TRACING GROUP FAIRNESS THROUGH TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

When a trained model performs much worse on some demographic groups than others, a natural question is: which training examples contribute to the disparity? Fairness-aware learning typically aims to reduce such gaps without attributing them to individual training examples, while standard data attribution methods focus on predictive performance rather than fairness. We propose FAIR-DS, which assigns each example a score for its accumulated contribution to a chosen group-aware objective in a single training run. Building on In-Run Data Shapley, the method uses group validation statistics to identify harmful examples under popular fairness criteria. Our analysis bounds the attribution error and connects the computed scores to changes in the selected fairness criteria. These guarantees clarify how individual examples contribute to fairness changes along the training trajectory. Empirical studies on label-bias detection and attribution-guided data repair demonstrate competitive performance on real-world tasks.

1 INTRODUCTION

Machine learning models can achieve high average accuracy while performing poorly on some demographic groups. For example, a health risk algorithm may underestimate the needs of Black patients (Obermeyer et al., 2019), while a face-analysis system may have higher error rates for darker-skinned women than for lighter-skinned men (Buolamwini & Gebru, 2018). Existing fairness methods aim to reduce such disparities through pre-processing (Louizos et al., 2015; Zemel et al., 2013; Calmon et al., 2017; Lum & Johndrow, 2016; Zhao et al., 2020; Creager et al., 2019; Tan et al., 2020), in-processing (Roh et al., 2020; Agarwal et al., 2018; Baharlouei et al., 2020; Lowy et al., 2022; Baharlouei et al., 2024; Donini et al., 2018; Prost et al., 2019), or post-processing (Hardt et al., 2016; Jang et al., 2021a). These methods improve fairness by changing the training data, the learning objective, or the model's predictions. However, reducing a disparity does not explain how individual training examples contributed to it. This raises a question: *which training examples contribute to group disparities?*

Answering this question can help us identify training examples that harm fairness and guide data inspection, removal, or reweighting. This is especially useful when training labels contain bias. Labels may reflect past decisions rather than unbiased outcomes, and learning from them can reproduce group-dependent errors (Wick et al., 2019; Zhang et al., 2024; Wang et al., 2021). However, a biased label does not necessarily harm every fairness criterion, and an example that improves average accuracy may still worsen group disparities. We therefore need fairness attribution to assess how individual examples affect a chosen fairness criterion, so that data interventions can target the outcome we want to improve.

Existing fairness-oriented data valuation methods use Shapley values or influence functions to assess how training examples affect fairness (Arnaiz-Rodríguez & Oliver, 2024; Wang et al., 2024; Yang et al., 2025). However, retraining-based Shapley estimation is expensive. While KNN-based methods reduce this cost by evaluating a surrogate learner, their scores may not reflect the behavior of the target model. Influence functions avoid retraining by approximating the effect of reweighting examples around fitted parameters, but this approximation can be unreliable for deep networks (Basu et al., 2021). These limitations motivate an alternative: attributing fairness changes along the model's actual training process. In-Run Data Shapley (Wang et al., 2025) provides a construction for allocating local objective changes to training examples within a single run. We build on this construction to attribute changes in group-aware objectives.

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.

However, this construction relies on gradients, whereas common fairness criteria such as demographic parity and equal opportunity are evaluated using hard predictions. These criteria have zero gradients almost everywhere and therefore cannot directly provide useful attribution scores. Worst-group risk poses a related difficulty because the maximum is nonsmooth when the worst-performing group changes. Using differentiable objectives allows us to compute the scores, but does not by itself establish that they explain changes in the original fairness criteria. We therefore need to control both the attribution approximation error and the discrepancy introduced by making these criteria differentiable. This connection is essential for interpreting the scores: accurately attributing a smooth objective does not necessarily mean accurately attributing the fairness criterion we want to assess.

To address these issues, we propose FAIR-DS, which combines gradient-based fairness attribution with error bounds relating the scores to the intended fairness criteria. At each step, it compares each training gradient with the gradient of a smooth group-aware validation objective. For worst-group risk, this gradient emphasizes groups with larger losses. For parity criteria, it uses signed weights that reflect each group's deviation from the mean rate. Positive scores indicate an estimated harmful contribution to the selected objective. We bound the error in attributing objective changes and quantify the gap between the smooth objectives and the original fairness criteria under stated conditions. These results clarify when the scores support the intended fairness interpretation. The scores describe contributions along the observed training run and their value for guiding data repair is evaluated through separate retraining experiments.

Our contributions are threefold. First, we propose FAIR-DS, which attributes fairness changes to individual training examples along a single training trajectory. We construct differentiable group-aware objectives for commonly used fairness criteria and explain how their group weights determine each example's contribution. Second, we bound the attribution approximation error and the gap between the smooth objectives and the original fairness criteria. These bounds clarify when the scores reflect the intended criteria, with an additional condition for ranking stability under worst-group smoothing. Third, we evaluate the method on label-bias detection and attribution-guided data repair, demonstrating competitive performance on both tasks.

2 RELATED WORK

Fairness-aware learning and data selection. Bias mitigation methods fall into three families. Pre-processing removes discriminatory information from the data before training, via fair representation learning (Louizos et al., 2015; Zemel et al., 2013; Lum & Johndrow, 2016; Creager et al., 2019), fair data generation (Jang et al., 2021b), data mapping (Calmon et al., 2017; Gordaliza et al., 2019), or group-level rebalancing such as resampling and reweighing. In-processing builds fairness into training itself: by imposing fairness constraints or regularizers on the objective (Roh et al., 2020; Baharlouei et al., 2024; Donini et al., 2018; Chiappa et al., 2020), by reweighting examples or groups during optimization (Sagawa et al., 2020), or by adversarially removing sensitive information from the representation (Zhang et al., 2018); these methods may be model-specific (Zafar et al., 2015; Calders et al., 2009) or model-agnostic (Agarwal et al., 2018; Lowy et al., 2022). Post-processing adjusts a trained model's outputs to satisfy a fairness criterion (Hardt et al., 2016). Data selection and reweighting also provide ways to address label bias (Zhang et al., 2024; Wick et al., 2019; Wang et al., 2024). These methods primarily aim to reduce disparities. Our goal is to explain how individual training examples contribute to a selected group-aware objective during training and use these scores to guide data inspection and repair.

Data attribution. Data Shapley allocates a utility to examples using average marginal contributions (Ghorbani & Zou, 2019). Its axioms characterize the allocation within the specified game. Influence functions estimate sensitivity to training-data perturbations (Koh & Liang, 2017), while TracIn uses gradient alignment at training checkpoints (Pruthi et al., 2020). A line of work has retargeted this machinery at fairness, valuing training points by their contribution to a group fairness metric rather than to accuracy, and then removing, reweighting, or acquiring data accordingly (Bhatt et al., 2021; Arnaiz-Rodríguez & Oliver, 2024; Wang et al., 2024; Yang et al., 2025; Pang et al., 2024). These approaches face different limitations. Retraining-based Shapley estimation is expensive, while KNN surrogates evaluate a different learner. Influence functions approximate reweighting effects around fitted parameters, but this approximation can be unreliable for deep networks (Basu et al., 2021). Conventional retraining-based Data Shapley also evaluates a learning

algorithm across data subsets, rather than explaining a particular realized training run (Wang et al., 2025). In-Run Data Shapley computes local allocations along a realized run and develops efficient gradient inner-product calculations (Wang et al., 2025). Building on this machinery, our analysis addresses how group-aware objectives, smooth approximations, and changing group risks affect the interpretation of its scores.

3 PRELIMINARIES

We consider a binary classification problem with demographic groups. The training set is defined as $\mathcal{D}^{\mathrm{tr}} = \{z_i\}_{i=1}^n$ with $z_i = (x_i, y_i, s_i)$ drawn from a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y} \times [G]$, where x_i denotes non-protected features, $y_i \in \{0, 1\}$ is a binary label, and $s_i \in [G] := \{1, \ldots, G\}$ records demographic group membership. A model f_θ with parameters $\theta \in \mathbb{R}^d$ outputs a logit $r_\theta(x) \in \mathbb{R}$, a predicted probability $p_\theta(x) := \sigma(r_\theta(x)) \in (0, 1)$ of $y = 1$ with σ as the logistic function, and a hard prediction $\hat{y}_\theta(x) := \mathbb{1}[r_\theta(x) \geq 0] = \mathbb{1}[p_\theta(x) \geq 1/2]$. It incurs pointwise training loss $\ell(z_i; \theta)$, and the goal is to minimize it via an iterative optimization procedure such as SGD. At step $t \in \{0, \ldots, T-1\}$, where T is the number of training steps, a batch $\mathcal{B}_t \subseteq \mathcal{D}^{\mathrm{tr}}$ is drawn, and the parameters are updated as follows:

$$\theta_{t+1} := \theta_t + \Delta_t(\mathcal{B}_t), \qquad \Delta_t(S) := -\eta_t \sum_{z_j \in S} \nabla_\theta \ell(z_j; \theta_t), \quad S \subseteq \mathcal{B}_t. \tag{1}$$

When training uses the mean batch loss, the factor $1/|\mathcal{B}_t|$ is absorbed into $\eta_t \geq 0$. Let $\mathcal{D}^{\mathrm{val}}$ be an independent validation set, partitioned by s into nonempty groups $\mathcal{D}_1^{\mathrm{val}}, \ldots, \mathcal{D}_G^{\mathrm{val}}$. The validation risk of group g is given as

$$L_g(\theta) := \frac{1}{\left|\mathcal{D}_g^{\mathrm{val}}\right|} \sum_{z^{\mathrm{val}} \in \mathcal{D}_g^{\mathrm{val}}} \ell(z^{\mathrm{val}}; \theta), \qquad g \in [G].$$

We use a scalar group-aware objective $F : \mathbb{R}^d \to \mathbb{R}$, with lower values preferred. Section 4.2 gives concrete choices of popular fairness criteria. We attribute changes in F along the realized training trajectory.

In-Run Data Shapley Data Shapley values a training example by its average marginal contribution to a chosen utility across data subsets (Ghorbani & Zou, 2019). Conventional utilities evaluate models retrained on those subsets. In-Run Data Shapley (Wang et al., 2025) instead attributes changes along one realized training run. Specifically, for a validation point z^{val}, its local utility is

$$U^t(S; z^{\mathrm{val}}) := \ell(z^{\mathrm{val}}; \theta_t + \Delta_t(S)) - \ell(z^{\mathrm{val}}; \theta_t), \qquad S \subseteq \mathcal{B}_t. \tag{2}$$

This measures the loss change when only S contributes to the current update, with the parameters and training randomness held fixed. We write $\widetilde{\theta}_{t+1}(S) := \theta_t + \Delta_t(S)$ for this virtual update. The full-run utility sums these local changes, and Shapley linearity allows each example's contributions to be accumulated over the steps in which it appears. With a first-order Taylor approximation, each example's contribution can be computed directly from its training gradient and the gradient of the validation loss, avoiding evaluations over different data subsets. We adopt this framework to trace how training examples affect group-aware objectives, with positive scores indicating an increase in the selected objective.

4 METHOD: FAIR-DS

We now present our method. We consider $\mathcal{D}^{\mathrm{tr}}$ with group-dependent label bias that can harm fairness, and assume access to a clean validation set $\mathcal{D}^{\mathrm{val}}$ with reliable labels. Our goal is to identify training examples that contribute to this harm through their effects on a group-aware validation objective. We first define the attribution objective and derive its first-order scores, then specify objectives for common fairness criteria.

4.1 GROUP-AWARE ATTRIBUTION ALONG TRAINING

Let $F(\theta)$ be a differentiable group-aware objective, where larger values indicate worse outcomes under the selected objective. Our goal is to attribute changes in F along the training trajectory to

individual instances. Replacing the validation loss in eq. (2) with F, the local utility of $S \subseteq \mathcal{B}_t$ at step t is

$$U_F^t(S) := F(\widetilde{\theta}_{t+1}(S)) - F(\theta_t). \tag{3}$$

A negative utility indicates that the update decreases F. For $S \subseteq \mathcal{D}^{\mathrm{tr}}$, define $U_F(S) := \sum_{t=0}^{T-1} U_F^t(S \cap \mathcal{B}_t)$. By Shapley linearity, an example's total contribution is the sum of its contributions over training: i.e., $\phi_{z_i}(U_F) = \sum_{t=0}^{T-1} \phi_{z_i}(U_F^t)$, where $\phi_{z_i}(U_F^t) = 0$ if $z_i \notin \mathcal{B}_t$. To compute these per-step contributions efficiently, we approximate F around the current parameters θ_t. For $S \subseteq \mathcal{B}_t$, a first-order Taylor expansion gives

$$U_F^t(S) = -\eta_t \sum_{z_j \in S} \langle \nabla F(\theta_t), \nabla_\theta \ell(z_j; \theta_t) \rangle + O(\|\Delta_t(S)\|^2). \tag{4}$$

The approximation is additive, so each example's Shapley value is its corresponding term in eq. (4).

Definition 4.1 (First-order in-run fairness attribution score) *For a training example $z_i \in \mathcal{D}^{\mathrm{tr}}$, a differentiable group-aware objective F, and a training trajectory $\{\theta_t\}_{t=0}^{T}$, the first-order in-run fairness attribution score is defined as*

$$\phi_{z_i}^{\mathrm{fair}} := \sum_{t:\, z_i \in \mathcal{B}_t} \left(-\eta_t \langle \nabla F(\theta_t), \nabla_\theta \ell(z_i; \theta_t) \rangle \right). \tag{5}$$

An example receives a positive contribution when its update increases F to first order, and a negative contribution when it decreases F. The accumulated score summarizes these contributions over training. The sum of all examples' scores therefore approximates the total change in F along the training trajectory:

$$\sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} = F(\theta_T) - F(\theta_0) + \sum_{t=0}^{T-1} O\big(\|\Delta_t(\mathcal{B}_t)\|^2\big).$$

Theorem 5.1 bounds this approximation error. Following Wang et al. (2025), a second-order Taylor approximation can further account for the curvature of F and interactions between examples in the same batch.

