WHEN EFFICIENT TRAINING BECOMES UNFAIR: AUDITING AND REPAIRING DATA PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data pruning reduces training cost by removing examples a model appears to have learned, often with little change in average performance. However, average metrics can conceal substantial subgroup harm. First, we conduct a systematic ***demographic-fairness audit*** of data pruning across static and dynamic criteria and find that pruning can significantly degrade subgroup and worst-group performance even when removal is demographically uniform. This harm is not explained by preferential deletion of minority data; instead, pruning reduces overall training exposure, disproportionately affecting data-scarce groups. Motivated by this finding, we propose ***EquiPrune***, a lightweight, modular wrapper that operates at the pruning decision stage to preserve sufficient training exposure. Across diverse datasets, architectures, and pruning strategies, EquiPrune restores baseline-level fairness in most affected settings while retaining substantial computational savings. These results suggest that efficient training should be evaluated on the fairness–efficiency frontier, not just accuracy–efficiency.

1 INTRODUCTION

Training modern deep neural networks is increasingly computationally expensive, motivating a growing body of work on reducing the cost of learning (Strubell et al., 2019; Schwartz et al., 2020; Sorscher et al., 2022). One practical direction is *data pruning*: reducing the amount of training data processed by identifying examples that contribute little additional learning signal. Static pruning methods score examples before training using criteria such as error norms or gradient magnitudes (Paul et al., 2021), forgetting events (Toneva et al., 2019), proxy-model confidence (Coleman et al., 2020), or coverage (Zheng et al., 2023), and train only on the retained subset. Dynamic methods instead adapt the retained data as optimization progresses (Raju et al., 2021; He et al., 2023). Some dynamic methods remove examples only temporarily; for example, InfoBatch (Qin et al., 2024) skips low-loss samples within an epoch but can reintroduce them later. Others make removal permanent; B-PAS (Alam et al., 2026), for instance, retires batches once their activation statistics stabilize. Permanent pruning can yield substantial computational savings because removed data never re-enters training, but it also makes pruning decisions irreversible. Across these methods, efficiency is typically evaluated by how much data or computation can be removed while preserving *average predictive performance*. Yet average performance alone does not reveal whether the resulting savings affect all demographic groups equally.

This concern is motivated by a broad literature showing that strong aggregate performance can conceal substantial subgroup disparities (Hardt et al., 2016; Buolamwini & Gebru, 2018; Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021; Zong et al., 2023), particularly when demographic groups are unevenly represented in the training data (Luo et al., 2024b). Similar effects have been observed for other efficiency mechanisms. Model compression can disproportionately degrade performance on rare or atypical examples (Hooker et al., 2019; 2020; Paganini, 2020), motivating fairness-aware weight pruning (Lin et al., 2022; Wu et al., 2022; Zayed et al., 2024). On the data side, prior work has considered fairness-constrained coresets (Huang et al., 2019), fair batch selection (Roh et al., 2021), fairness-aware deduplication (Slyman et al., 2024), and fairness-aware static data pruning (Zayed et al., 2023). More recent studies show that static pruning can amplify class imbalance (Vysogorets et al., 2025; Hassan et al., 2025) or alter robustness to spurious correlations even when average accuracy is preserved (Dharmasiri et al., 2025). Despite this progress, two important questions remain largely unexplored. ***First****, what are the demographic consequences of pruning training data itself?* Existing fairness-aware pruning studies have primarily focused on model parameters, while stud-

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.

ies of data pruning have largely examined class imbalance, spurious attributes, or fairness-aware subset construction rather than systematically auditing performance across protected demographic groups. ***Second**, what happens when pruning is dynamic and permanent, so that the effective training set evolves with the model and removal may become irreversible?* Existing fairness analyses of data pruning are predominantly static, leaving the subgroup consequences of dynamic, and particularly permanent, data removal poorly understood. These questions present two key challenges. *First*, subgroup degradation does not directly reveal its cause. A natural concern is that pruning disproportionately removes data from underrepresented groups, yet even demographically balanced pruning may affect groups unequally because data-scarce groups have less training signal to spare. Understanding the source of disparity requires examining group-wise training exposure throughout optimization, rather than only final subgroup performance. *Second*, mitigating the disparity should not negate the efficiency benefit of pruning. A practical solution should preserve the underlying pruning mechanism while preventing excessive loss of training exposure for vulnerable groups.

To address these challenges, we conduct a systematic demographic-fairness audit of data pruning and develop a mechanism-guided repair framework. Our *audit* spans both static and dynamic pruning settings and evaluates subgroup behavior across diverse datasets, architectures, optimizers, and pruning strategies. We find that pruning induces substantially larger degradation in worst-group and subgroup performance than in average predictive performance, with the severity of this imbalance varying across training regimes. Importantly, this degradation is not generally accompanied by preferential removal of particular demographic groups: group retention can remain approximately balanced even as subgroup performance deteriorates. To investigate this discrepancy, we introduce *retention accounting*, a per-epoch, per-group measure of the training exposure received under pruning. This analysis shows that the key failure mode is not simply who is removed more often, but how much useful training exposure remains available to each group. Aggressive pruning thins the overall training signal, and data-scarce groups are disproportionately sensitive to this reduction. This diagnosis motivates *EquiPrune*, a lightweight and modular wrapper for existing pruning methods. Rather than replacing the underlying pruning algorithm, EquiPrune preserves its scoring criterion and schedule while adding safeguards at the pruning and optimization stages to maintain sufficient training exposure and compensate groups that would otherwise become underexposed.

Our contributions are threefold. ***First**, we provide a systematic demographic-fairness audit of data pruning and introduce retention accounting to characterize its underlying failure mode.* Our analysis shows that pruning can preserve average performance while degrading subgroup and worst-group performance, and that this disparity can arise even without demographically imbalanced sample removal. ***Second**, we propose EquiPrune, a lightweight and modular fairness repair for dynamic data pruning.* EquiPrune preserves the underlying pruning criterion and schedule while introducing safeguards that maintain sufficient group-wise training exposure. ***Third**, we conduct extensive evaluations across datasets, architectures, optimizers, and pruning strategies.* The results show that EquiPrune consistently improves subgroup fairness while retaining substantial pruning-induced computational savings. Together, these findings motivate evaluating efficient-training methods on the *fairness–efficiency frontier*, in addition to the conventional accuracy–efficiency trade-off.

2 Background and Preliminaries

Setup and notation. Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$ be a training set where each example carries, besides input and label, protected attributes $a_i = (a_i^{(1)}, \ldots, a_i^{(J)})$ (e.g., race, gender). Attribute j partitions $\mathcal{D}$ into groups $\mathcal{G}^{(j)}$; each dataset designates one primary attribute, with groups $\mathcal{G}$ and group sizes n_g. Before training, $\mathcal{D}$ is partitioned once into K batches $\mathcal{B}_1, \ldots, \mathcal{B}_K$ of size b that are never reshuffled. We fix batch membership, following B-PAS (Alam et al., 2026), for two reasons: to enable comparisons of *the same batch* across epochs and to simplify per-group retention accounting by keeping each batch's demographic composition $n_{g,k}$ fixed throughout training. A batch-level dynamic pruner maintains an active set $\mathcal{A}_t \subseteq [K]$ over epochs $t = 1, \ldots, T$, and epoch t trains only on $\{\mathcal{B}_k : k \in \mathcal{A}_t\}$; the unpruned baseline keeps $\mathcal{A}_t = [K]$. A content-aware pruner gives each batch a time-varying utility signal, the activation-stability change $s_k(t)$ (the epoch-to-epoch change in mean activation dispersion), and compares it with a threshold $\delta(t)$ that decides when the batch counts as learned. Efficiency is measured by the data-savings index (DSI) (Alam et al., 2026), defined as the fraction of training data usage avoided relative to a full run, with DSI $\in [0, 1]$; higher values indicate greater savings.

Fixed membership with intra-batch shuffling. Following B-PAS (Alam et al., 2026), the batch partition is drawn once and kept fixed, while samples within each batch are reshuffled every epoch. Fixed membership ensures that "the same batch across epochs" is well defined, allowing stability signals to compare like with like, and enables exact, noise-free per-group accounting since each batch's composition is deterministic. Intra-batch shuffling preserves stochasticity and avoids overfitting to a fixed order. This provides exact bookkeeping, turning pruning from a black-box process into an interpretable mechanism and enabling the retention accounting used in our audit.

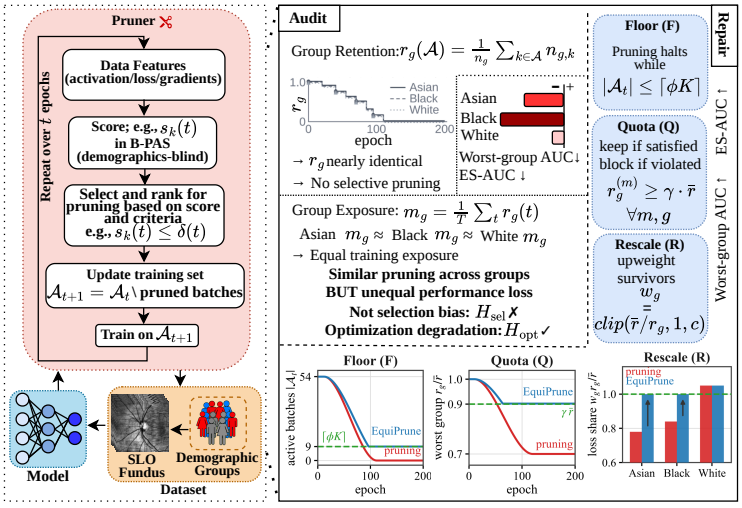


Figure 1: **The paper in one picture: a blind pruner, an audit, and a repair. Left, the pruner.** A dynamic pruning loop: each epoch scores every batch from its own training signal (activations, loss, or gradients), removes those falling under the schedule $\delta(t)$, and trains on what is left; in the permanent case we focus on, a removed batch never returns. The score never sees demographics, so nothing in the loop records who leaves. **Center top, the audit**. Retention $r_g(t)$ is the share of group g's data still in training at epoch t; exposure m_g is its average over the run. Across races, both come out almost identical, yet worst-group AUC and ES-AUC still fall. Deletion is even, but the damage is not, which rules out biased selection (H_{sel}) and points to optimization degradation (H_{opt}). **Right and center bottom, the repair**. EquiPrune wraps the same signal in three switchable guardrails at the removal decision: F stops pruning before the active set empties, Q blocks any removal that would push a group below $\gamma\,\bar{r}$, and R up-weights the survivors of an under-retained group. The three plots sketch the failure each one prevents and are illustrative.

Batch-level granularity. The pruning unit is the batch, as each batch is a random cross-section of the data, so removing it preserves class composition, whereas applying the same signal at the sample level induces class imbalance and causes a small drop in top-1 performance on ImageNet (Alam et al., 2026). Repeating the audit at the sample level (FairVLMed/ResNet-50+SGD) increases the harm, while our EquiPrune still repairs it (Appendix J); thus, the batch-level results are conservative and provide a lower bound on the harm at finer granularity.

Data pruning. We study three content-aware pruners spanning two axes: when removal is decided (dynamic vs. static) and whether it is reversible (temporary vs. permanent). B-PAS is dynamic and permanent, InfoBatch is dynamic and temporary, and EL2N is static and permanent. All rely on learned utility signals and are strictly group-blind. **B-PAS** (Alam et al., 2026) removes batches based on activation stability: when $s_k(t) < \delta(t)$, batch k is deemed learned and permanently removed, $\mathcal{A}_{t+1} = \mathcal{A}_t \setminus$ pruned batches. The schedule $\delta(t)$ governs pruning aggressiveness. **InfoBatch** (Qin et al., 2024) probabilistically skips low-loss batches (below the mean) each epoch and rescales retained losses to keep gradients unbiased. Skipped batches are revisited later, and pruning anneals off after 87.5% of training. **EL2N** (Paul et al., 2021) ranks batches by early-training error norm (epochs 2–10) and drops the lowest-scoring ones once (epoch 10). The pruning budget is fixed a priori; we use 35% DSI across all regimes, so savings are constant. All methods are implemented at batch granularity, enabling exact per-group accounting.

Fairness metrics. Following prior medical-AI fairness work (Luo et al., 2024b; Zong et al., 2023), we report overall accuracy and AUC, per-group AUCs $\{\mathrm{AUC}_g\}$, and worst-group AUC $\min_g \mathrm{AUC}_g$. Our primary metric is equity-scaled AUC, $ES\text{-}AUC = \frac{\mathrm{AUC}_{\text{overall}}}{1+\sum_g |\mathrm{AUC}_g - \mathrm{AUC}_{\text{overall}}|}$, which preserves overall performance while penalizing subgroup disparities. Larger deviations increase the denominator, lowering the score. We also report threshold-based parity metrics: demographic-parity distance (DPD) and equalized-odds distance (DEOdds). They complement ES-AUC, as they capture different failure modes and can dissociate under pruning. Two evaluation conventions are applied

throughout: (1) "Unknown" groups are excluded from gap computations to avoid instability from extremely small sample sizes, though their AUCs are still reported. (2) Changes in worst-group AUC are defined as $\Delta = \min_g \mathrm{AUC}_g^{\text{pruned}} - \min_g \mathrm{AUC}_g^{\text{base}}$, which measures the change in the worst-group performance. The identity of this group may differ across runs. This definition is conservative, as it may understate but does not overstate the degradation in the most affected group, while being more stable than fixed-group comparisons under a limited number of seeds.

3 Methodological Foundations

Audit. The audit follows a single principle: hold everything fixed and vary only which data is skipped. Within each regime (fixed dataset, architecture, optimizer), the baseline and all pruners (B-PAS, InfoBatch, EL2N) run under identical conditions: same model, optimizer, schedule, batch partition, seeds, and evaluation; so the pruning rule is the only degree of freedom, and any change in performance or fairness is attributable to pruning. Each experiment is read on two layers: the *outcome layer* evaluates subgroup harm via fairness metrics, while the *mechanism layer* explains it via retention accounting, which records per-group training exposure rather than final result. Both layers are pre-instrumented and reported in this order.

Retention accounting: measuring mechanism, not only outcome. The fairness metrics describe the model at the end of a run; they record nothing about the training it received along the way. Retention accounting captures this missing information through group retention: for any active set $\mathcal{A}$, $r_g(\mathcal{A}) = \frac{1}{n_g}\sum_{k\in\mathcal{A}} n_{g,k}, \bar{r}(\mathcal{A}) = \frac{|\mathcal{A}|}{K}$. The fraction of group-$g$ samples is $r_g(\mathcal{A})$ that contribute gradient updates under active set $\mathcal{A}$, and $1-\bar{r}(\mathcal{A}_t)$ is the per-epoch fraction of data skipped, which accumulates into DSI over training. We write $r_g(t) := r_g(\mathcal{A}_t)$ for the retention curve and $m_g = \frac{1}{T}\sum_t r_g(t) \in (0,1]$ for group g's total exposure over training. The audit panel in Figure 1 reports these two quantities. Retention accounting measures what a pruner removes, not what that removal costs, and no retention statistic by itself explains a change in ES-AUC. Retention accounting supports two complementary interpretations. First, the spread of r_g across groups tests $\mathrm{H_{sel}}$ (the following paragraph): preferential deletion is the channel through which the demographic composition of the removed data can affect group exposure, so uniform retention closes the channel by measurement empirically. Second, the level of r_g determines the exposure m_g in the scaling model of Equation 1, in which any remaining explanation of the harm must be expressed. This level also serves as the efficiency dial, $\mathrm{DSI} = 1-\bar{m}$, rather than as a fairness score: a run that reduces exposure equally across groups is efficient, but not necessarily fair.

Two hypotheses, and the instrumentation that separates them. If group-blind pruning harms fairness, two mechanisms could be responsible, and they imply different repairs. **$\mathbf{H_{sel}}$ (biased selection).** A pruning score aggregated over a batch is dominated by the batch's majority samples, so the signal can declare a batch "learned" while learning on its minority samples is still incomplete, and the pruner then deletes precisely the data the worst-served groups can least afford to lose. $\mathrm{H_{sel}}$ predicts: minority-heavy batches are pruned earlier; the pruning signal correlates with batch demographic composition; minority retention curves fall faster. **$\mathbf{H_{opt}}$ (regressive optimization degradation).** Pruning degrades the optimization process itself: fewer updates, on less data, with an earlier effective end of training, and a degraded process is not paid for equally: groups with abundant data have slack, groups with scarce data sit on a thin margin and break first. $\mathrm{H_{opt}}$ predicts: deletions are uncorrelated with composition and retention is demographically uniform, yet harm concentrates on datasets and lands on the smallest groups. **The instrumentation.** $\mathrm{H_{sel}}$ yields three testable predictions, each logged explicitly from the training record. Let π_k denote the minority fraction of batch k, fixed throughout training. All correlations are Spearman rank correlations over batches, pooled across seeds. **(D1) Prune time:** are minority-heavy batches removed first? ρ_{time} correlates π_k with each batch's removal epoch (survivors ranked last); $\mathrm{H_{sel}}$ predicts a significant $\rho_{\mathrm{time}} < 0$ (permanent pruners only: B-PAS, EL2N; 18 cells). **(D2) Signal–composition:** does the score mark minority-heavy batches as prunable? ρ_{sig} correlates π_k with the run-averaged B-PAS stability score s_k (lower is pruned sooner), on unpruned runs, over 32 regime–attribute pairs; $\mathrm{H_{sel}}$ predicts a significant $\rho_{\mathrm{sig}} < 0$. **(D3) Retention:** do minorities keep less data? We measure $m_g - \bar{m}$ and $\max_g r_g(t) - \min_g r_g(t)$ over 27 cells. $\mathrm{H_{sel}}$ predicts minorities below the mean, $\mathrm{H_{opt}}$ near-zero gaps despite harm. D1–D2 ask whether the pruner *chose* against a group, D3 whether the *outcome* was uneven; $\mathrm{H_{sel}}$ is rejected only if all three fail. Each hypothesis implies a different guardrail, evaluated independently by ablation.

Repair: EquiPrune, a Plug-in at the Pruning Decision. The audit yields a clear target: the pruner that saves the most also harms the most. B-PAS achieves the largest efficiency gains, but also induces the strongest fairness degradation. This makes it the natural repair point: if its failure can be corrected, one can retain high efficiency without sacrificing fairness. Crucially, the harm does not arise from its selection signal; B-PAS removes data demographically uniformly, but from how much is removed and the fact that removal is permanent. The signal therefore needs no redesign, only a bound. We propose *EquiPrune*, a decision-stage wrapper applied after the pruning scores and ordering are fixed. It leaves the pruning signal, threshold schedule, and removal order unchanged, intervening only in whether a proposed batch is actually dropped. *Three guardrails* enforce this constraint at the pruning decision, one corresponding to each panel of Figure 1.

(F) Absolute retention floor: *Never train on almost nothing.* The damage tracks how much data is gone rather than whose it was, and Proposition 1 explains why: an equal exposure cut multiplies every group's risk by the same factor, so whoever was already worst off loses the most points. The lever is therefore a hard bound on the size of the cut, and a permanent pruner cannot supply it. B-PAS stops removing when no batch looks stable enough, never when too little data is left, and because its threshold rises through training while the surviving batches keep settling, every batch eventually falls under it: in the worst regime the active set empties, later training with little or no data. F is the stopping rule the pruner is missing. Pruning halts while $|\mathcal{A}_t| \leq \lceil \phi K \rceil$ ($\phi = 0.15$). It uses no demographic information, bounding the worst case however aggressive the threshold schedule becomes.

(Q) Group-retention quota: *Proportional retention constraint.* The floor bounds the total but says nothing about how the remainder is shared, and uniform deletion is something we measured, not something any pruner promises: no activation-stability score, loss or error norm rules out a run in which one group's batches happen to settle first, and the smaller the group, the further a handful of batches moves its retained share. Q makes that uniformity hold by construction, tested at the only point where it is cheap to test, the instant before a batch is dropped: candidates are processed in the order the pruning signal induces, but a batch k may be pruned only if afterwards every group of every protected attribute retains at least γ times the overall retained fraction, $r_g^{(j)}(\mathcal{A}_t \setminus \{k\}) \;\geq\; \gamma\, \bar{r}(\mathcal{A}_t \setminus \{k\})$ for all $j \in [J],\ g \in \mathcal{G}^{(j)}$. A candidate that fails the test is simply kept and pruning moves on to the next, with all attributes checked simultaneously; if a future pruner does start selecting against a group, Q refuses the removal and the retention log records the refusal. With $\gamma = 0.9$ the rule is stricter than the four-fifths (80%) rule of disparate-impact analysis (Feldman et al., 2015), because dynamic pruning acts cumulatively: small imbalances repeated over 200 epochs compound into substantial under-exposure. In plain words, data may be discarded, but never proportionally more of one group's than of everyone else's.

