IDENTIFICATION IS NOT ENOUGH: WHICH MAJORITY EXAMPLES ARE FLAGGED CHANGES LAST-LAYER CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stratified last-layer correction fixes worst-group failures by upweighting examples a selector scores as likely minority, and selectors are compared by how well those flags recover the annotated minority. We test whether these summaries predict the outcome of the correction. Our matched control permutes, within each class, the scores of majority examples and leaves every minority example's score untouched. This provably fixes AUROC, precision, recall, the detected minority set and the expected weight each group receives, and changes only *which* majority examples carry the flagged weight. It needs group labels: a diagnostic, not a method. On CelebA, with two selectors (prediction confidence and a patch-blur score), random reassignment raises test worst-group accuracy on all five backbones under both balanced-sampling and expected-weight fitting. On the three original backbones the selectors' own assignments rank in the bottom 0–2.5% of 200 random reassignments on test, though higher (up to 44%) on two later backbones: the majority examples a selector flags, boundary cases frozen ERM often misclassifies, are worse to upweight than random ones. This is not only a threshold shift: after re-tuning each head's threshold on validation, reassignment still gains 2.7–3.2 worst-group accuracy points with no group losing more than 0.2, and a threshold-free worst-group-pair AUC agrees. Flagging the easiest majority examples helps about as much; fresh sampling seeds alone move accuracy under 1.4 points. The effect depends on the dataset: on Waterbirds and MetaShift, whose selectors flag few majority false positives, it is near zero under expected-weight fitting. At the default setting even a test-selected best reassignment trails frozen ERM; at validation-tuned settings random reassignment beats tuned class balancing in 7 of 10 backbone–selector pairs on test but only 5 of 10 on validation. Identification summaries and expected group weights are therefore not sufficient to predict the outcome of stratified last-layer correction.

1 INTRODUCTION

A classifier can learn to read a bird's class from its background, then fail when that association breaks. Pseudo-group correction addresses such failures without training group labels (Liu et al., 2021; Qiu et al., 2023; Han & Zou, 2024). A *selector* scores every training example, for instance by a frozen model's loss; the top-scoring fraction of each class is *flagged* as likely minority, and the last layer is refitted with every class-by-flag stratum balanced. We call this *stratified last-layer correction*. Selectors are compared by how well their flags identify the annotated minority: precision, recall and AUROC. Those summaries describe agreement with a group label, while the correction fits a loss on individual examples. Does anything else matter once these summaries and the weight each group receives are held equal?

Two selectors differ in how many examples they flag, which minority examples they recover and which majority examples they sweep in; we fix the first two and vary the third. Within each class we permute the scores of majority examples and leave every minority score attached to its own example. This keeps *exactly* the minority identities and their scores, the classwise flag counts, every true-group-by-flag count and therefore the expected weight each group receives, and with them AUROC, precision and recall. Only which majority examples receive the flagged and unflagged

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.

weights changes. Comparing two independently built selectors cannot isolate this, because they also differ in their score distributions and in the minority examples they find; keeping the same score multiset and minority assignments isolates where the majority weight goes. The control needs group labels, so it is an evaluation diagnostic for selectors, not a deployable algorithm.

On CelebA the effect has a consistent direction: **the majority examples a selector flags are worse to upweight than random ones.** We study two selectors, prediction confidence and a patch-blur score, on ResNet-50 backbones trained by empirical risk minimization (ERM), with two ways of fitting the corrected layer (§3; Figure 1). On the three original backbones the selector's own assignment ranks in the bottom 0–2.5% of 200 random reassignments on test; on two backbones trained later it ranks higher, up to 44%. On Waterbirds and MetaShift, whose selectors flag only 0–12 majority false positives per backbone, the response is small and depends on how the layer is fitted. On CelebA the gain is class-wide rather than targeted, blond females gaining more than the blond males who form the worst group, and it is carried by exchanges inside the non-blond class, which by construction leave every blond-class summary unchanged. It also survives re-tuning each corrected classifier's threshold on validation (§6).

The next question is **which majority examples should carry high weight**. The flagged ones sit on the hair-colour boundary: frozen ERM calls about half of the flagged non-blond, female-annotated images blond, against a tenth of their randomized replacements (Figure 2). We therefore compare reassignments that keep the flagged set's difficulty profile, measured by deciles of frozen-ERM loss, against exchanges that move the same number of flags. Holding difficulty fixed shrinks the gain on the original three backbones, but for confidence the sign reverses on the two new ones: difficulty is a plausible explanation, not an established mechanism.

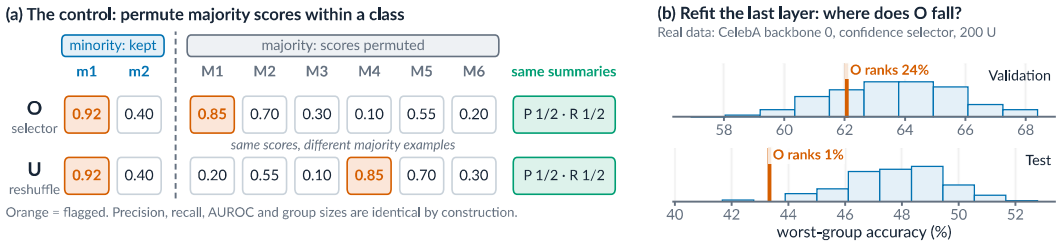


Figure 1: **The matched control.** (a) Schematic: majority scores are permuted within a class and minority scores stay with their examples, so every identification summary is unchanged. (b) Real data: the selector's own assignment O against 200 random reshuffles U (CelebA backbone 0, confidence). The diagnostic needs group labels; it is not an annotation-free correction method.

