 0$, $\Delta = 0$ and the bound does not provide nontrivial separation, consistent with the absence of group-dependent geometry in the assumed representation model.

B Expected Invariant-Margin Variance Analysis

We analyze D-CORE through an additive residual-bias model. Let $M_i^\star$ denote the invariant decision margin and suppose that

$$M(o_i) = M_i^\star + b_i + \varepsilon_i, \qquad \bar{\Delta}_i = b_i + \zeta_i, \tag{13}$$

where b_i denotes residual shortcut influence, ε_i the remaining prediction error, and ζ_i the error of the D-CORE proxy. Assume finite second moments and zero pairwise cross-moments, $\mathbb{E}[b_i\varepsilon_i] = \mathbb{E}[b_i\zeta_i] = \mathbb{E}[\varepsilon_i\zeta_i] = 0$. Define $\mathbb{E}[b_i^2] = \sigma_b^2$, $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2$, and $\mathbb{E}[\zeta_i^2] = \sigma_\zeta^2$.

Theorem B.1 (Expected invariant-margin variance)**.** *For the corrected margin $M_i(\alpha) := M(o_i) - \alpha\bar{\Delta}_i$, the expected variance from the invariant margin is*

$$\mathcal{Q}(\alpha) := \mathbb{E}\Big[(M_i(\alpha) - M_i^\star)^2\Big] = (1-\alpha)^2\sigma_b^2 + \sigma_\varepsilon^2 + \alpha^2\sigma_\zeta^2. \tag{14}$$

If $\sigma_b^2 > 0$, the theoretically optimal coefficient $\alpha^\star$ and its corresponding improvement over the uncorrected margin are, respectively, given by

$$\alpha^\star = \frac{\sigma_b^2}{\sigma_b^2+\sigma_\zeta^2}, \tag{15}$$

$$\mathcal{Q}(0) - \mathcal{Q}(\alpha^\star) = \frac{\sigma_b^4}{\sigma_b^2+\sigma_\zeta^2} > 0. \tag{16}$$

More generally, $\mathcal{Q}(\alpha) < \mathcal{Q}(0)$ whenever $0 < \alpha < 2\sigma_b^2/(\sigma_b^2+\sigma_\zeta^2)$.

Proof. Substituting equation 13 into the corrected margin gives $M_i(\alpha) - M_i^\star = (1-\alpha)b_i + \varepsilon_i - \alpha\zeta_i$. The assumed zero cross-moments eliminate all cross terms after squaring and taking expectation, yielding equation 14.

Differentiation gives $\mathcal{Q}'(\alpha) = -2(1-\alpha)\sigma_b^2 + 2\alpha\sigma_\zeta^2$. Setting this derivative to zero yields equation 15, while $\mathcal{Q}''(\alpha) = 2(\sigma_b^2+\sigma_\zeta^2) > 0$ establishes the unique minimum. Substitution of $\alpha^\star$ into equation 14 gives $\mathcal{Q}(\alpha^\star) = \sigma_\varepsilon^2 + \sigma_b^2\sigma_\zeta^2/(\sigma_b^2+\sigma_\zeta^2)$, from which Equation 16 follows. The stated interval for $\mathcal{Q}(\alpha) < \mathcal{Q}(0)$ follows directly from the quadratic difference $\mathcal{Q}(\alpha) - \mathcal{Q}(0)$. □

If the proxy is exact, $\sigma_\zeta^2 = 0$ and $\alpha^\star = 1$, and thus the residual-bias term is removed under the assumed model. However, the coefficient $\alpha^\star$ is a theoretically optimal quantity, since it depends on unobservable residual and proxy errors. In practice, α_{core} is selected by validation without estimating these quantities. The theorem therefore characterizes the correction mechanism under the stated assumptions; it does not imply that validation recovers the theoretically optimal coefficient or guaranties an improvement in the test WGA.

C Additional Experimental Details

C.1 Datasets and Evaluation Protocol

MIMIC-CXR+NIH. Following the source-bias construction of Li et al. (2023), we combine MIMIC-CXR (Johnson et al., 2019) and NIH ChestX-ray8 (Wang et al., 2017) for binary classification between *No Finding* and *Pneumonia*. Dataset source is treated as the spurious attribute. We

consider training bias ratios of 0.85, 0.90, and 0.95, while the validation and test sets are unbiased with respect to dataset source.