(R) Group-compensated loss rescaling: *Count what is left more heavily.* Retention is not the same as influence. When a batch is permanently retired, the groups inside it not only hold less data, they also carry less weight in every gradient step that follows, because the training loss is an average over whatever is left. Restoring the data is impossible once a permanent pruner has acted; restoring the weight is not, and it costs one scalar per group. R also covers Q's slack, since a group sitting at exactly γ times the overall retention satisfies the quota and is still under-retained. Surviving samples of group g (primary attribute) are up-weighted by $w_g(t) \;=\; \mathrm{clip}\Big(\frac{\bar{r}(\mathcal{A}_t)}{r_g(\mathcal{A}_t)},\ 1,\ c\Big), c = 5$. An under-retained group has $r_g(\mathcal{A}_t) < \bar{r}(\mathcal{A}_t)$, so the ratio exceeds one and its surviving samples count for more; clipping at 1 keeps any group from being down-weighted, and at c keeps the weights, and with them the gradient variance, from growing without limit.

The per-epoch loop preserves the pruner's order and adds two checks and one weight: candidates are processed in signal order, pruning stops at the floor $\lceil \phi K \rceil$ (F), any removal that would drop a group below $\gamma \bar{r}$ is skipped (Q), and group weights are set from final retention before training (R). Algorithm 1 in Appendix B.4 states it in full. We fix $(\phi, \gamma, c) = (0.15, 0.9, 5)$ across all datasets without tuning (Sensitivity analyses in Appendix K). We report four variants, all with F: EquiPrune (F+Q+R), EquiPrune-Q (F+Q; preserve group data), EquiPrune-R (F+R; reweight survivors), and EquiPrune-F (F only; no starvation). The floor-only arm is the key control: if F alone restores fairness, harm is optimization degradation ($\mathrm{H}_{\mathrm{opt}}$); if Q or R is needed, it indicates biased selection ($\mathrm{H}_{\mathrm{sel}}$). Comparing F+Q vs. F+R further separates preserving group data from reweighting, making the (F, Q, R) ablation both repair and diagnosis.

Theoretical implications: why demographics-blind pruning is regressive. The guardrails need no new theory to implement, but a simple model explains their shape and yields predictions the audit instrumentation can test. Assume the standard power-law relation between data quantity and excess risk (Sorscher et al., 2022): if group g's data receives exposure $m_g \in (0, 1]$ during training, its excess risk is

$$\varepsilon_g \;=\; C_g \,(m_g \, n_g)^{-\alpha}, \qquad \alpha > 0, \tag{1}$$

with C_g a group-difficulty constant and α the scaling exponent (empirically $\alpha \in [\frac{1}{2}, 1]$). The model is stylized, a lens rather than a performance guarantee, but it already separates the two hypotheses.

Proposition 1 (Uniform pruning is regressive)**.** *Suppose pruning reduces every group's exposure equally, $m_g = \bar{m} \in (0, 1)$ for all g. Then under Equation 1: (i) every group's excess risk is multiplied by the same factor $\bar{m}^{-\alpha} > 1$; (ii) the absolute increases $\Delta_g = (\bar{m}^{-\alpha} - 1)\, C_g \, n_g^{-\alpha}$ are largest for the groups with the largest baseline risk, which for comparable difficulty C_g means the smallest groups; (iii) the between-group risk gap is multiplied by $\bar{m}^{-\alpha}$, so gap-penalized metrics such as ES-AUC degrade although no group's data was preferentially deleted; (iv) a retention floor guaranteeing $m_g \geq \phi$ caps the multiplier at $\phi^{-\alpha}$.*

Proof. Substitute $m_g = \bar{m}$ (resp. $m_g \geq \phi$) into Equation 1; all four statements are immediate. □

Proposition 1 shows that *demographic blindness of the selection rule does not imply fairness-neutrality of its effect*. Scarcity converts even-handed deletion into uneven damage, its incidence falling on whoever holds the least data, which in imbalanced cohorts means the minority groups. This is why a demographics-blind bound on *total* deletion can be the binding remedy, and why we do not attempt to "de-bias the signal": under $\mathrm{H}_{\mathrm{opt}}$ there is no bias in the signal to remove. Q and R address the complementary risk that a real pruner need not remain uniform.

Predictions. Two falsifiable predictions are made. **(P1)** *Scale dependence*: harm concentrates where the data budget is smallest and vanishes where data is abundant, since $\Delta_g \propto n_g^{-\alpha}$. **(P2)** *Regressive incidence*: even under demographically uniform deletion, the largest loss falls on underrepresented (minority) groups rather than the majority. The experiments confirm both (Section 4).

4 Experiments

Our experiments address two questions: **R1** whether group-agnostic pruning harms subgroups, and **R2** whether the proposed guardrails repair this harm while preserving efficiency. We evaluate on six datasets spanning medical imaging and face attributes: Harvard-FairVLMed (Luo et al., 2024b), Harvard-FairVision (AMD, diabetic retinopathy, glaucoma) (Luo et al., 2024a), CheXpert (Irvin et al., 2019), and UTKFace (Zhang et al., 2017), covering variation in size and demographic imbalance. A regime is a (dataset, architecture, optimizer) triple, yielding nine regimes: FairVLMed under a 2×2 design (ResNet-50 (He et al., 2016) vs. ViT-B/16 (Dosovitskiy et al., 2021), SGD vs. AdamW) and five additional datasets under ResNet-50+SGD. Each audit regime is run with three pruners: B-PAS, InfoBatch, and EL2N, the unpruned baseline, and the repair is with four EquiPrune variants across three seeds. The two permanent pruners (B-PAS, EL2N) yield 18 (regime, pruner) pairs for temporal mechanism analysis. Protected attributes: four for FairVLMed/FairVision (race primary; others in Appendix F), one for CheXpert(gender)/UTKFace(race). All runs use 200 epochs, batch size 128, cosine decay with warmup and identical evaluation; details are in the appendix. **Fixed pre-registered decision rules.** To avoid post-hoc interpretation, all pass/fail criteria were fixed before running the experiments. All results are seed-paired against the unpruned baseline and reported as mean±std over three seeds. **R1** (harm): a method is deemed harmful in a regime if its mean ΔES-AUC or Δworst-group AUC is below -0.5. A regime supports R1 if at least two of the three audited pruners meet this criterion. **R2** (repair): the repair is successful if it achieves DSI $\geq 25\%$ while keeping both deltas ≥ -0.5 (strong when both are positive).

The audit: the harm is real (R1). We read the audit in the two layers of Section 3. The *outcome layer* (Table 1) asks whether pruning harms subgroups, and where. The *mechanism layer* (Figure 2 and diagnostics D1–D3) asks what the pruner did to cause it, and decides between $\mathrm{H}_{\mathrm{sel}}$ and $\mathrm{H}_{\mathrm{opt}}$.

Outcome: pruning harms subgroups in six of nine regimes. R1 holds on all four FairVLMed cells, FairVision-DR and FairVision-Glaucoma, and is rejected on FairVision-AMD, CheXpert and UTKFace. The largest harm is B-PAS on FairVLMed/ResNet-50+SGD: -4.08 ± 2.14 ES-AUC and -3.92 ± 2.37 worst-group AUC at 67% DSI. The most consistent harms across seeds are on FairVision-Glaucoma (B-PAS -3.26, p=0.008; EL2N -2.16, p=0.021). Appendix C gives the full reading; three patterns carry the argument. *(i) Harm follows how much is removed, and for*

Table 1: **The audit (R1).** Seed-paired Δ vs. the unpruned baseline, AUC points $\times 100$, mean$\pm$std over 3 seeds; negative = degraded. Shaded rows support R1 under the pre-registered rule of Section 4. Realized savings: B-PAS 52–73% DSI, InfoBatch 24–29%, EL2N 35% by construction. †underfit control cell. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t, n=3, descriptive).

	ΔES-AUC			Δworst-group AUC		
Regime	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	-4.08 ± 2.1	-0.84 ± 3.2	-1.45 ± 1.4	-3.92 ± 2.4	-0.56 ± 2.4	-0.58 ± 1.9
FairVLMed / R50+AdamW	-1.88 ± 3.4	$+2.20\pm1.1$	$-2.77\pm0.6^{*}$	-0.98 ± 3.0	$+2.77\pm2.8$	$-2.25\pm0.9^{*}$
FairVLMed / ViT+AdamW	-1.02 ± 4.1	-1.36 ± 2.7	$+0.11\pm2.7$	-0.39 ± 3.5	-1.19 ± 1.3	$+0.10\pm2.3$
FairVLMed / ViT+SGD†	$+1.55\pm9.2$	-1.82 ± 13.4	$-5.69\pm1.8^{*}$	$+1.50\pm11.4$	-2.40 ± 15.9	-6.14 ± 2.9
FairVision-AMD	-1.23 ± 1.5	$+0.55\pm1.9$	$+1.52\pm3.4$	-0.90 ± 1.6	$+1.71\pm2.4$	$+3.65\pm5.6$
FairVision-DR	-1.36 ± 6.8	-4.00 ± 4.3	-2.83 ± 2.3	-1.36 ± 6.4	-3.55 ± 3.6	-3.64 ± 2.0
FairVision-Glaucoma	$-3.26\pm0.5^{**}$	-1.44 ± 3.7	$-2.16\pm0.5^{*}$	$-3.39\pm0.2^{**}$	-1.09 ± 3.9	-1.34 ± 0.9
CheXpert	$+0.14\pm0.1$	-0.19 ± 0.2	-0.13 ± 0.1	$+0.08\pm0.1$	-0.21 ± 0.2	-0.16 ± 0.2
UTKFace	-0.42 ± 1.1	$+0.83\pm0.8$	$+0.36\pm0.8$	-0.12 ± 1.3	$+0.73\pm0.6$	$+0.11\pm0.3$

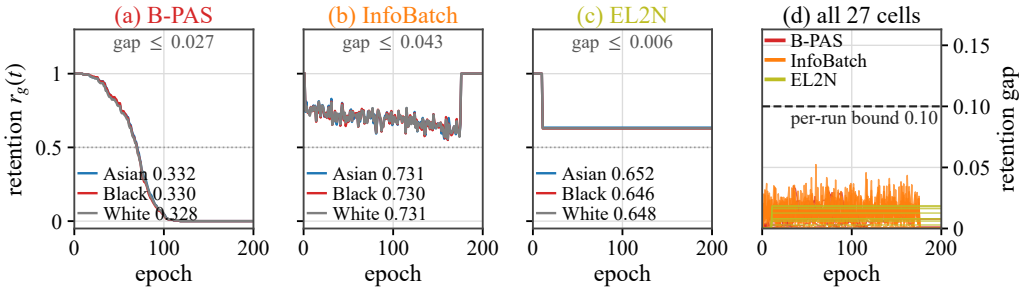


Figure 2: **Retention accounting: three pruners, three very different deletion schedules, all demographically even. (a–c)** Group retention curves $r_g(t)$ (Section 3) by race on FairVLMed/ResNet-50+SGD, the regime with the largest fairness harm in Table 1 (mean over 3 seeds; the unpruned baseline $r_g \equiv 1$ is omitted). Legends give each group's exposure m_g, and annotations the largest pointwise gap $\max_g r_g(t) - \min_g r_g(t)$ reached at any epoch. **(d)** The same gap for all 27 audited (regime, pruner) cells, on each regime's own primary attribute: it never leaves a band of width 0.052 for the whole of training.

how long. Outside the ViT+SGD cell, B-PAS averages -1.64 ES-AUC and -1.37 worst-group AUC, against -0.92 and -0.51 for EL2N and -0.53 and -0.17 for InfoBatch. The two permanent pruners lead, while InfoBatch, which returns all skipped data before training ends, trails. B-PAS also saves the most (52–73% DSI, against 24–29% and a fixed 35%). We read this as the finding, not a confound: savings and damage come from the same act, which is why the repair targets B-PAS. *(ii) The harm is not tied to one signal.* Activation stability, loss, and error norm each harm at least one regime, although they rank batches by different quantities. A selection bias would need all three to be skewed the same way by coincidence, a first strike against $\mathrm{H}_{\mathrm{sel}}$. *(iii) The safe regimes are the large ones.* No pruner harms CheXpert or UTKFace, the two largest and most balanced datasets (1,220 and 129 batches, against 46–54 for the medical sets), even though their retention falls as far and as fast as elsewhere. This is prediction (P1) of Proposition 1: the same cut costs little where every group keeps plenty of data. Fairness certified on a large benchmark therefore says nothing about the small, imbalanced clinical sets where pruning saves the most. FairVision-AMD, small and imbalanced, misses R1 by one pruner (only B-PAS harms it). *The ViT+SGD cell* underfits (baseline AUC 69.3 ± 4.0 against $\sim$80), which widens its spreads. We keep it because it completes the 2×2 design, which shows the harm is owned by neither architecture nor optimizer alone.

Mechanism: whose data was removed? The table shows the harm, not its cause. Under $\mathrm{H}_{\mathrm{sel}}$, the pruner removes minority data preferentially, which calls for a demographic quota. Under $\mathrm{H}_{\mathrm{opt}}$, removal is even across groups, but the shorter, thinner training hurts small groups most, which calls for a demographics-blind floor on how much is removed. **Retention is uniform across groups.** On the worst regime (Figure 2a–c) the three pruners follow very different schedules but treat groups alike. B-PAS decays steadily and empties the active set between epochs 101 and 118, so almost half of training sees no data, yet $m_{\text{Asian}} = 0.332$, $m_{\text{Black}} = 0.330$ and $m_{\text{White}} = 0.328$. InfoBatch oscillates and restores all data at epoch 176 (all $m_g \approx 0.73$); EL2N drops once at epoch 11 and stays flat ($m_g \approx 0.65$). Across all 27 audited cells (panel d), the gap between the best- and worst-retained

Table 2: **The repair (R2).** Seed-paired ΔES-AUC vs. the unpruned baseline (points ×100, mean±std, 3 seeds), with realized DSI in parentheses. **Bold** = the arm satisfying the pre-registered R2 criterion of Section 4; shaded = no arm passes. The rightmost column repeats the worst audited pruner from Table 1. †underfit control cell. $^*p < 0.05$.

Regime	EquiPrune (F+Q+R)	-Q (F+Q)	-R (F+R)	-F (floor only)	worst audited
FairVLMed / R50+SGD	−0.59±0.7 (56)	−2.60±3.4 (55)	−0.12±3.2 (56)	**+0.22**±2.2 (56)	−4.08
FairVLMed / R50+AdamW	−1.95±1.5 (64)	−0.70±4.1 (65)	−0.59±2.7 (66)	−1.59±2.1 (65)	−2.77
FairVLMed / ViT+AdamW	−0.60±3.1 (63)	−0.82±3.4 (65)	−0.51±2.3 (65)	**−0.35**±2.7 (65)	−1.36
FairVLMed / ViT+SGD†	**+4.43**±6.8 (61)	−1.84±12.3 (58)	**+3.90**±4.6 (67)	**+0.04**±11.9 (65)	−5.69
FairVision-AMD	**+2.58**±1.4 (56)	**+2.32**±4.6 (58)	**+1.21**±3.6 (59)	**+0.85**±3.3 (57)	−1.23
FairVision-DR	**+2.38**±5.5 (58)	**+0.59**±5.5 (56)	−0.97±2.3 (59)	−0.06±7.3 (57)	−4.00
FairVision-Glaucoma	−4.31±0.8* (58)	−2.87±1.7 (59)	−3.09±2.8 (57)	−1.81±1.5 (59)	−3.26
CheXpert	**−0.11**±0.2 (59)	**−0.13**±0.3 (59)	**−0.13**±0.2 (60)	**+0.08**±0.3 (60)	−0.19
UTKFace	**+0.47**±1.6 (47)	**+0.06**±0.8 (48)	−0.67±0.2 (43)	−1.16±0.5 (47)	−0.42

group never exceeds 0.052. Pruning changes *when* data is removed, not *whose*. **H_{sel} is rejected: all three diagnostics fail. (D1)** No permanent-pruner cell shows minority-heavy batches removed first: $|\rho_{time}| \leq 0.12$, none significant (smallest $p = 0.090$), signs split (7 of 18 negative). **(D2)** No pair shows the score favouring minority-heavy batches for removal: $|\rho_{sig}| \leq 0.19$ over 32 pairs, and the two nominal hits fail Bonferroni correction and are positive, retiring them *later*. **(D3)** Exposure is equal: $|m_g - \bar{m}| \leq 0.018$ in all 27 cells. **H_{opt} is accepted.** Exposure is equal; outcomes are not. In the worst regime, Black loses 5.07 AUC points against 3.29 for White, and across datasets the largest loss falls on a minority group (Figure 3, Appendix C.3), as prediction (P2) of Proposition 1 expects: an equal cut amplifies existing gaps. The level of exposure also orders the damage ($\bar{m} = 0.33$, 0.65, 0.73 for B-PAS, EL2N and InfoBatch, against −4.08, −1.45 and −0.84 ES-AUC), and replaying B-PAS's drop schedule on randomly chosen batches is no less harmful (Appendix K). The repair closes the loop: the demographics-blind floor alone, which bounds only how much is removed, passes R2 in five of nine regimes (Table 2). The harm is optimization degradation, not biased selection: how much is removed is first-order, when is second-order, and from whom is neither.

The Repair: EquiPrune and its Ablation (R2). EquiPrune repairs most of the damage while keeping most of the savings: at least one arm meets the pre-registered R2 criterion in seven of the nine regimes, at 47–67% DSI (Table 2). The worst-hit regime shows it best. On FairVLMed/R50+SGD, B-PAS loses 4.08 ES-AUC points at 67% DSI; with the guardrails the loss shrinks to -0.59 ± 0.7 (floor only: +0.22) at 56% DSI. The most efficient pruner in the study need not be paid for by its worst-served patients. In every regime, the best arm reduces the worst audited loss.

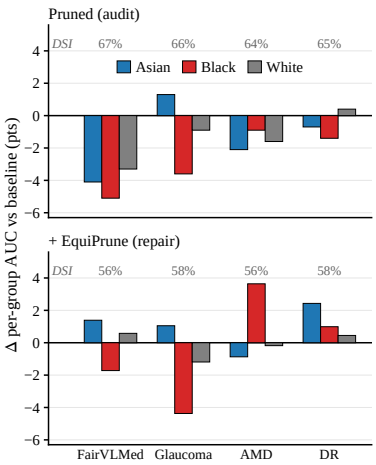


Figure 3: AUC change of each named race group against the unpruned baseline (ResNet-50+SGD, mean over 3 seeds; DSI above each panel). **Top:** audited pruner; **bottom:** the same runs with EquiPrune.

Who is compromised, and what comes back. Figure 3 follows each named group. Before repair, the deepest loss falls on a minority group on all four datasets (on FairVLMed, Black −5.1 against −3.3 for White), as prediction (P2) expects; the worst group is a minority in 167 of 168 runs (Appendix C.3). After repair, FairVLMed moves to Black −1.7 and White +0.6, and on FairVision-DR both minority groups turn positive. FairVision-Glaucoma is the exception (Black −3.6 to −4.4).

Ablation. F is the only demographics-free guardrail, and on its own it passes R2 in five of nine regimes, as often as any arm: bounding *how much* is removed is the main lever, as H_{opt} predicts. Q adds value where the floor falls short (FairVision-DR, UTKFace). Glaucoma confirms the ordering from the other side: the floor alone recovers about half of B-PAS's −3.26, and each group-aware component added on top gives it back (F −1.81, F+Q −2.87, F+R −3.09, F+Q+R −4.31), since Q and R guard against a selection bias the audit did not find. No arm passes on FairVLMed/R50+AdamW or FairVision-Glaucoma; in our sensitivity study, a larger floor ϕ raises ES-AUC at a cost in savings, the natural remedy there.

Table 3: **Pruning inside FairCLIP** (race; on FairVLMed, mean±std; both pruned rows at $\sim 35\%$ DSI). DPD and DEOdds are distances, lower is nominally better. **Bold** = best. All attributes: Table 17.

Method	AUC	ES-AUC	Worst-group AUC	DPD ↓	DEOdds ↓
Baseline (no pruning)	$53.91_{\pm 2.23}$	$51.47_{\pm 0.59}$	$52.23_{\pm 1.37}$	$20.44_{\pm 1.27}$	$23.22_{\pm 0.11}$
B-PAS	$53.75_{\pm 2.65}$	$50.58_{\pm 0.45}$	$51.06_{\pm 1.85}$	$18.19_{\pm 0.89}$	$\mathbf{19.41}_{\pm 2.81}$
EquiPrune	$54.39_{\pm 1.63}$	$\mathbf{51.53}_{\pm 1.62}$	$\mathbf{53.10}_{\pm 0.24}$	$\mathbf{17.86}_{\pm 2.31}$	$19.51_{\pm 4.12}$

The threshold-parity view, and why it is not the certificate. DPD and DEOdds, the parity metrics most often reported, miss the harm: 11 of the 16 cells harmful under R1 show a DPD improvement, and 8 pass as parity-neutral on both, because a model that predicts one class for everyone has perfect parity, so a degraded model can look fairer at every threshold from 0.3 to 0.7. We therefore treat ES-AUC and worst-group AUC as primary and parity as a diagnostic; the repair does not trade parity away, improving it slightly more than the pruners it wraps (Appendix G).