Contributions.

- **A matched control for selectors.** Permuting majority scores within class provably preserves AUROC, precision, recall, the minority set, every group-by-flag count and the expected weight of every group (§3, Proposition 1).
- **Evidence that these summaries do not predict the outcome of stratified last-layer correction.** On CelebA, reshuffling improves test worst-group accuracy on all five backbones, for both selectors and both estimators, beyond a threshold shift and a sampling placebo; on Waterbirds and MetaShift the response is small (§5; Table 3).
- **Where the effect comes from.** Selectors spend their false positives on the class boundary, the effect is carried by the non-blond class, and difficulty-matched controls attenuate it without settling the mechanism (§6).
- **An audit protocol** that reports a selector's rank, dose and classwise identification beside class-balanced and ERM references (§9), with anonymous code.

We make no claim about pseudo-group methods beyond stratified last-layer correction, use benchmarks we had used before, and report every adverse result. Table 1 defines the arms and symbols used in every table; U−O and H−D answer different questions and are never combined into a ratio.

Table 1: **Notation used in every table.** A *block* is one backbone with one selector; U has 5 reshuffles per block in Table 3 and 200 for ranks, H and D 20 each. All arms except N keep every quantity of Proposition 1 fixed.

Arm	Meaning	Symbol	Meaning
O	The selector's own flags	Sampled	Mean of 20 heads on balanced draws
U	Majority scores permuted within class	Expected	One head, draw-probability weights
H	U with loss-decile flag counts kept	U−O, H−D	Paired WGA difference (points)
D	H's control: same flag changes, dose	Rank	% of 200 U with lower WGA than O
E	Flags moved to the easiest majority	k, C	Flag fraction .10; L2 constant 1
N	Minority flags only (unmatched)	Dose T	TV distance between weights
CB, ERM	Class balancing; frozen ERM head	Orig./new	CelebA backbones 0–2 / 3–4

2 Related work

DFR establishes that frozen representations support robust last-layer retraining (Kirichenko et al., 2023). JTT upsamples early errors and AFR weights a head fit by prediction confidence (Liu et al., 2021; Qiu et al., 2023). SELF studies example choice once group balance is equalized (LaBonte et al., 2023); GIC shows that structured group-inference errors can hurt sampling-based correction yet help Mixup (Han & Zou, 2024); groups have also been inferred from twin networks' mistakes, a biased model's failures, subclass clustering or environment inference (Pezeshki et al., 2024; Nam et al., 2020; Sohoni et al., 2020; Creager et al., 2021; Zhang et al., 2022), and class balancing is a strong reference (Idrissi et al., 2022). These works already show that a partition can help training without recovering the annotated groups. Recent label-free methods balance class-conditional feature distributions (Zhao et al., 2025) or generate data for cluster-derived subpopulations (Luo et al., 2026), and a concurrent report flags samples whose predictions flip under blur, arguing that false positives are cheap because recall matters more than precision (Kumar, 2026). Our control bears on that premise without testing it directly: at matched precision and recall, which majority examples are flagged still moves worst-group accuracy.

D3M and Auto-D3M target the examples driving subgroup failure by data attribution (Koh & Liang, 2017; Ilyas et al., 2022; Jain et al., 2024); GSR optimizes last-layer example weights with influence functions and a group-labelled target set (Qiao et al., 2025); PPA and HSFM couple group inference or feature editing with head correction (Zhu et al., 2025; Parast et al., 2026). Our control complements these methods rather than proposing another weighting rule.

3 Matched-assignment control

Groups, flags and matching. The fitting pool D_r holds frozen features z_i, class labels y_i and attributes a_i; a true group is the pair $g_i = (y_i, a_i)$. Within each class the majority attribute is whichever value is most common in the *original training split*, not in the nearly balanced Waterbirds validation pool. On CelebA, female is modal in both hair classes, so the minority set includes non-blond males as well as blond males. A false positive is a majority example the selector flags as minority; the term does not mean its class label is wrong. An example is flagged when its score falls in the top k of its own class, $q_i = \mathbf{1}[s_i \geq Q_{1-k}(s \mid y_i)]$, with ties retained. We permute majority scores within class and hold minority scores fixed. This changes which example carries which weight, but leaves the classwise score multisets, the detected minority set and every group-by-flag count exactly as they were, and with them the weight each true group receives in expectation (its *expected group allocation*; Proposition 1; Appendix A.1). The question is whether identical identification evidence, through different per-example weights and hence a different head, can give a different worst-group accuracy.

The control is weaker than equating the whole training problem, since an exchanged majority example takes the other stratum's weight: it tests how weight is assigned within groups, not what adding false positives does.

