DETECTION IS NOT REPAIR: EVALUATING BUDGETED LABEL VERIFICATION FOR GROUP ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does uncovering more label errors necessarily improve worst-group accuracy after retraining? In budgeted data cleaning, standard practice prioritizes detection precision—identifying as many corruptions as possible within a fixed query budget. In this work, we investigate this premise when group annotations are inaccessible during data acquisition, and uncover a fundamental decoupling between error detection and downstream group robustness. Across extensive empirical evaluations on vision and text benchmarks, we document widespread empirical inversions where acquisition policies that correct substantially fewer errors achieve markedly superior worst-group accuracy compared to high-precision detectors. This divergence arises because high-precision selection systematically starves underrepresented, vulnerable subgroups of corrections, skewing the retrained decision boundary unfavorably. Mathematically, we formalize this structural failure: without group annotations, even an idealized oracle detector incurs a worst-case regret approaching 50 percentage points due to the inherent non-identifiability of latent group structure. Finally, we explore practical remedies in the absence of group metadata, showing that simple observed-class quotas safeguard against decision boundary destabilization, while tracking post-retraining clean validation improvement guides policy selection in balanced vision domains. Our results reposition data cleaning from passive noise filtering to an active intervention on classifier decision boundaries, underscoring that label verification must be evaluated by downstream subpopulation robustness rather than raw detection metrics.

1 INTRODUCTION

Correcting more label errors can leave the worst-performing subgroup worse off. Consider budgeted verification on CelebA: the gradient-attribution baseline TracIn-CP identifies label corruptions with an exceptional 99.38% precision, yet worst-group accuracy (WGA) drops by 4.92 percentage points (pp) after retraining. In contrast, under the exact same verification budget, simple training-loss ranking catches fewer corruptions overall, yet shifts the classifier's boundary favorably, boosting WGA by 5.21 pp. Thus, an acquisition rule can excel as an error detector while paradoxically degrading subpopulation robustness.

Budgeted label verification queries human or oracle annotators to rectify a limited subset of training labels under resource constraints. While classical data-cleaning pipelines treat error detection as the primary objective, the downstream utility of data cleaning depends fundamentally on how label rectifications reshape the decision boundary. Prior research has connected task-aware cleaning to average-case generalization (Krishnan et al., 2016; Bernhardt et al., 2022) and investigated data filtering for subpopulation robustness (Jain et al., 2024). In this work, we investigate the critical intersection of these two areas under a realistic operational constraint: when group annotations are inaccessible during acquisition, *to what extent does error detection precision govern post-retraining worst-group accuracy?*

To formalize this inquiry, we systematically decouple *detection precision* (the proportion of queried labels that are corrupted) from *downstream robustness recovery* (the change in WGA after retraining). In this paradigm, an acquisition policy operates strictly without group metadata during verifi-

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.

cation, and its downstream utility is assessed only after retraining on the partially verified dataset. Through controlled empirical benchmarks and theoretical analysis, we examine why high detection precision fails to translate into downstream robustness, trace how label corrections distribute across latent subpopulations, and establish actionable principles for selecting verification policies in practice.

Figure 1 summarizes the verification protocol and the empirical, mechanistic, and policy-selection analyses developed in this work.

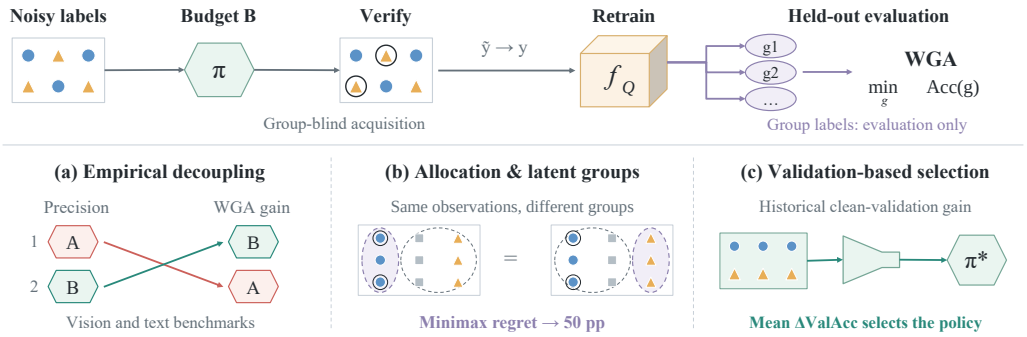

Figure 1: **Overview of budgeted label verification for group robustness.** The shared protocol verifies labels within a fixed budget, retrains on the partially corrected dataset, and evaluates held-out WGA. (a) Detection precision and downstream recovery can rank policies oppositely. (b) Correction placement matters; indistinguishable observations with different latent group partitions yield minimax regret approaching 50 pp in our theoretical construction, even with perfect error detection. Gray squares denote anchor examples. (c) Historical gains in clean, class-balanced validation accuracy guide policy selection. Group labels are withheld during acquisition and selection. Rings mark corrected labels. All geometry is schematic.

Our primary contributions are:

1. **Decoupling detection precision from downstream robustness.** Across standard vision and text benchmarks, we show that error detection precision routinely mispredicts downstream group recovery. Lower-precision policies consistently outperform high-precision baselines in WGA across both representation-learning regimes and real-world human annotation noise (Section 3).
2. **Mechanistic and theoretical limits of error allocation.** We demonstrate that raw correction count is insufficient: matched allocations in correction volume and flip direction yield divergent outcomes depending on minority-group coverage. Theoretically, we establish an information-theoretic minimax lower bound proving that without group annotations, even an idealized oracle detector incurs worst-case regret approaching 50 pp due to the non-identifiability of latent groups (Sections 4 and 5).
3. **Practical implications and exploratory group-blind remedies.** We explore how practitioners can navigate verification without demographic metadata. We show that simple class-balanced quotas (CB-Loss) prevent boundary destabilization across noise regimes, while monitoring clean validation recovery provides an actionable compass for model selection in vision domains (Section 6).[1]

2 Setting and Experimental Design

2.1 Budgeted Label Verification

Let $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ denote a training dataset with potentially corrupted labels. An acquisition policy selects a batch of examples $Q \subseteq [N]$ under a fixed query budget $|Q| \leq B$. Exact verifica-

[1]Code, configurations, and results: https://anonymous.4open.science/r/Budgeted-Label-Verification-for-Latent-Group-Robustness-10D8.

tion rectifies $\tilde{y}_i$ to its ground-truth label y_i for each $i \in Q$. Let f_Q be the classifier retrained on the partially verified dataset, and $f_\varnothing$ the baseline model trained on the unverified data. Given latent evaluation subgroups $\mathcal{G}$, we assess performance by worst-group accuracy (WGA) and the corresponding post-retraining utility gain:

$$\mathrm{WGA}(f) = \min_{g \in \mathcal{G}} \Pr(f(X) = Y \mid G = g), \qquad U(Q) = \mathrm{WGA}(f_Q) - \mathrm{WGA}(f_\varnothing). \tag{1}$$

Unlike data pruning, which discards suspicious examples, this framework preserves the training distribution size by rectifying verified corruptions.

An admissible acquisition policy constructs Q using only pre-verification information $\mathcal{F}$—including input representations, observed labels, training dynamics from probe models, and a modest clean validation set—without access to latent group identifiers. Subgroup annotations are held out strictly for post-hoc evaluation and theoretical analysis. Verification is conducted in a single batch, and querying an already-correct example consumes budget without modifying the label. Checkpoints are selected via clean-validation balanced accuracy.