Pruning inside FairCLIP. To check that the findings are not an artifact of supervised training, we prune the fine-tuning batches of **FairCLIP** (Luo et al., 2024b), contrastive fine-tuning of CLIP ViT-B/16 with a Sinkhorn fairness regularizer on clinical notes. Architecture, pretraining, objective and optimizer all change, while the pruning machinery and (ϕ, γ, c) stay fixed (setup and all four attributes in Appendix H). At $\sim 35\%$ DSI (Table 3), B-PAS preserves average AUC but degrades both ES-AUC and worst-group AUC, mirroring the audit: efficiency gains can mask subgroup harm. EquiPrune restores and slightly improves both metrics over the baseline, recovering fairness while retaining savings. This confirms that the observed failure mode is not specific to supervised training and that the wrapper transfers unchanged to fairness-regularized models.

5 Related Work

Dynamic data pruning. Prior work removes data during training using signals such as loss or uncertainty (Raju et al., 2021; He et al., 2023). InfoBatch (Qin et al., 2024) performs stochastic skipping with unbiased gradient rescaling, RCAP (Hassan et al., 2025) adapts class-wise retention, and B-PAS (Alam et al., 2026) uses activation stability for permanent pruning. These methods are evaluated mainly on average accuracy, without demographic fairness analysis. **Fairness of efficiency mechanisms.** Model compression can disproportionately harm minority groups while preserving average performance (Hooker et al., 2019; Tran et al., 2022), indicating that efficiency mechanisms may induce structured bias. We study this question on the data side. **Fairness-aware data selection.** Methods such as FairBatch (Roh et al., 2021) and group DRO (Sagawa et al., 2020) explicitly enforce fairness via sampling or reweighting, often trading off efficiency; we instead constrain an existing pruning signal at decision time. **Bias audits of data selection.** Existing audits focus on class-level or robustness effects (e.g., DRoP (Vysogorets et al., 2025), RCAP (Hassan et al., 2025), Dharmasiri et al. (2025)) and do not study demographic attributes or dynamic pruning.

6 Conclusion

Data pruning is sold as a free lunch: the same accuracy for less data. We asked who pays for it. Across three pruners, six datasets, two architectures and two optimizers, group-blind pruning harms subgroups in most regimes, most under permanent removal and most of all under B-PAS, the pruner that saves the most, and the loss lands on minority patients. Retention accounting shows this is not minority-first deletion: an even cut thins the optimization, and the groups with the least data pay the most, as Proposition 1 predicts. The repair is therefore cheap: EquiPrune keeps the pruner's signal and schedule, adds three guardrails at the removal decision, restores fairness in most regimes while keeping most of the savings, and carries over to FairCLIP fine-tuning; the demographics-blind floor alone does much of the work. Pruning is most attractive where compute is scarce, in smaller hospitals and low-resource labs, yet average accuracy says nothing about subgroups, and because the harm depends on a dataset's size and imbalance, fairness is a property of the regime, not of the pruner. We therefore recommend reporting pruning on the fairness–efficiency frontier, with per-group retention logs from every run. **Limitation and future work.** Q and R need demographic labels during training; an attribute-free guardrail over groups inferred from the model's representations would extend the frontier to data where protected attributes are never collected.

REPRODUCIBILITY STATEMENT

For reproducibility, we provide the codebase, detailed experimental configurations, and experiment instructions in the supplementary materials. Upon acceptance, we will publicly release the full codebase and documentation.

AI USE STATEMENT

Generative AI tools (LLMs and AI agents) were used to assist in manuscript preparation, including refining wording, improving clarity, and structuring sections, as well as supporting codebase organization. All AI-assisted outputs were carefully reviewed and verified by the authors.

REFERENCES

Md Mustakin Alam, Shaker Islam, and Aminul Islam. Batch pruning by activation stability. In *International Conference on Learning Representations (ICLR)*, 2026.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pp. 77–91, 2018.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Mitigating bias in dataset distillation. In *International Conference on Machine Learning (ICML)*, 2024.

Amaya Dharmasiri, William Yang, Polina Kirichenko, Lydia Liu, and Olga Russakovsky. The impact of coreset selection on spurious correlations and group robustness. *arXiv preprint arXiv:2507.11690*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 259–268, 2015.

Leonidas Gee, Andrea Zugarini, and Novi Quadrianto. Are compressed language models less subgroup robust? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15859–15868, 2023.

Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, and Sebastian Schelter. Automated data cleaning can hurt fairness in machine learning-based decision making. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Aviv Hadar, Tova Milo, and Kathy Razmadze. Interactive coreset selection for tabular data: Fairness-aware, explainable, and user-guided. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA)*, 2025.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Meraj Hashemizadeh, Juan Ramirez, Rohan Sukumaran, Golnoosh Farnadi, Simon Lacoste-Julien, and Jose Gallego-Posada. Balancing act: Constraining disparate impact in sparse models. In *International Conference on Learning Representations (ICLR)*, 2024.

Atif Hassan, Swanand Khare, and Jiaul H. Paik. Rcap: Robust, class-aware, probabilistic dynamic dataset pruning. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1648–1662. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/hassan25a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. *arXiv preprint arXiv:2306.05175*, 2023.

Feng Hong, Yueming Lyu, Jiangchao Yao, Ya Zhang, Ivor W. Tsang, and Yanfeng Wang. Diversified batch selection for training acceleration. In *International Conference on Machine Learning (ICML)*, 2024.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S. Kevin Zhou, and Xiaoxiao Li. FairMedFM: Fairness benchmarking for medical imaging foundation models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. FairGRAPE: Fairness-aware gradient pruning method for face attribute classification. In *European Conference on Computer Vision (ECCV)*, pp. 414–432, 2022.

Yan Luo, Muhammad Osama Khan, Yu Tian, Min Shi, Zehao Dou, Tobias Elze, Yi Fang, and Mengyu Wang. Fairvision: Equitable deep learning for eye disease screening via fair identity scaling, 2024a. URL https://arxiv.org/abs/2310.02492.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. FairCLIP: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12289–12301, 2024b.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. $\mathbb{D}^2$ pruning: Message passing for balancing diversity and difficulty in data pruning. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Michela Paganini. Prune responsibly. *arXiv preprint arXiv:2009.09936*, 2020.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 413–423, 2021.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. InfoBatch: Lossless training speed up by unbiased dynamic data pruning. In *International Conference on Learning Representations (ICLR)*, 2024.

Ravi S. Raju, Kyle Daruwalla, and Mikko Lipasti. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. FairBatch: Batch selection for model fairness. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

Eric Slyman, Stefan Lee, Scott Cohen, and Kushal Kafle. FairDeDup: Detecting and mitigating vision-language fairness disparities in semantic dataset deduplication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13905–13916, 2024.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Modar Sulaiman and Kallol Roy. GFLC: Graph-based fairness-aware label correction for fair classification. *arXiv preprint arXiv:2506.15620*, 2025.

Ki Hyun Tae, Hantian Zhang, Jaeyoung Park, Kexin Rong, and Steven Euijong Whang. Falcon: Fair active learning using multi-armed bandits. *Proceedings of the VLDB Endowment*, 17(5):952–965, 2024.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. FairSeg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yu Tian, Congcong Wen, Min Shi, Muhammad Muneeb Afzal, Hao Huang, Muhammad Osama Khan, Yan Luo, Yi Fang, and Mengyu Wang. FairDomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision (ECCV)*, 2024b.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sahil Verma, Michael Ernst, and Rene Just. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054*, 2021.

Artem Vysogorets, Kartik Ahuja, and Julia Kempe. DRoP: Distributionally robust data pruning. In *International Conference on Learning Representations (ICLR)*, 2025.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. FairPrune: Achieving fairness through pruning for dermatological disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 743–753, 2022.

Bobby Yan, Skyler Seto, and Nicholas Apostoloff. FORML: Learning to reweight data for fairness. *arXiv preprint arXiv:2202.01719*, 2022.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30(10):2838–2848, 2024.

Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. Deep learning on a healthy data diet: Finding important examples for fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14593–14601, 2023.

Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22484–22492, 2024.

Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, and Joey Tianyi Zhou. Spanning training progress: Temporal dual-depth scoring (TDDS) for enhanced dataset pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26223–26232, 2024.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5810–5818, 2017.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations (ICLR)*, 2023.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *International Conference on Learning Representations (ICLR)*, 2023.

A APPENDIX

How this appendix is organized. Appendix B gives the full experimental setup. Appendices C–E follow the order of Section 4: the audit (R1), the mechanism diagnostics D1–D3, and the repair (R2), each with the complete tables behind the main-text summaries. Appendix F repeats the audit and the repair on gender, ethnicity and language. Appendix G gives the threshold-parity view and Appendix H the FairCLIP study. Appendices I–K test how far the conclusions reach: other pruners, sample-level pruning, and hyperparameters. Appendix L is the extended related work. Throughout, as in the main text, all Δ values are seed-paired against the unpruned baseline (unless stated otherwise), in AUC points $\times 100$, reported as mean$\pm$std over three seeds, and a negative Δ on ES-AUC, worst-group AUC or AUC means degradation.

B EXPERIMENTAL SETUP

B.1 DATASETS AND REGIMES

We use six datasets from medical imaging and face analysis, chosen to vary in size and demographic imbalance, the two properties that Proposition 1 ties to the harm. Table 4 lists them. A *regime* is one (dataset, architecture, optimizer) training cell. Harvard-FairVLMed (Luo et al., 2024b) is run under a full 2×2 design (ResNet-50 or ViT-B/16, SGD or AdamW), which gives four regimes. Every other dataset is run once, under ResNet-50 with SGD. The nine regimes are FairVLMed/R50+SGD, FairVLMed/R50+AdamW, FairVLMed/ViT+AdamW, FairVLMed/ViT+SGD, FairVision-AMD, FairVision-DR, FairVision-Glaucoma, CheXpert and UTKFace.

Each dataset has one *primary* protected attribute. It is used for the headline fairness metrics and for the group rescaling (R). All other attributes are tracked as well: they enter the quota (Q), which checks every attribute at once, and the signal diagnostic D2 (Appendix D.3), and Appendix F reports the audit and repair on them. FairVLMed and the three FairVision tasks carry four attributes each (race, gender, ethnicity, language), CheXpert carries gender and age group, and UTKFace carries race and age group. This gives $7\times4 + 2\times2 = 32$ regime–attribute pairs.

Table 4: **The nine training regimes and their datasets.** A regime is one (dataset, architecture, optimizer) cell; the rightmost column lists each dataset's regimes, which sum to nine. K is the number of fixed batches, the pruning unit. The primary attribute is used for the headline metrics and for the rescaling (R). "Attrs" counts all tracked protected attributes, giving 32 regime–attribute pairs.

Dataset	Task	K	n_{test}	Primary attribute (groups)	Attrs	Regimes
Harvard-FairVLMed	glaucoma	54	2,000	race (Asian/Black/White)	4	4: R50+SGD, R50+AdamW, ViT+AdamW, ViT+SGD
FairVision-AMD	AMD	46	3,000	race (Asian/Black/White)	4	1: R50+SGD
FairVision-DR	diab. retinopathy	46	3,000	race (Asian/Black/White)	4	1: R50+SGD
FairVision-Glaucoma	glaucoma	46	3,000	race (Asian/Black/White)	4	1: R50+SGD
CheXpert	pleural effusion	1,220	44,991	gender (F/M)	2	1: R50+SGD
UTKFace	gender	129	4,741	race (5 groups)	2	1: R50+SGD
				totals	32	**9 regimes**

B.2 TRAINING PROTOCOL AND FIXED BATCHES

All models start from ImageNet-pretrained weights and train for 200 epochs with batch size 128 and `drop_last`. The learning rate follows cosine decay after a 10-epoch linear warmup. SGD runs use learning rate 0.15 (0.1 on CheXpert), Nesterov momentum 0.9 and weight decay 5×10^{-4}. AdamW runs use learning rate 10^{-4} and weight decay 0.05. Training uses 4-GPU distributed data parallel without mixed precision, on 224×224 images. Each configuration runs with seeds 42, 43 and 44. A seed fixes the initialization, the data order and the batch membership together. The checkpoint with the best validation AUC is evaluated once on the test set.

As described in Section 2, the training set is split once into K batches that stay fixed for the whole run, while the order of samples inside each batch is reshuffled every epoch. Fixed membership keeps each batch's group composition $n_{g,k}$ constant, so the retention curves $r_g(t)$ and exposures m_g of Section 3 are exact rather than estimated.

Algorithm 1 EquiPrune: one epoch t around the unmodified pruning signal

input: active set $\mathcal{A}$, pruning scores $\{s_k(t)\}$, threshold $\delta(t)$, floor ϕ, quota γ, clip c
$\mathcal{C} \leftarrow \{k \in \mathcal{A} : s_k(t) \le \delta(t)\}$, sorted most-stable-first ▷ B-PAS candidates, signal untouched
for all $k \in \mathcal{C}$ **do**
  if $|\mathcal{A}| \le \lceil \phi K \rceil$ **then break** ▷ **(F)** retention floor
  if $r_g^{(j)}(\mathcal{A} \setminus \{k\}) \ge \gamma\, \bar{r}(\mathcal{A} \setminus \{k\})$ for all $j,\ g$ **then** ▷ **(Q)** quota
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{k\}$ ▷ prune (permanent)
$w_g \leftarrow \mathrm{clip}\big(\bar{r}(\mathcal{A})/r_g(\mathcal{A}), 1, c\big)$ for each $g \in \mathcal{G}$ ▷ **(R)** rescaling
train one epoch on $\{\mathcal{B}_k\}_{k\in\mathcal{A}}$ with per-sample loss weights $w_{g(i)}(t)$

B.3 Pruner configurations

All three pruners run at batch level. In our code they are the batch ports InfoBatch and EL2N; we write InfoBatch and EL2N, as in the main text, and use EL2N-S only for the sample-level form in Appendix J.

B-PAS. The signal $s_k(t)$ is the epoch-over-epoch change in batch k's mean activation dispersion (the per-channel activation standard deviation, averaged over the monitored layers). A batch is retired for good once $s_k(t) < \delta(t)$. On ResNet backbones $\delta(t)$ rises exponentially from 4×10^{-5} to 2×10^{-3}. ViT activations live on a different scale, so on ViT backbones we use the scale-free relative mode, in which $\delta(t)$ is a rising fraction of the current epoch's median stability change. The threshold mode follows the architecture, not the optimizer.

InfoBatch. Each epoch, a batch whose loss score is below the mean of the current scores is skipped with probability 0.5, and the losses of the kept low-loss batches are scaled by $1/(1-0.5)$ so that the expected gradient is unchanged. Skipped batches keep their old score until they are trained again. Pruning is switched off for the last 12.5% of epochs, so all data returns at epoch 176.

EL2N. Each batch is scored by its mean error norm $\|p_i - y_i\|_2$ averaged over epochs 2–10, and the lowest-scoring batches are dropped once, at epoch $E = 10$. Unlike the other two pruners, EL2N's savings are a budget chosen in advance, not an outcome of the signal. We fix the target at $\mathrm{DSI} = 35\%$ for every dataset and regime, which is why its DSI is constant across the paper. The number of dropped batches follows from the target:

$$n_{\mathrm{drop}} \;=\; \Big\lceil \mathrm{DSI}_{\mathrm{target}} \cdot \frac{T\,K}{T-E} \Big\rfloor, \qquad \mathrm{DSI} \;=\; \frac{n_{\mathrm{drop}}\,(T-E)}{T\,K}, \tag{2}$$

where $\lceil\cdot\rfloor$ rounds to the nearest integer and $T = 200$. The factor $T/(T-E)$ pays for the scoring warmup: the first E epochs must train on all data, so saving 35% of the whole run means dropping a larger share of batches. On FairVLMed ($K = 54$) this gives $n_{\mathrm{drop}} = 20$, i.e. 37% of the batches for 35.19% DSI. Rounding makes the realized DSI differ slightly by dataset: 35.19% (K=54), 35.11% (K=46), 34.96% (K=1220) and 35.35% (K=129), and exactly 35.00% at sample level. The seed changes which batches are dropped but never how many, so EL2N's DSI has a standard deviation of zero. The target is not tuned; Appendix K sweeps it from 0.15 to 0.55.

B.4 EquiPrune configuration and cost

EquiPrune uses one fixed setting, $(\phi, \gamma, c) = (0.15, 0.9, 5)$, chosen before any experiment and shared by all datasets and pruners, with no per-dataset tuning. The quota (Q) checks every protected attribute of the dataset; the rescaling (R) uses the primary attribute. Algorithm 1 gives the per-epoch loop described in Section 3.

Cost. The guardrails add $O\big(\sum_j |\mathcal{G}^{(j)}|\big)$ counter updates per pruning decision and one multiplication per sample in the loss. There are no extra forward passes, no auxiliary model and no bilevel optimization.

Scope. Nothing in Algorithm 1 depends on activation stability. F and the quota test apply as written to any pruner that removes items in the order of a utility score; B-PAS is the flagship case, and Appendix I applies the same wrapper to EL2N and InfoBatch. When demographic labels are not

available, the floor F still works on its own, and the ablation of Section 4 shows it is the most reliable of the three guardrails.

B.5 METRICS AND STATISTICAL TESTING

Metrics. We report overall AUC, per-group AUCs $\{\mathrm{AUC}_g\}$, worst-group AUC $\min_g \mathrm{AUC}_g$, and the equity-scaled AUC of Section 2,

$$ES\text{-}AUC = \frac{\mathrm{AUC}_{\text{overall}}}{1 + \sum_g \left|\mathrm{AUC}_g - \mathrm{AUC}_{\text{overall}}\right|}.$$

We also report the demographic-parity distance (DPD) and the equalized-odds distance (DEOdds). In the main campaign both are read at decision threshold 0.5, and DEOdds is the sum of the true-positive-rate and false-positive-rate spreads across groups. In the FairCLIP study (Appendix H) DEOdds follows the FairCLIP convention, the larger of the two spreads. "Unknown" groups are excluded from every gap and from the worst group, but their AUCs are still computed. Δworst-group AUC is $\min_g \mathrm{AUC}_g^{\text{pruned}} - \min_g \mathrm{AUC}_g^{\text{base}}$, so the group attaining the minimum may differ between the two runs (Section 2). All metrics are computed on the held-out test set.

How a Δ and its $\pm$ are computed. Every Δ in this paper is formed by differencing *before* averaging. Fix a regime and a method m. For each seed $s \in \{42, 43, 44\}$ we take the two runs that share that seed (same initialization, same batch partition, same schedule, differing only in which data is skipped) and form the within-seed difference

$$\Delta_s(m) \;=\; \text{ES-AUC}_s(m) \;-\; \text{ES-AUC}_s(\text{baseline}). \tag{3}$$

The table entry is the mean of those $n = 3$ numbers, and the $\pm$ is their *sample* standard deviation:

$$\bar{\Delta} \;=\; \frac{1}{n}\sum_s \Delta_s, \qquad s_\Delta \;=\; \sqrt{\frac{1}{n-1}\sum_s \big(\Delta_s - \bar{\Delta}\big)^2}. \tag{4}$$

Worked example. For B-PAS on FairVLMed/R50+SGD, the largest harm in Table 1, the per-seed ES-AUC values are $73.88/78.43/76.66$ for the baseline and $72.22/72.71/71.80$ for B-PAS. This gives $\Delta_s = -1.66, -5.72, -4.86$, with mean -4.08. The deviations from the mean are $+2.42, -1.64, -0.78$, so $s_\Delta = \sqrt{(2.42^2 + 1.64^2 + 0.78^2)/2} = 2.14$ and the entry reads $-4.08_{\pm 2.14}$.

Why pairing matters. A seed that trains a strong baseline tends to train a strong pruned model too (here $r = +0.42$). Combining the two arms' own standard deviations, 2.29 for the baseline and 0.45 for B-PAS (Table 19), would give $\sqrt{2.29^2 + 0.45^2} = 2.34$. The paired value is 2.14; the difference is exactly the $-2r\, s_a s_b$ term. Pairing removes the shared seed-to-seed offset, gives the tighter and more honest interval, and matches the paired t-tests. Columns that are not differences (raw ES-AUC, AUC, worst-group AUC and DSI) are the mean and sample standard deviation of the three per-seed values.

Reference runs. The reference is the same-seed unpruned baseline in Tables 1 and 2 and in every companion table in this appendix. It is the same-seed *batch-level* baseline for both blocks of Table 19, which is why its sample-level spreads are wider. It is the wrapped pruner itself, not the baseline, in the generality study of Appendix I, so that the selection signal is held fixed.

Significance. Stars are one-sample paired t-tests of the three Δ_s against zero ($^*p < 0.05$, $^{**}p < 0.01$). With $n = 3$ these tests have little power, so they are descriptive. Our conclusions rest on the pre-registered effect-size rules of Section 4, applied the same way in every regime. For the same reason a three-seed s_Δ is itself noisy, and the sensitivity study reads effects against a pooled noise floor, $\sigma_\Delta = 2.20$ (Appendix K).