Two correction estimators. We fit the corrected head, a logistic last layer on frozen features, in two ways (the two *estimators*) and report them separately. *Sampled* DFR-style fitting draws $n_{\min}$ examples without replacement from every observed class-by-flag stratum and averages the coeffi-

cients and intercepts of 20 logistic heads, not their predictions; both flag states take part. Every assignment uses the same random seeds, but changed flags change the strata and hence the drawn examples; a placebo bounds this effect at under 1.4 points (Appendix D.5). Fitting uses L2 regularization with $C = 1$ and an unpenalized bias. As references, class-balanced fitting ignores the flags and true-group fitting uses the group labels. The original DFR code uses L1 regularization and a final refit, so we call ours DFR-style.

Expected-weight fitting replaces the draw by its expectation. For stratum size n_s and head $\theta = (W, b)$ it minimizes

$$\overline{F}_q(\theta) = \sum_i \pi_i(q)\ell_i(\theta) + \frac{\|W\|_F^2}{2C}, \qquad \pi_i(q) = \frac{n_{\min}}{n_{s_q(i)}}. \tag{1}$$

The weights sum to the draw size $Sn_{\min}$ over S strata; rescaling that total without rescaling C changes the effective regularization. Matching fixes how much weight each true group receives in expectation, not which losses that weight sits on, so two matched assignments differ only by $\sum_i[\pi_i(q') - \pi_i(q)]\ell_i(\theta)$. This is not the limit of averaging infinitely many sampled heads, since $\mathbb{E}[\arg\min F_{q,r}] \neq \arg\min \mathbb{E}[F_{q,r}]$ in general.

Response and dose. Worst-group accuracy (WGA) is accuracy in the least accurate observed group. For a fitting procedure $\mathcal{A}$ with randomness ξ, the paired response on a fixed evaluation cohort D_e is

$$\Delta_{\mathcal{A}}(q, q'; \xi) = \mathrm{WGA}[\mathcal{A}(D_r, q'; \xi); D_e] - \mathrm{WGA}[\mathcal{A}(D_r, q; \xi); D_e]. \tag{2}$$

A positive value means the reassignment gives higher WGA; expected-weight fitting has no randomness ξ. Accuracy is a percentage and differences are percentage points, and nested permutations and fits are averaged within each backbone before any summary. We measure how large an intervention is by the total-variation distance between normalized weights, $T(q, q') = \frac{1}{2}\sum_i |a_i(q') - a_i(q)|$ with $a_i = \pi_i / \sum_j \pi_j$, and refer to it as the *dose*. Equal dose need not mean the same change in feature geometry.

4 Experimental Design and Identification

Benchmarks. Waterbirds predicts bird class, with land/water background groups (Sagawa et al., 2019; Wah et al., 2011; Zhou et al., 2018); CelebA predicts hair color, grouped by its binary Male annotation (Liu et al., 2015); MetaShift predicts cat versus dog with indoor/outdoor context groups (Liang & Zou, 2022). Each has four class–attribute groups, and the CelebA minority set includes non-blond males as well as rare blond males. The non-blond fitting class is near-balanced by this attribute (about 4,200 each), yet holds 82–86% of the flags a CelebA reassignment changes, and §6 shows it carries the effect. These are dataset labels, not descriptions of identity.

Frozen features. The original backbones are three ResNet-50 (RN50) models (He et al., 2016) per dataset, trained by ERM from ImageNet1K V1 weights (Deng et al., 2009) with class-balanced sampling, 100 epochs on Waterbirds and 20 on CelebA. Each seed splits validation into the two pools in Table 2. Features are standardized with the held-out pool's mean and standard deviation (plus 10^{-6}); no test statistics are used. Training, fitting and test images have disjoint IDs, but CelebA contains some exact duplicate images across splits; removing the 13 affected test images moves the means by under 0.18 points (Appendix B.2). Two later CelebA backbones (seeds 3 and 4) repeat the recipe.

Table 2: **Benchmark protocol.** Image counts; all four groups enter WGA. Waterbirds and CelebA split validation into the fitting and held-out pools; MetaShift fits on 20% of its training split and holds out its validation split. Draw: the balanced sample, four strata of $n_{\min}$ each.

Dataset	Classes / groups	Training	Fitting	Held-out	Test	Draw size
Waterbirds	2 / 4	4,795	599	600	5,794	52–56
CelebA	2 / 4	162,770	9,933	9,934	19,962	608–624
MetaShift	2 / 4	1,821	455	349	874	84

Identification regimes. Original and reassigned ROC curves agree pointwise in all 60 original–reassigned pairs (Appendix Figure 5). CelebA's weak pooled AUROC is misleading because it averages two very different classes. In the blond class, which contains the worst group, identification is informative: .727–.741 for patch-blur and .849–.880 for confidence; in the non-blond class it falls below chance. A pooled AUROC therefore describes no class the correction actually fits. Each class carries its own threshold, so the flag rule is not a single pooled-ROC operating point (Appendices B.1 and C.1).