4.2 From Group Statistics to Example Contributions

A training update may improve performance for one group while worsening it for another. Whether this update is harmful depends on the selected criterion. For example, worst-group risk emphasizes groups with higher loss, while parity-based criteria measure differences in prediction rates across groups. We express each example's fairness attribution score as a weighted sum of its effects on the relevant validation-group statistics.

Let $R(\theta) = (R_1(\theta), \ldots, R_m(\theta))$ collect the relevant validation statistics, where m denotes the number of statistics used by the selected criterion (i.e., $m = G$ for worst-group risk, demographic parity, and equal opportunity; $m = 2G$ for equalized odds). We write $F(\theta) = \mathcal{A}(R(\theta))$, where $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}$ is an aggregation function that combines the validation statistics into the selected objective. Define $q_j(\theta) := \partial \mathcal{A}(R(\theta))/\partial R_j$. The chain rule gives $\nabla F(\theta) = \sum_{j=1}^{m} q_j(\theta) \nabla R_j(\theta)$, so eq. (5) becomes

$$\phi_{z_i}^{\mathrm{fair}} = -\sum_{t:\, z_i \in \mathcal{B}_t} \eta_t \sum_{j=1}^{m} q_j(\theta_t) \langle \nabla R_j(\theta_t), \nabla_\theta \ell(z_i; \theta_t) \rangle. \tag{6}$$

The term $-\eta_t \langle \nabla R_j(\theta_t), \nabla_\theta \ell(z_i; \theta_t) \rangle$ gives the first-order change in R_j due to the update from z_i, while $q_j(\theta_t)$ weights its contribution to F. Contributions through different R_j can have opposite signs and cancel in $\phi_{z_i}^{\mathrm{fair}}$. The attribution score therefore depends on how the selected criterion combines these effects, so the same example can receive different scores under different fairness objectives.

Computing these attribution scores requires a differentiable objective. Group parity statistics based on hard predictions have zero gradients almost everywhere, so we compute these statistics using predicted probabilities $p_\theta(x)$ (Zafar et al., 2019; Bendekgey & Sudderth, 2021; Donini et al., 2018). For worst-group risk, we use a smooth approximation to the maximum group loss. Table 1 summarizes the resulting objectives.

Table 1: Fairness objectives and their group statistics. All statistics are computed on $\mathcal{D}^{\mathrm{val}}$.

Criterion	Group statistic	Objective $F(\theta)$	Reported metric
Worst-group risk	L_g	$\mathrm{lse}_\tau(L_1, \ldots, L_G)$	WGA
Demographic parity	C_g	$\mathrm{Var}(C)$	DP gap
Equal opportunity	C_g^+	$\mathrm{Var}(C^+)$	EO gap
Equalized odds	C_g^+, C_g^-	$\mathrm{Var}(C^+) + \mathrm{Var}(C^-)$	EOdds gap

Worst-group performance. Worst-group accuracy (WGA) is based on hard predictions and has zero gradients almost everywhere, making it unsuitable for direct gradient-based attribution. We therefore use worst-group risk, $L_{\mathrm{wg}}(\theta) := \max_g L_g(\theta)$, as a loss-based proxy for worst-group performance. Since the maximum can be nondifferentiable when groups tie for the highest loss, we use the smooth approximation $F_{\mathrm{wg}}(\theta) := \mathrm{lse}_\tau(L_1(\theta), \ldots, L_G(\theta)) = \tau^{-1} \log \sum_{g=1}^G \exp(\tau L_g(\theta))$, where $\tau > 0$ controls the sharpness of the approximation. The coefficients in eq. (6) are $q_g(\theta) := \exp(\tau L_g(\theta)) / \sum_{h=1}^G \exp(\tau L_h(\theta))$. These nonnegative weights sum to one and give greater weight to groups with higher loss. An example can therefore receive a positive score when its update increases loss in high-loss groups, even if it reduces loss elsewhere. The weights change as the group risks change during training. Theorem 5.2 bounds the difference between the summed attribution scores and the change in worst-group risk, accounting for both smoothing and attribution errors. Worst-group risk measures performance in the worst-performing group rather than the gap between groups, so reducing it does not necessarily reduce group disparities.

Group parity. Parity objectives measure differences in prediction statistics across demographic groups. Demographic parity requires equal positive prediction rates across groups, equal opportunity requires equal true positive rates, and equalized odds requires both equal true positive rates and equal false positive rates. Let C_g be the mean predicted probability of the positive class on $\mathcal{D}_g^{\mathrm{val}}$. We use C_g^+ and C_g^- for the corresponding means among examples with $y = 1$ and $y = 0$, respectively, and assume that all required subgroups are nonempty. We measure differences across groups using $\mathrm{Var}(v) := G^{-1} \sum_{g=1}^G (v_g - \bar{v})^2$, where $v \in \mathbb{R}^G$ and $\bar{v} := G^{-1} \sum_{g=1}^G v_g$ (Williamson & Menon, 2019; Speicher et al., 2018; Berk et al., 2017). Writing $C = (C_1, \ldots, C_G)$ and $C^\pm = (C_1^\pm, \ldots, C_G^\pm)$, the soft objectives are $F_{\mathrm{dp}} := \mathrm{Var}(C)$, $F_{\mathrm{eo}} := \mathrm{Var}(C^+)$, and $F_{\mathrm{eodds}} := \mathrm{Var}(C^+) + \mathrm{Var}(C^-)$. For each variance term, R_g is the corresponding statistic in Table 1, with coefficient $2(R_g - \bar{R})/G$, where $\bar{R} := G^{-1} \sum_{g=1}^G R_g$. Unlike the nonnegative worst-group weights, these coefficients are positive above the mean and negative below it. So an update can reduce disparity among positive-label examples while increasing it among negative-label examples. Equal opportunity considers only the former, whereas equalized odds considers both. Appendix B details each attribution objective.

Relation to the original criteria. Our scores attribute changes in the smooth objectives, which approximate the original fairness criteria. For worst-group risk, $L_{\mathrm{wg}} \le F_{\mathrm{wg}} \le L_{\mathrm{wg}} + \log G/\tau$. A larger τ gives a closer approximation to the maximum, but can increase the error of first-order attribution. For parity objectives, equal mean predicted probabilities across groups do not necessarily imply equal proportions of positive predictions. Theorems 5.2 and 5.3 bound these approximation errors and relate the summed scores to changes in the original criteria along training. We evaluate whether correcting or removing highly ranked examples improves fairness by retraining the model. The empirical results are in Section 6.2, Section 6.3 and Section D.3.

5 Theoretical Analysis

We analyze two sources of error: approximating the attribution scores and using smooth approximations to the original fairness criteria. All results concern a fixed training trajectory and validation set. Proofs and supporting results are provided in Appendix C.

Theorem 5.1 (Approximate efficiency) *Suppose F is β-smooth on each segment $[\theta_t, \theta_{t+1}]$ of the realized SGD trajectory. Then*

$$\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(F(\theta_T) - F(\theta_0)\big) \right| \leq \frac{\beta}{2} \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2. \tag{7}$$

The summed scores account for the change in the selected smooth objective, with error controlled by curvature and update size. Theorem C.2 further bounds each step's error relative to its exact local Shapley value. To connect the scores to the original fairness criteria, we next account for the error introduced by smoothing.

Theorem 5.2 (Fidelity of worst-group risk) *Let $G \geq 1$ and $\tau > 0$. If $F = F_{\mathrm{wg}}$ satisfies the assumptions of Theorem 5.1, then*

$$\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(L_{\mathrm{wg}}(\theta_T) - L_{\mathrm{wg}}(\theta_0)\big) \right| \leq \frac{\beta}{2} \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2 + \frac{\log G}{\tau}. \tag{8}$$

The first term in eq. (8) bounds the first-order attribution error, while the second bounds the error from approximating the maximum with LSE. Increasing τ gives a closer approximation to the maximum, but may increase β and hence the attribution error bound. Appendix C.6 analyzes this trade-off. With binary cross-entropy loss, reducing worst-group risk tightens an upper bound on worst-group classification error (Appendix C.2). We evaluate whether attribution-guided repair improves worst-group accuracy in our experiments.

Theorem 5.3 (Parity fidelity) *Let F be a soft parity objective and H its hard-prediction counterpart, obtained by replacing $p_\theta(x)$ with $\hat{y}_\theta(x)$ in the group statistics. For demographic parity, define*

$$B_{\mathrm{dp}}(\theta) := \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \frac{1}{|\mathcal{D}_g^{\mathrm{val}}|} \sum_{z^{\mathrm{val}} \in \mathcal{D}_g^{\mathrm{val}}} \sigma(-|r_\theta(x)|).$$

For equal opportunity, define B_{eo} using the positive-label subgroups. For equalized odds, let B_{eodds} be the sum of the corresponding positive- and negative-label bounds. Suppose all relevant validation subgroups are nonempty and F satisfies the assumptions of Theorem 5.1. Then, with B chosen for the selected criterion,

$$\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(H(\theta_T) - H(\theta_0)\big) \right| \leq \frac{\beta}{2} \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2 + B(\theta_0) + B(\theta_T). \tag{9}$$

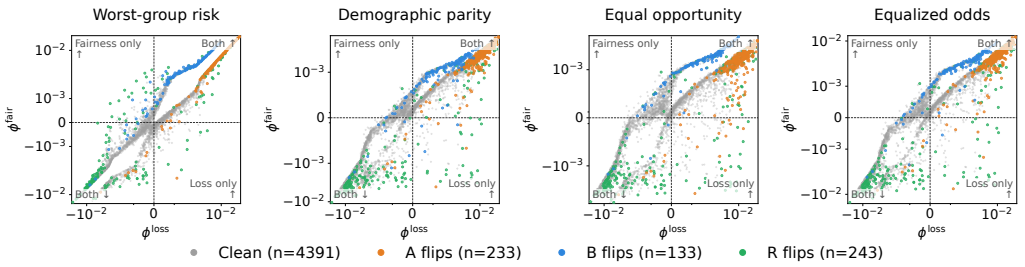
The bound separates first-order attribution error from the discrepancy between objectives computed using predicted probabilities and hard predictions. The latter is bounded by $B(\theta_0) + B(\theta_T)$, which depends on prediction confidence at the start and end of training (larger absolute logits give smaller bounds). When both error terms are small, the summed scores closely approximate the change in H. Here, H measures the variance of group prediction statistics, rather than the unsquared group gaps reported in our experiments. The proof combines Proposition C.1 with eq. (7).

6 Experiments

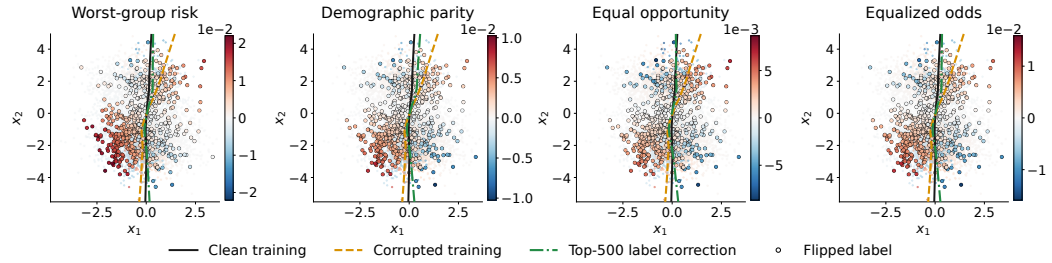
In this section, we evaluate two tasks: injected label bias detection and attribution-guided data repair (here, label correction). The first tests whether our attribution scores identify harmful examples, while the second examines whether correcting labels among highly ranked examples improves accuracy and fairness after retraining.

6.1 Experimental Setup

For each dataset, we inject group-dependent label bias (Wick et al., 2019) with $p(\tilde{y} = 1 \mid y = 0, s = 1) = p(\tilde{y} = 0 \mid y = 1, s = 2) = \rho$, where y and $\tilde{y}$ are the original and corrupted labels. We set $\rho = 0.4$ to evaluate performance under severe label bias. Results are reported as mean$\pm$std over 10 seeds. Model and training details are in Appendix D.2.



(a) Fairness attribution versus loss attribution.



(b) Effects of correcting top-ranked examples.

Figure 1: Attribution analysis on synthetic data. *A* and *B* denote group-dependent label flips, and *R* denotes additional random flips; ϕ^{loss} uses the average validation loss as the utility. **(a)** Most *A*/*B* flips lie in the upper quadrants, indicating positive fairness scores. **(b)** Correcting labels within the top-500 ranked examples moves the decision boundary closer to the clean model.

Dataset and Baselines. We use Adult Census Income (Adult Census) (Kohavi, 1996) to predict whether annual income exceeds \$50K, with gender as the sensitive attribute. On UTKFace (Zhang et al., 2017) and FairFace (Karkkainen & Joo, 2021), we predict the recorded binary gender label, with race as the sensitive attribute. We compare against fairness-oriented methods, *FairShap* (Arnaiz-Rodríguez & Oliver, 2024), *FairIF* (Wang et al., 2024), and *IFFair* (Yang et al., 2025). We also include performance-oriented attribution methods: *KNN-Shapley* (Jia et al., 2019), *Influence Function* (Koh & Liang, 2017), and first- and second-order *In-Run Shapley* (Wang et al., 2025). Simple baselines include *Train Loss*, *Gradient norm*, and *Random*, which rank examples by final training loss, final per-example gradient norm, and random order, respectively.