2.2 Detection Precision vs. Downstream Group Performance

Let $Z_i = \mathbf{1}\{\tilde{y}_i \neq y_i\}$ indicate label corruption. The number of rectified errors is $C(Q) = \sum_{i \in Q} Z_i$, yielding detection precision $D(Q) = C(Q)/|Q|$. Under a fixed query budget, error count and precision induce identical method rankings.

To evaluate how verification queries distribute across subpopulations, we identify the worst-performing subgroup under the unverified baseline model, $g_s^\star \in \arg\min_g \mathrm{Acc}_g(f_{\varnothing,s})$, and define its error recall as:

$$R_s(Q) = \frac{\sum_{i \in Q} Z_i \mathbf{1}\{g_i = g_s^\star\}}{\sum_{i=1}^{N} Z_i \mathbf{1}\{g_i = g_s^\star\}}. \tag{2}$$

Tracking both baseline subgroup recall and post-retraining WGA accounts for potential shifts where retraining causes a different subgroup to become the performance bottleneck.

2.3 Datasets, Noise Regimes, and Query Policies

Benchmarks and representations. We evaluate on Waterbirds, CelebA, and CivilComments (Sagawa et al., 2020; Liu et al., 2015; Borkan et al., 2019), where subgroups are formed by target–spurious attribute intersections (e.g., *Eyeglasses* vs. *Male* on CelebA; *Toxicity* vs. *Black identity* on CivilComments). For feature-based studies, representations are extracted from an ImageNet-pretrained ResNet-50 for vision benchmarks and all-MiniLM-L6-v2 for text. We examine two canonical corruption regimes: *symmetric label noise* (20% uniform label flips) and *subgroup-skewed noise* (40% corruption on the minority subgroup, 10% on the majority). In addition, we evaluate on CIFAR-10N (Wei et al., 2022) to test generalization under real-world human annotation noise using worst-class accuracy (WCA) and end-to-end ResNet-18 retraining.

Verification policies. We benchmark standard error detection and data selection paradigms: (1) *Heuristic scores*: Loss ranking, predictive entropy, and uniform random sampling; (2) *Cleanlab*: Out-of-fold confident learning; (3) *Attribution and active cleaning*: Adaptations of ActiveClean, Active Label Cleaning, and D3M; (4) *Composite noise scores*: NoiseScore (combining normalized ranks of loss, confidence, entropy, and probe disagreement), its class-prior-weighted variant ClassPrior, and the risk-gated variant GatedCoverage; (5) *Validation-guided ranking*: RepairValue (ranking candidates by validation gradient projection) and its anchor-gated variant GatedRepairValue; and (6) *Class-balanced selection*: CB-Loss, which allocates queries equally across observed classes.

Evaluation and statistical rigor. All comparisons are strictly paired across identical random seeds, dataset splits, and retraining trajectories. We report 95% confidence intervals throughout (paired Student's t for fixed representations and bootstrap intervals for end-to-end retraining), defining an *empirical inversion* whenever an acquisition policy corrects fewer training errors than standard loss ranking yet achieves a superior gain in post-retraining WGA (controlling for multiple testing; see Appendix A).

3 Detection and Downstream Performance Decouple in Practice

3.1 Empirical Inversions Across Benchmarks

Across 510 paired comparisons on frozen representations, empirical inversions are widespread: in 80 instances, an acquisition policy that uncovers fewer label corruptions than standard loss ranking achieves a significantly higher post-retraining WGA (Table 1; 76 remain significant under multiple-testing control).

The CelebA benchmark offers a salient illustration of this decoupling. TracIn-CP achieves a near-flawless 99.38% precision in uncovering corrupted labels, yet retraining on its verified set consistently degrades worst-group performance by an average of 4.92 pp across ten random seeds. In contrast, standard training-loss ranking uncovers fewer errors overall but steers the decision boundary favorably, lifting WGA by +5.21 pp. As visualized in Appendix Figure 5, this divergence is not an isolated anomaly, but a pervasive phenomenon across diverse verification heuristics. Figure 2 separates the detection and post-retraining evaluation paths in this example.

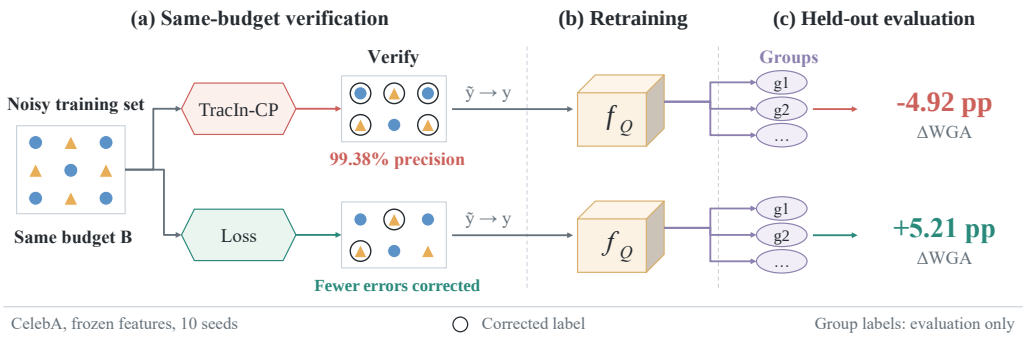

Figure 2: **Detection precision versus downstream repair.** Under the same verification budget, the CelebA frozen-representation experiment yields 99.38% precision but −4.92 pp WGA gain for TracIn-CP, whereas Loss ranking corrects fewer errors and yields +5.21 pp. The two paths distinguish verified-error detection from post-retraining evaluation. Sample and correction-ring counts are schematic, not measured proportions; group annotations are unavailable during acquisition.

Table 1: Empirical inversions in the original frozen-feature experiments. An empirical inversion occurs when an acquisition policy corrects fewer label corruptions than standard Loss ranking yet achieves a significantly superior post-retraining WGA (paired 95% CI strictly above zero). The last column applies Benjamini–Hochberg FDR control at $q < 0.05$ across all 510 eligible comparisons.

Dataset	Comparisons	Unadjusted	BH, $q < 0.05$
Waterbirds	190	29	26
CelebA	180	36	35
CivilComments	140	15	15

3.2 Generalization to End-to-End Representation Learning

A central question is whether end-to-end representation learning mitigates this misallocation through feature adaptation. To test this hypothesis, we evaluate policies under full end-to-end retraining across twenty held-out seeds, comparing policies chosen for high detection precision (π_{prec}) against those chosen for validation WGA improvement (π_{WGA}).

As reported in Table 2, feature adaptation does not resolve the tension. On CelebA, the precision-driven policy (TracIn-CP) achieves 99.76% precision but decreases WGA by 0.32 pp. Conversely, the utility-driven policy (RepairValue) operates at only 27.12% precision yet consistently improves group robustness, outperforming TracIn-CP by +0.76 pp across 18 of the 20 seeds. On Waterbirds,

Table 2: End-to-end evaluation of methods selected from development runs: 20% uniform noise, 10% query budget, twenty additional seeds per dataset. P and W denote the methods selected by mean detection precision (m^D) and WGA gain (m^U), respectively. Gains are relative to the uncorrected model for the same seed; W − P is their paired WGA difference. Brackets give 95% seed-bootstrap CIs (pp).