Assignments and fitting. For each original backbone we build the original assignment and five reassignments from the confidence score $1 - p(y \mid x)$ and the patch-blur score $1 - p(y \mid x_{\text{perturbed}})$. Patch-blur replaces the image patches that a gradient estimate marks as most important with a blurred copy; correction always uses features of the clean image (Appendix A.2). At $k = .10, C = 1$, fitting seeds 0–4 are crossed with all six assignments, each averaging 20 fits. All 60 pairs pass score, identity and allocation checks, and every head was fixed and hashed before test evaluation (Appendix B.1).

5 Does matched reassignment change correction?

What counts as independent evidence. The unit of replication is the trained backbone: three per dataset on Waterbirds and MetaShift, and five on CelebA (three original, plus two trained after the original results, which are not a confirmatory replication), plus three ViT backbones per dataset (§7.1). Reshuffles and fitting seeds are nested within a backbone and averaged before any summary, so they add precision, not replication. All CelebA backbones are evaluated on one test set of 19,962 images, whose worst group holds only 180, and are fitted on halves of one validation set. Seed intervals therefore rest on three or five backbone means ($df = 2$ or 4); test-resampling intervals hold the fitted heads fixed and add no training runs; and ranks among reshuffles are conditional descriptions. Holm adjustment over the eight headline CelebA effects is in Appendix B.2.

For backbone b, permutation j and fitting seed r, write $d_{bjr} = \mathrm{WGA}(q_{bj}; r) - \mathrm{WGA}(q_{b0}; r)$. Each reassignment stays paired with the original at the same fitting seed; we average fitting seeds and permutations within a backbone, then summarize backbone means. Table 3 gives O, U and U−O for every dataset, selector and estimator.

Table 3: **The selector's own assignment (O) against matched reshuffles (U), test WGA (%).** U averages five reshuffles per backbone (200 on MetaShift). U−O is in points, mean ± SD of backbone means; n_+: backbones with U>O, sampled/expected. Rank: O's percentile among 200 reshuffles under expected weighting, range over backbones (—: not run). Shaded: the headline CelebA rows; all CelebA backbones share one test set. Validation results are in Appendix B.1.

		Sampled			Expected weight				
Data (backbones)	Selector	O	U	U−O	O	U	U−O	n_+	Rank
Waterbirds (3)	Patch-blur	80.8	83.9	3.09 ± 3.23	86.6	86.6	-0.03 ± 0.20	2/1	—
	Confidence	73.4	78.9	5.57 ± 4.75	84.6	85.1	0.42 ± 0.23	3/3	—
CelebA, original (3)	Patch-blur	65.4	72.5	$\mathbf{7.05} \pm 4.60$	46.7	51.1	$\mathbf{4.44} \pm 0.38$	3/3	0–2.5
	Confidence	66.8	75.4	$\mathbf{8.58} \pm 2.52$	43.7	49.5	$\mathbf{5.78} \pm 1.35$	3/3	0–1.0
CelebA, all (5)	Patch-blur	66.0	73.1	$\mathbf{7.12} \pm 3.29$	48.1	51.5	$\mathbf{3.36} \pm 1.52$	5/5	0–44.3
	Confidence	67.5	75.6	$\mathbf{8.08} \pm 1.96$	45.6	50.3	$\mathbf{4.76} \pm 1.76$	5/5	0–16.5
MetaShift (3)	Patch-blur	72.2	74.7	2.45 ± 2.01	76.4	75.7	-0.72 ± 1.48	3/1	15.5–97.5
	Confidence	68.0	73.6	5.62 ± 1.20	75.4	75.9	0.50 ± 1.86	3/2	0–98.5

Sampled correction. Every CelebA backbone responds positively under both selectors. Waterbirds and MetaShift respond too, but Waterbirds varies far more: its sampled draws hold only 13–14 examples per stratum, a permutation changes at most eight flags (1,270–1,400 on CelebA), and for patch-blur on backbone 1 all 61 flags are true minorities, so the response is exactly zero and stays in the summary. Of the four original three-backbone seed intervals only CelebA confidence excludes zero, $[2.31, 14.84]$ (Appendix B.1).

Expected-weight correction. Expected-weight fitting is deterministic, uses the same $C = 1$, unpenalized bias and total weight as the sampled draws, and passes an independently implemented sta-

tionarity check, so its responses cannot come from fitting randomness. On Waterbirds and MetaShift the responses fall to within a point of zero. On CelebA they remain positive on every backbone for both selectors. Subsampling is therefore not needed for a sizeable matched response on CelebA, whereas elsewhere the response largely disappears without it. This does not attribute the sampled response to sampling variance: optimizing an expected loss and averaging fitted parameters are different estimators, and refitting O with fresh sampling seeds alone moves WGA by under 1.4 points (Appendix D.5).

Test resampling and granularity. Paired test-resampling intervals (2,000 group-stratified resamples) exclude zero for CelebA under both estimators and scores. **On validation, every interval across the original three backbones includes zero** (Appendices B.2 and B.1). CelebA's worst test group holds 180 images, so a 4.44-point response is eight images: effects of a few points show direction more than size.