C THE AUDIT (R1): FULL RESULTS

This appendix gives the full reading of the audit summarized in Section 4 (Table 1), adds the average-AUC view of the same runs, and identifies the group that worst-group AUC lands on.

C.1 READING THE AUDIT TABLE

Attribution and size of the harm. Within a regime, everything except the choice of skipped data is held fixed, so every entry of Table 1 is caused by pruning. The largest harm, B-PAS on FairVLMed/R50+SGD (-4.08 ± 2.14 ES-AUC and -3.92 ± 2.37 worst-group AUC at 67% DSI), is about a sixth of the distance between our baseline ES-AUC and chance. The most consistent harms across seeds are on FairVision-Glaucoma, where B-PAS costs -3.26 ± 0.5 (p=0.008) and EL2N -2.16 ± 0.5 (p=0.021), and on FairVLMed/R50+AdamW, where EL2N costs -2.77 ± 0.6 (p=0.015).

Harm by pruner. Over the eight regimes outside the underfit control cell, B-PAS averages -1.64 ES-AUC and -1.37 worst-group AUC, against -0.92 and -0.51 for EL2N and -0.53 and -0.17 for InfoBatch. B-PAS is negative on ES-AUC in seven of the eight and meets the pre-registered harm rule in six regimes, against five for each of the other two (sixteen harmful cells in total). The two permanent pruners lead and the temporary pruner, which returns all skipped data before training ends, trails. Part of this is dose, since B-PAS also removes the most. We do not treat dose as a confound: a practitioner does not choose between equal savings, they choose the largest savings at the lowest cost, and the fairness cost comes with them. This is why B-PAS is the pruner most worth repairing.

The harm is not tied to one signal. The three regimes that reject R1 do so with at most one pruner meeting the harm rule: none on CheXpert or UTKFace, and only B-PAS on FairVision-AMD. The three pruners rank batches by different quantities (activation stability, loss, error norm), so a selection pathology would be signal-specific. The harm instead follows how much each pruner removes and whether it gives any of it back.

The safe regimes are the large ones. CheXpert ($|\Delta| \leq 0.21$ everywhere) and UTKFace have no harmful pruner. Their safety comes from the data, not from gentler pruning: their retention falls as far and as fast as elsewhere (Table 10). Under Proposition 1 the risk multiplier grows as a group's remaining data shrinks, so a group with ample data absorbs a 60% cut that a group with a few hundred samples cannot (prediction P1). FairVision-AMD is not in this category: it is small and imbalanced, and B-PAS does harm it, one pruner short of the R1 rule.

The ViT+SGD control cell. FairVLMed/ViT+SGD trains ViT-B/16 with plain SGD at learning rate 0.15 and underfits (baseline AUC 69.3 ± 4.0 against about 80 in the other FairVLMed cells), which inflates every standard deviation in its row. We keep it because it completes the 2×2 design, which shows that the harm belongs to neither the architecture nor the optimizer alone. No conclusion rests on this row, it is left out of the pruner averages above, and it is marked wherever it appears.

C.2 AVERAGE AUC UNDERSTATES THE HARM

Efficiency methods are usually certified on average accuracy. Table 5 scores the same 27 audited cells as Table 1 on overall AUC, and Table 6 puts the three metrics side by side.

Average AUC moves less than the fairness metrics. Over the 16 harmful cells, overall AUC falls by 1.63 points on average, against 2.32 for ES-AUC and 2.11 for worst-group AUC. In 13 of the 16 cells the AUC loss is the smallest of the three. The three exceptions are B-PAS on FairVision-AMD (-1.77 AUC against -1.23 ES-AUC) and EL2N and InfoBatch in the underfit ViT+SGD cell, where the whole model degrades and every metric falls with it.

An average-AUC audit would miss a quarter of the harm. Applying the same -0.5 rule to overall AUC alone flags 12 of the 16 harmful cells. It misses four (marked §): B-PAS on FairVision-DR (-0.09 AUC against -1.36 on both fairness metrics), EL2N on FairVision-Glaucoma (-0.30 AUC against -2.16 ES-AUC, one of the most consistent harms in the audit, p=0.021), InfoBatch on FairVLMed/R50+SGD (-0.25 against -0.84), and B-PAS on FairVLMed/ViT+AdamW, where AUC even rises ($+0.43$) while ES-AUC falls by 1.02. It also raises one false alarm, EL2N on FairVision-AMD (-1.20 AUC while ES-AUC gains 1.52). The two kinds of metric are related (Spearman $\rho = 0.67$ between ΔAUC and ΔES-AUC over the 24 non-control cells) but they are not the same measurement.

The gap is largest for the pruner that removes the most. For B-PAS the fairness losses are 1.5 to 1.8 times the AUC loss (-1.64 ES-AUC and -1.37 worst-group against -0.89 AUC). For EL2N

Table 5: **Average AUC of the audit**, the companion to Table 1 over the same 27 (regime, pruner) cells. "Base" is the unpruned baseline AUC; the other columns are seed-paired ΔAUC against it (mean±std, 3 seeds). Shaded rows support R1 in Table 1. §A cell that is harmful under R1 on ES-AUC or worst-group AUC but would pass an average-AUC audit with the same -0.5 rule. †underfit control cell. $^*p < 0.05$ (paired t, n=3, descriptive).

		ΔAUC (overall)		
Regime	Base	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	80.41	$-3.35\pm0.8^*$	$-0.25\pm1.8^§$	-1.01 ± 0.6
FairVLMed / R50+AdamW	79.39	-1.04 ± 1.0	$+0.60\pm1.2$	-0.79 ± 0.6
FairVLMed / ViT+AdamW	80.67	$+0.43\pm0.8^§$	-0.70 ± 0.4	$+0.30\pm0.3$
FairVLMed / ViT+SGD†	69.28	$+0.22\pm10.4$	-2.97 ± 13.6	-7.31 ± 4.0
FairVision-AMD	85.26	-1.77 ± 1.1	-0.37 ± 1.0	$-1.20\pm0.3^*$
FairVision-DR	78.56	$-0.09\pm0.7^§$	$-2.42\pm0.7^*$	-2.28 ± 3.0
FairVision-Glaucoma	78.98	-1.07 ± 1.0	-1.14 ± 1.1	$-0.30\pm1.0^§$
CheXpert	86.29	$+0.02\pm0.2$	-0.23 ± 0.2	-0.18 ± 0.3
UTKFace	97.58	-0.28 ± 0.3	$+0.19\pm0.1^*$	-0.13 ± 0.1

Table 6: **Average AUC against the two fairness metrics.** Mean seed-paired Δ per pruner over the eight regimes outside the ViT+SGD control cell, and over the 16 audited cells that are harmful under R1. The last row counts, among those 16 cells, how many a -0.5 rule on each metric alone would flag.

	Mean over 8 regimes			
Metric	B-PAS	InfoBatch	EL2N	16 harmful cells
ΔAUC (overall)	-0.89	-0.54	-0.70	-1.63
ΔES-AUC	-1.64	-0.53	-0.92	-2.32
Δworst-group AUC	-1.37	-0.17	-0.51	-2.11
Harmful cells flagged by this metric alone				AUC 12, ES-AUC 16, worst 15

ES-AUC again falls more than AUC (-0.92 against -0.70), while its worst-group loss is smaller (-0.51). For InfoBatch, which returns all skipped data, AUC and ES-AUC move together (-0.54 and -0.53) and the worst group barely moves. So average AUC is not always the gentler reading, but it is clearly gentler for the permanent pruners, and most of all for B-PAS.

Why. Overall AUC is dominated by the majority group, while the worst group is a minority holding 7–15% of the test set (Appendix C.3). A large loss on a small group moves the overall AUC only a little. On FairVLMed/R50+SGD, B-PAS costs the Black group 5.07 points, the Asian group 4.1 and the White majority 3.29 (Figure 3). Overall AUC falls by 3.35, close to the loss of the majority, which holds 77% of the test set, while the widening gaps pull ES-AUC down by 4.08. ES-AUC falls both when AUC falls and when group gaps widen, so it registers exactly the part of the damage that the average hides.

C.3 Who the worst group is

Worst-group AUC is a fairness statistic only if the group it lands on is one fairness cares about, and nothing in $\min_g \mathrm{AUC}_g$ guarantees that. Figure 4 and Table 7 check it over the 168 runs on FairVLMed and FairVision, the seven imbalanced regimes and the ones where the audit finds harm. For each run we take the primary attribute (race throughout), find the worst group $\arg\min_g \mathrm{AUC}_g$, and read off its share of the test set. Shares come from the per-sample labels stored with each run's test predictions, so they are the exact evaluation-set composition. "Unknown" samples are excluded, as everywhere else.

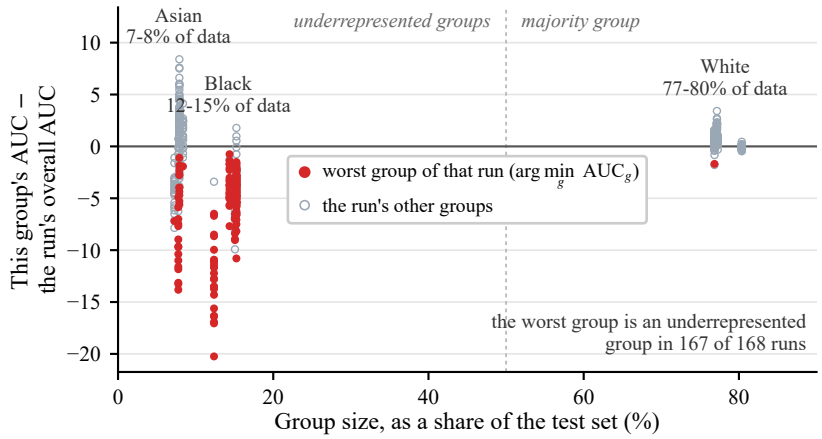


Figure 4: **The worst group is an underrepresented minority group.** Each point is one racial group in one run: its share of the test set against its AUC relative to that run's overall AUC. Filled points are the run's worst group, $\arg\min_g \mathrm{AUC}_g$; open points are the other two groups. All 504 points from the 168 FairVLMed and FairVision runs are shown. Every filled point but one belongs to Asian or Black patients, who hold 7–8% and 12–15% of the data. The single filled point on the majority side is one EL2N run in the underfit ViT+SGD control cell, where the model is near chance on the White group itself.

Table 7: **The worst group is an underrepresented group in every regime.** For each FairVLMed and FairVision regime: the largest group and its test-set share, and the group attaining $\min_g \mathrm{AUC}_g$ over that regime's 24 runs (8 methods $\times$ 3 seeds), with how often each group is the worst. ΔAUC is that group's mean $\mathrm{AUC}_g - \mathrm{AUC}_{\mathrm{overall}}$ over the same runs. The attribute is race throughout. The one appearance of the majority group is a single EL2N run in the underfit ViT+SGD cell.

Regime	Largest group (share)	ΔAUC	Worst group over 24 runs	ΔAUC
FairVLMed / R50+SGD	White (76.8%)	+0.71	Black (15.2%) 15/24; Asian (7.9%) 9/24	−3.64
FairVLMed / R50+AdamW	White (76.8%)	+0.29	Black (15.2%) 24/24	−3.70
FairVLMed / ViT+AdamW	White (76.8%)	+0.79	Black (15.2%) 24/24	−4.87
FairVLMed / ViT+SGD	White (76.8%)	+0.56	Black (15.2%) 21/24; Asian (7.9%) 2/24; White (76.8%) 1/24	−4.40
FairVision-AMD	White (80.4%)	−0.08	Black (12.4%) 23/24; Asian (7.3%) 1/24	−12.47
FairVision-DR	White (77.2%)	+1.97	Asian (7.8%) 15/24; Black (15.1%) 9/24	−8.77
FairVision-Glaucoma	White (77.3%)	+0.45	Black (14.4%) 23/24; Asian (8.4%) 1/24	−3.73

The worst group is a minority. It is Black in 139 of the 168 runs, Asian in 28 and White in 1. The two minority groups hold 7.3–8.4% and 12.4–15.2% of each test set, the White majority 76.8–80.4%. So in 167 of 168 runs the group that defines worst-group AUC is one of the two groups with the least data, and in none of the 21 unpruned baseline runs is it the majority. The separation is wide: averaged over all runs, the Black group sits 5.47 AUC points below its run's overall AUC and the Asian group 0.56 below, while the White group sits 0.67 above. The single exception is one EL2N run in the ViT+SGD control cell, where the model is close to chance on the majority group itself (AUC 53.7).

Two things follow. Every worst-group entry in Tables 1 and 2 is a measurement on Black or Asian patients, so both the damage and the recovery land on the groups with the least representation. And worst-group AUC and ES-AUC agree on who is at stake, which is the agreement the threshold-parity metrics fail to give (Appendix G).

The worst group is not simply the rarest group. We claim that the worst group is underrepresented, not that harm is monotone in group size. Black patients, at 12–15%, hold the position far more often than Asian patients at 7–8%: in all 24 runs of two regimes and in 23 of 24 in two more. Only on FairVision-DR does the rarest group hold it more often than not (15 of 24). Over the

504 group observations, the rank correlation between a group's share and its AUC deviation is only $+0.19$. Representation predicts which side of the majority–minority divide the worst group falls on; it does not rank the groups within the minority side, and Proposition 1 does not require it to.

D THE MECHANISM: DIAGNOSTICS D1–D3 IN FULL

Section 3 defines three diagnostics, one for each prediction of the biased-selection hypothesis H_{sel}, and Section 4 reports that all three fail. This appendix gives the design details, every per-cell value, the statistical power of each test, and what the diagnostics can and cannot establish.

D.1 WHAT THE DIAGNOSTICS MEASURE

The composition variable. For an attribute, π_k is the minority fraction of batch k: the share of its samples outside that attribute's majority group. Batch membership is fixed, so π_k is a constant of the run. H_{sel} says that a batch-level score, being dominated by the batch's majority samples, can call a batch "learned" while its minority samples are still being learned. If that happened systematically, a batch's fate would depend on π_k, and the three diagnostics look for that dependence at three different points.

- **D1 (prune time)** asks whether minority-heavy batches are *removed first*. It needs a removal epoch, which only the two permanent pruners have.
- **D2 (signal and composition)** asks whether the *score itself* marks minority-heavy batches as prunable, before any removal happens.
- **D3 (retention)** asks whether the *result* is uneven: whether some group keeps less of its data, at any epoch or on average.

D1 and D2 test whether the pruner chose against a group; D3 tests whether the outcome was uneven even without such a choice. H_{sel} needs only one of them to hold, so it is rejected only if all three fail.

Tests and power. D1 and D2 are Spearman rank correlations over batches, pooled across the three seeds (each seed draws its own partition). This gives $n = 162$ batches on FairVLMed ($K = 54$), 138 on each FairVision task ($K = 46$), 387 on UTKFace and 3,660 on CheXpert. At $n = 138$–162, a two-sided test at $\alpha = 0.05$ detects a correlation of about $|\rho| \approx 0.16$–0.17; on UTKFace and CheXpert the detectable size is about 0.10 and 0.03. A selection bias of moderate size would therefore show up. D2 runs 32 tests, one per regime–attribute pair, and with that many tests a few would reach $p < 0.05$ by chance alone even if no effect existed. We therefore also apply the Bonferroni correction, which divides the significance level by the number of tests: a D2 correlation counts as significant only if $p < 0.05/32 \approx 0.0016$. This keeps the chance of even one false positive across all 32 tests below 5%.

A caveat on range. Each batch is a random cross-section of the data, so π_k varies over a narrow range: 0.12–0.31 for race on the four medical datasets (median 0.20–0.23), 0.27–0.57 for gender on CheXpert, and 0.44–0.69 for race on UTKFace. A narrow range limits how strongly composition could steer any batch-level score, which cuts both ways: it is part of why batch-level pruning deletes evenly, and it is why we repeat the audit with single-image units in Appendix J.

D.2 D1: REMOVAL TIME

ρ_{time} correlates π_k (primary attribute) with the epoch at which batch k is permanently removed. Batches that survive to the end are ranked after every removed batch (censored at $T + 1$). H_{sel} predicts $\rho_{\text{time}} < 0$: minority-heavy batches leave first. For B-PAS the removal epoch is the retirement epoch. For EL2N all removals happen at epoch 11, so ρ_{time} reduces to asking whether the dropped batches (20 of 54 on FairVLMed) are more minority-heavy than the kept ones.

Result. Table 8 (left) gives all 18 cells. Every correlation satisfies $|\rho_{\text{time}}| \leq 0.12$, below the detectable size on the medical datasets, and none is significant even before any correction (smallest $p = 0.090$, B-PAS on UTKFace). The signs are split: seven of the 18 are negative and the rest are zero or positive. The two largest values ($+0.12$, B-PAS on FairVLMed/R50+SGD and EL2N on

Table 8: **D1 and the other two pruning signals.** Left: D1, the Spearman correlation between a batch's minority fraction π_k (primary attribute) and its removal epoch, for the two permanent pruners (survivors censored at $T + 1$; batches pooled over 3 seeds). H_{sel} predicts a *negative* value. Right: the D2 test applied to the signals of InfoBatch and EL2N on the unpruned baseline runs, the correlation between π_k and the early-training loss and the EL2N score (epochs 2–10). A *positive* value means minority-heavy batches have higher loss or error, so both pruners keep them longer. p-values in parentheses; none is below 0.05. †underfit control cell.

	D1: $\rho(\pi_k,$ removal epoch)		Other signals: $\rho(\pi_k, \cdot)$ on baseline runs	
Regime	B-PAS	EL2N	early loss (InfoBatch)	EL2N score
FairVLMed / R50+SGD	+0.12 (0.117)	0.00 (0.993)	+0.07 (0.404)	+0.09 (0.280)
FairVLMed / R50+AdamW	−0.01 (0.857)	+0.01 (0.915)	+0.02 (0.791)	+0.03 (0.692)
FairVLMed / ViT+AdamW	−0.08 (0.292)	−0.02 (0.821)	−0.01 (0.947)	+0.06 (0.422)
FairVLMed / ViT+SGD†	−0.01 (0.919)	+0.12 (0.118)	−0.02 (0.781)	+0.12 (0.127)
FairVision-AMD	−0.08 (0.376)	−0.03 (0.751)	−0.10 (0.231)	0.00 (0.981)
FairVision-DR	+0.07 (0.448)	+0.11 (0.211)	+0.16 (0.060)	+0.09 (0.269)
FairVision-Glaucoma	+0.06 (0.506)	+0.06 (0.519)	−0.03 (0.720)	−0.02 (0.793)
CheXpert	+0.00 (0.992)	+0.00 (0.931)	−0.01 (0.706)	−0.01 (0.477)
UTKFace	+0.09 (0.090)	−0.05 (0.320)	−0.01 (0.798)	−0.02 (0.736)

Table 9: **D2 for every protected attribute.** Spearman correlation between a batch's minority fraction π_k for that attribute and its run-averaged B-PAS stability score, on unpruned baseline runs (batches pooled over 3 seeds), with p in parentheses. H_{sel} predicts a *negative* value (minority-heavy batches look more stable and are retired sooner). Blank cells: attribute not recorded for that dataset. $^*p < 0.05$ before correction; no value survives Bonferroni correction over the 32 tests ($\alpha = 0.0016$). †underfit control cell.

Regime	race	gender	ethnicity	language	age group
FairVLMed / R50+SGD	+0.07 (0.394)	+0.01 (0.932)	−0.04 (0.594)	−0.10 (0.224)	
FairVLMed / R50+AdamW	+0.02 (0.793)	−0.01 (0.888)	+0.04 (0.594)	+0.04 (0.611)	
FairVLMed / ViT+AdamW	+0.19* (0.014)	+0.02 (0.753)	−0.09 (0.263)	+0.02 (0.789)	
FairVLMed / ViT+SGD†	−0.01 (0.852)	+0.04 (0.650)	+0.05 (0.552)	+0.01 (0.930)	
FairVision-AMD	+0.18* (0.034)	−0.02 (0.814)	+0.11 (0.216)	+0.10 (0.234)	
FairVision-DR	+0.10 (0.263)	−0.07 (0.433)	−0.07 (0.411)	+0.10 (0.228)	
FairVision-Glaucoma	+0.04 (0.646)	+0.02 (0.838)	−0.06 (0.512)	−0.02 (0.827)	
CheXpert		+0.01 (0.460)			0.00 (0.902)
UTKFace	+0.01 (0.813)				+0.01 (0.820)

ViT+SGD) are positive, so if anything minority-heavy batches left *later*. Removal order does not depend on batch composition.

D.3 D2: DOES THE SIGNAL TRACK COMPOSITION?

ρ_{sig} correlates π_k with the B-PAS stability score s_k averaged over the run. It is measured on *unpruned* baseline runs, so that the score is not shaped by the removals it would itself cause. A lower score means a batch looks more stable and is retired sooner, so H_{sel} predicts $\rho_{sig} < 0$. Because the quota (Q) protects every attribute, D2 is run on every attribute of every regime, 32 pairs in total.