Dataset	Selected method	Prec. (%)	ΔWGA [95% CI]	W − P [95% CI]
Waterbirds	Loss (P)	91.61	$+4.68$ $[+2.33, +7.18]$	
	NoiseScore (W)	71.41	$+5.91$ $[+3.03, +9.12]$	$+1.23$ $[-0.38, +2.78]$
CelebA	TracIn-CP (P)	99.76	-0.32 $[-0.66, +0.02]$	
	RepairValue (W)	27.12	$+0.44$ $[+0.16, +0.72]$	$+0.76$ $[+0.51, +1.04]$

both policies deliver positive gains, with the lower-precision composite score NoiseScore yielding the higher recovery ($+5.91$ pp vs. $+4.68$ pp). These results confirm that representation learning does not automatically compensate for precision-driven error misallocation.

3.3 Robustness Under Real-World Human Annotation Noise

Table 3: CIFAR-10N, 2% budget. Ten training seeds share one human-noise realization. The metric is worst-class accuracy (WCA). Gains are relative to the uncorrected baseline, and paired differences compare each method with Loss. Brackets are 95% t intervals.

Method	Precision (%)	Corrected	ΔWCA (pp)	Paired vs Loss (pp)
Loss	99.70	997.0	$+1.10$ $[-2.10, +4.29]$	reference
NoiseScore	97.28	972.8	$+3.24$ $[-0.13, +6.60]$	$+2.14$ $[-0.06, +4.34]$
ClassPrior	97.35	973.5	$+2.95$ $[-0.61, +6.50]$	$+1.85$ $[+0.50, +3.19]$
Random	40.70	407.0	$+1.89$ $[-1.38, +5.15]$	$+0.79$ $[-1.12, +2.69]$

We further examine whether this decoupling persists under natural annotation errors using the CIFAR-10N benchmark under authentic human noise, evaluating worst-class accuracy (WCA) under end-to-end retraining. Under a 2% query budget, class-prior reweighting (ClassPrior) detects fewer errors than standard loss ranking (973.5 vs. 997.0) but achieves a $+1.85$ pp higher WCA (Table 3).

Crucially, verification efficacy exhibits substantial dataset dependence: under identical 20% noise and a 2% budget, NoiseScore achieves lower precision than Loss on both CelebA and CivilComments, yet its WGA gap relative to Loss is $+2.07$ pp on CelebA versus -2.00 pp on CivilComments (Appendix Table 9). This divergent behavior confirms that neither error count nor raw precision provides a reliable surrogate for subgroup performance, motivating an examination of sample-level allocation mechanisms.

4 Mechanisms of Downstream Performance Recovery

4.1 High Precision Can Starve Vulnerable Subgroups

Table 4: Corrections by group on CelebA, evaluation seeds 1000–1019, 10% budget. The initial worst group is identified from each uncorrected model. Recall is the mean within-seed error recall for this group; pooled counts are corrected/available training-label errors. Initial-group gain and WGA gain are in pp.

Method	Prec. (%)	Recall (%)	Pooled counts	Initial-group gain	ΔWGA
RepairValue	27.12	98.70	16913/17178	$+0.555$	$+0.440$
Loss	99.62	46.23	8412/17178	-0.317	-0.539
TracIn-CP	99.76	0.00	0/17178	-0.145	-0.319

Inspecting how verified examples distribute across subpopulations explains why high detection precision fails to translate into downstream robustness. On CelebA, TracIn-CP concentrates its queries almost entirely on overrepresented majority slices, correcting exactly zero errors within the baseline's worst-performing subgroup across all twenty seeds. In contrast, the gradient-guided rule RepairValue allocates queries to uncover 98.70% of the errors within this vulnerable subpopulation (Table 4).

Consequently, near-perfect precision can coincide with the complete omission of critical minority errors. By directly rectifying corruptions within the worst-off subpopulation, RepairValue improves that subgroup's accuracy by $+0.55$ pp, leading to a net WGA gain of $+0.44$ pp despite operating at a substantially lower overall precision.

4.2 Controlled Impact of Sample-Level Allocation

To isolate the effect of subpopulation allocation from raw verification volume, we construct strictly matched query sets on Waterbirds that hold the total number of rectified corruptions, the label transition directions ($0 \to 1$ vs. $1 \to 0$), and redundant clean queries constant, varying solely whether corrections fall inside or outside the baseline worst-performing subgroup.

Targeting corrections to the vulnerable subgroup yields a substantial $+3.24$ pp higher WGA compared to applying identical corrections elsewhere (Figure 3), with positive margins across 17 of the 20 seeds. This controlled comparison demonstrates that identical verification counts produce starkly divergent downstream models depending on the subpopulation placement and boundary alignment of the verified instances.

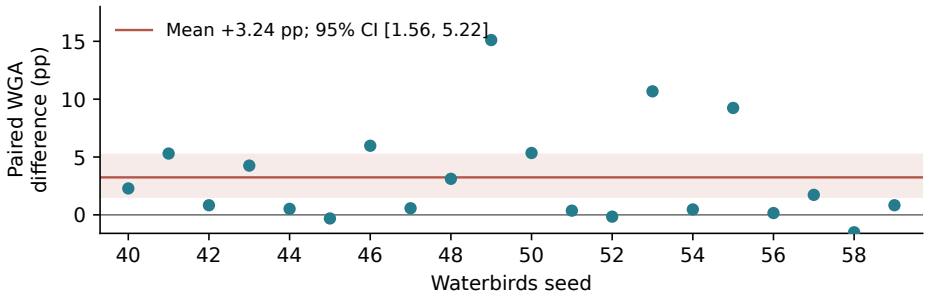

Figure 3: **Controlled verification allocation.** Post-retraining WGA differences between applying corrections inside versus outside the initial bottleneck subgroup on Waterbirds. Query sets strictly match correction counts, transition directions, and redundant clean queries. Points reflect seed averages; the shaded band depicts the 95% bootstrap interval.

4.3 First-Order Influence and Validation Gradients

To analytically characterize sample-level effects, we formalize the relationship between label corrections and parameter updates under regularized logistic retraining. A first-order Taylor expansion reveals that the predicted reduction in subgroup loss decomposes into the product of two factors: the total number of rectified errors and the average inner product between the corrected samples and the subgroup loss gradient (Appendix E.2). This decomposition elucidates why precision alone is structurally insufficient: precision controls the volume factor, while leaving the directional gradient alignment completely unconstrained.

Empirically, while validation-gradient projections (RepairValue) effectively steer queries on CelebA, their efficacy depends on the fidelity of the validation signal. On CelebA, injecting a modest 5% label noise into the scoring validation set reduces RepairValue's advantage over Loss from $+3.47$ pp to -1.39 pp (Appendix D.3). Thus, while gradient projections align query directions effectively when clean validation anchors exist, first-order approximations remain inherently sensitive to validation label noise. Crucially, downstream recovery is not governed by minority recall alone: because retraining solves a joint regularized objective across all data, unconstrained error harvesting

concentrated within a single class slice can destabilize global decision boundaries (motivating the balanced allocation remedies explored in Section 6).

5 Perfect Detection Still Leaves an Allocation Dilemma

Our empirical findings establish that error detection and downstream group recovery are fundamentally distinct objectives. We now ask whether an idealized, error-free detector could resolve this dilemma. By constructing two classification environments that share identical observable data distributions and verified label responses but diverge in their latent subgroup partitions, we establish an exact information-theoretic limit for group-blind verification.

For environment $w \in \{0, 1\}$, let $U_w(Q)$ denote the post-retraining WGA gain, $U_w^\star = \max_{|A| \le B} U_w(A)$ the group-aware oracle optimum, and $\mathrm{Reg}_w(Q) = U_w^\star - U_w(Q)$ the policy regret. Theorem 1 characterizes the exact minimax regret across randomized and sequential acquisition policies.