Matched summaries do not determine correction. On CelebA the response is positive on every backbone under both estimators; on Waterbirds and MetaShift it depends on the estimator.

6 Who is flagged, conditional ranks and training replication

Selectors spend their false positives on the label boundary. The majority examples a selector flags are not a random draw from the majority, and a reshuffle replaces them with typical ones (Figure 2). Averaged over five backbones, frozen ERM calls 45–60% of the flagged non-blond, female-annotated images blond, against 10% of their replacements, and under 1% of them carry CelebA's black-hair annotation, against 19% of the class. In the blond class ERM misclassifies 29–69% of the flagged images against 5% (Appendix C.2). For confidence, which flags the highest-loss examples, this is true by construction; what does not follow is that upweighting them is harmful. Upweighting majority examples near the hair-colour boundary plausibly pulls the decision boundary against blond images, and randomization removes that pull. Annotations describe the images; they do not establish label error.

Not only a threshold shift. On the five backbones' expected-weight heads, one intercept shift chosen on validation lifts mean test WGA of the O, U and class-balanced heads from 46–51% to 74–78%, above ERM. After this shift U–O is still 3.24 $[1.78, 4.70]$ (patch-blur) and 2.69 $[0.85, 4.52]$ (confidence), and no group loses more than 0.2 points. Without any threshold, the worst AUC over pairs of groups from the two classes improves by 0.73 $[0.38, 1.08]$ and 1.49 $[1.03, 1.95]$. Giving the flags to the easiest majority examples instead, itself one of the allowed reshuffles, gains 3.67 and 2.33 points over O after the same shift (Appendix D.5).

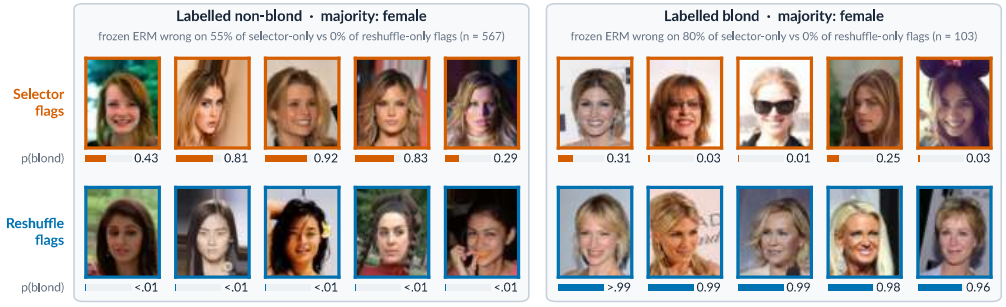


Figure 2: **Whom the selector flags, and whom a reshuffle flags instead.** Majority (female) examples flagged by confidence on CelebA backbone 0 but not by reshuffle 0 (orange), and the reverse (blue): the first five of each in stored order, not curated. Bars are frozen-ERM $p(\text{blond})$; headers give ERM's error rate over all such examples. Rates over all five backbones are in Appendix C.2.

How unusual is the selector's own assignment? For each CelebA backbone and selector we fit expected-weight heads under 200 unrestricted reshuffles at $k = .10$, $C = 1$. On the original three backbones' test split the selector's own assignment ranks in the bottom 0–2.5%: 194–200 of the

200 reshuffles strictly improve WGA in every block. On the two backbones trained later it ranks much higher, at 30.00/44.25% for patch-blur and .75/16.50% for confidence, so across all five the ranges are 0–44.25% and 0–16.5%, not universally the bottom 0–2.5%. Validation is less uniform: patch-blur on backbone 0 ranks at the 80th percentile. These are conditional descriptive ranks, given the trained backbones, not randomization p-values or fresh training runs. Figure 3 shows the distributions for the original three backbones; Appendix C.1 states how ties are counted.

The non-blond class carries the effect. Refitting the same 200 reassignments with exchanges confined to one class (4,000 heads, protocol frozen before fitting) gives test responses of 2.97 $[0.67, 5.27]$ and 4.32 $[1.47, 7.16]$ points for non-blond-only exchanges (patch-blur/confidence, five backbones), against 0.02 $[-1.50, 1.53]$ and 1.10 $[0.75, 1.44]$ for blond-only exchanges at similar dose; the two restricted responses add up to the unrestricted one within error, and validation shows the same ordering (Appendix Table 23). The effect also appears when blond males are treated as the only minority. Exchanges inside the non-blond class then leave pooled and blond-class AUROC, precision, recall and per-class flag counts unchanged by construction (verified on all 2,000 assignments), yet changing which non-blond females carry the flag still raises blond accuracy, by 5.5/7.8 points for blond females and 3.0/4.3 for blond males, while non-blond groups move by under two points: the worst group improves as part of a class-wide shift, not through targeted help.

Constraining difficulty needs a same-size control. We build 20 reshuffles that stay inside the frozen-ERM loss deciles of each majority class (H), each paired before any outcome is seen with an unrestricted reshuffle D that moves *the same number of flags in each class, at the same dose*. Deciles, not exact loss ranks, leave something to exchange for confidence, which is a monotone function of loss; D keeps a smaller intervention from being mistaken for an effect of difficulty. H–D asks whether the constraint matters at equal weight.