Evaluation Metrics. For detection, we rank examples by decreasing ϕ^{fair} and report AUROC, average precision (AP), and Precision@K for $K \in \{100, 1000, 2000\}$ where feasible. For attribution-guided label correction, we additionally report *No correction*, which retains all corrupted labels, and an *Oracle* that restores all corrupted labels without a selection-budget constraint. All comparisons use the same architecture, optimizer, and paired random seeds. Downstream metrics include test accuracy, WGA, DP gap, and EO gap. Further experimental details are in Appendix D.2.

6.2 Detection of Injected Label Bias

Fairness attribution and data repair. We generate 8,000 synthetic examples with two group-conditional Gaussian features. We use 5,000 examples for training and 1,500 each for validation and testing. Training labels undergo majority-group $0 \to 1$ flips (*A*, probability 0.15), minority-group $1 \to 0$ flips (*B*, probability 0.35), and random flips among the remaining unflipped examples (*R*, probability 0.05). Appendix D.1 provides generation details. Across all four objectives, over 90% of both *A* and *B* flips receive positive fairness scores (Figure 1a). Under equal opportunity, 11.3% of *B* flips receive positive fairness scores but non-positive loss scores, showing that loss attribution can miss biased examples. Correcting labels within the top-500 examples improves accuracy, reduces DP, TPR, and FPR gaps, and brings predictions closer to the clean model under all four objectives (Figure 1b). Removal results are in Appendix D.3.

Table 2: Evaluation of injected label bias detection. Higher is better for all metrics. FAIR-DS uses an LSE approximation to worst-group risk with $\tau = 0.01$. The best results are **bold**, and the second-best results are underlined.

Method	AUROC	AP	Precision@100	Precision@1000	Precision@2000
			Adult Census		
Random	0.501±0.005	0.201±0.004	0.198±0.019	0.199±0.015	0.199±0.011
KNN-Shapley	0.881±0.002	0.655±0.005	0.852±0.021	0.790±0.006	0.772±0.007
Train Loss	0.847±0.005	0.583±0.008	0.848±0.022	0.740±0.012	0.706±0.017
Gradient norm	0.832±0.004	0.474±0.012	0.472±0.045	0.474±0.029	0.509±0.030
FairShap	0.460±0.002	0.199±0.003	0.142±0.029	0.239±0.011	0.246±0.004
FairIF	0.583±0.009	0.359±0.007	0.822±0.038	0.617±0.019	0.542±0.013
IFFair	0.658±0.006	0.551±0.008	**0.966±0.015**	0.874±0.012	0.829±0.011
Influence Function	0.874±0.036	0.689±0.052	0.896±0.037	0.854±0.030	0.838±0.035
In-Run Shapley (1)	**0.891±0.003**	0.739±0.018	0.895±0.070	0.867±0.040	0.867±0.030
In-Run Shapley (2)	**0.891±0.003**	0.739±0.018	0.895±0.070	0.867±0.040	0.866±0.030
FAIR-DS	**0.891±0.003**	**0.743±0.017**	0.916±0.079	**0.881±0.036**	**0.876±0.026**
			UTKFace		
Random	0.499±0.005	0.195±0.004	0.208±0.021	0.191±0.011	0.195±0.010
KNN-Shapley	**0.857±0.002**	0.615±0.005	0.901±0.019	0.756±0.007	0.683±0.007
Train Loss	0.736±0.086	0.418±0.109	0.557±0.198	0.540±0.170	0.499±0.146
Gradient norm	0.727±0.083	0.367±0.087	0.389±0.126	0.422±0.114	0.427±0.106
FairShap	0.439±0.004	0.173±0.003	0.196±0.030	0.171±0.010	0.168±0.006
FairIF	0.587±0.062	0.331±0.068	0.512±0.141	0.497±0.114	0.454±0.099
IFFair	0.683±0.067	0.322±0.069	0.381±0.185	0.379±0.119	0.371±0.100
Influence Function	0.584±0.087	0.338±0.074	0.568±0.167	0.512±0.106	0.459±0.088
In-Run Shapley (1)	0.810±0.011	0.654±0.020	0.917±0.041	0.874±0.015	0.792±0.020
In-Run Shapley (2)	0.842±0.012	0.665±0.023	0.892±0.043	0.851±0.022	0.771±0.019
FAIR-DS	0.835±0.008	**0.700±0.017**	**0.924±0.035**	**0.900±0.018**	**0.843±0.020**
			FairFace		
Random	0.499±0.002	0.191±0.001	0.177±0.028	0.191±0.012	0.190±0.007
KNN-Shapley	0.728±0.002	0.377±0.003	0.695±0.028	0.553±0.009	0.514±0.010
Train Loss	0.741±0.051	0.386±0.076	0.750±0.093	0.582±0.043	0.523±0.076
Gradient norm	0.732±0.047	0.352±0.053	0.593±0.077	0.506±0.048	0.462±0.031
FairShap	0.545±0.001	0.214±0.002	0.467±0.037	0.305±0.014	0.273±0.007
FairIF	0.553±0.026	0.279±0.023	0.573±0.060	0.462±0.069	0.445±0.061
IFFair	0.660±0.018	0.294±0.027	0.437±0.105	0.386±0.062	0.369±0.065
Influence Function	0.581±0.022	0.300±0.026	0.509±0.081	0.468±0.077	0.463±0.070
In-Run Shapley (1)	0.756±0.003	0.582±0.007	0.851±0.052	0.869±0.020	0.862±0.015
In-Run Shapley (2)	0.754±0.003	0.569±0.009	0.803±0.059	0.845±0.018	0.840±0.019
FAIR-DS	**0.791±0.009**	**0.627±0.009**	**0.950±0.026**	**0.949±0.014**	**0.931±0.013**

Comparison results. Table 2 shows that FAIR-DS achieves the highest average precision and Precision@1000 and @2000 on all three datasets. On Adult Census, its AUROC is similar to In-Run Shapley, but its average precision and top-K precision are higher. This suggests that FAIR-DS identifies more corrupted examples within the same inspection budget. The distinction is clearer on UTKFace: KNN-Shapley achieves higher AUROC, whereas FAIR-DS leads in average precision and all three top-K metrics. On FairFace, FAIR-DS also performs best across all metrics. These results suggest that fairness-based attribution helps identify group-dependent label bias that loss-based attribution may miss. FairShap performs below random on Adult Census and UTKFace, implying that optimizing a single fairness metric can fail to identify truly harmful examples. Additional results are provided in Appendix D.4.

6.3 DOWNSTREAM DATA REPAIR

For the downstream task, we inspect the top-4000 ranked examples and restore the original labels only for those known to have been synthetically flipped. Selected clean examples remain unchanged, and all training examples are retained for retraining to evaluate accuracy and fairness. Table 3 shows that the choice of attribution objective affects the repair outcome. The worst-group risk variant

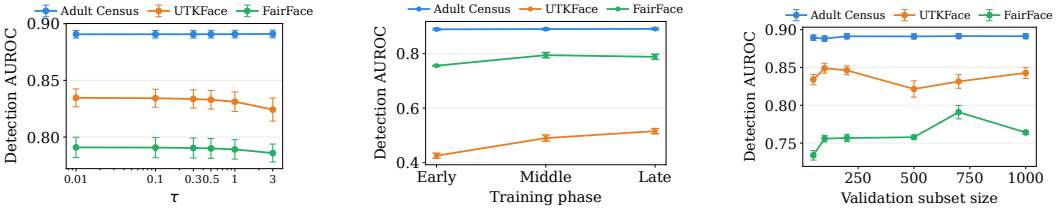
Table 3: Downstream performance after top-4000 label correction on the FairFace dataset. All training examples are retained for retraining. LSE denotes the variant using an LSE approximation to worst-group risk; DP, EO, and EOdds denote demographic parity, equal opportunity, and equalized odds, respectively.

Method	True corrected	Precision	Errors left	TestAcc↑	WGA ↑	DP Gap ↓	EO Gap ↓
No correction	0.0±0.0	/	16664.7±105.7	0.733±0.037	0.717±0.043	0.167±0.029	0.133±0.072
FAIR-DS$_{\text{LSE}}$	3487.7±48.7	0.872±0.012	13177.0±124.4	0.764±0.019	0.750±0.026	0.140±0.023	0.125±0.059
FAIR-DS$_{\text{DP}}$	2191.6±65.0	0.548±0.016	14473.1±140.1	0.747±0.035	0.734±0.040	0.105±0.024	0.041±0.009
FAIR-DS$_{\text{EO}}$	2085.2±26.1	0.521±0.007	14579.5±105.8	0.743±0.024	0.729±0.027	0.121±0.024	0.050±0.012
FAIR-DS$_{\text{EOdds}}$	2467.8±137.2	0.617±0.034	14196.9±207.8	0.761±0.033	0.751±0.037	0.087±0.027	0.038±0.009
Oracle	16664.7±105.7	1.000±0.000	0.0±0.0	0.818±0.018	0.815±0.019	0.018±0.017	0.026±0.020

corrects the most flipped labels and achieves the highest mean accuracy, but yields only a small reduction in the mean EO gap. The parity objectives correct fewer labels yet reduce group disparities more. Equalized odds achieves the lowest DP and EO gaps, reducing them from 0.167 to 0.087 and from 0.133 to 0.038, respectively, while also improving accuracy and WGA. These results suggest that fairness improvements depend on which examples are corrected, not only on how many label errors are corrected. Detailed results for all baselines are provided in Appendix D.4.

6.4 Ablation Study

We study FAIR-DS's sensitivity to the smoothing parameter, training phase, and validation set size, reporting detection AUROC in Figure 2. We first vary $\tau \in \{0.01, 0.1, 0.3, 0.5, 1.0, 3.0\}$ to assess sensitivity to the smooth approximation of worst-group risk. We then examine how attribution depends on the stage of training by accumulating scores separately over early ($0 \le t < T/4$), middle ($T/4 \le t < 3T/4$), and late ($3T/4 \le t < T$) steps. Finally, we vary the validation set size over $\{50, 100, 200, 500, 700, 1000\}$ to assess how much validation data the method needs for reliable detection (The validation subsets are resampled independently). Detection is stable across τ on Adult Census and declines slightly at larger τ on the image datasets, where scores from the middle training stage perform best. Larger validation sets improve detection on FairFace but provide little benefit on Adult Census and no consistent gain on UTKFace.



(a) LSE smoothing parameter. (b) Training phase. (c) Validation set size.

Figure 2: Ablation studies on the smoothing parameter, training phase, and validation set size across different datasets. Error bars show standard deviations over 10 seeds.

7 Conclusions

We proposed FAIR-DS, a unified framework that identifies which training examples harm fairness and traces how their contributions accumulate during training. The framework supports a broad range of popular fairness criteria within a single training run. Our theoretical analysis bounds the attribution error and the gap between smooth objectives and the original fairness criteria, establishing when the summed attribution scores explain changes in fairness. Empirical results demonstrate its effectiveness in detecting injected label bias and guiding label correction. These results suggest that FAIR-DS can help prioritize data inspection and identify repair targets, helping practitioners address fairness issues in training data. A limitation is the need for clean validation data with adequate group coverage. Future work will explore attribution with limited or noisy validation data.

AI USE STATEMENT

We used AI tools to improve the clarity and organization of the manuscript, assist in checking and revising mathematical derivations and proofs, provide feedback on experimental design, and support the implementation and debugging of experimental code. The authors take full responsibility for the final manuscript, theoretical claims, code, and reported results. Also, the authors reviewed the AI-assisted text and proofs and tested the experimental code.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018.

Adrián Arnaiz-Rodríguez and Nuria Oliver. Towards algorithmic fairness by means of instance-level data re-weighting based on shapley values. In *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR)*, 2024. Method known as FairShap; extended version at arXiv:2303.01928.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference of Learning Representations*, 2020.

Sina Baharlouei, Shivam Patel, and Meisam Razaviyayn. f-FERM: A scalable framework for robust fair empirical risk minimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

Harry Bendekgey and Erik B. Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017.

Umang Bhatt, Isabel Chien, Muhammad Bilal Zafar, and Adrian Weller. DIVINE: Diverse influential training points for data visualization and model refinement. *arXiv preprint arXiv:2107.05978*, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, Dec 2009. doi: 10.1109/ICDMW.2009.83.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, pp. 3992–4001. 2017.

Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. A general approach to fairness with optimal transport. In *AAAI*, pp. 3633–3640. AAAI Press, 2020.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. *CoRR*, abs/1906.02589, 2019.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2796–2806, Red Hook, NY, USA, 2018.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019.

Paula Gordaliza, Eustasio Del Barrio, Fabrice Gamboa, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2357–2365. PMLR, 09–15 Jun 2019.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification, 2021a.

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7908–7916, May 2021b. doi: 10.1609/aaai.v35i9.16965.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gürel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 12(11):1610–1623, 2019. doi: 10.14778/3342263.3342637.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 202–207. AAAI Press, 1996.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The Variational Fair Autoencoder. *arXiv e-prints*, art. arXiv:1511.00830, Nov 2015.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Expert Certification.

Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv e-prints*, art. arXiv:1610.08077, Oct 2016.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Jinlong Pang, Jialu Wang, Zhaowei Zhu, Yuanshun Yao, Chen Qian, and Yang Liu. Fairness without harm: An influence-guided active sampling approach. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61513–61548, 2024.