Result. Table 9 gives all 32 pairs. The largest correlation is $|\rho_{sig}| = 0.19$ and the median is 0.04. Only two pairs reach $p < 0.05$: race on FairVLMed/ViT+AdamW (+0.19, p=0.014) and race on FairVision-AMD (+0.18, p=0.034). Neither survives Bonferroni correction, and both are *positive*: minority-heavy batches look *less* stable and would be retired later, the opposite of what H_{sel} needs. The only composition signal in the data is therefore mildly protective. Negative values are fewer and small: ten of the 32, none beyond −0.10.

Table 10: **D3 for all 27 audited cells** (primary attribute, mean over 3 seeds). $\bar{m}$ is the mean exposure ($= 1-$DSI). "Exposure gap" is $\min_g m_g - \bar{m}$, how far the least-exposed group sits below the mean over the run. "Pointwise gap" is the largest value of $\max_g r_g(t) - \min_g r_g(t)$ over all epochs of the seed-averaged curves (panel d of Figure 2). $\mathrm{H}_{\mathrm{sel}}$ predicts clearly negative exposure gaps for minority groups; $\mathrm{H}_{\mathrm{opt}}$ predicts both gaps near zero. †underfit control cell.

	Mean exposure $\bar{m}$			Exposure gap $\min_g m_g - \bar{m}$			Pointwise gap (max over t)		
Regime	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	0.330	0.731	0.649	−0.004	−0.001	−0.006	0.027	0.043	0.006
FairVLMed / R50+AdamW	0.275	0.724	0.649	−0.002	−0.002	−0.015	0.035	0.039	0.016
FairVLMed / ViT+AdamW	0.292	0.733	0.648	−0.009	−0.002	−0.003	0.030	0.040	0.003
FairVLMed / ViT+SGD†	0.418	0.761	0.654	−0.002	−0.009	−0.018	0.023	0.036	0.019
FairVision-AMD	0.357	0.725	0.648	−0.003	−0.001	−0.008	0.030	0.052	0.008
FairVision-DR	0.346	0.711	0.651	−0.004	−0.001	−0.017	0.026	0.037	0.018
FairVision-Glaucoma	0.339	0.729	0.650	−0.003	−0.002	−0.008	0.029	0.046	0.008
CheXpert	0.322	0.737	0.650	−0.000	−0.000	−0.000	0.002	0.004	0.000
UTKFace	0.476	0.718	0.645	−0.001	−0.001	−0.012	0.015	0.026	0.013

The other two signals. D2 in the main text tests the B-PAS score. Table 8 (right) applies the same test to the quantities the other two pruners rank by: the early-training loss, which drives InfoBatch's skipping, and the EL2N score. Both are measured on the unpruned baseline runs for the primary attribute. Both pruners drop the *lowest*-scoring batches, so $\mathrm{H}_{\mathrm{sel}}$ would need a negative correlation. None of the 18 values reaches $p < 0.05$; the largest is $+0.16$ (early loss on FairVision-DR, p=0.060), again in the protective direction. None of the three signals singles out minority-heavy batches, which matches the observation in Appendix C.1 that harm does not depend on which signal is used.

D.4 D3: RETENTION AND EXPOSURE

D3 looks at the outcome of the removals rather than at the choices. It uses two statistics, because they can disagree. The *exposure gap* $m_g - \bar{m}$ compares each group's exposure over the whole run with the mean. The *pointwise gap* $\max_g r_g(t) - \min_g r_g(t)$ compares the best- and worst-retained group at each epoch, and so also catches a group that is starved early and compensated late. A pointwise gap of 0.05 means that two groups differ by five samples in every hundred at that epoch. Exposure is averaged over the full 200-epoch budget, counting epochs with an empty active set as zero retention, so that $\bar{m} = 1 - \mathrm{DSI}$.

Result. Table 10 gives all 27 cells. The least-exposed group is never more than 0.018 below the mean ($|m_g - \bar{m}| \leq 0.018$), while the mean exposure itself ranges from 0.28 to 0.76. The pointwise gap never exceeds 0.052 on the seed-averaged curves, and never exceeds 0.097 on any single run. Each pruner leaves its own signature. InfoBatch has the largest pointwise gaps (0.026–0.052 outside CheXpert), because it skips batches at random every epoch, but small exposure gaps (≤ 0.009), because that randomness averages out over the run. EL2N is the reverse: its retention is flat after a single drop, so its pointwise gaps are the smallest (≤ 0.019), but that one draw fixes the composition for the rest of the run, so it has the largest exposure gaps (up to 0.018), still under two points. B-PAS has intermediate pointwise gaps (0.015–0.035) and small exposure gaps (≤ 0.009). On CheXpert, with 1,220 batches, both gaps are essentially zero.

No consistent direction. Under $\mathrm{H}_{\mathrm{sel}}$ the least-retained group should be a minority. Under B-PAS it is not: it is the White majority on FairVLMed/R50+SGD and ViT+SGD, Black on R50+AdamW, Asian on ViT+AdamW and the three FairVision tasks, Male on CheXpert and White on UTKFace. Which group ends up slightly behind changes from cell to cell, as expected from noise.

Other attributes. On the non-primary attributes the pointwise gap stays below 0.10 in 260 of the 288 run–attribute pairs of the audit. The 28 exceptions are language (25) and ethnicity (3), whose smallest groups are so small that moving a single batch shifts their retained fraction a long way.

Retention schedules in detail. On FairVLMed/R50+SGD (Figure 2a–c), B-PAS decays steadily, crosses 0.5 near epoch 70 and reaches zero between epochs 101 and 118 depending on the seed, with

$m_{\text{Asian}} = 0.332$, $m_{\text{Black}} = 0.330$, $m_{\text{White}} = 0.328$ and a largest pointwise gap of 0.027. InfoBatch oscillates between 0.55 and 0.80 and restores the full data at epoch 176, leaving all exposures near 0.73 and a gap of 0.043 that behaves like noise rather than a trend. EL2N drops once at epoch 11 to about 0.63 and stays flat, with exposures of 0.652, 0.646 and 0.648 and a gap of 0.006.

D.5 What the diagnostics show, and what they do not

$\mathrm{H}_{\mathrm{sel}}$ is rejected. All three predictions fail in every cell: removal order ignores composition (D1), none of the three pruning signals prefers minority-heavy batches (D2), and every group keeps the same share of its data to within two points (D3). The only statistically notable composition effects point the protective way.

$\mathrm{H}_{\mathrm{opt}}$ is what remains, with a stated scope. $\mathrm{H}_{\mathrm{opt}}$ survives because the alternative fails, and it matches the observed incidence: equal exposure, unequal outcomes, with the loss on minority groups (Appendix C.3). Two further results support it from independent directions. Replaying B-PAS's exact drop schedule on randomly chosen batches is no less harmful than B-PAS itself (Appendix K), and the demographics-blind floor F alone is the most reliable repair (Section 4). But the path from lost exposure to lost subgroup AUC is not measured directly, and the amount of exposure is not the only thing that matters. On FairVLMed/R50+AdamW, EL2N keeps $2.4\times$ more data than B-PAS ($\bar{m} = 0.65$ against 0.28) and still costs more (-2.77 against -1.88 ES-AUC), because it removes its share at epoch 11 and the remaining 95% of training runs without it. How much is removed is first-order, when it is removed second-order, and from whom neither.

Limits of a batch-level test. The diagnostics are exact at batch level, but the narrow range of π_k (Appendix D.1) means they cannot rule out a selection effect at a finer grain. Pruning individual images, where each unit belongs to one group, is harsher (Appendix J), so the batch-level harm reported here is best read as a lower bound.

E The repair (R2): full results

This appendix gives the full reading of the repair summarized in Section 4 (Table 2). Table 2 reports ES-AUC; the R2 rule also requires worst-group AUC, which Table 11 reports.

E.1 Reading the repair table

The worst-hit regime. On FairVLMed/R50+SGD, B-PAS loses 4.08 ES-AUC and 3.92 worst-group AUC while saving 67% of the data. The full EquiPrune (F+Q+R) brings the ES-AUC loss down to -0.59 ± 0.7 at 56% DSI, and the floor alone recovers it fully ($+0.22 \pm 2.2$). Most of the savings survive (56% against 67% DSI), and so does accuracy: overall AUC goes from 77.07 ± 1.01 under B-PAS to 80.72 ± 0.75, against 80.41 ± 0.69 unpruned. In this regime the guardrails give back what the pruner took, in fairness and in average accuracy, and still deliver more than half the data savings.

Roles of the three guardrails. F is the only guardrail that uses no demographic information, so we read it first. Q and R only ever run on top of it, since neither means much once the active set has collapsed. Q acts on which data is kept, R on how much the kept data counts. Counting the arms that meet the R2 rule on both ES-AUC and worst-group AUC (bold in Table 11), the floor alone passes in five regimes, F+Q+R in five, F+Q in four and F+R in three. The floor already recovers much of the loss, so bounding how much is removed is the main lever, as $\mathrm{H}_{\mathrm{opt}}$ predicts. Q adds value where the floor falls short (FairVision-DR and UTKFace) without driving the overall effect. In practice Q is rarely binding at batch level: with $\gamma = 0.9$ it vetoes 3–17 B-PAS candidates per run, against about 500 vetoes by the floor on FairVLMed/R50+SGD (Appendix I).

Where no arm passes. On FairVision-Glaucoma the arms order by how much demographic machinery they carry: floor only -1.81, F+Q -2.87, F+R -3.09 and all three -4.31 ES-AUC. The floor alone recovers about half of B-PAS's -3.26, and each group-aware part added on top gives some of that back. This fits $\mathrm{H}_{\mathrm{opt}}$: the floor addresses starvation, while Q and R guard against a selection effect that the audit did not find. FairVLMed/R50+AdamW also fails R2 for every arm, but only on ES-AUC (Table 11 shows its worst-group AUC is held). In both regimes the default

Table 11: **Worst-group AUC of the repair**, the companion to Table 2 for the same four EquiPrune arms in the same nine regimes. Seed-paired Δworst-group AUC against the unpruned baseline (mean±std, 3 seeds); DSI is as in Table 2. **Bold** = the arm meets the full R2 rule (DSI $\geq 25\%$ and both ΔES-AUC and Δworst-group AUC ≥ -0.5). The rightmost column is the most negative Δworst-group AUC among the three audited pruners (Table 1). Shaded = no arm passes. †underfit control cell. $^{**}p < 0.01$ (paired t, n=3, descriptive).

Regime	EquiPrune (F+Q+R)	-Q (F+Q)	-R (F+R)	-F (floor only)	worst audited
FairVLMed / R50+SGD	$-0.98_{\pm 1.3}$	$-2.46_{\pm 2.2}$	$-0.69_{\pm 3.4}$	$\mathbf{+0.28}_{\pm 1.0}$	-3.92
FairVLMed / R50+AdamW	$-0.13_{\pm 2.5}$	$+0.08_{\pm 3.6}$	$+0.09_{\pm 1.9}$	$-0.59_{\pm 3.6}$	-2.25
FairVLMed / ViT+AdamW	$-0.28_{\pm 2.6}$	$-0.42_{\pm 2.5}$	$-0.59_{\pm 1.5}$	$\mathbf{-0.44}_{\pm 1.9}$	-1.19
FairVLMed / ViT+SGD†	$\mathbf{+5.27}_{\pm 6.9}$	$-2.27_{\pm 14.4}$	$\mathbf{+3.04}_{\pm 6.0}$	$\mathbf{-0.34}_{\pm 17.5}$	-6.14
FairVision-AMD	$\mathbf{+2.41}_{\pm 3.3}$	$\mathbf{+4.21}_{\pm 4.9}$	$\mathbf{+2.37}_{\pm 3.8}$	$\mathbf{+2.17}_{\pm 5.9}$	-0.90
FairVision-DR	$\mathbf{+0.73}_{\pm 5.1}$	$\mathbf{+0.31}_{\pm 6.1}$	$-1.64_{\pm 3.4}$	$-0.77_{\pm 5.7}$	-3.64
FairVision-Glaucoma	$-4.15_{\pm 0.3}^{**}$	$-2.59_{\pm 1.8}$	$-2.55_{\pm 3.5}$	$-1.71_{\pm 1.6}$	-3.39
CheXpert	$\mathbf{-0.19}_{\pm 0.3}$	$\mathbf{-0.20}_{\pm 0.4}$	$\mathbf{-0.10}_{\pm 0.2}$	$\mathbf{-0.03}_{\pm 0.3}$	-0.21
UTKFace	$\mathbf{+0.60}_{\pm 1.0}$	$\mathbf{+0.37}_{\pm 0.7}$	$-0.43_{\pm 0.6}$	$-0.29_{\pm 0.4}$	-0.12

floor seems to leave too little data. The sensitivity study (Appendix K) shows the floor-only arm gaining ES-AUC as ϕ rises, at a cost in savings, so a larger ϕ is the natural remedy there.

E.2 Worst-group AUC of the repair

The best arm protects the worst group in every regime. In all nine regimes the best arm's worst-group change is above the worst audited pruner's: on FairVLMed/R50+SGD $+0.28$ against -3.92, on FairVision-DR $+0.73$ against -3.64, on FairVLMed/R50+AdamW $+0.09$ against -2.25, and even on FairVision-Glaucoma, where no arm passes, -1.71 against -3.39. Worst-group AUC is the metric that tracks a single, named minority group (Appendix C.3), so this is the most direct evidence that the repair reaches the patients who paid for the savings.

Where the two fairness metrics disagree. On FairVLMed/R50+AdamW every arm holds worst-group AUC within 0.6 points of baseline, three of them at or above -0.13, yet every arm fails on ES-AUC (-0.59 to -1.95). The remaining loss there sits in the gaps between groups other than the worst one, which ES-AUC counts and worst-group AUC does not. On FairVision-Glaucoma the two metrics agree: the full arm loses 4.15 ± 0.3 worst-group points (p=0.001), the largest and most consistent loss in the table, and the floor alone loses least (-1.71). On FairVision-AMD all four arms raise worst-group AUC, by 2.2 to 4.2 points.

E.3 Who pays, and what comes back

Two estimands. Figure 3 tracks a fixed, named group across each pair of runs, $\mathrm{AUC}_g^{\mathrm{method}} - \mathrm{AUC}_g^{\mathrm{base}}$. This differs from the Δworst-group AUC of Tables 1 and 2, whose minimizing group may change between the two runs and which therefore reads more conservatively. Under B-PAS the deepest loss belongs to a minority group on all four datasets: Black on FairVLMed (-5.1, against -3.3 for the White majority), on FairVision-Glaucoma (-3.6) and on FairVision-DR (-1.4), and Asian on FairVision-AMD (-2.1). On FairVision-DR the majority is not harmed at all and gains $+0.4$, so the same deletion that costs a minority group points leaves the majority slightly better off. What separates these groups is how much data they have, not how the pruner treated them, which is what Proposition 1 predicts and $\mathrm{H}_{\mathrm{sel}}$ cannot produce. The two estimands agree on who pays: over the 168 FairVLMed and FairVision runs the worst group is an underrepresented minority in 167 (Appendix C.3).

What the repair returns per group. With the full EquiPrune, the FairVLMed harm is largely undone: Black moves from -5.1 to -1.7 and White from -3.3 to $+0.6$. On FairVision-DR both minority groups turn positive, Asian from -0.7 to $+2.4$ and Black from -1.4 to $+1.0$. Both cost seven to eleven points of DSI. FairVision-Glaucoma moves the other way, Black falling from -3.6 to -4.4, which is the shaded row of Table 2 read per group.

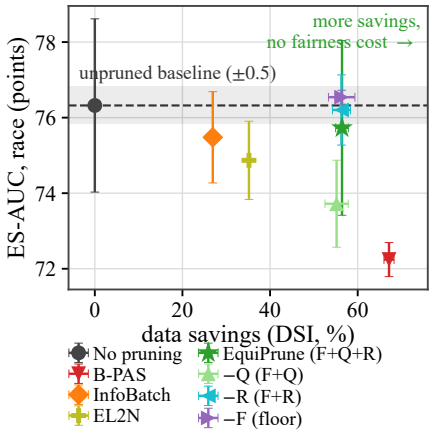


Figure 5: **The fairness–efficiency frontier** (FairVLMed / R50+SGD, race; mean±std over 3 seeds). Data savings against ES-AUC; up and to the right is better. The horizontal line is the unpruned baseline and the band its ± 0.5 equivalence tolerance. The audited pruners lie below the band; the EquiPrune arms keep 55–56% savings at or near baseline-level ES-AUC.

E.4 The fairness–efficiency frontier

Figure 5 draws the worst-hit regime, FairVLMed/R50+SGD, as a frontier: data savings on the horizontal axis, race ES-AUC on the vertical, and the unpruned baseline with its ± 0.5 equivalence band. The three audited pruners all fall below the band and trade one axis for the other: B-PAS reaches 67% DSI at a cost of 4.1 ES-AUC points, InfoBatch stays nearest the baseline but saves only 27%, and EL2N sits in between on both. The EquiPrune arms sit away from that curve, at 55–56% DSI. The floor-only arm ($+0.22$) and F+R (-0.12) are inside the band, F+Q+R sits just below its edge (-0.59), and F+Q stays well below (-2.60). Against B-PAS, the guarded runs give up about a sixth of the savings and recover most or all of the ES-AUC loss, which is the trade the guardrails are designed to make. The upper-right region is reachable, and in this regime only the guarded arms reach it.

F Other protected attributes

The main text reports race, the primary attribute of FairVLMed and the three FairVision tasks. These datasets also record gender, ethnicity and language. This appendix repeats the audit (R1) and the repair (R2) on those three attributes for the seven FairVLMed and FairVision regimes. It uses exactly the same runs as Tables 1 and 2, so the savings are unchanged; only the grouping of the test set changes. As before, R compensates the primary attribute (race), while Q checks all four attributes at once (Appendix B.4).

The groups are small. Gender is close to balanced (on FairVLMed, 1,096 female and 904 male test patients). Ethnicity is not: Hispanic patients number 76 of 2,000 on FairVLMed and 102–113 of 3,000 on each FairVision task, about 4%. Language is the most imbalanced: Spanish speakers number 31 and other non-English speakers 11 on FairVLMed, and 44–63 and 174–215 on FairVision. "Unknown" records are excluded, as everywhere else. For FairVLMed language we use the corrected code mapping (one of the post-hoc axes of Appendix K). With groups this small, a few patients can move a group AUC by several points, so the language cells in particular have wide spreads and are best read as patterns.

The audit holds on every attribute. Tables 12 and 13 apply the R1 rule attribute by attribute. At least two pruners harm gender in 5 of the 7 regimes, ethnicity in 6 and language in 6, against 6 of these 7 for race. Of the 21 (regime, pruner) cells per attribute, 14 are harmful for gender and 16 each for ethnicity and language, as for race. B-PAS harms each of the three attributes in five of the seven regimes. The ranking of pruners is not fixed, though: on ethnicity EL2N has the largest mean loss

Table 12: **The audit on the other attributes: ES-AUC.** Seed-paired ΔES-AUC against the unpruned baseline for gender, ethnicity and language, in the same runs as Table 1 (mean±std, 3 seeds; negative = degraded). †underfit control cell. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t, n=3, descriptive).

	Gender			Ethnicity			Language		
Regime	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	−3.89±1.8	+1.38±1.4	−0.14±1.8	−3.40±1.2*	−0.04±2.8	−0.87±2.8	−16.87±6.5*	−2.21±2.9	−5.31±5.8
FairVLMed / R50+AdamW	−2.38±0.7*	−0.24±0.7	−0.63±0.8	−0.61±1.4	−2.77±2.7	−1.29±2.4	−2.38±4.9	+0.02±4.6	−1.01±3.4
FairVLMed / ViT+AdamW	+0.33±0.6	−0.85±0.4	+0.45±0.6	+1.69±1.6	+1.86±1.1	+1.10±0.3*	−0.71±0.9	−2.19±2.1	−0.39±2.1
FairVLMed / ViT+SGD†	+0.16±10.6	−3.54±14.0	−7.58±4.6	+0.81±7.9	−3.17±10.3	−8.18±0.8**	+8.88±11.3	−2.06±19.8	−3.38±3.8
FairVision-AMD	−1.06±0.7	+0.45±1.3	−1.26±0.6	−1.97±4.8	−1.43±3.1	−4.90±5.2	+3.91±3.7	−0.48±7.2	−3.47±5.1
FairVision-DR	−0.64±3.2	−1.99±1.6	−1.12±3.4	−0.81±3.2	−2.80±2.2	−3.64±6.1	−1.80±7.0	−2.77±6.0	−1.26±11.5
FairVision-Glaucoma	−1.03±1.8	−1.50±0.3**	−0.58±1.9	−2.50±3.9	−1.53±0.9	−2.10±1.0	+0.56±4.5	+1.42±2.9	+2.07±2.9

Table 13: **The audit on the other attributes: worst-group AUC.** Seed-paired Δworst-group AUC for the same cells as Table 12 (mean±std, 3 seeds). †underfit control cell. $^{*}p < 0.05$ (paired t, n=3, descriptive).