Table 4: **Constraining difficulty attenuates test gains beyond dose.** Expected-weight CelebA WGA response (points), mean ± SD of three backbone means. U: 200 unrestricted reshuffles; H: 20 constrained to loss deciles; D: 20 controls matched to H's classwise flag count and dose. H–D is paired before aggregation. The confidence contrast does not persist on the two further backbones.

Score	Split	U	H	D	H–D
Patch-blur	Val.	1.96 ± 3.45	0.10 ± 2.38	0.75 ± 2.38	-0.65 ± 0.13
	Test	4.12 ± 1.39	0.40 ± 0.33	2.02 ± 1.36	-1.62 ± 1.07
Confidence	Val.	3.65 ± 2.61	0.32 ± 1.77	0.13 ± 2.30	0.19 ± 0.53
	Test	6.04 ± 2.38	-0.18 ± 1.29	1.67 ± 1.17	-1.84 ± 0.97

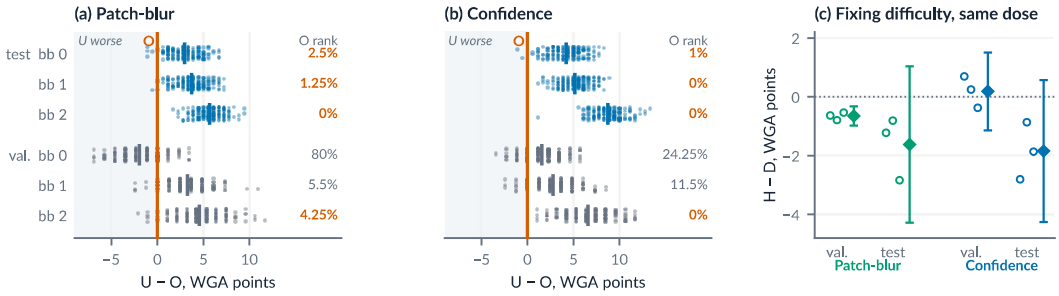


Figure 3: **Where the selector's own assignment ranks, and what constraining difficulty does.** (a,b) Each dot is one of 200 unrestricted reshuffles U on one backbone (bb), placed by its WGA gain over the selector's own assignment O (orange line); bars mark the mean. O rank is the share of reshuffles worse than O; orange marks 5% or less. (c) H–D over 20 matched pairs: seed points, mean and a 95% t interval ($df = 2$). Ranks are conditional descriptions, not randomization p-values.

On the original three backbones the H–D test contrasts (Table 4) are negative on every backbone for both scores, but their 95% intervals across backbones still include zero; patch-blur is negative on every validation seed, while confidence does not. Test-resampling intervals, $[-2.58, -0.66]$ and $[-2.79, -0.95]$, exclude zero but condition on fixed heads, and both validation resampling intervals include zero (Appendix C.1). Restricted to one class, the H–D test contrast sits mainly in the non-blond class, not the blond class that contains the worst group (Appendix C.3).

Two additional training seeds: persistence and reversal. Figure 9 and Table 27 in Appendix C.5 report the original three backbones, the two new ones and all five together. Responses over the five standard reshuffles stay positive on test under both estimators. The 200-reshuffle effect is much smaller on the two new patch-blur backbones, 1.072 and .169 points, consistent with their higher conditional ranks. Patch-blur H–D stays negative on both new seeds and its five-seed test interval excludes zero; confidence H–D turns positive on both, so its five-backbone interval includes zero. The H and D arms match in achieved dose and class composition on all five backbones, so this reversal is not a design artifact (Appendix B.4). Validation shows the same split. A bounded continuous-loss control gives the same signs, with every interval including zero (Appendix C.4).

At the fixed setting, low rank is not useful absolute correction. On CelebA, expected weighting lowers original WGA from 65.44 to 46.67% (patch-blur), and the best of $C \in \{.1, 1, 10\}$ reaches only 50.19% (Appendix Table 15). Table 5 places O and the 200 reshuffles on the original three backbones beside expected-weight class-balanced (CB) heads at two fixed weight scales. Even the retrospective, test-selected best U trails ERM and sampled CB at this fixed setting; a validation-chosen intercept alone lifts these heads above ERM (§6). The two CB scales differ despite identical relative class weights, because the total weight changes effective regularization at $C = 1$.

Table 5: **Absolute utility at the fixed setting (CelebA, expected weight, original three backbones).** Test WGA (%). Left: O, the mean over 200 reshuffles and the test-selected best reshuffle (an oracle) for each selector. Right: references shared by both selectors, mean ± SD over backbones. Bold: references that beat every reassignment.

Assignment	Patch-blur	Confidence	Reference	WGA
Original (O)	46.67	43.70	CB, natural total	47.78 ± 2.22
Mean U	50.79	49.74	CB, matched total	53.89 ± 1.67
Best U (oracle)	54.44	53.89	Frozen ERM	$\mathbf{67.41} \pm 1.16$
			Sampled CB	$\mathbf{70.04} \pm 2.29$

Constraints attenuate patch-blur, not confidence. Confidence H–D reverses on both new backbones, and at the fixed setting low ranks give no useful absolute utility.