Theorem 1 (Information-theoretic minimax regret) *For integers $B \ge 2$ and $M \ge \max\{B + 1, B(B-1)\}$, there exist two binary classification environments and an ℓ_2-regularized logistic retraining objective with a unique global minimizer. Both environments share identical noisy training distributions, clean validation data, test marginals (X, Y), and label verification responses to every query. They differ only in their latent evaluation subgroup partitions, with identical group proportions across environments. Each environment admits an oracle utility $U_w^\star = M/(M + B)$. Let Π_B include all randomized and sequential group-blind acquisition policies querying at most B samples. Then:*

$$\inf_{\pi \in \Pi_B} \max_{w \in \{0,1\}} \mathbb{E}[\mathrm{Reg}_w(Q_\pi)] = \frac{M - B}{2(M + B)}. \tag{3}$$

Furthermore, this lower bound holds even if the acquisition policy is endowed with oracle knowledge of the corruption status of every training example.

Mechanism of non-identifiability. Each environment contains two disjoint blocks of B corrupted candidates. In environment 0, the first candidate block forms the performance-limiting subgroup, while the second block shares a subgroup with correctly classified anchor instances; environment 1 symmetrically inverts these subgroup assignments. Consequently, allocating the entire budget to one block is optimal in one environment but yields zero utility gain for the worst-off group in the other.

Because all observable features and verification responses are identical across environments, class-label queries cannot identify which block determines worst-group performance. Splitting the budget equally in expectation between the blocks attains the optimal worst-case expected utility of $1/2$ (Appendix G).

Fundamental limits of pure detection. As $M \to \infty$, oracle utility approaches 1, while every group-blind policy has expected utility at most $1/2$ in at least one of the two environments. The exact minimax regret therefore approaches 50 pp. Every candidate-only query has 100% detection precision, yet its utility depends on which block forms the worst group. Knowing all error locations does not identify the allocation that maximizes group performance.

6 Toward Practical Remedies: Exploring Group-Blind Proxies

If detection precision systematically misranks acquisition policies, how should practitioners approach verification when group annotations are inaccessible? While Theorem 1 establishes that no group-blind policy can eliminate worst-case regret in the worst case, empirical proxy signals and structural constraints can still offer actionable guidance. In this section, we explore practical remedies along two complementary axes: direct query allocation via observed-class quotas, and cross-budget policy selection via downstream clean validation monitoring.

Observed-class quota allocation (CB-Loss). A simple observed-class quota provides an effective first-line heuristic: Class-Balanced Loss (CB-Loss) allocates equal verification budgets to each observed class before ranking candidates by training loss. On CelebA and CivilComments, CB-Loss

consistently outperforms both standard Loss and NoiseScore across all eight evaluated noise and budget regimes (Appendix E.1), delivering advantages of up to $+13.79$ pp over Loss on CelebA and $+11.35$ pp over NoiseScore on CivilComments (with a $+2.45$ pp gain over NoiseScore on CelebA; Table 18). By distributing verification queries across classes, observed-class quotas help mitigate severe single-class boundary collapse (though like other heuristics, they can still yield negative repair utility under certain vision schedules; Appendix E.1), establishing a robust baseline especially in skewed text domains where unconstrained error harvesting concentrates exclusively in majority slices.

Validation-guided policy selection. Beyond fixed allocation rules, we investigate *downstream validation improvement*—the gain in clean, class-balanced validation accuracy after relabeling and retraining—as an exploratory compass to select policies across budgets without demographic labels. We evaluate this signal across twelve candidate methods under a leave-one-budget-and-seed-out cross-validation protocol: for any target budget and seed, policies are ranked by average validation gains across remaining budgets and disjoint seeds, directly testing cross-scale transfer without target-budget data leakage (Figure 4).

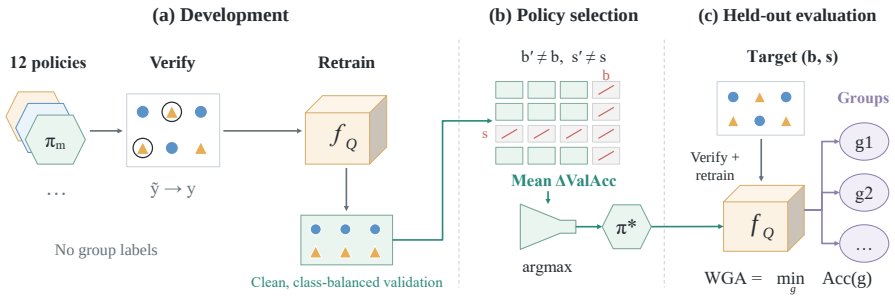

Figure 4: **Historical clean-validation recovery drives group-blind policy selection.** Candidate policies are scored by post-retraining gain in clean, class-balanced validation accuracy. Cross-budget selection excludes the target budget and seed from all source histories. Held-out WGA is evaluated strictly post-hoc.

Table 5: Validation-selected minus precision-selected held-out WGA (pp). Within-budget selection holds out the evaluation seed; cross-budget selection also holds out its budget and excludes that seed at every source budget. Each dataset averages ten fixed noise–budget conditions. Brackets are exploratory 95% seed-cluster bootstrap intervals (10,000 draws).

Dataset	Within budget	Cross budget
Waterbirds	$+11.33$ $[+9.33, +13.38]$	$+11.00$ $[+9.47, +12.78]$
CelebA	$+9.20$ $[+8.16, +9.91]$	$+8.81$ $[+7.96, +9.58]$
CivilComments	$+0.295$ $[+0.246, +0.344]$	-0.030 $[-0.041, +0.000]$

Cross-budget transfer and boundary conditions. Across held-out budgets, selecting policies by validation gain decisively outperforms precision selection on vision benchmarks, delivering an 11.00 pp WGA advantage on Waterbirds and 8.81 pp on CelebA (Table 5). Validation tracking consistently identifies acquisition strategies aligned with robust decision boundaries (e.g., selecting adapted D3M on Waterbirds, outperforming fixed NoiseScore by $+7.11$ pp; Appendix Table 24) because monitoring parameter behavior across classes penalizes harvesting errors exclusively from majority slices. In contrast, on CivilComments, validation selection performs comparably to precision selection (-0.03 pp margin). This distinction illuminates an important boundary condition: extreme text class imbalance (toxic comments comprise only $\sim 11\%$ of clean data, causing observed positive labels to suffer a 66% corruption rate; Appendix F) is hypothesized to increase validation evaluation variance across budgets, though the precise mechanism governing cross-budget transfer failure on skewed text benchmarks remains an open question (Appendix F).

Practical implications. Together, these exploratory findings establish practical guideposts for data verification pipelines: observed-class budgeting (CB-Loss) provides an immediate, zero-cost safeguard against single-class boundary collapse in skewed domains, while clean validation monitoring reliably steers higher-order policy choices in balanced vision applications.

7 Related Work

Sample attribution and data cleaning. Classical data cleaning prioritizes human review to optimize model improvements under budget constraints (Krishnan et al., 2016; Bernhardt et al., 2022). Corrupted samples are identified via learning dynamics, such as confident learning (Northcutt et al., 2021), forgetting events (Toneva et al., 2019), area under the margin (Pleiss et al., 2020), and dataset cartography (Swayamdipta et al., 2020). In parallel, data attribution methods—including influence functions (Koh & Liang, 2017), TracIn (Pruthi et al., 2020), and datamodels (Ilyas et al., 2022)—estimate instance influence on downstream predictions. Our work bridges these scoring signals with subpopulation robustness, investigating how instance attribution translates into post-retraining worst-group accuracy when subgroup memberships are unobserved.