Gender-biased Pneumothorax (GbP). Following Luo et al. (2022), the task is constructed from NIH ChestX-ray images to distinguish *No Finding* from *Pneumothorax*, with patient gender serving as the spurious attribute. Case 1 associates *Pneumothorax* predominantly with male patients and *No Finding* with female patients, whereas Case 2 reverses this association. The validation and test sets are unbiased with respect to gender.

CelebA. We use the official CelebA splits (Liu et al., 2015), with `Blond_Hair` as the target attribute and `Male` as the spurious attribute.

Waterbirds. Following Sagawa et al. (2020), the task distinguishes landbirds from waterbirds, while background type constitutes the spurious attribute. The training set exhibits a 0.95 correlation between bird category and background type, and the official dataset splits are retained.

Throughout training and model selection, spurious-attribute and target–attribute group annotations are unavailable to the proposed method. They are used only for final group-wise evaluation and post-hoc diagnostic analysis.

C.2 Implementation and Model Selection

We use XRV DenseNet-121 for both GbP configurations, ResNet-18 for MIMIC-CXR+NIH and CelebA, and ResNet-50 for Waterbirds. All input images are resized to 224×224. Experiments are implemented in Python 3.10 using PyTorch 2.1.2 and CUDA 11.8 and are conducted on an NVIDIA GeForce RTX 3090 Ti GPU.

Model checkpoint selection, decision-threshold selection, and D-CORE parameter selection are performed exclusively on the validation set using worst-class accuracy (WCA). For target classes $y \in \{1, \ldots, C\}$, WCA is defined as the minimum classification accuracy across target classes and therefore requires only target labels rather than spurious-attribute or group annotations. This criterion ensures that true bias information is not introduced during either training or model selection.

For final evaluation, we report mean classification accuracy and worst-group accuracy (WGA). WGA is defined as the minimum accuracy across target–attribute groups. True spurious-attribute annotations are accessed only after training and model selection, for final group-wise evaluation and post-hoc diagnostic analysis. They do not participate in pseudo-group construction, optimization, checkpoint selection, decision-threshold selection, or D-CORE parameter selection.

D Additional Ablation Studies

We provide a fine-grained analysis of the principal training-stage design choices on MIMIC-CXR+NIH at a bias ratio of 0.90. Specifically, we examine the effects of multi-view evaluation, selective pseudo-group assignment, sample-wise reliability estimation, and invariant-prototype regularization. D-CORE is applied to all configurations in Table 3 and is therefore held fixed throughout these comparisons. For configurations using uniform reliability, κ_i is fixed to 1, removing sample-wise reliability differentiation while preserving the corresponding pseudo-group construction procedure.

Pseudo-group construction. The first three configurations isolate the effect of pseudo-group construction under uniform reliability and fixed prototype regularization. Replacing the single-view binary rule with multi-view evaluation increases WGA from 40.80% to 48.80%, indicating that repeated evaluation of the geometric criterion provides a more effective basis for pseudo-group inference than a single representation. Introducing selective A/C/U assignment further raises WGA to 55.80%. Since this comparison retains the same multi-view evaluation and uniform reliability, the additional improvement supports explicitly separating geometrically ambiguous samples rather than forcing all samples into a binary pseudo-group partition.

Reliability formulation. We next retain selective multi-view assignment and prototype regularization while varying the reliability formulation. Using either cross-view consistency c_i or within-class rank r_i alone yields WGA values of 42.88% and 44.48%, respectively, both below the uniform-weighting configuration at 55.80%. In contrast, the joint formulation $\kappa_i = c_i^{\gamma} r_i^{\eta}$ achieves 66.24% in

Table 3: **Fine-grained ablation of pseudo-group discovery and reliability-aware optimization on MIMIC-CXR+NIH at a bias ratio of 0.90.** A checkmark indicates that the corresponding component is enabled. "Uniform" denotes $\kappa_i = 1$. D-CORE is applied to all configurations. Results are reported as mean $\pm$ standard deviation over five random seeds; higher is better.