7 Scope: Dose, Selection and Representations

The response across flagging fractions. Refitting the expected-weight control at six flagging fractions from 2% to 50% on all eight backbones, the five-backbone CelebA response falls steadily as k grows while the number of flags moved rises, and is exactly zero at $k = .50$, where expected weights barely depend on the flags; its intervals exclude zero for both scores from $k = .02$ to .10 (Table 6). Waterbirds stays within $[-.34, .42]$ throughout (Appendix B.5). A larger intervention is not a larger effect; at fixed k, reassignments of growing TV dose and a regularization sweep show no dose–response trend either (Appendix B.3). A synthetic score family with exchangeable majority scores gives at most .50 points at any AUROC from .30 to .90, so the pipeline does not manufacture a response, though this cannot show that identification quality is irrelevant (Appendix B.6).

Table 6: **Matched response across the flagging fraction (CelebA, five backbones).** Expected-weight test WGA response in points, backbone mean with a descriptive 95% t interval ($df = 4$); bold: interval excludes zero. Changed: flags moved per permutation (patch-blur/confidence). At $k = .50$ (shaded) expected weights barely depend on the flags, so the response is zero.

k	Patch-blur	95% CI	Confidence	95% CI	Changed
.02	**6.18**	$[1.31, 11.04]$	**10.40**	$[8.31, 12.49]$	324/336
.05	**5.04**	$[1.83, 8.26]$	**8.53**	$[4.40, 12.67]$	745/774
.10	**3.36**	$[1.61, 5.10]$	**4.76**	$[2.74, 6.78]$	1326/1356
.20	1.93	$[-.30, 4.17]$	**2.98**	$[1.73, 4.22]$	2112/2155
.30	1.38	$[-.18, 2.94]$	.64	$[-.71, 2.00]$	2579/2602
.50	.00	$[0, 0]$	.00	$[0, 0]$	2791/2737

7.1 Does the finding extend to selected settings and transformers?

At the RN50 settings chosen by validation in the original study, reused without retuning or forcing pseudo-groups on (Appendix Table 33), four of the six CelebA selector–backbone choices fall back to class balancing, which ignores the flags and responds by exactly zero; every selected-setting three-seed interval includes zero. The one CelebA backbone whose selection kept the flags gives 3.00/1.02 points sampled and 4.33/2.00 expected-weight. Refitting with flags forced on and k, C tuned on validation WGA gives sampled heads that beat frozen ERM in 9 of 10 CelebA blocks ($+3.47$ $[1.88, 5.06]$ points) and match tuned class balancing. At these tuned settings the reshuffle response is $+2.16$ $[-0.59, 4.92]$ on test and 0.00 on validation, carried by the four blocks where tuning chose $k \leq .20$ (a grouping chosen after seeing the results; Appendix D.3). Reshuffled heads beat tuned class balancing in 7 of 10 test blocks ($+1.61$) but 5 of 10 validation blocks ($+0.34$); heads fitted on the selector's own assignment do not (-0.55).

Six ViT-B/16 checkpoints (Dosovitskiy et al., 2021) repeat the control at the fixed $k = .10$, $C = 1$: CelebA effects stay positive on every backbone for both scores and estimators (6.89–10.97 points), and Waterbirds stays within $[-0.18, 0.30]$. The checkpoints differ in pretraining, recipe and preprocessing, so this replicates the pipeline, not the architecture (Appendix D.2). On MetaShift (Liang & Zou, 2022), selectors flag only 5–12 majority false positives per backbone; sampled responses exceed a sampling placebo, expected-weight ones stay near zero (Table 3; Appendix D.4).

8 What predicts the size of the response?

Two exploratory analyses ask what predicts the response (Appendix E). Of six refit-level descriptors, only correction influence (Koh & Liang, 2017), a curvature-adjusted first-order estimate of how reweighting moves worst-group loss, explains a material share of the worst-group margin change: $R^2 = .195$ within settings, against .026 for dose, loss displacement, feature-centroid and covariance shifts and the size of the head move together. It predicts how large the response is ($\rho = .91$ across settings, and within each dataset), but on CelebA, where a permutation moves about 1,320 flags and a first-order expansion is far from accurate (Bae et al., 2022), it barely orders individual reassignments (median per-setting R^2 .02–.09, against .47–.58 on Waterbirds; Appendix Figure 11). The backbone's reliance on the shortcut does not predict the response either, in a 36-backbone background-by-marker factorial.

9 Discussion and conclusion

A reporting procedure, not a replacement metric. With identification held fixed, the central 90% of 200 reassignments spans 5.4 test WGA points on average over the ten CelebA backbone–selector blocks (7.0 on validation). The selectors' own assignments lie in the tail of this range. The range shows how much WGA identification leaves unconstrained, and correlations between identification and response across reshuffles supply no causal or predictive rule. We therefore propose an audit rather than a scalar: refit the correction under matched reassignments and report the selector's rank, the dose and the classwise identification beside class-balanced and ERM references at a stated weight scale. Sensitivity needs false positives that are both on the boundary and numerous: Waterbirds' and MetaShift's number 0–12 (Appendix C.2).