Group robustness and data selection. Approaches targeting subpopulation shifts without explicit group labels typically exploit loss distributions or latent clustering (Hashimoto et al., 2018; Lahoti et al., 2020; Liu et al., 2021; Sohoni et al., 2020; Creager et al., 2021). In data selection, D3M removes training examples that drive model failures on minority groups (Jain et al., 2024), while Panaitescu-Liess et al. (2024) highlight trade-offs between poisoning defense and minority protection. Unlike data pruning or deletion, our framework investigates targeted label verification where all training samples are retained and rectified.

Unobserved subgroup structure. Prior literature has examined the identifiability of fairness metrics when protected attributes are unobserved (Kallus et al., 2022). We extend this perspective to active data acquisition, proving that even with clean validation data and exact label feedback, latent subgroup geometry imposes fundamental minimax limits on group-level utility recovery.

8 Discussion and Conclusion

This study establishes a core principle of data-centric robustness: *error detection precision does not govern downstream performance recovery*. Across empirical benchmarks and theoretical constructions, we demonstrate that verification policies cannot be judged by error counts alone. Rectifying high-confidence errors in majority subpopulations often starves vulnerable minority groups of corrections, shifting the retrained decision boundary unfavorably. Purely prioritizing detection accuracy thus creates a false sense of security in safety-critical domains, demonstrating that data cleaning acts as a direct intervention on classifier decision boundaries.

Future directions. Promising extensions include: (1) *sequential adaptive verification*, where intermediate model updates dynamically steer subsequent query batches; (2) *robustness under non-expert verifiers*, characterizing tolerance to imperfect oracle feedback; and (3) *latent group discovery*, incorporating self-supervised clustering priors to circumvent the group-blind lower bound. Ultimately, future research in budgeted data verification should systematically evaluate post-retraining group robustness alongside classical detection metrics.

AI use statement

Generative AI assisted with manuscript editing, analysis implementation, and consistency checks. The authors reviewed the work and take responsibility for the manuscript and results.

Ethics statement

The study uses public datasets and collects no new human-subject data. Group annotations are used for offline evaluation and reference method selection, not by the acquisition rules. Dataset licenses continue to apply.

Reproducibility statement

The source package provides per-seed results, configurations, analysis scripts, and checksums, including development choices and all twenty additional evaluation seeds per image dataset. Appendix I lists the materials and the dependencies needed for training-level reproduction.

References

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem Can Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13:1161, 2022. doi: 10.1038/s41467-022-28818-3. URL https://doi.org/10.1038/s41467-022-28818-3.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. Association for Computing Machinery, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

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

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587. PMLR, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Improving subgroup robustness via data selection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94490–94511, 2024. doi: 10.52202/079017-2995. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/abbbb25cddb2c2cd08714e6bfa2f0634-Abstract-Conference.html.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022. doi: 10.1287/mnsc.2020.3850. URL https://doi.org/10.1287/mnsc.2020.3850.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. ActiveClean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9(12):948–959, 2016. doi: 10.14778/2994509.2994514. URL https://doi.org/10.14778/2994509.2994514.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Wook Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 728–740, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425. URL https://doi.org/10.1109/ICCV.2015.425.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/JAIR.1.12125. URL https://doi.org/10.1613/JAIR.1.12125.

Michael-Andrei Panaitescu-Liess, Yigitcan Kaya, Sicheng Zhu, Furong Huang, and Tudor Dumitras. Like oil and water: Group robustness methods and poisoning defenses may be at odds. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ff7c9c90030f1eb3cbf5c81b6fbd9a05-Abstract-Conference.html.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17044–17056, 2020.

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

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. URL https://openreview.net/forum?id=ryxGuJrFvS.

Nimit Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Mariya Toneva, Alessandro Sordoni, Rémi 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. URL https://openreview.net/forum?id=BJlxm30cKm.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

A Training Protocols and Query Scores

This appendix specifies the training blocks, paired comparisons, and query rules used throughout the paper. The following appendices give the frozen-feature comparisons, end-to-end results, allocation experiments, theoretical analysis, and method-selection procedures.

Table 6: Experiment blocks. End-to-end (E2E) retraining updates all parameters from pretrained initialization. Method comparisons are paired within each block and seed.

Experiment	Seeds; budget	Model and comparison
Original frozen-feature	0–9; five budgets	ResNet-50 / MiniLM features; method comparisons
Extended frozen-feature	0–9; five budgets	30-epoch heads; method selection
CelebA TracIn-CP / RepairValue	10–19 / 20–29; 10%	ResNet-50 E2E; comparisons with Loss
Waterbirds TracIn-CP / RepairValue	10–19 / 20–39; 10%	ResNet-50 E2E; comparisons with Loss
Waterbirds Gated-RepairValue / random fill	40–59; 10%	E2E; gradient ranking vs. random fill
Additional-seed evaluation	1000–1019; 10%	E2E; methods chosen on development runs
Validation sensitivity	0–9 / 0–19	Frozen heads; scoring set and loss-tail fraction
CIFAR-10N	0–9; 1%, 2%	ResNet-18 E2E; human label noise

Training. ResNet-50 RepairValue and additional-seed experiments use ImageNet initialization, 224-pixel inputs, data augmentation, batch size 48, and AdamW with learning rate 3×10^{-5} and weight decay 10^{-4}. Training lasts at most ten epochs with early-stopping patience three. Checkpoints maximize clean-validation balanced accuracy.

CIFAR-10N uses pretrained ResNet-18, 64-pixel inputs, batch size 256, learning rate 10^{-4}, weight decay 10^{-4}, twenty epochs, and patience five. The supplementary configurations specify the remaining experiment settings.

Paired comparisons. The original frozen-feature analysis uses ten complete seed pairs with Loss for each eligible method–condition combination. For $d_s = U_{m,s} - U_{\mathrm{Loss},s}$, the paired 95% interval is $\bar{d} \pm t_{.975,9} s_d / \sqrt{10}$. Of 580 combinations, 30 are oracle results, 30 are Loss references, and ten lack complete pairing, leaving 510 comparisons.

An empirical inversion is confirmed when an acquisition policy achieves fewer mean corrections while maintaining a strictly positive lower endpoint for the paired WGA difference. Across the 510 pairwise comparisons, controlling the false discovery rate via the Benjamini–Hochberg procedure retains 76 of the 80 empirical inversions ($q < 0.05$). For allocation comparisons with potential positive test dependencies, the conservative Benjamini–Yekutieli procedure is applied across the 24 tests.

End-to-end evaluations report 50,000 seed-level percentile bootstrap draws, incorporating ties and deterministic resampling sequences documented in the supplementary archive. Appendix C and Appendix H detail the confirmatory evaluation and policy-selection bootstrap procedures, respectively.

A.1 Evaluation Protocols and Metrics

Comparisons within a training schedule. Each comparison pairs methods on the same dataset split, noise realization, budget, seed, probe, initialization, retraining schedule, and checkpoint rule. We use complete seed pairs within an experiment block. Absolute utility compares a corrected model with the uncorrected model trained under that schedule.

Group measurements. We identify the initial worst group from the uncorrected model, breaking ties deterministically, and retain that identity when measuring its recall and accuracy change. Final

WGA is recomputed over all groups. We separately count seeds in which another group becomes worst-performing.

Mean recall averages the seed-level ratios. Pooled recall instead sums corrected and available errors before taking their ratio, weighting seeds by their error counts. A seed with no errors in its initial worst group has undefined recall and is omitted from the mean; we report the eligible seed count.