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching, 2019.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. FR-train: A mutual information-based approach to fair and robust training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8147–8157. PMLR, 13–18 Jul 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 (ICLR)*, 2020.

Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018. doi: 10.1145/3219819.3220046.

Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pp. 155–166. PMLR, 2020.

Haonan Wang, Ziwei Wu, and Jingrui He. Fairif: Boosting fairness in deep learning via influence functions with validation set sensitive attributes. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pp. 721–730, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635844.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *International Conference on Learning Representations*, 2025.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 526–536, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Jingran Yang, Min Zhang, Lingfeng Zhang, Zhaohui Wang, and Yonggang Zhang. IFFair: Influence function-driven sample reweighting for fair classification. *arXiv preprint arXiv:2512.07249*, 2025.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, art. arXiv:1507.05259, Jul 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3278721.3278779.

Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. Mitigating label bias in machine learning: Fairness through confident learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16917–16925, Mar. 2024. doi: 10.1609/aaai.v38i15.29634.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020.

A Implementation and cost

At each attribution step, we compute $v_t := \nabla F(\theta_t)$ on $\mathcal{D}^{\mathrm{val}}$ and take its inner product with each training gradient in $\mathcal{B}_t$. These inner products can be computed directly or using the ghost dot-product technique of Wang et al. (2025), which avoids storing per-example gradients for supported layers. Algorithm 1 summarizes this procedure. Our method requires only one training run, but adds computation for the validation objective and its gradient. Therefore, the additional cost depends on the validation set size, attribution frequency, model architecture, and implementation.

Algorithm 1: FAIR-DS: first-order fairness attribution

Input: Training set $\mathcal{D}^{\mathrm{tr}}$, validation set $\mathcal{D}^{\mathrm{val}}$, objective F, initial parameters θ_0, training steps T, effective step sizes $\{\eta_t\}_{t=0}^{T-1}$ as in eq. (1).

Output: Attribution scores $\{\phi_{z_i}^{\mathrm{fair}}\}_{i=1}^{n}$.

```
Initialize φ_{z_i}^fair ← 0 for every z_i ∈ D^tr;
for t = 0, ..., T − 1 do
    Obtain a training batch B_t from D^tr;
    Compute the gradients ∇_θ ℓ(z_i; θ_t) for z_i ∈ B_t, with training randomness fixed;
    Compute group statistics on D^val and get ∇F(θ_t);
    foreach z_i ∈ B_t do
        φ_{z_i}^fair ← φ_{z_i}^fair − η_t ⟨∇F(θ_t), ∇_θ ℓ(z_i; θ_t)⟩;
    θ_{t+1} ← θ_t − η_t Σ_{z_i ∈ B_t} ∇_θ ℓ(z_i; θ_t);
return {φ_{z_i}^fair}_{i=1}^n;
```

Attribution frequency. In the experiment, we obtain the attribution score at every training step. To reduce computation, we may instead score a subset of steps $\mathcal{I}$, giving

$$\phi_{z_i}^{\mathrm{fair},\mathcal{I}} := -\sum_{t\in\mathcal{I}:\, z_i\in\mathcal{B}_t} \eta_t \left\langle \nabla F(\theta_t), \nabla_\theta \ell(z_i;\theta_t)\right\rangle .$$

This score includes only the contributions from the selected steps. Its difference from the full score satisfies

$$\left|\phi_{z_i}^{\mathrm{fair}} - \phi_{z_i}^{\mathrm{fair},\mathcal{I}}\right| \le \sum_{t\notin\mathcal{I}:\, z_i\in\mathcal{B}_t} \eta_t \|\nabla F(\theta_t)\| \|\nabla_\theta \ell(z_i;\theta_t)\|.$$

The omitted contributions must therefore be accounted for when relating the summed attribution scores to the total objective change. Scaling the attribution scores by the sampling interval does not generally recover the full scores.

Validation computation. We compute the group statistics on the full validation set and obtain $\nabla F(\theta_t)$ by automatic differentiation. If the validation set is processed in batches, we can keep θ_t fixed and combine the batch statistics before applying the fairness objective. Computing the objective separately on each batch and averaging the results generally will give a different objective value and gradient from those computed on the full validation set because the aggregation is nonlinear. Our theoretical results assume the same fixed validation set throughout attribution.

B Objective Details

This section details how the group statistics and attribution scores in Section 4.2 are computed. For training with the mean batch loss, η_t is the learning rate divided by the batch size, as in eq. (1). For each demographic group g, define

$$C_g(\theta) := \frac{1}{|\mathcal{D}_g^{\mathrm{val}}|} \sum_{z^{\mathrm{val}}\in\mathcal{D}_g^{\mathrm{val}}} p_\theta(x).$$

Let $\mathcal{D}^{\mathrm{val}}{}_g^{+} := \{z^{\mathrm{val}} \in \mathcal{D}^{\mathrm{val}}{}_g : y = 1\}$ and $\mathcal{D}^{\mathrm{val}}{}_g^{-} := \{z^{\mathrm{val}} \in \mathcal{D}^{\mathrm{val}}{}_g : y = 0\}$. The conditional soft rates are

$$C_g^{+}(\theta) := \frac{1}{|\mathcal{D}^{\mathrm{val}}{}_g^{+}|} \sum_{z^{\mathrm{val}}\in\mathcal{D}^{\mathrm{val}}{}_g^{+}} p_\theta(x), \qquad C_g^{-}(\theta) := \frac{1}{|\mathcal{D}^{\mathrm{val}}{}_g^{-}|} \sum_{z^{\mathrm{val}}\in\mathcal{D}^{\mathrm{val}}{}_g^{-}} p_\theta(x).$$

We compute each statistic by averaging within its group, so larger groups do not automatically receive greater weight. Parity objectives give equal weight to each group's squared deviation from the mean, while the worst-group objective gives more weight to groups with higher loss. Demographic parity uses group membership and predicted probabilities without requiring validation labels while worst-group risk, equal opportunity, and equalized odds require validation labels.

For any $v \in \mathbb{R}^G$,

$$\mathrm{Var}(v) = \frac{1}{G}\sum_{g=1}^{G}(v_g - \bar{v})^2 = \frac{1}{G^2}\sum_{g<h}(v_g - v_h)^2.$$

It is zero exactly when all group statistics agree. For two groups, it equals one quarter of their squared difference. In particular, the soft parity objectives are variance objectives rather than the unsquared hard-prediction gaps reported in the experiments.

Worst-group risk. For $\tau > 0$, the log-sum-exp approximation satisfies

$$L_{\mathrm{wg}}(\theta) \le \frac{1}{\tau}\log\sum_{g=1}^{G}\exp(\tau L_g(\theta)) \le L_{\mathrm{wg}}(\theta) + \frac{\log G}{\tau}.$$

Differentiating F_{wg} gives

$$\nabla F_{\mathrm{wg}}(\theta) = \sum_{g=1}^{G} q_g(\theta)\nabla L_g(\theta), \qquad q_g(\theta) := \frac{\exp(\tau L_g(\theta))}{\sum_{h=1}^{G}\exp(\tau L_h(\theta))}.$$

Its attribution score is therefore

$$\phi_{z_i}^{\mathrm{fair,wg}} = -\sum_{t:\, z_i\in\mathcal{B}_t}\eta_t\sum_{g=1}^{G} q_g(\theta_t)\left\langle \nabla L_g(\theta_t), \nabla_\theta \ell(z_i;\theta_t)\right\rangle.$$

When the validation loss is binary cross-entropy and the prediction threshold is $1/2$, we have

$$1 - \mathrm{WGA}(\theta) \le \frac{L_{\mathrm{wg}}(\theta)}{\log 2}.$$

Here, WGA is evaluated on the same validation set as L_{wg}. Thus, reducing worst-group risk tightens an upper bound on worst-group classification error, but does not necessarily improve WGA.

Demographic parity. For $F_{\mathrm{dp}}(\theta) := \mathrm{Var}(C(\theta))$, we define

$$\bar{C}(\theta) := \frac{1}{G}\sum_{g=1}^{G} C_g(\theta), \qquad w_g(\theta) := \frac{2}{G}\big(C_g(\theta) - \bar{C}(\theta)\big).$$

Then

$$\phi_{z_i}^{\mathrm{fair,dp}} = -\sum_{t:\, z_i\in\mathcal{B}_t}\eta_t\sum_{g=1}^{G} w_g(\theta_t)\left\langle \nabla C_g(\theta_t), \nabla_\theta \ell(z_i;\theta_t)\right\rangle.$$

The coefficient w_g is positive when $C_g > \bar{C}$, negative when $C_g < \bar{C}$, and zero when $C_g = \bar{C}$. Thus, its sign indicates whether the group's mean predicted probability is above or below the average across groups.

Equal opportunity and equalized odds. Define

$$\bar{C}^{\pm}(\theta) := \frac{1}{G}\sum_{g=1}^{G} C_g^{\pm}(\theta), \qquad w_g^{\pm}(\theta) := \frac{2}{G}\big(C_g^{\pm}(\theta) - \bar{C}^{\pm}(\theta)\big).$$

The equal opportunity score is

$$\phi_{z_i}^{\mathrm{fair,eo}} = -\sum_{t:\, z_i\in\mathcal{B}_t}\eta_t\sum_{g=1}^{G} w_g^{+}(\theta_t)\left\langle \nabla C_g^{+}(\theta_t), \nabla_\theta \ell(z_i;\theta_t)\right\rangle.$$

For equalized odds, since $F_{\text{eodds}} = \text{Var}(C^+) + \text{Var}(C^-)$, its first-order score is

$$\phi_{z_i}^{\text{fair,eodds}} = \phi_{z_i}^{\text{fair,eo}} - \sum_{t:\, z_i \in \mathcal{B}_t} \eta_t \sum_{g=1}^{G} w_g^-(\theta_t) \left\langle \nabla C_g^-(\theta_t), \nabla_\theta \ell(z_i; \theta_t) \right\rangle.$$

These weights correspond to the coefficients in eq. (6): $q_g = w_g$ for demographic parity and $q_g = w_g^+$ for equal opportunity. For equalized odds, we list the G positive-label statistics first, followed by the G negative-label statistics. Thus, group g's positive-label statistic has coefficient $q_g = w_g^+$, while its negative-label statistic occupies position $G + g$ and has coefficient $q_{G+g} = w_g^-$.

B.1 Decomposition by Validation Cell

Each attribution score combines an example's effects on different validation groups. To see whether these effects come from positive- or negative-label validation examples, we further divide each group by label. We call each group–label subset a validation cell and express the total score as the sum of its contributions through these cells. For group g and label $a \in \{0, 1\}$, let $\mathcal{D}^{\text{val}}{}_g^{(a)} := \{z^{\text{val}} \in \mathcal{D}^{\text{val}}{}_g : y = a\}$. We assume that all cells are nonempty. Define the mean loss and mean predicted probability within each cell, and its proportion within group g, as

$$L_g^{(a)}(\theta) := \frac{1}{|\mathcal{D}^{\text{val}}{}_g^{(a)}|} \sum_{z^{\text{val}} \in \mathcal{D}^{\text{val}}{}_g^{(a)}} \ell(z^{\text{val}}; \theta),$$

$$C_g^{(a)}(\theta) := \frac{1}{|\mathcal{D}^{\text{val}}{}_g^{(a)}|} \sum_{z^{\text{val}} \in \mathcal{D}^{\text{val}}{}_g^{(a)}} p_\theta(x), \qquad \pi_g^{(a)} := \frac{|\mathcal{D}^{\text{val}}{}_g^{(a)}|}{|\mathcal{D}^{\text{val}}{}_g|}.$$

Thus, $L_g = \sum_a \pi_g^{(a)} L_g^{(a)}$ and $C_g = \sum_a \pi_g^{(a)} C_g^{(a)}$. Note that $C_g^{(1)} = C_g^+$ and $C_g^{(0)} = C_g^-$. The cell statistic and coefficient for each objective are

Objective	Cell statistic $\Psi_g^{(a)}$	Coefficient $b_g^{(a)}$
F_{wg}	$L_g^{(a)}$	$q_g \pi_g^{(a)}$
F_{dp}	$C_g^{(a)}$	$w_g \pi_g^{(a)}$
F_{eo}	$C_g^{(a)}$	$\mathbb{1}[a = 1] w_g^+$
F_{eodds}	$C_g^{(a)}$	$\mathbb{1}[a = 1] w_g^+ + \mathbb{1}[a = 0] w_g^-$

The contribution through validation cell (g, a) is

$$\phi_{z_i;g,a}^{\text{fair}} := - \sum_{t:\, z_i \in \mathcal{B}_t} \eta_t b_g^{(a)}(\theta_t) \left\langle \nabla \Psi_g^{(a)}(\theta_t), \nabla_\theta \ell(z_i; \theta_t) \right\rangle.$$

By the chain rule and linearity,

$$\phi_{z_i}^{\text{fair}} = \sum_{g=1}^{G} \sum_{a \in \{0,1\}} \phi_{z_i;g,a}^{\text{fair}}.$$

Summing the contributions from all validation cells recovers the original first-order attribution score without any additional approximation.

We use this decomposition to construct the heatmaps in Figure 6. Each row contains training examples with the same sensitive group and observed label, and each column corresponds to a validation cell. An entry shows the mean contribution through that cell for the training examples in the row. Summing across a row therefore gives their mean total attribution score.

C Proofs

We use the same notation as the main text. For brevity, write $u_{i,t} := \nabla_\theta \ell(z_i; \theta_t)$ and $v_t := \nabla F(\theta_t)$. All gradients at step t are evaluated at θ_t, before the model update. The effective step size η_t already includes the mean-batch normalization, so no additional factor $1/|\mathcal{B}_t|$ appears in the proofs.