	Gender			Ethnicity			Language		
Regime	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	−3.74±1.5	+0.76±1.6	−0.50±1.4	−3.63±1.5	−0.05±3.3	−0.93±3.6	−29.39±12.8	−3.09±4.6	−1.02±3.9
FairVLMed / R50+AdamW	−1.95±0.8*	+0.11±0.8	−0.71±0.7	−0.63±2.2	−4.17±3.7	−1.48±3.8	+3.23±5.0	+2.19±2.8	−1.52±2.0
FairVLMed / ViT+AdamW	+0.38±0.7	−0.80±0.3*	+0.38±0.4	+2.08±1.9	+2.68±1.5	+1.37±0.5*	+0.11±1.7	−0.86±2.7	−2.28±0.5*
FairVLMed / ViT+SGD†	+0.23±10.6	−3.36±14.1	−7.61±4.6	+1.72±5.9	−2.11±9.4	−8.78±2.2*	+8.67±7.0	+6.06±12.1	−8.76±18.7
FairVision-AMD	−1.21±0.8	+0.32±1.1	−1.27±0.4*	−2.16±6.1	−1.79±4.0	−6.37±7.0	+0.98±4.2	−1.06±6.9	−3.88±4.6
FairVision-DR	−0.57±2.3	−2.16±1.3	−1.53±3.1	+0.11±4.1	−2.35±3.1	−2.85±7.5	−1.39±7.7	−1.55±6.1	−2.02±12.8
FairVision-Glaucoma	−1.08±1.3	−1.31±0.3*	−0.45±1.5	−3.28±4.9	−1.63±1.1	−2.95±0.9*	−1.00±4.4	−0.29±3.6	+1.06±4.2

outside the control cell (-1.95 ES-AUC, against -1.27 for B-PAS and -1.12 for InfoBatch). This matches pattern (ii) of the main audit: the harm is not tied to one signal.

The smaller the group, the larger the loss. Averaged over the six regimes outside the ViT+SGD control cell, B-PAS costs gender 1.45 ES-AUC and 1.36 worst-group AUC, ethnicity 1.27 and 1.25, and language 2.88 and 4.58. Language, with the smallest groups, loses the most, and its worst case is extreme: on FairVLMed/R50+SGD, B-PAS lowers language ES-AUC by 16.87 and worst-group AUC by 29.39. This is prediction (P1) of Proposition 1 at the level of groups: the same cut costs most where a group holds the least data. It is also where the minority carries the loss, as (P2) expects: at baseline the worst ethnicity group on FairVLMed is Hispanic in all 12 runs, and the worst language group is Spanish or Other in 9 of 12. Gender, the one balanced attribute, is still harmed in the worst regime (-3.89 ES-AUC under B-PAS), which is part (iii) of the proposition: a gap that exists at baseline widens under an equal cut, whatever the group sizes.

The repair carries over to gender and ethnicity. Table 14 shows the two headline arms; we count a regime as repaired on an attribute when any of the four arms meets R2 on it. Gender is repaired in 5 of the 7 regimes and ethnicity in 5, the same count as race in these regimes. In the worst-hit regime, FairVLMed/R50+SGD, the full arm turns B-PAS's gender loss of -3.89 into a gain of $+1.56$ and its ethnicity loss of -3.40 into -0.09 (worst-group $+1.17$), at the same 56% DSI. FairVision-Glaucoma stays hard, as on race, and shows the same ordering: on ethnicity the floor alone loses 1.27 while the full arm loses 3.77.

Language is the limit. Language is repaired in only 3 of 7 regimes, and in none of the three well-trained FairVLMed regimes, where the smallest group has 11 test patients. Even there the guardrails shrink the loss sharply: on FairVLMed/R50+SGD, B-PAS's -16.87 becomes -5.98 with the full arm and -2.45 with F+R. A batch-level repair can bound how much of each group is removed, but it cannot make a group of a few dozen patients well estimated.

Off the primary attribute, the group-aware parts help more. On race the floor alone is the most reliable arm (Section 4). Over the 21 attribute–regime cells here, the ranking changes: the full arm meets R2 in 12, F+Q and F+R in 6 each, and the floor alone in 4. Q is the only guardrail that checks every attribute, and these groups are small enough that a handful of batches can move their retained share, which is the case the quota was designed for (Section 3). The floor remains the demographics-free core; Q and R matter more as the protected groups get smaller.

Table 14: **The repair on the other attributes.** For each attribute, the full EquiPrune (F+Q+R) and the floor-only arm (F), seed-paired against the unpruned baseline: ΔES-AUC (mean±std, 3 seeds) with Δworst-group AUC in parentheses. Same runs and DSI as Table 2. **Bold** = the arm meets the R2 rule on that attribute (both deltas ≥ -0.5). †underfit control cell. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t on ΔES-AUC, n=3, descriptive).

	Gender		Ethnicity		Language	
Regime	F+Q+R	F only	F+Q+R	F only	F+Q+R	F only
FairVLMed / R50+SGD	$+\mathbf{1.56}_{\pm 0.7}$ ($+1.09$)	$+\mathbf{1.10}_{\pm 1.8}$ ($+0.73$)	$-\mathbf{0.09}_{\pm 1.6}$ ($+1.17$)	$-2.24_{\pm 3.5}$ (-3.19)	$-5.98_{\pm 1.0}^{**}$ (-6.83)	$-6.81_{\pm 3.0}$ (-6.83)
FairVLMed / R50+AdamW	$-1.30_{\pm 0.9}$ (-1.18)	$-1.68_{\pm 1.7}$ (-1.26)	$+\mathbf{0.07}_{\pm 3.4}$ ($+0.30$)	$-2.03_{\pm 4.1}$ (-2.62)	$-4.55_{\pm 11.3}$ (-7.10)	$-3.23_{\pm 6.5}$ (-2.59)
FairVLMed / ViT+AdamW	$-\mathbf{0.09}_{\pm 1.1}$ (-0.05)	$-0.63_{\pm 0.8}$ (-0.63)	$+\mathbf{2.82}_{\pm 2.5}$ ($+3.66$)	$-1.76_{\pm 3.0}$ (-2.34)	$-3.76_{\pm 5.8}$ (-4.63)	$-6.89_{\pm 2.4}^{*}$ (-8.10)
FairVLMed / ViT+SGD†	$+\mathbf{4.82}_{\pm 5.3}$ ($+4.68$)	$+\mathbf{0.30}_{\pm 10.5}$ ($+0.52$)	$+\mathbf{1.76}_{\pm 5.2}$ ($+0.62$)	$-0.27_{\pm 7.8}$ (-0.76)	$+\mathbf{10.38}_{\pm 3.1}^{*}$ ($+5.26$)	$+\mathbf{6.55}_{\pm 12.2}$ ($+3.69$)
FairVision-AMD	$-0.73_{\pm 1.6}$ (-0.90)	$-0.62_{\pm 1.3}$ (-0.74)	$-3.36_{\pm 5.2}$ (-4.17)	$-1.60_{\pm 2.1}$ (-1.85)	$-2.30_{\pm 8.3}$ (-2.80)	$-1.61_{\pm 2.2}$ (-2.75)
FairVision-DR	$+\mathbf{1.06}_{\pm 2.5}$ ($+0.54$)	$+\mathbf{0.36}_{\pm 1.3}$ ($+0.26$)	$-\mathbf{0.03}_{\pm 3.4}$ (-0.46)	$-1.59_{\pm 3.3}$ (-0.58)	$+\mathbf{2.49}_{\pm 5.4}$ ($+1.85$)	$+0.72_{\pm 5.1}$ (-0.69)
FairVision-Glaucoma	$-2.12_{\pm 2.5}$ (-1.89)	$-1.48_{\pm 2.2}$ (-1.18)	$-3.77_{\pm 1.3}^{*}$ (-4.79)	$-1.27_{\pm 3.0}$ (-1.67)	$-\mathbf{0.26}_{\pm 0.7}$ (-0.46)	$-1.43_{\pm 2.1}$ (-2.64)

Summary. The race findings carry over. Pruning harms every attribute in most imbalanced regimes, the loss is largest for the smallest groups, and EquiPrune recovers gender and ethnicity about as often as race. Language, whose groups hold a few dozen patients, marks the limit of what a batch-level guardrail can guarantee.

G THE THRESHOLD-PARITY VIEW OF THE AUDIT AND THE REPAIR

This appendix holds the full DPD and DEOdds results summarized in Section 4. Table 15 is the parity companion to Table 1, and Table 16 the companion to Table 2: the same runs and the same seed-pairing, scored on the threshold-parity metrics instead of ES-AUC and worst-group AUC. **In this appendix only, a negative number is a nominal improvement**, because DPD and DEOdds are disparity distances.

Parity misses the harm. Table 15 scores the same 27 audited cells as Table 1. Sixteen of them are harmful under the R1 rule. In **11** of those **16**, ΔDPD is negative: the parity metric does not just miss the damage, it reports an improvement. With the same ± 0.5 tolerance we apply to ES-AUC, **12** of the 16 harmful cells would pass as parity-neutral or better on DPD, and **8** would pass on DPD *and* DEOdds together, including all three pruners on FairVLMed/R50+SGD, the regime with the largest subgroup damage. The most consistent parity movement in the whole audit is B-PAS's -5.11 ± 0.67 (p=0.006) on that regime: a 43% drop in measured disparity from a baseline DPD of 11.8, obtained while losing 4.08 points of ES-AUC.

The two families are decoupled, not opposed. Across the 27 cells the Spearman correlation between ΔES-AUC and ΔDPD is $+0.08$, and between ΔES-AUC and ΔDEOdds it is -0.07. Knowing what pruning did to parity says almost nothing about what it did to subgroup AUC. The two parity metrics do not even agree with each other: they disagree on the direction of change in 9 of the 27 cells. The sign is not stable within one pruner either: B-PAS's ΔDPD runs from -5.22 on FairVLMed/ViT+SGD to $+2.58$ on FairVLMed/ViT+AdamW, with no relation to where it damages ES-AUC. The largest single case has a clear cause: a degraded model can compress its predicted-positive rates toward zero, and a model that predicts one class for everyone has perfect parity. That only happens when pruning is aggressive enough, so it is a mechanism rather than a rule, which is why the pooled correlation is flat rather than positive. An anti-correlated metric could at least be read backwards; an uninformative one cannot.

Consequences for reporting. Adding a parity metric to an average-accuracy report does not fix it: a DPD-only audit would have cleared 12 of the 16 harmful cells, and adding DEOdds would still clear 8. The failure also concentrates where the stakes are highest. Every entry of Table 15 below -2.5, i.e. every large nominal parity "gain", falls in a shaded row where Table 1 records real subgroup harm, while the two safe, balanced regimes, CheXpert and UTKFace, show parity movements of $|\Delta| \leq 2.0$ that say nothing either way. We therefore draw the fairness–efficiency frontier in ES-AUC and worst-group AUC and report parity alongside it as a diagnostic, not as a certificate.

The effect does not depend on the 0.5 threshold. Axis P6 of the sensitivity study (Appendix K) re-scores the saved predictions of FairVLMed/R50+SGD at five decision thresholds. B-PAS's mean

Table 15: **The parity view of the audit**, the companion to Table 1 over the same 27 (regime, pruner) cells. Seed-paired Δ against the unpruned baseline (mean$\pm$std, 3 seeds), with each regime's baseline level in the "base" columns. **The sign convention is the reverse of Table 1**: DPD and DEOdds are distances, so a negative entry means measured disparity *shrank*. Shaded rows are the six regimes that support R1 in Table 1, i.e. where subgroup AUC was actually damaged. †underfit control cell. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t, n=3, descriptive).

	ΔDPD ↓				ΔDEOdds ↓			
Regime	base	B-PAS	InfoBatch	EL2N	base	B-PAS	InfoBatch	EL2N
FairVLMed / R50+SGD	11.8	$-5.11_{\pm 0.7}^{**}$	$-1.03_{\pm 3.7}$	$-4.96_{\pm 1.9}^{*}$	15.5	$-0.64_{\pm 2.0}$	$-3.84_{\pm 7.0}$	$-6.71_{\pm 4.2}$
FairVLMed / R50+AdamW	10.3	$-0.95_{\pm 3.9}$	$-2.93_{\pm 5.7}$	$-3.52_{\pm 1.1}^{*}$	13.3	$+1.69_{\pm 4.2}$	$-0.31_{\pm 4.9}$	$+5.19_{\pm 3.5}$
FairVLMed / ViT+AdamW	6.8	$+2.58_{\pm 3.8}$	$+2.88_{\pm 4.2}$	$+0.49_{\pm 5.3}$	14.4	$+7.37_{\pm 4.3}$	$+0.97_{\pm 3.4}$	$-4.08_{\pm 2.1}$
FairVLMed / ViT+SGD†	10.4	$-5.22_{\pm 8.1}$	$-5.84_{\pm 9.8}$	$+3.81_{\pm 6.0}$	17.0	$-5.85_{\pm 14.4}$	$-5.52_{\pm 13.2}$	$+13.52_{\pm 12.4}$
FairVision-AMD	22.5	$-1.31_{\pm 1.8}$	$+1.36_{\pm 0.5}^{*}$	$-0.21_{\pm 2.0}$	24.6	$+4.52_{\pm 11.5}$	$+1.75_{\pm 9.7}$	$+1.12_{\pm 0.5}$
FairVision-DR	2.6	$-0.47_{\pm 1.5}$	$+1.23_{\pm 1.8}$	$-0.25_{\pm 2.2}$	25.7	$-1.63_{\pm 9.6}$	$-6.06_{\pm 10.6}$	$-6.59_{\pm 15.0}$
FairVision-Glaucoma	10.5	$+0.09_{\pm 2.6}$	$-1.73_{\pm 0.9}$	$-0.66_{\pm 2.0}$	12.6	$+2.03_{\pm 5.2}$	$-4.93_{\pm 3.7}$	$-0.16_{\pm 5.7}$
CheXpert	2.6	$-0.87_{\pm 1.0}$	$+0.14_{\pm 1.6}$	$-0.10_{\pm 0.9}$	3.9	$-1.68_{\pm 2.2}$	$+0.38_{\pm 3.4}$	$-0.28_{\pm 2.0}$
UTKFace	10.7	$+1.10_{\pm 1.7}$	$+1.65_{\pm 1.9}$	$+2.00_{\pm 0.6}^{*}$	16.4	$-0.15_{\pm 2.8}$	$-1.30_{\pm 0.7}$	$-0.67_{\pm 0.3}$

ΔDPD is -5.77, -4.70, -4.82, -3.86 and -3.00 at thresholds 0.3, 0.4, 0.5, 0.6 and 0.7, and EL2N's is negative at all five (-2.91 to -1.90). The corresponding ΔDEOdds values are negative at $5/5$ thresholds for EL2N and InfoBatch and at $4/5$ for B-PAS. The nominal parity gain is a property of the pruned model, not of where we read it.

The repair does not trade parity for ES-AUC. Since the two families move independently, a gain in ES-AUC could in principle come at the cost of parity. Table 16 checks this for the same four arms in the same nine regimes. Over all 36 arm–regime cells the mean ΔDPD is -1.37 and the mean ΔDEOdds -0.55, both toward less measured disparity, and 29 of the 36 keep DPD within the $+0.5$ tolerance. The audited pruners, over their 27 cells, average -0.66 and -0.44 and keep DPD within tolerance in 19 of 27. So the guardrails move parity slightly further in the nominally good direction than the pruners they wrap, while also being the only methods that hold ES-AUC and worst-group AUC. They do this without any parity term: F, Q and R constrain retention and reweight the loss, and none of them sees a decision threshold.

The worst-hit regime shows the point in two rows. B-PAS reports $\Delta\text{DPD} = -5.11$ while losing 4.08 points of ES-AUC. EquiPrune-F reports $\Delta\text{DPD} = -0.98$ and $\Delta\text{DEOdds} = -2.74$ while *gaining* 0.22 points of ES-AUC and 0.28 of worst-group AUC. The repair keeps parity below baseline and gives back the subgroup performance, so the 5.1-point parity "gain" of B-PAS was never worth having. In the seven regimes where some arm meets R2, the best passing arm also keeps DPD within tolerance in five. The two exceptions are FairVLMed/ViT+AdamW ($+1.16$) and FairVision-DR ($+1.07$, next to a -10.91 movement in DEOdds in the same cell). Our claim is the narrow one the data support: the guardrails do not trade the parity family for the equity-scaled family. They do not optimize parity either, and should not: a method that improved DPD by underfitting would reproduce exactly the failure described above.

H Pruning inside FairCLIP: setup and full results

Setup. The main campaign trains supervised classifiers. To check that the findings are not tied to that recipe, we move the same pruning methods into FairCLIP (Luo et al., 2024b): contrastive vision–language fine-tuning of CLIP ViT-B/16 on FairVLMed images and clinical notes, with the Sinkhorn fairness regularizer. We prune its fine-tuning batches. This changes the architecture, the pretraining, the objective and the optimizer (Adam, learning rate 10^{-5}, 20 epochs, batch 32, K=218), while the pruning machinery and (ϕ, γ, c) stay the same. Each protected attribute is fine-tuned as its own target. DPD and DEOdds need a decision threshold that turns predicted probabilities into yes/no predictions. CLIP's probabilities are not centered at 0.5, so a fixed 0.5 threshold would label almost every patient the same way. We instead choose the threshold on the validation set by Youden's J, which picks the threshold maximizing $J = \text{TPR} - \text{FPR}$ (sensitivity + specificity -1), i.e. the

Table 16: **The parity view of the repair**, the companion to Table 2 for the same four EquiPrune arms in the same nine regimes. Seed-paired Δ against the unpruned baseline (mean$\pm$std, 3 seeds). As in Table 15, **negative means measured disparity shrank**, and baseline levels are those of Table 15. Arm labels follow Table 2 ($-$Q = F+Q, $-$R = F+R, $-$F = floor only). †underfit control cell. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t, n=3, descriptive).

	ΔDPD ↓				ΔDEOdds ↓			
Regime	F+Q+R	$-$Q	$-$R	$-$F	F+Q+R	$-$Q	$-$R	$-$F
FairVLMed / R50+SGD	$-1.38_{\pm 5.2}$	$-3.31_{\pm 1.9}$	$-2.60_{\pm 4.1}$	$-0.98_{\pm 2.4}$	$+1.22_{\pm 7.7}$	$-0.79_{\pm 2.7}$	$-1.94_{\pm 7.2}$	$-2.74_{\pm 2.3}$
FairVLMed / R50+AdamW	$-4.66_{\pm 3.3}$	$-3.20_{\pm 0.6}^{*}$	$-4.20_{\pm 5.1}$	$-1.92_{\pm 8.0}$	$+1.18_{\pm 7.2}$	$+1.28_{\pm 8.2}$	$+1.14_{\pm 7.5}$	$+9.27_{\pm 8.8}$
FairVLMed / ViT+AdamW	$+1.15_{\pm 1.3}$	$+0.22_{\pm 0.6}$	$+0.65_{\pm 4.5}$	$+1.16_{\pm 1.0}$	$+0.25_{\pm 7.9}$	$+0.08_{\pm 9.0}$	$-1.60_{\pm 3.7}$	$+2.95_{\pm 7.1}$
FairVLMed / ViT+SGD†	$-5.69_{\pm 6.2}$	$-8.80_{\pm 6.9}$	$-5.87_{\pm 10.6}$	$-2.49_{\pm 9.4}$	$+2.28_{\pm 10.3}$	$-8.65_{\pm 11.9}$	$+2.10_{\pm 11.3}$	$-0.21_{\pm 22.3}$
FairVision-AMD	$+1.04_{\pm 0.02}^{**}$	$-0.97_{\pm 3.5}$	$+1.82_{\pm 2.0}$	$+0.92_{\pm 3.5}$	$+6.99_{\pm 13.5}$	$-2.14_{\pm 16.2}$	$+5.48_{\pm 7.3}$	$+5.54_{\pm 17.1}$
FairVision-DR	$+1.07_{\pm 2.4}$	$+0.00_{\pm 1.3}$	$-0.71_{\pm 1.6}$	$+0.34_{\pm 1.3}$	$-10.91_{\pm 15.2}$	$-7.38_{\pm 8.1}$	$-8.42_{\pm 3.1}^{*}$	$-5.31_{\pm 11.2}$
FairVision-Glaucoma	$-0.60_{\pm 0.8}$	$-2.86_{\pm 2.5}$	$-0.33_{\pm 1.0}$	$-2.96_{\pm 2.5}$	$+0.89_{\pm 2.8}$	$+0.57_{\pm 4.3}$	$+3.17_{\pm 5.4}$	$-5.24_{\pm 6.8}$
CheXpert	$-0.77_{\pm 1.1}$	$-0.91_{\pm 1.7}$	$-0.85_{\pm 1.0}$	$-1.21_{\pm 0.8}$	$-1.75_{\pm 2.1}$	$-1.04_{\pm 2.3}$	$-1.59_{\pm 1.9}$	$-2.20_{\pm 1.0}$
UTKFace	$-0.11_{\pm 1.7}$	$-0.23_{\pm 2.3}$	$+0.13_{\pm 1.1}$	$-0.06_{\pm 1.9}$	$-2.45_{\pm 1.3}$	$-0.32_{\pm 0.3}$	$-0.29_{\pm 2.7}$	$+0.92_{\pm 0.6}$

Table 17: **Pruning inside FairCLIP** (CLIP ViT-B/16, Sinkhorn fine-tuning on Harvard-FairVLMed, $\times 100$, mean$\pm$std over seeds). Each block fine-tunes with the fairness term targeting that attribute and reports that attribute's metrics. B-PAS and EquiPrune both realize about 35% DSI, so they are compared at equal savings. DPD and DEOdds follow the FairCLIP convention (DEOdds = the larger of the TPR and FPR spreads); both are distances, so lower is nominally better. **Bold** = best entry in each block.