Query Budget and Computational Scope. The query budget B strictly accounts for all selected examples, including instances where queried labels are already correct. Historical policy selection additionally leverages validation histories from prior development runs, which are evaluated strictly out-of-sample on held-out seeds. Appendix I details the dataset splits, checkpoints, and artifact mappings in the supplementary repository.

A.2 Diagnostic Pipeline and Stage Metrics

Table 7 outlines the progressive stages of our diagnostic framework, connecting instance acquisition to post-retraining group robustness. We report absolute gains relative to unverified baselines alongside paired policy differences under identical retraining schedules.

Table 7: Diagnostic metrics tracking the verification pipeline from candidate selection to downstream group robustness.

Pipeline Stage	Diagnostic Metric	Role in Empirical Analysis
Detection	Precision and rectified error count	Quantifies error-identification fidelity
Allocation	Subgroup error recall (R_s)	Tracks query distribution across latent slices
Repair	Post-retraining WGA gain and group shifts	Evaluates downstream decision boundary recovery
Policy selection	Held-out cross-budget WGA	Assesses group-blind acquisition transfer

A.3 Query Scores

Loss, Random, and NoiseScore. Loss ranks examples by cross-entropy under the observed label; Random samples uniformly. For probe probabilities q_i, define loss $\ell_i = -\log q_{i\tilde{y}_i}$, observed-label confidence $c_i = q_{i\tilde{y}_i}$, entropy $h_i = -\sum_k q_{ik} \log q_{ik}$, and disagreement $d_i = \mathbf{1}\{\arg\max_k q_{ik} \neq \tilde{y}_i\}$. NoiseScore averages their normalized ranks within each observed class:

$$\rho_i = \tfrac{1}{4}[\mathrm{nr}_{\tilde{y}_i}(\ell_i) + \mathrm{nr}_{\tilde{y}_i}(1 - c_i) + \mathrm{nr}_{\tilde{y}_i}(h_i) + \mathrm{nr}_{\tilde{y}_i}(d_i)]. \tag{4}$$

Ranks use average ranks for ties. Loss and low confidence induce the same order, so that order receives twice the weight of entropy or disagreement.

Class weighting and coverage. ClassPrior is the class-prior balanced allocation variant of NoiseScore. It multiplies ρ_i by a class weight based on relative mean loss, shrunk toward one according to the agreement between within-class and global loss ranks, and clipped to $[0.5, 2]$. It uses neither the cluster-coverage component nor the fallback rules of the combined variants.

GatedCoverage is the noise-risk-gated coverage variant. It uses NoiseScore directly for budgets up to 2%, or when the mean loss of the top-B loss-ranked examples exceeds four times the overall mean loss. Otherwise it combines NoiseScore ranking with cluster coverage. CB-Loss imposes equal observed-class query quotas (Appendix E.1).

Validation-gradient scores. For each class c, RepairValue uses the highest-loss fraction τ of validation examples, denoted T_c. Example j receives weight $w_j = 1/(C|T_{y_j}|)$ if selected, and zero otherwise, where C is the number of classes. For probe features $\phi(x)$, define

$$g_c = \sum_j w_j(q_{jc} - \mathbf{1}\{y_j = c\})\phi(x_j), \qquad \hat{v}_i = (1 - 2\tilde{y}_i)\phi(x_i)^\top(g_0 - g_1). \tag{5}$$

RepairValue converts $\hat{v}_i$ to rank $r_i \in [0, 1]$ and selects the largest products $\rho_i r_i$. Ties are resolved by NoiseScore, then Loss. The ResNet-50 experiments use $\tau = 0.20$.

Gated-RepairValue fixes the top $\lceil B/2 \rceil$ NoiseScore examples as anchor queries. It fills the remaining slots by $\hat{v}_i$ from the top $\min(N, \lceil 2B \rceil)$ NoiseScore candidates, excluding those anchors.

Adapted baselines. OOF Cleanlab scores each example using a probe trained on the other four of five noisy-label folds. The ActiveClean adaptation uses expected model change; Active Label Cleaning uses cross-entropy minus entropy. AUTO-D3M-Q adapts datamodel alignment to label verification using one trial and 64-dimensional projections. Group-aware balanced queries utilize ground-truth subgroup annotations to provide an empirical upper reference.

B Frozen-Feature Comparisons

The extended sweep compares the same query policies across datasets, noise settings, and budgets. Table 8 summarizes their WGA gains, and Table 9 pairs detection precision with WGA in selected conditions. The group-aware balanced rule serves as an empirical reference ceiling.

Table 8: All thirteen methods in the extended frozen-feature experiments. Each entry is mean detection precision (%) / WGA gain over the uncorrected baseline (pp), equally averaging the ten corruption–budget conditions after averaging ten seeds per condition. Method-selection comparisons use results within each condition. The oracle is excluded from method selection.

Method	Waterbirds	CelebA	CivilComments
AUTO-D3M-Q (adapt.)	$62.73/+11.84$	$38.33/+8.07$	$19.35/+5.46$
Active Label Cleaning (adapt.)	$98.05/+0.46$	$97.66/+3.32$	$91.56/+7.59$
ActiveClean (EMC adapt.)	$95.98/-4.35$	$97.80/+0.96$	$91.56/+7.59$
Cleanlab (OOF)	$97.42/+0.24$	$98.66/+1.98$	$93.03/+7.58$
Cleanlab (in-sample)	$98.64/+0.67$	$98.24/+3.44$	$91.56/+7.59$
Entropy	$33.88/+3.94$	$20.50/+2.30$	$15.61/+1.03$
Loss	$98.64/+0.67$	$98.24/+3.44$	$91.56/+7.59$
GatedCoverage	$45.92/+5.02$	$35.80/+7.46$	$22.16/+6.16$
NoiseScore	$49.59/+4.73$	$34.43/+7.63$	$22.06/+6.27$
Oracle group-balanced	$70.96/+8.83$	$79.93/+8.95$	$44.59/+3.24$
RepairValue	$8.92/+4.13$	$14.25/+3.02$	$11.34/+2.92$
Random	$15.75/+1.04$	$15.22/+0.05$	$15.04/+0.26$
TracIn-CP (multi)	$96.23/-3.41$	$98.64/-2.12$	$65.91/+1.38$

Table 9: Method comparisons in the extended frozen-feature experiments: 20% uniform noise, 2% budget, ten seeds. AUTO-D3M-Q is adapted to label verification. WGA gains are relative to the uncorrected baseline; paired differences compare each method with Loss. Brackets are 95% CIs. All methods are summarized in Appendix B.

Dataset	Method	Precision (%)	ΔWGA (pp)	Paired vs Loss (pp)
CelebA	Loss	98.67	$+4.11$ $[+2.52, +5.70]$	reference
CelebA	NoiseScore	34.37	$+6.18$ $[+4.52, +7.85]$	$+2.07$ $[+0.95, +3.19]$
CelebA	AUTO-D3M-Q	34.21	$+6.21$ $[+4.86, +7.57]$	$+2.10$ $[+0.80, +3.41]$
CivilComments	Loss	97.31	$+9.36$ $[+9.19, +9.52]$	reference
CivilComments	NoiseScore	27.12	$+7.36$ $[+7.19, +7.53]$	-2.00 $[-2.15, -1.85]$
CivilComments	AUTO-D3M-Q	22.47	$+5.93$ $[+5.73, +6.12]$	-3.43 $[-3.60, -3.26]$

Figure 5 visualizes the complete detection–repair landscape across the original sweep. Of the 164 comparisons where lower mean precision coincides with higher mean WGA, 80 exhibit paired 95% confidence intervals strictly above zero, establishing statistically confirmed empirical inversions rather than heuristic quadrant positioning.