C.1 Proof of Theorem 5.1

Proof. Let $\delta_t := \Delta_t(\mathcal{B}_t) = \theta_{t+1} - \theta_t$. The fundamental theorem of calculus gives

$$F(\theta_{t+1}) - F(\theta_t) = \langle v_t, \delta_t \rangle + r_t,$$
$$r_t = \int_0^1 \langle \nabla F(\theta_t + u\delta_t) - v_t, \delta_t \rangle \, du.$$

The segment smoothness assumption and Cauchy–Schwarz imply $|r_t| \le \int_0^1 \beta u \|\delta_t\|^2 du = \beta\|\delta_t\|^2/2$. Then by eq. (5), we have

$$\sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} = -\sum_{t=0}^{T-1} \eta_t \left\langle v_t, \sum_{z_i \in \mathcal{B}_t} u_{i,t} \right\rangle = \sum_{t=0}^{T-1} \langle v_t, \delta_t \rangle .$$

Summing over training steps cancels the intermediate objective values, giving

$$F(\theta_T) - F(\theta_0) = \sum_{t=0}^{T-1} \langle v_t, \delta_t \rangle + \sum_{t=0}^{T-1} r_t.$$

Since the first sum equals the total attribution score,

$$\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(F(\theta_T) - F(\theta_0)\big) \right| \le \sum_{t=0}^{T-1} |r_t| \le \frac{\beta}{2} \sum_{t=0}^{T-1} \|\delta_t\|^2,$$

which proves eq. (7). If the smoothness constant varies across steps, the same argument gives the bound $\frac{1}{2}\sum_{t=0}^{T-1} \beta_t \|\delta_t\|^2$. □

C.2 Proof of Theorem 5.2

Proof. By the definition of the log-sum-exp objective, we have

$$F_{\mathrm{wg}}(\theta) - L_{\mathrm{wg}}(\theta) = \frac{1}{\tau} \log \sum_{g=1}^{G} e^{\tau(L_g(\theta) - L_{\mathrm{wg}}(\theta))}.$$

At least one term in the sum equals one, and all terms are at most one. Therefore, the sum lies between 1 and G. Hence, writing $e(\theta) := F_{\mathrm{wg}}(\theta) - L_{\mathrm{wg}}(\theta)$, we have $0 \le e(\theta) \le \log G/\tau$. It follows that

$$\big|\big(F_{\mathrm{wg}}(\theta_T) - F_{\mathrm{wg}}(\theta_0)\big) - \big(L_{\mathrm{wg}}(\theta_T) - L_{\mathrm{wg}}(\theta_0)\big)\big| = |e(\theta_T) - e(\theta_0)| \le \frac{\log G}{\tau}.$$

The smoothing error in the objective change is $e(\theta_T) - e(\theta_0)$. Since both errors lie between 0 and $\log G/\tau$, their difference has absolute value at most $\log G/\tau$.

Combining this inequality with Theorem 5.1 gives

$$\begin{aligned}
&\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(L_{\mathrm{wg}}(\theta_T) - L_{\mathrm{wg}}(\theta_0)\big) \right| \\
&\le \left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(F_{\mathrm{wg}}(\theta_T) - F_{\mathrm{wg}}(\theta_0)\big) \right| + |e(\theta_T) - e(\theta_0)| \\
&\le \frac{\beta}{2} \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2 + \frac{\log G}{\tau},
\end{aligned}$$

which proves eq. (8). □

For binary cross-entropy with natural logarithms and $\hat{y}_\theta(x) = \mathbb{1}[r_\theta(x) \ge 0]$, every misclassified example has loss at least $\log 2$. Therefore,

$$\mathbb{1}[\hat{y}_\theta(x) \ne y] \le \frac{\ell(z;\theta)}{\log 2}.$$

Averaging within each validation group and taking the maximum over groups gives

$$1 - \mathrm{WGA}(\theta) \le \frac{L_{\mathrm{wg}}(\theta)}{\log 2}.$$

Thus, reducing worst-group risk tightens an upper bound on worst-group classification error. The fidelity theorem connects the summed attribution scores to changes in this risk objective. The actual change in accuracy after data repair is assessed via retraining empirical evaluation.

C.3 Soft–Hard Parity Bounds

We first bound the difference between group statistics computed using predicted probabilities and hard predictions, then combine this result with Theorem 5.1.

Proposition C.1 (Soft and hard-prediction parity) *For each nonempty validation group $\mathcal{D}_g^{\mathrm{val}}$, define*

$$\mathrm{PPR}_g(\theta) := \frac{1}{|\mathcal{D}_g^{\mathrm{val}}|} \sum_{z^{\mathrm{val}} \in \mathcal{D}_g^{\mathrm{val}}} \hat{y}_\theta(x),$$

$$\epsilon_g(\theta) := \frac{1}{|\mathcal{D}_g^{\mathrm{val}}|} \sum_{z^{\mathrm{val}} \in \mathcal{D}_g^{\mathrm{val}}} \sigma(-|r_\theta(x)|).$$

For the soft rates C_g defined in Section 4.2, $|C_g(\theta) - \mathrm{PPR}_g(\theta)| \le \epsilon_g(\theta)$.

Suppressing the argument θ, let $\mathrm{PPR} = (\mathrm{PPR}_1, \dots, \mathrm{PPR}_G)$ and $\mathrm{gap}(v) := \max_g v_g - \min_g v_g$. Then

$$||C_g - C_h| - |\mathrm{PPR}_g - \mathrm{PPR}_h|| \le \epsilon_g + \epsilon_h,$$

$$|\mathrm{Var}(C) - \mathrm{Var}(\mathrm{PPR})| \le \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \epsilon_g,$$

$$|\,\mathrm{gap}(C) - \mathrm{gap}(\mathrm{PPR})| \le 2 \max_g \epsilon_g.$$

The same bounds hold on the positive- and negative-label subgroups, provided the relevant subgroups are nonempty. For equalized odds, the variance error is bounded by the sum of the positive- and negative-label variance bounds.

Proof. For any real logit r, including $r = 0$, we have

$$|\sigma(r) - \mathbb{1}[r \ge 0]| = \sigma(-|r|).$$

By averaging over $\mathcal{D}_g^{\mathrm{val}}$ and applying the triangle inequality, we get

$$|C_g(\theta) - \mathrm{PPR}_g(\theta)| \le \epsilon_g(\theta).$$

Then, the reverse triangle inequality gives

$$||C_g - C_h| - |\mathrm{PPR}_g - \mathrm{PPR}_h|| \le |(C_g - C_h) - (\mathrm{PPR}_g - \mathrm{PPR}_h)| \le \epsilon_g + \epsilon_h.$$

For any $v \in \mathbb{R}^G$,

$$\mathrm{Var}(v) = \frac{1}{G} \sum_{g=1}^{G} (v_g - \bar{v})^2 = \frac{1}{G^2} \sum_{g<h} (v_g - v_h)^2.$$

Since all soft and hard rates lie in $[0, 1]$, we can obtain

$$\begin{aligned} \left|(C_g - C_h)^2 - (\mathrm{PPR}_g - \mathrm{PPR}_h)^2\right| &\le 2|(C_g - C_h) - (\mathrm{PPR}_g - \mathrm{PPR}_h)| \\ &\le 2(\epsilon_g + \epsilon_h). \end{aligned}$$

Each group appears in $G - 1$ pairs, so each ϵ_g appears $G - 1$ times in the sum. Therefore,

$$|\mathrm{Var}(C) - \mathrm{Var}(\mathrm{PPR})| \le \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \epsilon_g.$$

The maximum and minimum are both 1-Lipschitz under the infinity norm. Thus,

$$|\operatorname{gap}(C) - \operatorname{gap}(\mathrm{PPR})| \le 2\|C - \mathrm{PPR}\|_\infty \le 2 \max_g \epsilon_g.$$

Repeating the argument using only positive-label validation examples gives the bound for equal opportunity. For equalized odds, we also apply the argument to negative-label validation examples and add the two bounds. All required validation subgroups are assumed nonempty. □

C.4 Proof of Theorem 5.3

We define the corresponding fairness objectives using hard predictions. Let PPR^+ and PPR^- denote the vectors of group true positive rates and false positive rates, respectively. Then

$$\begin{aligned} H_{\mathrm{dp}}(\theta) &:= \mathrm{Var}(\mathrm{PPR}(\theta)), \\ H_{\mathrm{eo}}(\theta) &:= \mathrm{Var}(\mathrm{PPR}^+(\theta)), \\ H_{\mathrm{eodds}}(\theta) &:= \mathrm{Var}(\mathrm{PPR}^+(\theta)) + \mathrm{Var}(\mathrm{PPR}^-(\theta)). \end{aligned}$$

These objectives use the same variance formula as F, but compute group statistics from binary predictions rather than predicted probabilities.

For nonempty class-conditional subgroups, define

$$\epsilon_g^{\pm}(\theta) := \frac{1}{|\mathcal{D}^{\mathrm{val}}{}_g^{\pm}|} \sum_{z^{\mathrm{val}} \in \mathcal{D}^{\mathrm{val}}{}_g^{\pm}} \sigma(-|r_\theta(x)|).$$

The corresponding soft–hard error bounds are

$$\begin{aligned} B_{\mathrm{dp}}(\theta) &:= \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \epsilon_g(\theta), \\ B_{\mathrm{eo}}(\theta) &:= \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \epsilon_g^+(\theta), \\ B_{\mathrm{eodds}}(\theta) &:= \frac{2(G-1)}{G^2} \sum_{g=1}^{G} \big(\epsilon_g^+(\theta) + \epsilon_g^-(\theta)\big). \end{aligned}$$

Proof. For each selected criterion, Proposition C.1 gives $|F(\theta) - H(\theta)| \le B(\theta)$. Applying this bound at the two endpoints yields

$$\begin{aligned} &\big|\big(F(\theta_T) - F(\theta_0)\big) - \big(H(\theta_T) - H(\theta_0)\big)\big| \\ &\le |F(\theta_T) - H(\theta_T)| + |F(\theta_0) - H(\theta_0)| \\ &\le B(\theta_T) + B(\theta_0). \end{aligned}$$

Combining this inequality with Theorem 5.1 gives

$$\begin{aligned} &\left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(H(\theta_T) - H(\theta_0)\big) \right| \\ &\le \left| \sum_{z_i \in \mathcal{D}^{\mathrm{tr}}} \phi_{z_i}^{\mathrm{fair}} - \big(F(\theta_T) - F(\theta_0)\big) \right| + B(\theta_T) + B(\theta_0) \\ &\le \frac{\beta}{2} \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2 + B(\theta_T) + B(\theta_0), \end{aligned}$$

which proves eq. (9). □

The difference between predicted probabilities and binary predictions becomes smaller as probabilities approach 0 or 1. Specifically, if every relevant validation example satisfies $|r_\theta(x)| \ge \kappa$ at the start or end of training, then $\epsilon_g(\theta) \le \sigma(-\kappa)$. The same bound applies when statistics are computed separately for positive and negative labels. Thus, larger absolute logits at both stages give a smaller bound on the discrepancy between F and H. The first-order attribution error is bounded separately along the training trajectory.

C.5 Local Shapley Approximation

Theorem C.2 (Local Shapley approximation) *Let $\phi_{z_i,t}^{\text{exact}} := \phi_{z_i}(U_F^t)$ denote the exact Shapley value of z_i for the utility in eq.* (3)*, without a Taylor approximation. Assume that F is β_t-smooth on a convex region containing all parameters obtained by updating θ_t with any subset of $\mathcal{B}_t$. Then, for every $z_i \in \mathcal{B}_t$,*

$$\left|\phi_{z_i,t}^{\text{exact}} + \eta_t \left\langle \nabla F(\theta_t), \nabla_\theta \ell(z_i;\theta_t)\right\rangle\right| \le \frac{\beta_t \eta_t^2}{2} \|\nabla_\theta \ell(z_i;\theta_t)\| \sum_{z_j \in \mathcal{B}_t} \|\nabla_\theta \ell(z_j;\theta_t)\|. \tag{10}$$

Proof. Consider a training step t and an example $z_i \in \mathcal{B}_t$. The parameter update from z_i alone is $\delta_i := -\eta_t \nabla_\theta \ell(z_i;\theta_t)$. For a subset $S \subseteq \mathcal{B}_t \setminus \{z_i\}$, the combined update from its examples is $\delta_S := -\eta_t \sum_{z_j \in S} \nabla_\theta \ell(z_j;\theta_t)$. Define the exact marginal

$$m_i(S) := F(\theta_t + \delta_S + \delta_i) - F(\theta_t + \delta_S).$$

Both $\theta_t + \delta_S$ and $\theta_t + \delta_S + \delta_i$ lie in the assumed convex region, so every point between them also lies in it. Since θ_t is in this region as well, the smoothness bound applies when comparing the gradient at $\theta_t + \delta_S + u\delta_i$ with that at θ_t. Hence,

$$\begin{aligned} |m_i(S) - \langle v_t, \delta_i\rangle| &= \left| \int_0^1 \langle \nabla F(\theta_t + \delta_S + u\delta_i) - v_t, \delta_i\rangle \, du \right| \\ &\le \beta_t \|\delta_i\| \int_0^1 \|\delta_S + u\delta_i\| \, du \\ &\le \beta_t \|\delta_i\| \|\delta_S\| + \frac{\beta_t}{2}\|\delta_i\|^2. \end{aligned}$$