Attribute	Method	AUC	ES-AUC	Worst-group AUC	DPD ↓	DEOdds ↓
race	Baseline (no pruning)	$53.91_{\pm 2.23}$	$51.47_{\pm 0.59}$	$52.23_{\pm 1.37}$	$20.44_{\pm 1.27}$	$23.22_{\pm 0.11}$
	B-PAS	$53.75_{\pm 2.65}$	$50.58_{\pm 0.45}$	$51.06_{\pm 1.85}$	$18.19_{\pm 0.89}$	$\mathbf{19.41}_{\pm 2.81}$
	EquiPrune	$\mathbf{54.39}_{\pm 1.63}$	$\mathbf{51.53}_{\pm 1.62}$	$\mathbf{53.10}_{\pm 0.24}$	$\mathbf{17.86}_{\pm 2.31}$	$19.51_{\pm 4.12}$
gender	Baseline (no pruning)	$\mathbf{59.47}_{\pm 4.15}$	$\mathbf{56.68}_{\pm 3.64}$	$\mathbf{57.36}_{\pm 3.83}$	$5.01_{\pm 1.35}$	$8.88_{\pm 1.44}$
	B-PAS	$55.68_{\pm 4.72}$	$54.34_{\pm 3.56}$	$54.62_{\pm 3.86}$	$5.37_{\pm 0.38}$	$\mathbf{7.03}_{\pm 1.68}$
	EquiPrune	$57.57_{\pm 4.84}$	$55.92_{\pm 3.23}$	$56.26_{\pm 3.72}$	$\mathbf{5.04}_{\pm 1.15}$	$7.61_{\pm 2.20}$
ethnicity	Baseline (no pruning)	$54.27_{\pm 1.28}$	$\mathbf{51.37}_{\pm 3.34}$	$\mathbf{48.73}_{\pm 6.46}$	$\mathbf{2.35}_{\pm 2.62}$	$9.49_{\pm 6.08}$
	B-PAS	$\mathbf{54.68}_{\pm 1.88}$	$50.24_{\pm 3.33}$	$46.06_{\pm 5.24}$	$3.65_{\pm 2.51}$	$\mathbf{7.25}_{\pm 3.17}$
	EquiPrune	$53.67_{\pm 3.15}$	$50.04_{\pm 3.45}$	$46.78_{\pm 6.25}$	$5.58_{\pm 2.58}$	$9.72_{\pm 2.83}$
language	Baseline (no pruning)	$55.10_{\pm 0.82}$	$44.32_{\pm 3.30}$	$54.77_{\pm 0.78}$	$21.48_{\pm 5.48}$	$26.81_{\pm 10.57}$
	B-PAS	$54.25_{\pm 3.16}$	$40.49_{\pm 4.26}$	$41.18_{\pm 20.67}$	$\mathbf{14.40}_{\pm 10.17}$	$34.78_{\pm 22.02}$
	EquiPrune	$\mathbf{56.38}_{\pm 0.58}$	$\mathbf{46.48}_{\pm 5.15}$	$\mathbf{56.10}_{\pm 0.73}$	$14.53_{\pm 11.12}$	$\mathbf{15.44}_{\pm 14.12}$

point of the ROC curve farthest above the chance diagonal. The threshold is then frozen and applied unchanged to the test set.

Harm. Against the unpruned baseline in each block, B-PAS costs race 0.89 ES-AUC and 1.17 worst-group AUC, gender 2.34 ES-AUC and 2.74 worst-group AUC, ethnicity 1.13 and 2.67, and language 3.83 and 13.59. The harm is largest where group sizes are most skewed, and on language the seed spread of worst-group AUC widens to ± 20.7.

Repair. At the same $\sim$35% savings, EquiPrune improves worst-group AUC over B-PAS on all four attributes (race 53.10 against 51.06, gender 56.26 against 54.62, ethnicity 46.78 against 46.06, language 56.10 against 41.18). On race and language it ends above the unpruned baseline ($+0.06$ and $+2.16$ ES-AUC; $+0.87$ and $+1.33$ worst-group AUC). On gender it recovers most of B-PAS's loss (-0.76 ES-AUC, -1.10 worst-group AUC). Ethnicity is the one block it does not repair on ES-AUC (-1.33 against B-PAS's -1.13), although it still narrows the worst-group loss from -2.67 to -1.95.

Parity dissociation. The pattern of Appendix G reappears. B-PAS posts the ethnicity block's *best* DEOdds (7.25 against the baseline's 9.49) in the same block where it costs 2.7 points of worst-group AUC, and the language block's best DPD (14.40 against 21.48) where it costs 13.6. A fairness-regularized objective does not protect a model from group-blind pruning, and the guardrails carry over to it unchanged.

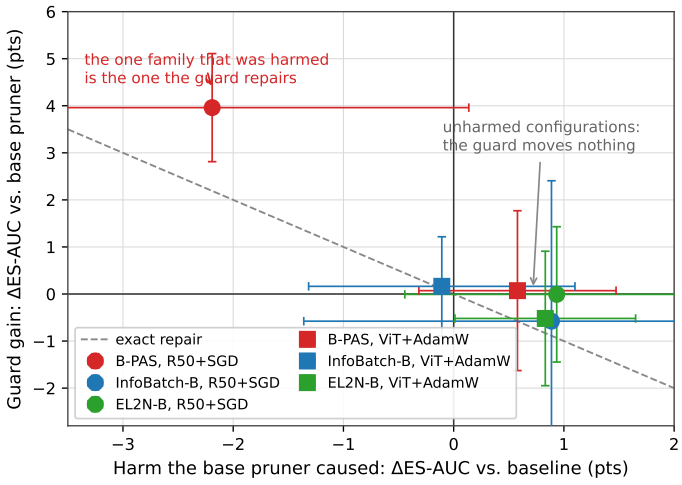


Figure 6: **The benefit of the guardrails tracks the harm** (FairVLMed, race, 3 seeds, ΔES-AUC). Horizontal axis: what the base pruner did to fairness relative to the unpruned baseline. Vertical axis: what wrapping it in EquiPrune recovered, paired against that same pruner so the selection signal is held fixed. The dashed line marks exact repair. Only B-PAS on ResNet-50 was harmed, and it is the only point far from the origin; the five unharmed configurations cluster at $(0, 0)$. In the four configurations with a budgeted pruner (InfoBatch, EL2N) the floor recorded *zero* vetoes, so there it was not ineffective but inactive. Error bars are ± 1 s.d.

I GENERALITY: THE GUARDRAILS ON OTHER PRUNERS

EquiPrune reads only a pruner's ordered removal proposal, so in principle it wraps any pruner. We test this in a separate 132-run study that puts the identical wrapper, with identical (ϕ, γ, c), around each of the three pruning signals in two regimes, FairVLMed/R50+SGD and FairVLMed/ViT+AdamW, i.e. $3 \times 2 = 6$ (pruner, regime) configurations. To isolate the wrapper, each arm is paired against *the pruner it wraps*, not against the baseline, so the selection signal is held fixed and only F, Q and R change. The quota arms are also run at a stricter $\gamma = 0.98$, because at batch level $\gamma = 0.9$ rarely binds. These are independent runs of the two regimes, so their levels differ from Table 1: here B-PAS costs -2.19 ± 2.33 ES-AUC on FairVLMed/R50+SGD, against -4.08 ± 2.14 in the main campaign, in line with the run-to-run variation measured in Appendix K.

The benefit tracks the harm. Figure 6 shows the whole study. Each point is one of the six configurations: its horizontal position is what the pruner did to fairness, its vertical position is what wrapping it recovered, and the dashed line is exact repair. A wrapper that helped regardless would lift every point; one that did nothing would flatten them onto $y{=}0$. Neither happens. The single harmed configuration sits high on the left near the exact-repair line, and the five unharmed ones cluster at the origin.

The two permanent pruners. Table 18 gives the numbers for B-PAS and EL2N, the two permanent pruners, in both regimes. On FairVLMed/R50+SGD, B-PAS is the only pruner that damages fairness (-2.19 ES-AUC, -2.67 worst-group AUC, -3.29 AUC, $p{=}0.048$). Every arm reverses it: the full arm by $+3.96 \pm 1.15$ ES-AUC ($p{=}0.027$), the floor alone by $+4.66 \pm 2.29$, with worst-group gains of 3.4 to 4.9 points and AUC gains of 3.2 to 3.6 points, while keeping 54–56% DSI. Against the unpruned baseline, all four arms end above it on ES-AUC ($+1.77$ to $+2.59$) and worst-group AUC ($+0.70$ to $+2.22$). EL2N was not harmed in this regime ($+0.93$ ES-AUC), and the wrapper moves it by -0.60 to $+0.22$, which is indistinguishable from zero. On FairVLMed/ViT+AdamW neither permanent pruner is harmed (B-PAS $+0.58$, EL2N $+0.83$), and no arm changes ES-AUC by more than 1.7 points in either direction. The one nominally significant change, F+R on EL2N (-1.04 ± 0.36, $p{=}0.038$), is within the seed-noise floor $\sigma_\Delta = 2.20$ and is not repeated by the other arms.

Table 18: **The guardrails on the two permanent pruners**, B-PAS and EL2N, in the 132-run generality study (FairVLMed, race, mean±std over 3 seeds, $\gamma = 0.9$). The first row of each block is the pruner against the unpruned baseline. The arm rows are paired against *the wrapped pruner*, so they show what the wrapper changes with the selection signal held fixed. "Floor vetoes" is the range over seeds of candidate removals blocked by F per run. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t, n=3, descriptive).

Regime	Row (reference)	ΔES-AUC	Δworst-group AUC	ΔAUC	DSI (%)	Floor vetoes
FairVLMed / R50+SGD	B-PAS (vs. baseline)	$-2.19_{\pm 2.3}$	$-2.67_{\pm 4.0}$	$-3.29_{\pm 1.3}^{*}$	$65.6_{\pm 2.3}$	n/a
	F+Q+R (vs. B-PAS)	$+3.96_{\pm 1.2}^{*}$	$+3.37_{\pm 2.9}$	$+3.17_{\pm 1.0}^{*}$	$54.0_{\pm 1.7}$	507–539
	F+Q (vs. B-PAS)	$+4.78_{\pm 4.9}$	$+4.89_{\pm 5.8}$	$+3.36_{\pm 1.7}$	$53.7_{\pm 1.2}$	488–539
	F+R (vs. B-PAS)	$+4.52_{\pm 2.3}$	$+4.37_{\pm 4.6}$	$+3.44_{\pm 1.6}$	$53.7_{\pm 1.9}$	493–514
	F only (vs. B-PAS)	$+4.66_{\pm 2.3}$	$+4.35_{\pm 3.6}$	$+3.55_{\pm 0.9}^{*}$	$56.0_{\pm 2.9}$	462–504
	EL2N (vs. baseline)	$+0.93_{\pm 1.4}$	$-0.32_{\pm 2.2}$	$-0.76_{\pm 1.1}$	$35.2_{\pm 0.0}$	n/a
	F+Q+R (vs. EL2N)	$-0.01_{\pm 1.4}$	$+0.61_{\pm 1.9}$	$+0.56_{\pm 1.1}$	$35.2_{\pm 0.0}$	0
	F+Q (vs. EL2N)	$-0.60_{\pm 0.6}$	$+0.12_{\pm 1.8}$	$+0.29_{\pm 0.8}$	$35.2_{\pm 0.0}$	0
	F+R (vs. EL2N)	$+0.05_{\pm 2.3}$	$+0.01_{\pm 2.7}$	$+0.35_{\pm 1.4}$	$35.2_{\pm 0.0}$	0
	F only (vs. EL2N)	$+0.22_{\pm 2.8}$	$+1.05_{\pm 3.2}$	$+0.06_{\pm 0.7}$	$35.2_{\pm 0.0}$	0
FairVLMed / ViT+AdamW	B-PAS (vs. baseline)	$+0.58_{\pm 0.9}$	$+1.32_{\pm 2.0}$	$+0.64_{\pm 0.05}^{**}$	$73.3_{\pm 4.1}$	n/a
	F+Q+R (vs. B-PAS)	$+0.07_{\pm 1.7}$	$-0.10_{\pm 1.6}$	$-0.47_{\pm 0.7}$	$63.5_{\pm 2.5}$	169–181
	F+Q (vs. B-PAS)	$-0.66_{\pm 1.0}$	$-1.26_{\pm 1.8}$	$-1.23_{\pm 0.7}$	$63.6_{\pm 0.6}$	165–176
	F+R (vs. B-PAS)	$-0.43_{\pm 1.8}$	$-0.82_{\pm 2.2}$	$-0.84_{\pm 1.0}$	$64.7_{\pm 2.5}$	178–183
	F only (vs. B-PAS)	$+0.43_{\pm 0.2}$	$+0.28_{\pm 0.7}$	$-0.30_{\pm 0.2}$	$65.7_{\pm 1.2}$	170–191
	EL2N (vs. baseline)	$+0.83_{\pm 0.8}$	$+1.21_{\pm 0.8}$	$-0.10_{\pm 0.7}$	$35.2_{\pm 0.0}$	n/a
	F+Q+R (vs. EL2N)	$-0.52_{\pm 1.4}$	$-1.44_{\pm 2.6}$	$-0.95_{\pm 0.5}$	$35.2_{\pm 0.0}$	0
	F+Q (vs. EL2N)	$-0.11_{\pm 1.0}$	$-0.25_{\pm 2.2}$	$-0.31_{\pm 0.1}$	$35.2_{\pm 0.0}$	0
	F+R (vs. EL2N)	$-1.04_{\pm 0.4}^{*}$	$-1.36_{\pm 1.5}$	$-0.08_{\pm 0.5}$	$35.2_{\pm 0.0}$	0
	F only (vs. EL2N)	$-1.67_{\pm 2.4}$	$-1.39_{\pm 1.7}$	$-0.37_{\pm 1.2}$	$35.2_{\pm 0.0}$	0

Why B-PAS and EL2N respond so differently. Both are permanent, but only B-PAS has no stopping rule. The floor blocks 462–539 B-PAS removals per run on ResNet-50 and 165–191 on ViT-B/16, because B-PAS would otherwise keep retiring batches until none are left. EL2N stops by itself at its 35% budget, far above the floor, so the floor blocks *zero* EL2N removals in both regimes and the wrapper leaves EL2N's savings exactly unchanged (ΔDSI $= 0$): a fixed-budget pruner reaches its target whichever candidates are vetoed. The quota is almost as quiet on EL2N, with 0–2 vetoes per run at $\gamma = 0.9$. So on EL2N the wrapper is not failing; it is mostly not acting. This is a built-in control: the effect on B-PAS appears exactly where the guardrails act and nowhere else, which ties it to the guardrails rather than to incidental changes in the run.

InfoBatch. InfoBatch, the temporary pruner, was not harmed in either regime ($+0.89$ and -0.11 ES-AUC against baseline). The floor never fires on it either. The quota does veto some InfoBatch candidates (37–117 per run at $\gamma = 0.9$, several hundred at $\gamma = 0.98$), but the resulting changes are small and go both ways (-0.82 to $+0.99$ ES-AUC on FairVLMed/R50+SGD, at most 1.98 in size on ViT+AdamW). Savings are preserved throughout: EquiPrune[B-PAS] keeps 49–66% DSI across all arms and both γ values, EquiPrune[InfoBatch] 19–27%, and EquiPrune[EL2N] 35.2%.

What this is and is not. This is not "three of three pruners repaired". A rule that demanded a positive change on an unharmed pruner would ask a repair to improve something that was not broken. The claim is narrower and better supported: the wrapper composes with all three pruners without harming them, and its benefit appears where, and only where, there is harm to repair.

J GRANULARITY: BATCH VERSUS SAMPLE LEVEL

We test whether the conclusions depend on the pruning unit by repeating the audit at sample level on FairVLMed/R50+SGD, with everything else fixed. Table 19 reports both granularities against a common baseline.

Sample-level pruning is more harmful. At matched savings, pruning single images does more damage. EL2N goes from -1.45 to -7.82 ES-AUC points at the same $\sim$35% DSI, and its overall

Table 19: **Batch-level versus sample-level pruning**, FairVLMed / R50+SGD (race, $\times 100$, mean$\pm$std over 3 seeds). Δ is the seed-paired change in ES-AUC against the unpruned batch-level baseline, which is the reference for both blocks, so its row is repeated. At batch level the audited EL2N is the batch port; at sample level it is the original sample-level EL2N-S, and the pruning unit is one image rather than a batch of 128. EL2N is five times more harmful at sample level, while the EquiPrune arms keep ΔES-AUC within $[-0.56, +0.22]$ at about 70% DSI.

	Batch-level (unit = 128 images)				Sample-level (unit = 1 image)			
Method	DSI	AUC	ES-AUC	ΔES-AUC	DSI	AUC	ES-AUC	ΔES-AUC
Baseline (no pruning)	0.00	80.41±0.69	76.32±2.29	(ref.)	0.00	80.41±0.69	76.32±2.29	(ref.)
B-PAS	67.15±1.15	77.07±1.01	72.24±0.45	−4.08±2.14	78.01±0.16	76.45±1.89	72.33±2.24	−3.99±4.40
EL2N	35.19±0.00	79.40±0.26	74.87±1.04	−1.45±1.44	35.00±0.00	74.18±0.83	68.50±3.30	**−7.82**±5.33
EquiPrune (F+Q+R)	56.42±1.45	80.72±0.75	75.73±2.31	−0.59±0.67	69.62±0.08	79.64±0.66	75.77±1.59	−0.56±0.89
-Q (F+Q)	55.19±2.66	79.84±0.37	73.72±1.15	−2.60±3.38	69.80±0.13	80.59±0.13	76.54±1.08	+0.22±3.31
-R (F+R)	56.30±2.06	79.34±0.55	76.20±0.93	−0.12±3.17	69.66±0.02	79.32±0.37	76.03±2.21	−0.29±0.78
-F (floor only)	56.34±3.01	80.53±0.15	76.54±0.18	+0.22±2.23	69.79±0.13	80.39±0.24	76.14±1.10	−0.18±3.37

AUC drops from 79.40 to 74.18. B-PAS costs about the same ES-AUC at both granularities (-4.08 and -3.99) and a similar worst-group loss (-3.92 and -4.06), but it prunes further at sample level (78% against 67% DSI) and loses a little more overall AUC (76.45 against 77.07). Where granularity changes the outcome, it changes it toward more damage.

Interpretation. At batch level every unit is a random mix of groups, so pruning acts as nearly uniform data removal (Appendix D.4). At sample level every unit belongs to one group, so a group-blind signal is no longer forced to delete evenly, and the removal itself can become uneven. The batch-level results are therefore conservative: they measure the harm of uniform thinning alone.

Why the main campaign uses batches. First, it isolates a clean mechanism: retention stays demographically uniform, so the harm can be attributed to optimization degradation rather than to selection. Second, it matches the pruning unit for which B-PAS was designed and reported. Sample-level pruning adds variance and does worse without making the mechanism clearer.

The repair still transfers. All four EquiPrune arms keep ΔES-AUC between -0.56 and $+0.22$ at about 70% DSI, far from the losses of the audited pruners. Two arms meet the full R2 rule at sample level, including worst-group AUC measured against the same batch-level baseline: F+Q ($+0.22$ ES-AUC, -0.49 worst-group) and F+R (-0.29 and $+0.13$). The full arm (-0.56 and -0.73) and the floor alone (-0.18 and -0.60) fall just short. The repair is not tied to batch granularity.

K Sensitivity analysis: are the conclusions hyperparameter artifacts?

A fairness result at one operating point invites the objection that the point was chosen to produce it. We therefore ran a pre-registered sensitivity campaign of 129 additional training runs (3 seeds per point) that varies, one at a time, every free hyperparameter of the audited pruners and of EquiPrune, plus nine post-hoc analysis axes that need no training. The reference point is an *independent replication* of FairVLMed/R50+SGD with the identical configuration and seeds, run separately; its 24 runs are the centre of every sweep. We report the comparison openly, because it is the honest measure of run-to-run variability: the replication reproduces the sign and the ordering of every method but not the size, with B-PAS at -0.79 ± 1.54 here against -4.08 ± 2.14 in Table 1 (mean absolute difference over the 24 paired runs: 1.98 ES-AUC points). Training is not bit-deterministic across launches, so a fixed seed pins the initialization and batch membership but not the full trajectory. We draw two consequences: sweep points are only compared *within* this replication, against its own baselines, and the noise floor below is estimated from these runs rather than assumed. Every range brackets the reference on both sides and includes values that switch a mechanism off ($\phi = 0.05$, $\gamma = 1.0$, $c = 1$), because a sweep that cannot show a mechanism mattering cannot show it working either.