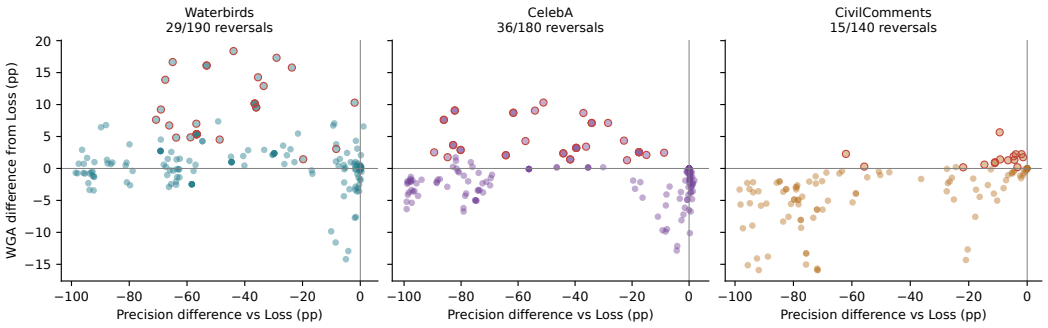

Figure 5: Detection precision and WGA differences from Loss in the original sweep. Each point is a method–condition comparison. The upper-left quadrant contains 164 comparisons with lower mean precision and higher mean WGA. Red outlines mark the 80 comparisons whose paired 95% WGA intervals lie above zero.

C End-to-End Results

Development and evaluation. We use development results to choose a precision-maximizing and a WGA-maximizing method for each image dataset, then compare their retrained models on additional seeds. Waterbirds development seeds 20–39 select Loss by precision and NoiseScore by WGA. CelebA development seeds 20–29 select TracIn-CP and RepairValue, respectively. Each choice is a unique maximizer, and CelebA includes Loss as a fixed reference.

Evaluation uses seeds 1000–1019 in two batches of ten, with the chosen methods, settings, source code, and dataset splits fixed across batches. Within each seed, methods share a probe and retrain from the same original ImageNet initialization. Training follows Appendix A, utilizing automatic mixed precision and deterministic cuDNN operations for reproducible retraining.

Statistical estimation. We estimate confidence distributions via 50,000 bootstrap resamples over the twenty evaluation seeds with the selected methods held fixed. Resampling sequences and seed-level partitions are deterministically recorded in the replication repository. Tables 10 and 11 report individual seed breakdowns and batch aggregations.

CelebA. The advantage of RepairValue over TracIn-CP appears in both batches, with mean differences of 0.80 and 0.71 pp. Across all twenty seeds, RepairValue wins in eighteen and ties in two (two-sided sign-test $p = 7.63 \times 10^{-6}$). Relative to the uncorrected model, RepairValue improves WGA in fourteen seeds and reduces it in six. Loss has mean absolute gain -0.54 pp (95% CI $[-0.84, -0.24]$), and its paired difference from TracIn-CP is -0.22 pp ($[-0.50, 0.05]$).

Waterbirds. The Waterbirds comparison varies more across seeds. NoiseScore exceeds Loss in thirteen seeds, falls below it in six, and ties in one (two-sided sign-test $p = 0.167$). Batch mean differences are 1.05 and 1.40 pp, while individual differences range from -6.85 to 6.54 pp. Both methods often improve over the uncorrected model: NoiseScore does so in seventeen seeds and Loss in fifteen.

Allocation on CelebA. Table 4 uses the same twenty evaluation seeds. Across their initial worst groups, TracIn-CP corrects none of 17,178 available errors, while RepairValue corrects 16,913. The reported 98.70% recall is the mean of seed-level ratios; the pooled counts weight seeds differently, as defined in Appendix A.1. Per-seed group accuracies and correction counts are included in the supplement.

C.1 Development-Block Comparisons

Table 12 reports each development block against its paired reference. Seeds and reference methods are identified row by row. In particular, the CelebA TracIn-CP and RepairValue rows use different seed sets; their comparison is paired on the additional evaluation seeds above.

Table 10: Per-seed WGA gains over the uncorrected baseline (pp). Precision-selected P (m^D) / WGA-selected W (m^U): Loss/NoiseScore (WB), TracIn-CP/RepairValue (CA). Gap: W − P.

Seed	WB: P	WB: W	WB: gap	CA: P	CA: W	CA: gap
1000	+5.499	+1.907	−3.592	+0.795	+1.104	+0.309
1001	−0.779	−0.467	+0.312	−0.131	+1.262	+1.393
1002	+1.558	+1.410	−0.148	−1.262	−0.158	+1.104
1003	+2.181	+1.246	−0.935	+0.823	+0.823	+0.000
1004	+5.919	+9.991	+4.072	−1.893	+0.631	+2.524
1005	−2.648	+1.246	+3.894	+0.000	+0.480	+0.480
1006	+5.763	+5.763	+0.000	−0.631	−0.631	+0.000
1007	+16.667	+23.209	+6.542	−0.898	−0.267	+0.631
1008	+13.551	+18.404	+4.853	−0.206	+1.194	+1.399
1009	+5.543	+1.020	−4.523	−0.158	+0.048	+0.206
1010	+10.125	+13.905	+3.781	−0.412	−0.103	+0.309
1011	−0.156	+4.673	+4.829	−0.315	+0.315	+0.631
1012	−3.427	+1.402	+4.829	+1.207	+1.262	+0.055
1013	+1.434	−1.404	−2.838	−0.631	+0.315	+0.946
1014	+7.788	+14.060	+6.272	−1.303	−0.412	+0.891
1015	+1.090	+1.130	+0.040	+0.000	+0.946	+0.946
1016	+1.713	+3.427	+1.713	+0.000	+1.262	+1.262
1017	−1.125	−1.081	+0.044	−0.891	−0.576	+0.315
1018	+11.215	+13.446	+2.231	−1.104	+0.206	+1.310
1019	+11.682	+4.829	−6.854	+0.631	+1.097	+0.466

Table 11: WGA differences W − P by evaluation batch (pp; 95% seed-bootstrap CIs). P (m^D) and W (m^U) denote precision-selected and WGA-selected methods. Signs count positive/negative/tied differences across the two ten-seed batches.

Dataset	Seeds	W − P [95% CI]	Signs
Waterbirds	1000–1009	$+1.05$ $[-1.12, +3.20]$	5/4/1
Waterbirds	1010–1019	$+1.40$ $[-1.09, +3.60]$	8/2/0
CelebA	1000–1009	$+0.80$ $[+0.37, +1.30]$	8/0/2
CelebA	1010–1019	$+0.71$ $[+0.46, +0.96]$	10/0/0

Table 12: End-to-end retraining under 20% uniform noise and a 10% query budget. WGA gains are relative to the uncorrected baseline; paired differences use the reference method and seeds listed in each row. Brackets are 95% CIs from 50,000 seed-bootstrap draws. Signs count positive/negative/tied paired differences.