Configuration	Multi-view	Selective A/C/U	Reliability weighting	Prototype	WGA ↑
Single-view Binary			Uniform	✓	$40.80_{\pm 2.85}$
Multi-view Binary	✓		Uniform	✓	$48.80_{\pm 1.73}$
Selective Multi-view	✓	✓	Uniform	✓	$55.80_{\pm 3.17}$
$\kappa_i = c_i$	✓	✓	c_i	✓	$42.88_{\pm 1.65}$
$\kappa_i = r_i$	✓	✓	r_i	✓	$44.48_{\pm 2.80}$
Without Prototype	✓	✓	$c_i^\gamma r_i^\eta$		$45.60_{\pm 1.13}$
Full Training	✓	✓	$c_i^\gamma r_i^\eta$	✓	$\mathbf{66.24_{\pm 0.93}}$

Table 4: **Post-hoc evaluation of pseudo-group discovery quality over five random seeds.** Results are reported as mean $\pm$ standard deviation. Higher values are better; the best results for directly comparable metrics are shown in **bold**.

Method	Precision ↑	F1 ↑	Error Rejection Rate ↑	Correction Rate ↑
NSF	$42.42_{\pm 8.48}$	$56.69_{\pm 8.03}$	–	–
Ours	$\mathbf{89.94_{\pm 2.02}}$	$\mathbf{91.18_{\pm 3.15}}$	$44.98_{\pm 1.50}$	$53.46_{\pm 1.30}$

the complete training configuration. This non-monotonic behavior indicates that the improvement does not arise from sample-wise weighting alone. Rather, the result suggests that cross-view consistency and within-class position provide complementary information when jointly used to determine the reliability of retained pseudo-conflicting samples.

Prototype regularization. Finally, we isolate the contribution of invariant-prototype regularization under selective multi-view assignment and the joint reliability formulation. Removing the prototype term reduces WGA from 66.24% to 45.60%. The substantial degradation indicates that reliability-aware weighting alone is insufficient to reproduce the performance of the complete training configuration. Prototype regularization therefore provides an additional representation-level constraint that complements selective pseudo-group assignment and sample-wise reliability estimation.

Overall, the fine-grained ablations show that the individual design choices do not produce uniformly monotonic improvements when considered in isolation. Multi-view evaluation is most effective when coupled with selective pseudo-group assignment, while the reliability terms are most effective when combined in the joint formulation. The strongest performance is obtained when these components are further integrated with invariant-prototype regularization, supporting the use of the complete training configuration rather than any individual mechanism alone.

D.1 Post-hoc Pseudo-Group Quality Analysis

We further evaluate the quality of the discovered pseudo groups by comparing NSF with the proposed selective GNI procedure over five random seeds. This analysis is conducted strictly post hoc: true target–attribute group annotations are used only to assess pseudo-group quality and are not involved in pseudo-group construction, training, model selection, or inference. Precision and F1 evaluate the identification quality of pseudo-conflicting samples. The Error Rejection Rate is computed as the proportion of samples misassigned by NSF that are redirected to the uncertain group by the proposed procedure, whereas the Correction Rate is computed as the proportion of NSF misassignments that the proposed method reassigns to their correct pseudo groups.

As shown in Table 4, the proposed method substantially improves pseudo-conflicting sample identification. Precision increases from 42.42% for NSF to 89.94%, corresponding to an absolute improvement of 47.52 percentage points. F1 similarly increases from 56.69% to 91.18%, corresponding to an absolute improvement of 34.49 percentage points. These results indicate that the proposed

multi-view selective procedure yields substantially more accurate pseudo-conflicting assignments than the deterministic binary assignment used by NSF.

Beyond improving the retained pseudo-group assignments, the proposed procedure also explicitly handles errors produced by the NSF binary rule. An Error Rejection Rate of 44.98% indicates that nearly 45% of the samples misassigned by NSF are redirected to the uncertain group rather than retained as pseudo-group-specific supervision. In addition, a Correction Rate of 53.46% shows that more than half of the NSF misassignments are reassigned to their correct pseudo groups. Together, these two outcomes demonstrate that the proposed selective procedure does not rely solely on rejecting ambiguous samples, but also corrects a substantial proportion of erroneous pseudo-group assignments when the multi-view geometric evidence is sufficiently discriminative.

Taken together, the substantial improvements in precision and F1, together with the rejection and correction of erroneous NSF assignments, provide direct post-hoc evidence that the proposed selective multi-view procedure improves the quality of pseudo-group discovery while reducing the propagation of unreliable pseudo-group supervision into subsequent debiasing.