Randomly order the examples in $\mathcal{B}_t$, with all orderings equally likely, and let S contain the examples appearing before z_i. Averaging the contribution of adding z_i to S over these orderings gives its exact local Shapley value. Each other example has probability $1/2$ of appearing before z_i and therefore belonging to S, so

$$\mathbb{E}\|\delta_S\| \le \eta_t \sum_{z_j \in \mathcal{B}_t \setminus \{z_i\}} \Pr(z_j \in S)\|u_{j,t}\| = \frac{\eta_t}{2} \sum_{z_j \in \mathcal{B}_t \setminus \{z_i\}} \|u_{j,t}\|.$$

Taking expectations of the marginal error, and using $\phi_{z_i,t}^{\text{exact}} = \mathbb{E} m_i(S)$, this yields

$$\left|\phi_{z_i,t}^{\text{exact}} + \eta_t \langle v_t, u_{i,t}\rangle\right| \le \frac{\beta_t \eta_t^2}{2} \|u_{i,t}\| \sum_{z_j \in \mathcal{B}_t} \|u_{j,t}\|,$$

which is eq. (10). The argument also covers a singleton batch. □

Exact accounting and accumulated individual error. Set $\phi_{z_i,t}^{\text{exact}} = 0$ for examples not in $\mathcal{B}_t$. A uniform permutation of $\mathcal{D}^{\text{tr}}$ induces a uniform relative order on $\mathcal{B}_t$. Thus adding inactive null players does not change the local values. Shapley linearity gives

$$\phi_{z_i}(U_F) = \sum_{t=0}^{T-1} \phi_{z_i,t}^{\text{exact}}.$$

Since $U_F^t(\varnothing) = 0$ and $U_F^t(\mathcal{B}_t) = F(\theta_{t+1}) - F(\theta_t)$, exact efficiency gives

$$\sum_{z_i \in \mathcal{D}^{\text{tr}}} \phi_{z_i}(U_F) = F(\theta_T) - F(\theta_0).$$

Applying the local error bound and summing over the steps in which z_i appears gives

$$|\phi_{z_i}(U_F) - \phi_{z_i}^{\text{fair}}| \le E_i, \quad E_i := \frac{1}{2} \sum_{t: z_i \in \mathcal{B}_t} \beta_t \eta_t^2 \|u_{i,t}\| \sum_{z_j \in \mathcal{B}_t} \|u_{j,t}\|.$$

The exact in-run score of z_i lies within E_i of $\phi_{z_i}^{\text{fair}}$. Its sign is therefore guaranteed to match the computed score when $|\phi_{z_i}^{\text{fair}}| > E_i$. Similarly, if $\phi_{z_i}^{\text{fair}} - \phi_{z_j}^{\text{fair}} > E_i + E_j$, then the exact score of z_i is also larger than that of z_j. These guarantees apply to attribution along the observed training run and they do not establish whether a label is corrupted or whether removing an example improves the retrained model.

C.6 Curvature and Temperature

If $\|\nabla^2 L_g\|_{\mathrm{op}} \le \beta_0$ and $\|\nabla L_g\| \le M$ on the relevant convex regions, then

$$\|\nabla^2 F_{\mathrm{wg}}\|_{\mathrm{op}} \le \beta_0 + \tau M^2. \tag{11}$$

Corollary 1 (Choosing τ) *For a fixed trajectory, write $S_\Delta = \sum_{t=0}^{T-1} \|\Delta_t(\mathcal{B}_t)\|^2$. Suppose the smoothness constants satisfy $\beta(\tau) \le \beta_0 + c\tau$ for all $\tau > 0$, with $\beta_0 \ge 0$ and $c > 0$ independent of τ (eg., $c = M^2$ by eq. (11)). If $G > 1$ and $S_\Delta > 0$, the upper bound $\frac{1}{2}(\beta_0 + c\tau)S_\Delta + \log G/\tau$ is minimized at*

$$\tau^\star = \sqrt{\frac{2\log G}{cS_\Delta}}, \qquad \min_{\tau>0}\left\{\frac{(\beta_0 + c\tau)S_\Delta}{2} + \frac{\log G}{\tau}\right\} = \frac{\beta_0 S_\Delta}{2} + \sqrt{2cS_\Delta \log G}.$$

Differentiating the softmax mixture gives

$$\nabla^2 F_{\mathrm{wg}} = \sum_g q_g \nabla^2 L_g + \tau\left(\sum_g q_g \nabla L_g(\theta)\nabla L_g(\theta)^\top - \nabla F_{\mathrm{wg}}(\theta)\nabla F_{\mathrm{wg}}(\theta)^\top\right).$$

For any unit vector u, the quadratic form of the matrix in parentheses is $\mathrm{Var}_{g\sim q}(u^\top \nabla L_g(\theta)) \le \sum_g q_g (u^\top \nabla L_g(\theta))^2 \le M^2$. That matrix is positive semidefinite, so its operator norm is at most M^2. The triangle inequality and the group Hessian bound give eq. (11), where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. On a convex region, the Hessian norm bound implies the stated gradient Lipschitz bound.

Proof. For Corollary 1, the part of the upper bound that depends on τ is $h(\tau) = cS_\Delta\tau/2 + \log G/\tau$. Under $G > 1$, $S_\Delta > 0$, and $c > 0$,

$$h'(\tau) = \frac{cS_\Delta}{2} - \frac{\log G}{\tau^2}, \qquad h''(\tau) = \frac{2\log G}{\tau^3} > 0.$$

Solving $h'(\tau) = 0$ gives $\tau^\star = \sqrt{2\log G/(cS_\Delta)}$. Since $h''(\tau) > 0$, this is the unique minimizer. Substituting $\tau^\star$ into h gives $h(\tau^\star) = \sqrt{2cS_\Delta \log G}$. If $G = 1$, there is no smoothing error. If $G > 1$ and $S_\Delta = 0$, the bound decreases toward zero as τ increases, so there is no finite minimizer. This choice of τ minimizes the error bound for a fixed training trajectory, and it does not necessarily give the best detection or repair performance. □

C.7 Ranking Stability

Proposition C.3 (Ranking stability under worst-group smoothing) *Let $G > 1$, $\|\nabla L_g(\theta_t)\| \le M$, and choose $g_t^* \in \arg\max_g L_g(\theta_t)$. Let γ_t be the difference between the largest and second-largest group risks, and define*

$$\widetilde{\phi}_{z_i} := -\sum_{t: z_i \in \mathcal{B}_t} \eta_t \left\langle \nabla L_{g_t^*}(\theta_t), \nabla_\theta \ell(z_i; \theta_t)\right\rangle.$$

Then, for the attribution scores using F_{wg},

$$\begin{aligned} &|\phi_{z_i}^{\mathrm{fair}} - \widetilde{\phi}_{z_i}| \le D_i(\tau), \\ &\qquad D_i(\tau) := 2M \sum_{t: z_i \in \mathcal{B}_t} \eta_t \|\nabla_\theta \ell(z_i; \theta_t)\| \min\{1, (G-1)e^{-\tau\gamma_t}\}. \end{aligned} \tag{12}$$

If $\widetilde{\phi}_{z_i} - \widetilde{\phi}_{z_j} > D_i(\tau) + D_j(\tau)$, then $\phi_{z_i}^{\mathrm{fair}} > \phi_{z_j}^{\mathrm{fair}}$.

Proof. Fix a step t and write $q_g := q_g(\theta_t)$ for simplicity. Since $\sum_g q_g = 1$,

$$\nabla F_{\mathrm{wg}}(\theta_t) - \nabla L_{g_t^*}(\theta_t) = \sum_{g \ne g_t^*} q_g \left(\nabla L_g(\theta_t) - \nabla L_{g_t^*}(\theta_t)\right).$$

The gradient norm bound implies

$$\left\|\nabla F_{\mathrm{wg}}(\theta_t) - \nabla L_{g_t^*}(\theta_t)\right\| \le 2M(1 - q_{g_t^*}).$$

Let

$$A_t := \sum_{g \neq g_t^*} \exp\left[-\tau\left(L_{g_t^*}(\theta_t) - L_g(\theta_t)\right)\right].$$

Since g_t^* is a maximizer, every risk gap in this sum is at least $\gamma_t \geq 0$, with $\gamma_t = 0$ in the case of a tie. Consequently,

$$1 - q_{g_t^*} = \frac{A_t}{1 + A_t} \leq \min\{1, A_t\} \leq \min\{1, (G-1)e^{-\tau\gamma_t}\}.$$

Subtracting the two score definitions, applying Cauchy–Schwarz at each sampled step, and summing proves eq. (12). Finally,

$$\phi_{z_i}^{\text{fair}} - \phi_{z_j}^{\text{fair}} \geq \widetilde{\phi}_{z_i} - \widetilde{\phi}_{z_j} - D_i(\tau) - D_j(\tau) > 0$$

under the stated condition. This proves ranking stability. □

D Additional Experiments

D.1 Synthetic Data

To evaluate the effectiveness of FAIR-DS, we construct a two-dimensional synthetic classification task with known ground-truth labels. Each example has a binary sensitive attribute $s \sim \text{Bernoulli}(0.3)$. For this synthetic task, we encode s as 0 and 1 instead of using $[G]$. We draw a latent qualification $u \sim \mathcal{N}(\mu_s, 1)$, with $\mu_0 = 0$ and $\mu_1 = -0.5$, and set $y = \mathbb{1}[u + \xi_y > 0]$, where $\xi_y \sim \mathcal{N}(0, 0.3^2)$. The model receives two features: $x_1 = u + \xi_1$ and $x_2 = 1.5(2s - 1) + \xi_2$, where $\xi_1 \sim \mathcal{N}(0, 0.5^2)$ and $\xi_2 \sim \mathcal{N}(0, 1)$ are independent noise terms. The model does not receive s directly. The feature x_2 provides a group proxy, while the group-dependent qualification distributions also make x_1 informative about group membership.

We generate 8,000 examples and split them into 5,000 training, 1,500 validation, and 1,500 test examples. Validation and test labels remain clean, and label bias is injected only into the training set. For *B* flips, we change the labels of qualified minority examples ($s = 1, y = 1$) to negative with probability 0.35. For *A* flips, we change the labels of unqualified majority examples ($s = 0, y = 0$) to positive with probability 0.15. We then independently flip each remaining unflipped label with probability 0.05, regardless of group membership, and denote these additional random flips by *R*. The resulting training set contains 233 *A* flips, 133 *B* flips, 243 *R* flips, and 4,391 clean examples. We compute loss attribution scores $\phi_{z_i}^{\text{loss}}$ using the average validation loss and fairness attribution scores using the four corresponding group-aware objectives.

For the synthetic experiments, we use an MLP with a single hidden layer of 128 units and ReLU activation and trained with SGD for 65 epochs, with learning rate 0.005, batch size 128, and zero momentum and weight decay. The worst-group attribution plots use $\tau = 5$ and demographic groups defined by s. The quadrant, residual, decision-boundary, trajectory, and validation-cell plots use training seed 0. The deletion curves use five paired training seeds on the same generated dataset, split, and corruption mask. Within each seed, retraining uses the same initialization and epoch ordering, with removed examples filtered out. All models are trained for the same number of epochs, so removal reduces the number of updates. For the deletion curves, the equalized-odds gap is the maximum of the TPR and FPR gaps.

D.2 Experimental Setup

For Adult Census, we use an MLP with a single hidden layer of 128 units and ReLU activation. For UTKFace and FairFace, we use a pretrained ResNet-50 with an MLP classifier, keeping the ResNet-50 parameters fixed. All images are resized to 224×224. For the image datasets, θ denotes the trainable classifier parameters, and all attribution gradients are taken with respect to them. For validation, we sample 500 clean examples from the original Adult Census test split and 700 from the official FairFace validation split. For UTKFace, we first split the data into 80% training and 20% test data, then sample 700 clean examples from the test split for validation. These validation sets remain fixed during training and are excluded from test evaluation. These validation sizes apply to the main experiments. For the validation-size ablation, we sample subsets of different sizes from

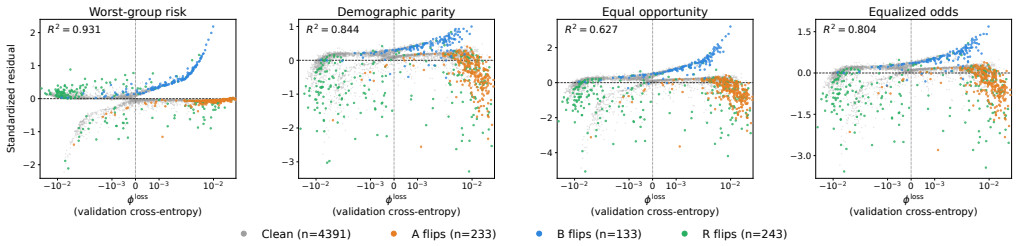


Figure 3: Residuals after removing a linear scaling of loss attribution. Residuals are computed as $\phi_{z_i}^{\text{fair}} - \hat{b}\phi_{z_i}^{\text{loss}}$ and divided by the standard deviation of the fairness scores, where $\hat{b}$ is the fitted coefficient relating loss attribution to fairness attribution. Positive residuals indicate fairness scores above the fitted linear relation.

the same sources: the test splits of Adult Census and UTKFace, and the official validation split of FairFace, excluding the selected examples from test evaluation. Across all datasets, we use SGD with batch size 128, zero momentum, and zero weight decay, and repeat each experiment over 10 random seeds. We train for 65 epochs with learning rate 0.005 on Adult Census, 73 epochs with learning rate 0.08 on UTKFace, and 70 epochs with learning rate 0.0135 on FairFace. It is worth noting that, for all detection and repair comparisons, we orient scores so that larger values indicate greater estimated harm, negating scores from baselines where larger values indicate benefit.