Limits. The diagnostic needs group labels, and all backbones of a dataset are evaluated on the same test set. Treating backbones as units, all eight headline five-backbone CelebA test effects survive Holm adjustment (adjusted $p \leq .049$); none of their validation counterparts does, with under 100 images in the validation worst group (Appendix B.2). Protocols were fixed before new fits but designed after earlier results, and no experiment establishes the mechanism. No real matched-selector pair, text benchmark or hardness-structured synthetic family is tested.

Conclusion. Standard identification summaries do not determine the outcome of stratified last-layer correction: on CelebA the majority examples a selector flags are worse to upweight than random or the easiest ones, beyond a threshold shift, while on Waterbirds and MetaShift the response is small. Tuned random reassignment beats class balancing on test only; the mechanism remains open.

Ethics statement

We use only public benchmarks (Waterbirds, built from CUB and Places; CelebA; MetaShift; and Places365 backgrounds) and collect no new data from people. CelebA contains face images with annotated attributes, including a binary sex annotation; we use these only as the benchmark's group labels. The face composites in Appendix E are programmatic edits of benchmark images made for diagnosis, not demographic editing tools. The matched control needs group labels and is an evaluation diagnostic; our results do not certify equitable deployment.

Reproducibility statement

Appendix A defines the matching, both correction estimators and both selector scores, and proves the matching invariants (Proposition 1). Appendices B–D give data splits, backbone training, fitting settings, seeds and every per-backbone result, and Appendix F describes the saved records and what was verified. An anonymous implementation is provided as supplementary material. It covers the flag rule and the matching-invariant checks, the O, U, H, D and E arms and the N and class-balanced references, both estimators, the confidence and patch-blur scores, recalibrated WGA and the worst-pair AUC, the influence audit, ResNet-50 training on Waterbirds, CelebA and MetaShift, and scripts from backbone training to the rank summaries, with unit tests. The ViT pipeline and the background-by-marker factorial are not included. A separate replay package with de-identified predictions (Appendix F.2) passes local replay but is not released, because redistribution requires license review.

AI use statement

Generative AI assisted literature retrieval, design feedback, coding, debugging, analysis, mathematical formulation and proof checking, figure preparation, drafting and editing. No generated images are used: the photographic composites are programmatic edits of benchmark images. The authors reviewed all AI-assisted work: code was checked by unit tests and by independently written verification programs (Appendix F), and the proofs and reported numbers were checked against the saved results. Computational checks are not independent human review. The authors take full responsibility for the content of this paper, including text, claims and artifacts produced with the aid of generative AI.

References

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *NeurIPS*, 2022.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. Environment inference for invariant learning. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009.

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 *ICLR*. OpenReview.net, 2021.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17480–17504. PMLR / OpenReview.net, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE Computer Society, 2016.

Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 2022.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *CoRR*, abs/2202.00622, 2022.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Improving subgroup robustness via data selection. In *NeurIPS*, 2024.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. *CoRR*, abs/2306.11957, 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*. OpenReview.net, 2023.

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

Nilesh Kumar. Perturbation sensitivity at convergence: A simple signal for identifying spuriously correlated samples. *CoRR*, abs/2608.05419, 2026.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *NeurIPS*, 2023.

Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *ICLR*. OpenReview.net, 2022.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pp. 3730–3738. IEEE Computer Society, 2015.

Yiming Luo, Rongqiang Zhao, and Jie Liu. Sage: Subpopulation-aware generative enhancement for mitigating spurious correlations, 2026. URL https://arxiv.org/abs/2609.01051.

Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *CoRR*, abs/2007.02561, 2020.

Jun Hyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *ICLR*. OpenReview.net, 2022.

Aryan Yazdan Parast, Khawar Islam, Soyoun Won, Basim Azam, and Naveed Akhtar. HSFM: hard-set-guided feature-space meta-learning for robust classification under spurious correlations. *CoRR*, abs/2603.29313, 2026.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with XRM. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40551–40569. PMLR / OpenReview.net, 2024.

Rui Qiao, Zhaoxuan Wu, Jingtan Wang, Pang Wei Koh, and Bryan Kian Hsiang Low. Group-robust sample reweighting for subpopulation shifts via influence functions. In *ICLR*. OpenReview.net, 2025.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28448–28467. PMLR, 2023.

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. *CoRR*, abs/1911.08731, 2019.

Nimit Sharad Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020.

Saeid Asgari Taghanaki, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *NeurIPS*, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*, July 2011.

Shirley Wu, Mert Yüksekgönül, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37765–37786. PMLR, 2023.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961. PMLR, 2024.

Michael Zhang, Nimit Sharad Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 2022.

Miaoyun Zhao, Qiang Zhang, and Chenrong Li. Class-conditional distribution balancing for group robust classification. *CoRR*, abs/2504.17314, 2025.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464, 2018.

Beier Zhu, Jiequan Cui, Hanwang Zhang, and Chi Zhang. Project-probe-aggregate: Efficient fine-tuning for group robustness. In *CVPR*, pp. 25487–25496. Computer Vision Foundation / IEEE, 2025.