What counts as an effect. All sweeps are read against a noise floor estimated from the seed replicates and fixed before any sweep ran: $\sigma_\Delta = 2.20$ ES-AUC points for a within-seed delta, so the

Table 20: **Sensitivity of the headline result to every swept hyperparameter**, on the FairVLMed R50+SGD replication, 3 seeds per point. Swing is the range of mean ΔES-AUC across the sweep, in units of the seed-noise floor $\sigma_\Delta = 2.20$ points; a swing below 1.0 cannot be told apart from re-running with another seed. "Conclusion holds" gives the sub-range over which the paper's claim for that arm survives (harm, for the audited pruners; savings without fairness loss, for EquiPrune). Rows are ordered by swing within each block.

Parameter	Arm	Range swept	Swing (σ_Δ)	Conclusion holds
Pruning intensity and schedule (audited methods)				
δ_{end}	B-PAS	5×10^{-4}–8×10^{-3}	1.6	entire range
target DSI	EL2N	0.15–0.55	1.1	DSI $\geq 35\%$ (not 46%)
δ_{start}	B-PAS	10^{-5}–1.6×10^{-4}	1.0	entire range
anneal	InfoBatch	0.75–1.00	0.9	not resolvable
prune prob.	InfoBatch	0.25–0.75	0.8	DSI $\geq 41\%$
score epoch	EL2N	5–40	0.7	not resolvable
EquiPrune guardrail hyperparameters				
rescale attribute	EquiPrune	4 attrs. / all	1.2	any single attribute
floor ϕ	EquiPrune-F	0.05–0.40	1.1	all but ϕ=0.10
clip c	EquiPrune	1–100	1.0	$c \geq 2$
quota γ	EquiPrune	0.70–1.00	1.0	$\gamma \leq 0.95$
floor ϕ	EquiPrune	0.05–0.40	0.9	$\phi \geq 0.15$
Selection controls (no claim attached)				
target DSI	Random (DSI-matched)	0.15–0.55	0.4	n/a
trajectory	Random (traj.-matched)	B-PAS drop curve	n/a	n/a

standard error of a three-seed mean is 1.27 (a paired test-set bootstrap gives a similar $\sigma = 1.43$). Each parameter's *swing*, the range of its mean ΔES-AUC across the sweep, is reported in units of σ_Δ, and a swing below $1\sigma_\Delta$ is treated as not resolvable rather than as a trend. This convention makes it harder, not easier, for us to call a hyperparameter inert. Table 20 summarizes all thirteen swept axes.

The audit holds at every pruning intensity (R1). B-PAS is the most heavily swept method and the most stable conclusion. It lowers ES-AUC at all five threshold endpoints, from -0.79 at the default to -4.33 at the most aggressive (-2.40 to -2.57 at the other three), and at all three start points. Its swing is resolvable ($1.6\sigma_\Delta$ for δ_{end}), but it is a swing in size, not in sign: the harm never disappears, and it is largest at the most aggressive setting. The two budgeted pruners behave as the regime-dependence of Section 4 suggests: both are harmless at low savings and become harmful only once the budget grows, InfoBatch past about 41% DSI and EL2N at 35% and 55% (its 46% point reverts to -0.09, within noise of zero, so the boundary is ragged rather than sharp). This does not weaken the audit; it locates it. Demographic harm is a dose effect of how much training data is discarded, and a practitioner who prunes gently is genuinely safe. When the EL2N score is measured does not matter (swing $0.7\sigma_\Delta$ over an $8\times$ range of scoring epochs), so the harm is not an artifact of a noisy early score.

The EquiPrune hyperparameters need no tuning. Table 21 reports (ϕ, γ, c) in full. Two of the three are flat to within the noise floor across their plausible range: the quota γ swings $0.7\sigma_\Delta$ over 0.70–0.95, and the clip c cannot be told apart from its default anywhere in $[2, 100]$, a $50\times$ range, with 53–57% DSI throughout. The claim fails only at the settings that switch a guardrail off: $c = 1$ disables compensation, and $\gamma = 1.0$ demands exactly proportional retention, vetoes every candidate and collapses savings to 0%. These are not robustness failures; they show that the mechanisms carry weight, which is why they were placed inside the ranges. γ and c can therefore be fixed in advance, as they are here.

The floor ϕ behaves as Proposition 1 predicts. The one guardrail setting with a resolvable effect is the floor, and it moves in the predicted direction. Proposition 1(iv) says that a floor guaranteeing $m_g \geq \phi$ caps every group's risk multiplier at $\phi^{-\alpha}$, so a lower ϕ allows more damage. That is what we see: the floor-only arm rises by about 4.8 ES-AUC points per unit of ϕ (Table 21, upper left), sits above baseline at $\phi \in \{0.15, 0.25, 0.40\}$ and loses that margin at $\phi = 0.10$. The trend is not

Table 21: **Full sweeps of the three EquiPrune hyperparameters** (ϕ, γ, c), fixed in advance at $(0.15, 0.9, 5)$ for every dataset. ΔES-AUC is the mean within-seed change against that seed's own unpruned baseline, $\pm$ s.d. over 3 seeds; † marks the default. The claim fails where a guardrail is switched off (c=1: no compensation; γ=1.0: no pruning at all) or the floor is set below the default; the floor-only arm at ϕ=0.05 is the one exception, landing inside the equivalence band.

Setting	ΔES-AUC	DSI	Holds	Setting	ΔES-AUC	DSI	Holds
Floor ϕ (F only)				*Quota γ (F+Q+R)*			
$\phi = 0.05$	-0.04 ± 2.20	63.0%	✓	$\gamma = 0.70$	$+1.29 \pm 0.77$	56.8%	✓
$\phi = 0.10$	-0.80 ± 0.78	58.7%	×	$\gamma = 0.80$	$+0.35 \pm 0.94$	54.5%	✓
$\phi = 0.15^{\dagger}$	$+1.50 \pm 2.20$	56.0%	✓	$\gamma = 0.90^{\dagger}$	$+1.39 \pm 0.47$	53.6%	✓
$\phi = 0.25$	$+0.07 \pm 2.39$	50.4%	✓	$\gamma = 0.95$	$+1.94 \pm 1.73$	53.9%	✓
$\phi = 0.40$	$+1.67 \pm 1.10$	42.7%	✓	$\gamma = 1.00$	-0.30 ± 1.65	0.0%	×
Floor ϕ (F+Q+R)				*Clip c (F+Q+R)*			
$\phi = 0.05$	-0.67 ± 1.60	58.4%	×	$c = 1$	-0.62 ± 3.64	54.7%	×
$\phi = 0.15^{\dagger}$	$+1.39 \pm 0.47$	53.6%	✓	$c = 2$	$+0.44 \pm 2.37$	56.6%	✓
$\phi = 0.40$	$+0.55 \pm 2.74$	41.9%	✓	$c = 5^{\dagger}$	$+1.39 \pm 0.47$	53.6%	✓
				$c = 10$	$+1.67 \pm 2.48$	54.3%	✓
				$c = 100$	$+0.87 \pm 0.94$	54.4%	✓

perfectly monotone: $\phi = 0.05$ lands at -0.04 ± 2.20, inside the equivalence band, so we report the trend together with its one ragged point. The sweep turns $\phi = 0.15$ into a reported point on a known trade-off: raising ϕ to 0.40 leaves the arm at $+1.67$ ES-AUC against $+1.50$ at the default, while savings fall from 56% to 43% DSI. A practitioner who wants more safety knows which dial to turn and what it costs.

One attribute at a time is the right design for R. Compensating any single protected attribute works about equally well: race, gender, ethnicity and language all leave race ES-AUC at or above baseline (spread 1.1 points, within noise), which reproduces the cross-attribute spillover that motivated the design. Compensating all four at once is the only variant that fails (-1.27 ± 0.62): four sets of weights multiply into a high-variance objective that over-constrains training. R is therefore a single-attribute mechanism, and its benefit does not depend on choosing the right attribute in advance. This matters in practice, since the attribute one most wants to protect is not always the one that is best measured.

Pooled over all sweep runs, EquiPrune flattens the fairness–efficiency slope. Regressing ΔES-AUC on realized DSI with seed fixed effects over all 66 content-aware pruning runs gives a slope of -0.620 points per $+10$ points of DSI (95% CI $[-0.869, -0.372]$, $p < 10^{-4}$): fairness degrades steadily with how much data is discarded. The same regression over the 66 EquiPrune runs gives $+0.061$ per $+10$% DSI (95% CI $[-0.266, +0.387]$, $p = 0.71$), an order of magnitude smaller and not different from zero. At matched savings the gap grows with intensity: at 60% DSI the fitted loss is -2.90 points for content-aware pruning against -0.33 for EquiPrune. This is the paper's central claim drawn as a curve rather than a point, estimated from 132 runs spanning 0–79% DSI.

The selection controls support $\mathbf{H}_{\text{opt}}$. Two control arms that exist only in this study separate *what* is deleted from *how much*. Random pruning matched to EL2N's DSI at five budgets cannot be told apart from content-aware pruning (difference in slopes -0.105 per $+10$% DSI, 95% CI $[-0.592, +0.382]$). Random pruning that replays B-PAS's exact drop schedule with randomly chosen batches is, if anything, *worse* than B-PAS (-2.71 ± 0.56, $p = 0.014$, against -0.79 ± 1.54 at the same DSI). Demographic harm does not need a biased selection signal; removing data is enough. This supports the retention accounting of Section 4 and Appendix D from an independent direction, and explains why a demographics-blind floor is enough to repair it.

Nor is the result an artifact of an analysis choice. Nine post-hoc axes re-score the saved predictions under other conventions at no training cost: the two ES-AUC definitions in circulation, the two equalized-odds definitions, including or excluding missing-data groups, the corrected versus README FairVLMed language encoding, five minimum-group-size floors, five parity decision thresholds, three model-selection rules, and a 1000-fold paired bootstrap of the test set. The sign

of every arm is stable across all of them: B-PAS, InfoBatch and EL2N are negative under $5/5$ ES-AUC variants, and EquiPrune, EquiPrune-Q and EquiPrune-F positive under $5/5$. The two ES-AUC definitions differ only in how strict they are on this arm: under the ES-AUC of Section 2, the one we report, EquiPrune's gain in this replication is resolvable ($+1.39 \pm 0.47$, $p < 0.05$), while under the max–min definition of earlier work it has the same sign but is not resolvable ($+0.59 \pm 0.71$). Sweeping the decision rules themselves, the equivalence tolerance ($0.25/0.5/1.0$ points) crossed with the minimum-savings floor ($10/15/25\%$ DSI), leaves the EquiPrune verdict unchanged in all nine combinations. The R1 verdict of this single replication regime holds at tolerances of 0.25 and 0.5 points but not at 1.0, where only EL2N's harm still clears the bar. The main campaign's R1 verdict, which rests on nine regimes rather than one, is unaffected.

Scope. This is a one-at-a-time study around a single regime. It shows that our conclusions do not depend on the hyperparameters we chose, not that no interactions exist. Robustness across architecture, optimizer and dataset comes from the nine-regime main campaign (Section 4), and robustness across pruning signals from the generality study (Appendix I); together they cover the axes a reader would reasonably ask about. Full per-point tables, the swing ranking, the dose–response fits and all nine post-hoc grids are released with the code.

L Extended related work

Our work sits where four literatures meet: data pruning for efficient training, the fairness of efficiency mechanisms, fairness-aware data selection, and fairness evaluation in medical imaging. We review each and then state what remains open.

L.1 Data pruning and coreset selection

Static selection. The dominant approach scores every training example once, usually after a short warmup run, and trains on the highest-scoring subset. Scores include forgetting events (Toneva et al., 2019), gradient norm and error ℓ_2-norm (GraNd and EL2N; Paul et al., 2021), proxy-model uncertainty (Coleman et al., 2020), distance to a class prototype in embedding space (Sorscher et al., 2022), and estimated influence on generalization (Yang et al., 2023). Because top-k selection loses diversity at aggressive pruning rates, later work adds coverage, graph or distributional constraints (Zheng et al., 2023; Maharana et al., 2024; Hong et al., 2024), or aggregates scores over the training trajectory rather than at one checkpoint (Zhang et al., 2024). All of these methods commit to a single subset before most of training happens.

Dynamic pruning. A second line prunes *during* training, so the kept set changes with the model. Raju et al. (2021) periodically re-select on uncertainty or ε-greedy scores, and He et al. (2023) track dynamic uncertainty at scale. InfoBatch (Qin et al., 2024) temporarily skips a random half of the below-mean-loss samples each epoch and rescales the survivors' gradients by $1/(1-r)$ so that the expected update is unbiased; our rescaling guardrail R lifts this idea from samples to demographic groups. RCAP (Hassan et al., 2025) sets *class-wise* keep ratios from aggregated class loss and resamples within each class. Batch Pruning by Activation Stability (B-PAS; Alam et al., 2026) is the most aggressive member of this family and the state of the art in reported savings. It fixes batch membership for the whole run, monitors the epoch-over-epoch change in each batch's internal activation statistics, and *permanently retires* a batch once that change falls below a rising threshold, reporting up to 57% fewer samples processed and 61% fewer GPU node-hours at no cost in average accuracy. Two properties make B-PAS the sharpest object of study for us. Its signal is purely internal, using no labels, no losses and no demographic information, and its deletions are irreversible, so a group's data, once removed, never returns. Across this literature the reported metric is average accuracy; to our knowledge no dynamic pruner has been evaluated on subgroup performance.

L.2 Fairness of efficiency mechanisms

The closest warning for our setting comes from *model* compression. Hooker et al. (2019; 2020) showed that pruned and quantized networks keep top-line accuracy by sacrificing a small set of atypical, long-tail and often minority-group examples. Paganini (2020) documented the same effect as per-class performance imbalance across roughly 10^5 pruned image classifiers, and argued for

population-aware reporting of accuracy–efficiency trade-offs. Tran et al. (2022) traced pruning's disparate impact to group differences in gradient norm and distance to the decision boundary, with a "rich get richer" effect in which already-accurate classes gain and already-weak classes lose. Gee et al. (2023) found that the worst-group robustness of compressed language models depends on *which* compression method is used rather than on model size alone, an early case of the setting-dependence we observe on the data side. A repair literature followed. FairGRAPE (Lin et al., 2022) preserves each group's share of total weight importance when pruning face-attribute models; FairPrune (Wu et al., 2022) uses the fact that a parameter's importance differs by group to prune toward a smaller accuracy gap in dermatological diagnosis; Zayed et al. (2024) prune attention heads that harm fairness while keeping those that matter for performance; and Hashemizadeh et al. (2024) cast the problem as constrained optimization that bounds the per-subgroup accuracy change between the dense and the sparse model. Bias amplification has also been documented for dataset distillation, the synthetic counterpart of data pruning (Cui et al., 2024).

All of these results concern *which weights* an efficiency method removes. The complementary question, what happens when an efficiency method removes *data*, adaptively, while training is still running, is the subject of this paper. The analogy is suggestive but not decisive: weight pruning removes capacity that all groups share, while data pruning removes supervision that is already unequally distributed, through a signal that is itself shaped by how well the current model fits the majority.

L.3 Fairness-aware data selection and reweighting

A large body of work *imposes* fairness constraints on which data a model sees. Classical results give coresets for clustering with fairness constraints and provable approximation guarantees (Huang et al., 2019). FairBatch (Roh et al., 2021) treats per-group sampling probabilities as the outer variable of a bilevel program and re-balances minibatch composition to minimize a fairness loss. Data-diet approaches select for fairness rather than accuracy: Zayed et al. (2023) prune factual and counterfactual examples to maximize demographic parity and equalized odds, and Verma et al. (2021) remove training points identified as sources of individual discrimination, improving both fairness and accuracy. FairDeDup (Slyman et al., 2024) makes semantic deduplication of web-scale vision–language corpora group-aware. Instance reweighting is the soft analogue: FORML (Yan et al., 2022) meta-learns per-sample weights against an equality-of-opportunity objective, and group distributionally robust optimization (Sagawa et al., 2020) up-weights the worst-performing group directly. In medical imaging, Puyol-Antón et al. (2021) trace racial bias in cardiac MR segmentation to training-data imbalance and reduce it with stratified batch sampling. This is the closest existing analogue to a group-aware constraint on batch composition, although it balances which data enters a batch rather than limiting which batches are deleted. Recent work extends fairness-aware curation to label correction (Sulaiman & Roy, 2025) and to human-in-the-loop coreset selection over tabular data, where users can audit and adjust which regions a coreset drops (Hadar et al., 2025). A parallel line studies fairness under a *labeling* budget rather than a compute budget: fair active learning chooses which unlabeled points to annotate so as to close group gaps (Tae et al., 2024). The difference matters here. Active learning decides what to *acquire* from an unlabeled pool, while dynamic pruning decides what to *discard* from data already collected and labeled, so the question is not which examples are worth buying but which already-paid-for examples a group can afford to lose.

These methods are the natural comparison for our repair, but they differ in objective. They start from a fairness goal and design a selection rule to serve it, accepting whatever efficiency follows. We start from an existing, efficiency-driven selection rule that practitioners have strong reasons to use, leave its signal and schedule untouched, and ask what minimal constraint at the moment of deletion restores equity. The guardrails of EquiPrune are correspondingly cheap: no auxiliary model, no bilevel optimization, no extra forward passes, and demographic labels are used only for retention bookkeeping, never by the pruning signal.

L.4 Bias audits of data selection

Three recent studies audit data selection for bias and are the closest prior work to our audit. DRoP (Vysogorets et al., 2025) is the first systematic study of data pruning's effect on *classification bias*: it shows that standard one-shot pruning can yield highly biased classifiers, and proposes distribu-

tionally robust class-wise pruning ratios with random pruning within each class. RCAP (Hassan et al., 2025) targets worst-*class* accuracy under dynamic pruning with adaptive class-wise keep ratios. Dharmasiri et al. (2025) show that coresets with the same average accuracy can differ sharply in robustness to *spurious correlations* and in worst-group accuracy on group-shift benchmarks. Further afield, Guha et al. (2024) audit a different automated data operation, error detection and cleaning, over more than $26{,}000$ model evaluations. They find that it is more likely to worsen fairness than to improve it, and that whether a technique helps depends on the fairness metric and on whether groups are single-attribute or intersectional. This is the closest methodological precedent for our finding that pruning harm depends on the regime, although it concerns a preprocessing step rather than a mechanism inside the training loop. Two gaps remain. First, the audited groups are class labels (Vysogorets et al., 2025; Hassan et al., 2025) or synthetic spurious attributes (Dharmasiri et al., 2025), not *real protected attributes* such as race, gender, ethnicity, language and age, where the imbalance is severe, the groups cut across classes, and the harm is a deployment risk rather than a benchmark artifact. Second, apart from RCAP, the audited setting is *static* one-shot selection, where the pruning decision is made once and does not interact with training. The failure mode we study, a group-blind internal signal declaring a batch finished and then deleting it permanently, can only arise when pruning is dynamic.

L.5 Fairness in medical imaging

Medical imaging is where our audit matters most, because demographic imbalance there is structural rather than incidental. Seyyed-Kalantari et al. (2021) showed that chest-radiograph classifiers systematically underdiagnose under-served populations, with the largest error rates for intersectional subgroups. Yang et al. (2024) showed across radiology, dermatology and ophthalmology that models encoding demographic shortcuts can be locally fair on the training distribution yet lose that fairness under deployment shift, a reminder that fairness is a property of a training *setting*, not a certificate carried by a method. MEDFAIR (Zong et al., 2023) benchmarks bias-mitigation methods across eleven medical datasets and finds that no method dominates and that model-selection criteria matter as much as the mitigation algorithm. The Harvard Ophthalmology AI Lab has released a series of fairness datasets and methods: Harvard-FairVLMed with FairCLIP (Luo et al., 2024b), which regularizes group-conditional and overall feature distributions with a Sinkhorn distance and introduces the equity-scaled AUC (ES-AUC) we use as our primary metric; FairVision (Luo et al., 2024a) with fair identity scaling for three eye-disease screening tasks; FairSeg (Tian et al., 2024a), which adds fair error-bound scaling for segmentation; and FairDomain (Tian et al., 2024b), which studies fairness under domain shift. FairMedFM (Jin et al., 2024) extends fairness benchmarking to twenty foundation models over seventeen datasets and reports that disparities are consistent across models on the same data and that existing mitigation methods help only a little. This literature treats fairness as a property of *objectives, architectures and pretraining*, with the training data held fixed. We hold the objective and architecture fixed and vary only which data is skipped, which isolates data selection as a separate source of fairness risk, one that a practitioner adopting a pruner to save cost would not think to check.