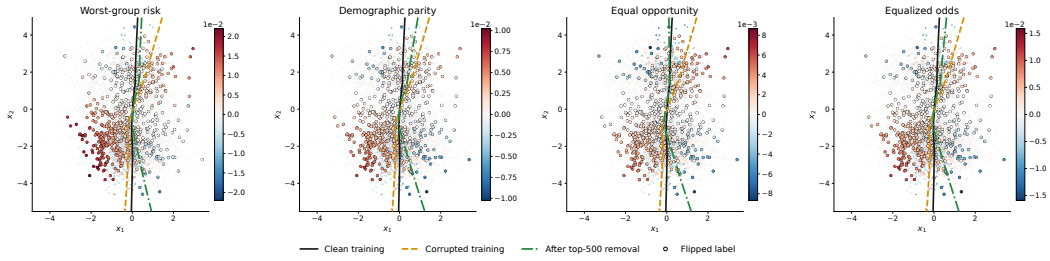
D.3 Analysis of fairness attribution

Fairness attribution beyond loss attribution. To examine whether fairness attribution simply rescales loss attribution, we fit $\phi_{z_i}^{\text{fair}} \approx b\phi_{z_i}^{\text{loss}}$ by least squares without an intercept, using all training examples and no corruption labels. Figure 3 plots the residuals, normalized by the standard deviation of the corresponding fairness scores. Across the four objectives, 92.5%–100% of *B* flips retain positive residuals, showing that their fairness scores are higher than a linear scaling of loss attribution would predict. Under equal opportunity, *B* flips have a mean standardized residual of 1.00, compared with -0.51 for *A* flips and -0.80 for *R* flips. Thus, fairness attribution assigns different relative importance to these corruption types that a single rescaling of loss attribution cannot capture. Residual signs indicate deviations from the fitted relation rather than harm, and nonlinear dependence on loss attribution may remain.

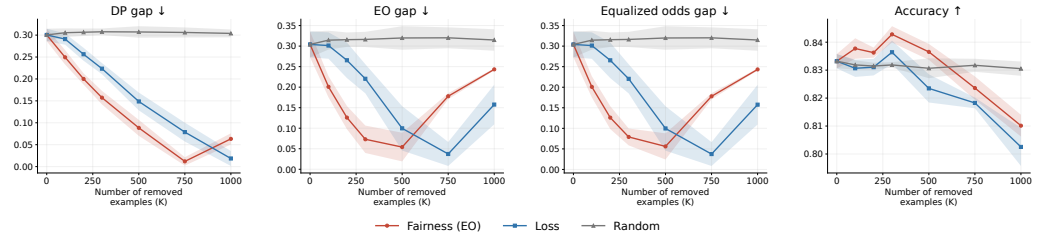
Effects of data removal. Figure 4a shows that removing the top-500 examples reduces DP, TPR, and FPR gaps under all four attribution objectives, although the decision boundaries do not fully recover the clean boundary. Compared with label correction, removal yields smaller fairness gaps but greater prediction disagreement with the clean model in this experiment. Across five runs, Figure 4b further shows that EO-based attribution reduces disparities faster than loss-based or random removal at small budgets. At $K = 500$, it achieves a mean EO gap of 0.054 versus 0.100 for loss attribution, with higher accuracy (0.837 versus 0.823). Further removal increases the EO gap and reduces accuracy, highlighting the importance of the removal budget.

Contributions along training. Figure 5 shows that *A* and *B* flips accumulate positive contributions as training proceeds, while clean examples have much smaller mean scores. Their relative contributions depend on the objective: *B* flips have a larger mean score under equal opportunity, whereas *A* flips contribute more under the other three objectives. *R* flips have negative mean scores under the parity objectives but positive mean scores under worst-group risk. These trajectories show how the effects of different corruption types accumulate throughout training, providing information that a final attribution score alone does not reveal.

Contributions from validation subgroups. The decomposition shows that a training example can contribute in opposite directions through different validation subgroups. Its final score combines these contributions according to the selected fairness objective. Equal opportunity uses only the positive-label validation cells, while equalized odds adds contributions from the negative-label cells.



(a) Decision boundaries after top-500 removal.



(b) Test performance across removal budgets.

Figure 4: Effects of data removal. **(a)** Colors show attribution scores; black outlines mark label flips. Black, orange, and green boundaries correspond to clean training, corrupted training, and retraining after top-500 removal, respectively. **(b)** Deletion curves compare EO-based attribution, loss attribution, and random removal. Shaded bands show one standard deviation over five runs.

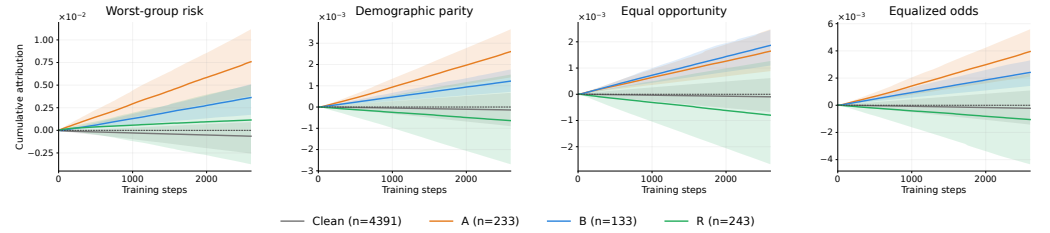


Figure 5: Cumulative contributions along training. Lines show the mean cumulative attribution per example for each corruption type in a single training run. Shaded bands show the 25th–75th percentiles within each type. Scores are recorded at the end of each epoch. Positive values indicate accumulated contributions that increase the selected objective.

This makes the source of each score explicit and helps explain why the same training example can receive different scores under different fairness criteria.

D.4 Additional Comparison Results

Detection at small inspection budgets. Table 4 reports detection results at smaller inspection budgets. Because some baselines are computationally expensive (e.g., those requiring retraining), these results are computed on a random subset of 5,000 training examples rather than the full training set used in Table 2. The two tables are therefore not directly comparable. On UTKFace and FairFace, FAIR-DS achieves the highest mean precision at $K = 100$, 300, and 500. Its Precision@100 is 94.7% on both datasets, compared with 92.9% and 87.0% for the strongest baselines, respectively. On Adult Census, In-Run Shapley achieves slightly higher AUROC and Precision@500, while FAIR-DS achieves higher AP and Precision@300. On UTKFace, KNN-Shapley achieves the highest AUROC but lower top-K precision, showing that better overall separation does not necessarily identify more corrupted examples within a small inspection budget.

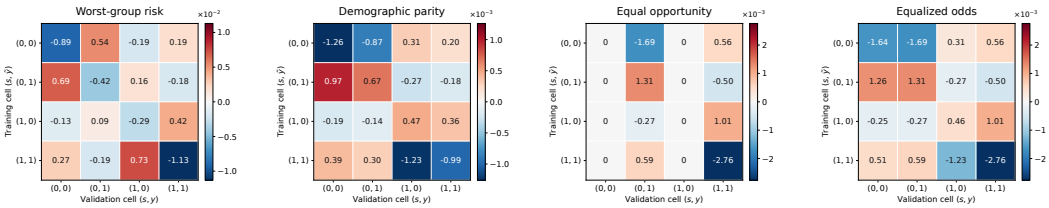


Figure 6: Attribution decomposed by validation subgroup. Rows group training examples by sensitive attribute and observed training label $(s, \tilde{y})$; columns group validation examples by sensitive attribute and clean label (s, y). Each entry is the mean cumulative contribution from the corresponding validation cell, including the weights of the selected fairness objective. Red indicates positive contributions and blue indicates negative contributions. Summing each row recovers the mean attribution score of that training group. Color scales differ across panels.

Detection under parity objectives. Table 5 shows that the parity objectives concentrate corrupted examples near the top of the ranking. Precision@100 reaches 96.4% for DP, 94.7% for EO, and 96.9% for equalized odds, the highest means within their respective comparisons. All three variants also have the highest mean Precision@500, although the margin is small for EO. This advantage does not extend to the full ranking: In-Run Shapley achieves higher AUROC, AP, and precision at $K = 1000$ and 2000. In particular, the EO variant has an AUROC of 0.346 despite its high Precision@100. Thus, the scores can identify a small set of corrupted examples without reliably separating all corrupted examples from clean ones. This is consistent with their purpose: the scores measure the harm to the selected fairness objective, so biased examples that do not increase this objective may receive low scores.

Comparison of downstream repair. Table 6 compares label correction under the same inspection budget. The worst-group variant corrects the most labels, but FairIF and Influence Function achieve smaller fairness gaps. The DP variant achieves the lowest mean EO gap in its comparison (0.041), while FairIF achieves a smaller DP gap (0.093 versus 0.105). Under EO, our method and FairIF have similar mean EO gaps (0.050 and 0.049). Under equalized odds, our method achieves the lowest mean DP and EO gaps among the budgeted methods (0.087 and 0.038), while its accuracy is lower than Influence Function (0.761 versus 0.769). These results show that no method is best on every metric. Correcting more labels does not necessarily reduce group disparities more; the selected examples and the attribution objective both matter.

Table 4: Detection of injected label bias at small inspection budgets ($K \in \{100, 300, 500\}$). Higher is better for all metrics. FAIR-DS uses an LSE approximation to worst-group risk with $\tau = 0.01$. Results in this table are computed on a random subset of 5,000 training examples, whereas Table 2 uses the full training set. Results are reported as mean±std over 10 seeds. The best results are **bold**, and the second-best results are underlined.

Method	AUROC	AP	Precision@100	Precision@300	Precision@500
Adult Census					
Random	0.503±0.010	0.199±0.006	0.192±0.019	0.198±0.013	0.197±0.011
KNN-Shapley	0.877±0.003	0.644±0.012	0.824±0.034	0.752±0.010	0.725±0.016
Train Loss	0.840±0.005	0.562±0.015	0.732±0.036	0.679±0.027	0.691±0.023
Gradient norm	0.826±0.005	0.456±0.019	0.394±0.046	0.488±0.035	0.506±0.031
FairShap	0.461±0.007	0.201±0.004	0.268±0.032	0.277±0.015	0.278±0.013
FairIF	0.588±0.007	0.368±0.008	0.681±0.055	0.572±0.015	0.513±0.016
IFFair	0.651±0.006	0.531±0.019	**0.892±0.029**	0.809±0.023	0.773±0.026
Influence Function	0.863±0.065	0.665±0.083	0.842±0.030	0.805±0.060	0.765±0.072
In-Run Shapley (1)	**0.889±0.005**	0.722±0.019	0.851±0.044	0.844±0.037	**0.813±0.020**
In-Run Shapley (2)	**0.889±0.005**	0.722±0.019	0.852±0.045	0.845±0.038	**0.813±0.020**
FAIR-DS	0.888±0.005	**0.726±0.017**	0.874±0.043	**0.852±0.039**	0.812±0.019
UTKFace					
Random	0.539±0.007	0.223±0.009	0.232±0.037	0.178±0.014	0.190±0.012
KNN-Shapley	**0.864±0.003**	0.624±0.008	0.811±0.019	0.759±0.012	0.706±0.009
Train Loss	0.737±0.085	0.417±0.111	0.547±0.175	0.535±0.166	0.502±0.151
Gradient norm	0.729±0.082	0.371±0.090	0.430±0.122	0.424±0.110	0.424±0.106
FairShap	0.441±0.009	0.176±0.005	0.166±0.036	0.183±0.013	0.176±0.010
FairIF	0.588±0.068	0.335±0.078	0.508±0.153	0.497±0.122	0.462±0.107
IFFair	0.685±0.067	0.323±0.071	0.365±0.130	0.381±0.119	0.373±0.102
Influence Function	0.584±0.090	0.341±0.077	0.531±0.153	0.512±0.106	0.471±0.092
In-Run Shapley (1)	0.813±0.013	0.666±0.023	0.929±0.020	0.880±0.016	0.816±0.025
In-Run Shapley (2)	0.845±0.012	0.675±0.023	0.906±0.025	0.848±0.022	0.794±0.024
FAIR-DS	0.839±0.010	**0.712±0.019**	**0.947±0.019**	**0.908±0.020**	**0.860±0.025**
FairFace					
Random	0.498±0.009	0.196±0.008	0.188±0.061	0.193±0.036	0.199±0.022
KNN-Shapley	0.719±0.003	0.376±0.010	0.514±0.042	0.459±0.026	0.433±0.015
Train Loss	0.740±0.052	0.397±0.078	0.548±0.083	0.480±0.113	0.454±0.121
Gradient norm	0.731±0.048	0.363±0.052	0.508±0.038	0.410±0.043	0.383±0.078
FairShap	0.548±0.004	0.216±0.004	0.217±0.028	0.223±0.017	0.209±0.015
IFFair	0.658±0.020	0.299±0.034	0.378±0.103	0.342±0.087	0.337±0.069
Influence Function	0.568±0.036	0.298±0.042	0.481±0.117	0.436±0.086	0.409±0.066
In-Run Shapley (1)	0.753±0.007	0.583±0.012	0.870±0.026	0.818±0.021	0.742±0.011
In-Run Shapley (2)	0.750±0.007	0.569±0.013	0.846±0.034	0.795±0.021	0.722±0.010
FAIR-DS	**0.792±0.012**	**0.632±0.012**	**0.947±0.018**	**0.867±0.008**	**0.783±0.015**

Table 5: Detection on FairFace under three group fairness objectives: *Demographic Parity*, *Equal Opportunity*, and *Equalized Odds*. We evaluate how well each method's ranking recovers the flipped-label examples; higher is better for all metrics. FairShap uses the official EO-oriented implementation and is not re-optimized for each attribution objective. Results are mean±std over 10 random seeds, with the best in **bold** and the second best underlined.

Method	AUROC	AP	Precision@100	Precision@500	Precision@1000	Precision@2000
Demographic parity						
Random	0.499±0.002	0.191±0.001	0.177±0.028	0.185±0.020	0.191±0.012	0.190±0.007
KNN-Shapley	0.728±0.002	0.377±0.003	0.695±0.028	0.590±0.018	0.553±0.009	0.514±0.010
Train Loss	0.741±0.051	0.386±0.076	0.750±0.093	0.635±0.033	0.582±0.043	0.523±0.076
Gradient norm	0.732±0.047	0.352±0.053	0.593±0.077	0.542±0.071	0.506±0.048	0.462±0.031
FairShap	0.545±0.001	0.214±0.002	0.467±0.037	0.358±0.018	0.305±0.014	0.273±0.007
FairIF	0.535±0.040	0.264±0.022	0.539±0.072	0.465±0.051	0.447±0.043	0.420±0.035
IFFair	0.662±0.017	0.298±0.023	0.492±0.063	0.433±0.026	0.417±0.032	0.392±0.043
Influence Function	0.581±0.022	0.300±0.026	0.509±0.081	0.473±0.072	0.468±0.077	0.463±0.070
In-Run Shapley (1)	**0.756±0.003**	**0.582±0.007**	0.851±0.052	0.863±0.028	**0.869±0.020**	**0.862±0.015**
In-Run Shapley (2)	0.754±0.003	0.569±0.009	0.803±0.059	0.836±0.029	0.845±0.018	0.840±0.019
FAIR-DS	0.490±0.026	0.279±0.010	**0.964±0.015**	**0.904±0.013**	0.830±0.007	0.705±0.009
Equal opportunity						
Random	0.499±0.002	0.191±0.001	0.177±0.028	0.185±0.020	0.191±0.012	0.190±0.007
KNN-Shapley	0.728±0.002	0.377±0.003	0.695±0.028	0.590±0.018	0.553±0.009	0.514±0.010
Train Loss	0.741±0.051	0.386±0.076	0.750±0.093	0.635±0.033	0.582±0.043	0.523±0.076
Gradient norm	0.732±0.047	0.352±0.053	0.593±0.077	0.542±0.071	0.506±0.048	0.462±0.031
FairShap	0.545±0.001	0.214±0.002	0.467±0.037	0.358±0.018	0.305±0.014	0.273±0.007
FairIF	0.513±0.015	0.249±0.017	0.599±0.121	0.489±0.090	0.440±0.072	0.404±0.062
IFFair	0.659±0.014	0.295±0.020	0.500±0.116	0.447±0.050	0.409±0.038	0.380±0.045
Influence Function	0.581±0.022	0.300±0.026	0.509±0.081	0.473±0.072	0.468±0.077	0.463±0.070
In-Run Shapley (1)	**0.756±0.003**	**0.582±0.007**	0.851±0.052	0.863±0.028	**0.869±0.020**	**0.862±0.015**
In-Run Shapley (2)	0.754±0.003	0.569±0.009	0.803±0.059	0.836±0.029	0.845±0.018	0.840±0.019
FAIR-DS	0.346±0.006	0.229±0.002	**0.947±0.020**	**0.866±0.014**	0.797±0.009	0.661±0.007
Equalized odds						
Random	0.499±0.002	0.191±0.001	0.177±0.028	0.185±0.020	0.191±0.012	0.190±0.007
KNN-Shapley	0.728±0.002	0.377±0.003	0.695±0.028	0.590±0.018	0.553±0.009	0.514±0.010
Train Loss	0.741±0.051	0.386±0.076	0.750±0.093	0.635±0.033	0.582±0.043	0.523±0.076
Gradient norm	0.732±0.047	0.352±0.053	0.593±0.077	0.542±0.071	0.506±0.048	0.462±0.031
FairShap	0.545±0.001	0.214±0.002	0.467±0.037	0.358±0.018	0.305±0.014	0.273±0.007
FairIF	0.534±0.030	0.268±0.025	0.580±0.087	0.512±0.072	0.475±0.071	0.443±0.066
IFFair	0.695±0.029	0.324±0.036	0.508±0.127	0.473±0.076	0.447±0.055	0.423±0.050
Influence Function	0.581±0.022	0.300±0.026	0.509±0.081	0.473±0.072	0.468±0.077	0.463±0.070
In-Run Shapley (1)	**0.756±0.003**	**0.582±0.007**	0.851±0.052	0.863±0.028	**0.869±0.020**	**0.862±0.015**
In-Run Shapley (2)	0.754±0.003	0.569±0.009	0.803±0.059	0.836±0.029	0.845±0.018	0.840±0.019
FAIR-DS	0.544±0.041	0.308±0.020	**0.969±0.013**	**0.907±0.015**	0.834±0.015	0.709±0.012

Table 6: Downstream performance on FairFace after correcting the top-4000 examples selected by each method under four objectives (worst-group risk, DP, EO, EOdds). FairShap targets a single fairness criterion (equal opportunity in its official implementation), so we use the same FairShap scores for all attribution objectives. All samples are kept for retraining. Mean±std over 10 seeds; best in **bold**, second best underlined.

Method	True corrected	Precision	Errors left	TestAcc↑	WGA ↑	DP Gap ↓	EO Gap ↓
Worst-group risk							
No correction	0.0±0.0	/	16664.7±105.7	0.733±0.037	0.717±0.043	0.167±0.029	0.133±0.072
Random	755.1±19.3	0.189±0.005	15909.6±112.6	0.743±0.032	0.725±0.037	0.154±0.025	0.125±0.067
KNN-Shapley	2520.0±25.8	0.630±0.006	14144.7±96.7	0.751±0.026	0.736±0.027	0.136±0.020	0.091±0.048
Train Loss	1906.9±420.5	0.477±0.105	14757.8±424.4	0.741±0.025	0.727±0.030	0.152±0.020	0.107±0.054
Gradient norm	1648.7±153.3	0.412±0.038	15016.0±162.7	0.742±0.028	0.727±0.032	0.143±0.023	0.095±0.052
FairShap	941.7±36.7	0.235±0.009	15640.0±83.5	0.751±0.013	0.732±0.020	0.159±0.041	0.134±0.093
FairIF	2417.5±315.2	0.604±0.079	14247.2±370.8	0.767±0.028	0.752±0.032	**0.125±0.014**	**0.066±0.044**
IFFair	1470.3±289.0	0.368±0.072	15194.4±312.0	0.745±0.033	0.730±0.041	0.138±0.029	0.094±0.063
Influence Function	2379.4±557.6	0.595±0.139	14285.3±551.0	**0.769±0.034**	**0.753±0.044**	0.129±0.029	0.104±0.063
In-Run Shapley (1)	3228.8±55.5	0.807±0.014	13435.9±121.0	0.757±0.029	0.743±0.035	0.140±0.021	0.124±0.060
In-Run Shapley (2)	3140.6±61.0	0.785±0.015	13524.1±123.7	0.755±0.028	0.741±0.036	0.144±0.014	0.122±0.056
FAIR-DS	**3487.7±48.7**	**0.872±0.012**	**13177.0±124.4**	0.764±0.019	0.750±0.026	0.140±0.023	0.125±0.059
Oracle	16664.7±105.7	1.000±0.000	0.0±0.0	0.818±0.018	0.815±0.019	0.018±0.017	0.026±0.020
Demographic parity							
No correction	0.0±0.0	/	16664.7±105.7	0.733±0.037	0.717±0.043	0.167±0.029	0.133±0.072
Random	755.1±19.3	0.189±0.005	15909.6±112.6	0.743±0.032	0.725±0.037	0.154±0.025	0.125±0.067
KNN-Shapley	2520.0±25.8	0.630±0.006	14144.7±96.7	0.751±0.026	0.736±0.027	0.136±0.020	0.091±0.048
Train Loss	1906.9±420.5	0.477±0.105	14757.8±424.4	0.741±0.025	0.727±0.030	0.152±0.020	0.107±0.054
Gradient norm	1648.7±153.3	0.412±0.038	15016.0±162.7	0.742±0.028	0.727±0.032	0.143±0.023	0.095±0.052
FairShap	941.7±36.7	0.235±0.009	15640.0±83.5	0.751±0.013	0.732±0.020	0.159±0.041	0.134±0.093
FairIF	1951.1±374.0	0.488±0.094	14713.6±408.9	0.743±0.044	0.730±0.048	**0.093±0.033**	0.049±0.043
IFFair	1504.8±202.2	0.376±0.051	15159.9±231.5	0.740±0.035	0.727±0.040	0.134±0.023	0.079±0.046
Influence Function	2379.4±557.6	0.595±0.139	14285.3±551.0	**0.769±0.034**	**0.753±0.044**	0.129±0.029	0.104±0.063
In-Run Shapley (1)	**3228.8±55.5**	**0.807±0.014**	**13435.9±121.0**	0.757±0.029	0.743±0.035	0.140±0.021	0.124±0.060
In-Run Shapley (2)	3140.6±61.0	0.785±0.015	13524.1±123.7	0.755±0.028	0.741±0.036	0.144±0.014	0.122±0.056
FAIR-DS	2191.6±65.0	0.548±0.016	14473.1±140.1	0.747±0.035	0.734±0.040	0.105±0.024	**0.041±0.009**
Oracle	16664.7±105.7	1.000±0.000	0.0±0.0	0.818±0.018	0.815±0.019	0.018±0.017	0.026±0.020
Equal opportunity							
No correction	0.0±0.0	/	16664.7±105.7	0.733±0.037	0.717±0.043	0.167±0.029	0.133±0.072
Random	755.1±19.3	0.189±0.005	15909.6±112.6	0.743±0.032	0.725±0.037	0.154±0.025	0.125±0.067
KNN-Shapley	2520.0±25.8	0.630±0.006	14144.7±96.7	0.751±0.026	0.736±0.027	0.136±0.020	0.091±0.048
Train Loss	1906.9±420.5	0.477±0.105	14757.8±424.4	0.741±0.025	0.727±0.030	0.152±0.020	0.107±0.054
Gradient norm	1648.7±153.3	0.412±0.038	15016.0±162.7	0.742±0.028	0.727±0.032	0.143±0.023	0.095±0.052
FairShap	941.7±36.7	0.235±0.009	15640.0±83.5	0.751±0.013	0.732±0.020	0.159±0.041	0.134±0.093
FairIF	1830.2±229.9	0.458±0.058	14834.5±263.3	0.746±0.031	0.732±0.036	**0.115±0.015**	**0.049±0.041**
IFFair	1472.1±177.8	0.368±0.044	15192.6±238.0	0.734±0.042	0.718±0.047	0.138±0.022	0.084±0.061
Influence Function	2379.4±557.6	0.595±0.139	14285.3±551.0	**0.769±0.034**	**0.753±0.044**	0.129±0.029	0.104±0.063
In-Run Shapley (1)	**3228.8±55.5**	**0.807±0.014**	**13435.9±121.0**	0.757±0.029	0.743±0.035	0.140±0.021	0.124±0.060
In-Run Shapley (2)	3140.6±61.0	0.785±0.015	13524.1±123.7	0.755±0.028	0.741±0.036	0.144±0.014	0.122±0.056
FAIR-DS	2085.2±26.1	0.521±0.007	14579.5±105.8	0.743±0.024	0.729±0.027	0.121±0.024	0.050±0.012
Oracle	16664.7±105.7	1.000±0.000	0.0±0.0	0.818±0.018	0.815±0.019	0.018±0.017	0.026±0.020
Equalized odds							
No correction	0.0±0.0	/	16664.7±105.7	0.733±0.037	0.717±0.043	0.167±0.029	0.133±0.072
Random	755.1±19.3	0.189±0.005	15909.6±112.6	0.743±0.032	0.725±0.037	0.154±0.025	0.125±0.067
KNN-Shapley	2520.0±25.8	0.630±0.006	14144.7±96.7	0.751±0.026	0.736±0.027	0.136±0.020	0.091±0.048
Train Loss	1906.9±420.5	0.477±0.105	14757.8±424.4	0.741±0.025	0.727±0.030	0.152±0.020	0.107±0.054
Gradient norm	1648.7±153.3	0.412±0.038	15016.0±162.7	0.742±0.028	0.727±0.032	0.143±0.023	0.095±0.052
FairShap	941.7±36.7	0.235±0.009	15640.0±83.5	0.751±0.013	0.732±0.020	0.159±0.041	0.134±0.093
FairIF	1998.9±320.8	0.500±0.080	14665.8±295.3	0.751±0.040	0.740±0.043	0.092±0.025	0.063±0.034
IFFair	1530.0±239.5	0.383±0.060	15134.7±267.6	0.745±0.031	0.728±0.037	0.138±0.019	0.094±0.060
Influence Function	2379.4±557.6	0.595±0.139	14285.3±551.0	**0.769±0.034**	**0.753±0.044**	0.129±0.029	0.104±0.063
In-Run Shapley (1)	**3228.8±55.5**	**0.807±0.014**	**13435.9±121.0**	0.757±0.029	0.743±0.035	0.140±0.021	0.124±0.060
In-Run Shapley (2)	3140.6±61.0	0.785±0.015	13524.1±123.7	0.755±0.028	0.741±0.036	0.144±0.014	0.122±0.056
FAIR-DS	2467.8±137.2	0.617±0.034	14196.9±207.8	0.761±0.033	0.751±0.037	**0.087±0.027**	**0.038±0.009**
Oracle	16664.7±105.7	1.000±0.000	0.0±0.0	0.818±0.018	0.815±0.019	0.018±0.017	0.026±0.020