Dataset, seeds	Method	Reference	ΔWGA (pp)	Difference (pp)	Prec. (%)	Seeds (+/-/0)
CelebA, 10–19	TracIn-CP	Loss	$+0.01$ $[-0.51, +0.40]$	$+0.16$ $[-0.34, +0.59]$	99.79	7/3/0
CelebA, 20–29	RepairValue	Loss	$+0.43$ $[+0.10, +0.79]$	$+0.99$ $[+0.62, +1.32]$	30.27	9/0/1
Waterbirds, 10–19	TracIn-CP	Loss	$+0.51$ $[-5.64, +5.68]$	-5.00 $[-9.27, -1.00]$	84.81	3/7/0
Waterbirds, 20–39	RepairValue	Loss	$+1.23$ $[-1.66, +3.87]$	-2.64 $[-5.08, -0.47]$	5.59	6/14/0
Waterbirds, 40–59	NoiseScore	Loss	$+5.71$ $[+2.55, +8.87]$	$+3.02$ $[+1.39, +4.70]$	66.74	15/4/1
Waterbirds, 40–59	Gated-RepairValue	NoiseScore	$+4.44$ $[+1.30, +7.71]$	-1.27 $[-3.09, +0.37]$	47.05	7/13/0

D Group Outcomes and Sensitivity Experiments

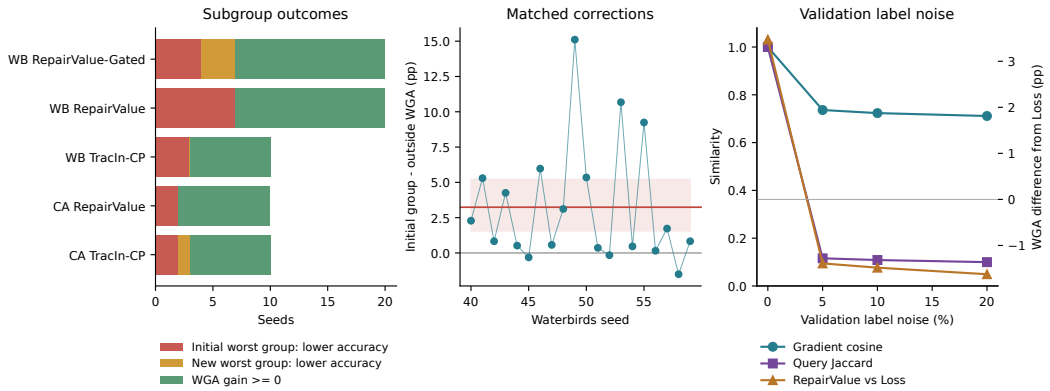

Figure 6: Group outcomes and score sensitivity. Left: per-seed outcomes in the development end-to-end blocks. Middle: the matched-correction comparison in Figure 3. Right: perturbing the validation labels used for RepairValue scoring. Panels refer to separate experiment blocks.

D.1 Group Accuracy and Matched Corrections

An increase in WGA can accompany a decrease in accuracy on another group. Table 13 reports these outcomes in five end-to-end comparisons, distinguishing the initial worst group, a newly worst-performing group, and any group. Measurements follow Appendix A.1.

Table 13: Subgroup accuracy in five end-to-end comparisons. WGA gain and initial-group gain are means relative to the uncorrected baseline (pp). LOO range gives the minimum and maximum leave-one-out mean WGA gains, not a confidence interval. Decrease counts refer to the initial worst group / a new worst group / any group. Changes count seeds with a different worst-performing group after retraining.

Comparison	Seeds	WGA gain	Initial-group gain	LOO range	Decrease counts	Changes
CelebA TracIn-CP	10	+0.01	+0.17	[−0.08, +0.23]	2/1/9	5
CelebA RV-Q	10	+0.43	+0.66	[+0.30, +0.51]	2/0/9	4
Waterbirds TracIn-CP	10	+0.51	+1.90	[−0.82, +2.87]	3/0/9	1
Waterbirds RV-Q	20	+1.23	+1.30	[+0.59, +2.08]	7/0/19	1
Waterbirds RV-Q-Gated	20	+4.44	+7.69	[+3.69, +5.13]	4/3/18	6

Matched query sets. To hold correction count and flip direction fixed, we use Waterbirds seeds 40–59, 20% uniform noise, budget $B = 480$, and cached ResNet-50 features. Linear heads use AdamW with batch size 4096, learning rate 10^{-3}, weight decay 10^{-4}, at most twenty epochs, and patience five. Checkpoints maximize clean-validation balanced accuracy.

For each seed, let E_{in} be the number of available errors in the initial worst group and E_{out} the number of matched errors outside it. Each query set corrects $k = \min\{40, E_{\text{in}}, E_{\text{out}}\}$ labels. We compare sets with 0%, 50%, or 100% of those errors in the initial worst group. Sets have equal correction counts, flip directions, and already-correct queries, but different sample identities. Three query-set randomizations are averaged within each seed before bootstrapping. The inside-minus-outside difference is 3.24 pp, with seventeen positive, three negative, and no tied seed differences. This comparison changes the identities of the corrected examples while matching their counts and flip directions; it estimates the effect of the resulting query sets.

D.2 Gradient Ranking and Random Fill

We test whether gradient ranking improves the choice of queries within a NoiseScore candidate pool. In the Waterbirds development block (seeds 40–59), Gated-RepairValue raises WGA by 3.181

Table 14: Waterbirds frozen-feature controls, seeds 40–59. Each condition averages three query-set randomizations per seed before bootstrapping. The matched-correction comparison holds the number of corrected labels fixed while changing the fraction in the initial worst group. RV-Q-Gated and random fill share a NoiseScore candidate pool. Signs count positive/negative/tied paired WGA differences.

Comparison	Seeds	WGA difference (pp)	Signs (+/-/0)
Initial group - outside	20	$+3.24$ $[+1.56, +5.22]$	17/3/0
Half initial group - outside	20	$+0.92$ $[+0.01, +2.03]$	14/6/0
RV-Q-Gated - random fill	20	-2.10 $[-4.04, +0.26]$	4/16/0
RV-Q-Gated - NoiseScore	20	$+0.37$ $[-1.85, +2.57]$	11/9/0

pp relative to RepairValue (95% CI $[0.198, 6.396]$), but its difference from NoiseScore is -1.266 pp ($[-3.110, 0.336]$), with seven positive seed differences. Its precision is 19.69 pp lower than NoiseScore ($[-24.20, -14.81]$ for the signed difference), reflecting a substantial reallocation of queries away from majority errors. Bootstrap resampling accounts for minor interval variations across experiment blocks in Table 12.

The end-to-end random-fill control uses the same seeds and probe. The top 240 NoiseScore examples are shared anchors in a pool of 960 candidates. Gated-RepairValue selects another 240 by gradient score; random fill draws 240 uniformly from the remaining 720, using deterministic pseudorandom sequences. Both models retrain from the same pretrained initialization and schedule. Table 14 also reports a frozen-feature version of this control.

Gated-RepairValue minus random fill yields a WGA difference of -0.50 pp (95% CI $[-2.38, 1.30]$). Random fill rectifies 68.4 more corruptions on average and degrades WGA relative to uncorrected baselines in fewer seeds (4/20 versus 7/20). Because the confidence interval encompasses zero, this comparison indicates that once a candidate pool is restricted to high-risk errors, secondary gradient ranking and uniform sampling achieve comparable downstream utility.

Table 15: End-to-end Waterbirds comparison with random fill: 20% uniform noise, seeds 40–59, 10% query budget. Both methods use the same 240 anchor queries and 960-example candidate pool. Brackets are 95% CIs from 50,000 paired seed-bootstrap draws. WGA gains over the uncorrected baseline and differences from random fill are in pp.

Method	Precision (%)	Corrected	WGA gain	Diff. vs random fill
Gated-RepairValue	47.05	225.85	$+4.44$ $[+1.30, +7.71]$	-0.50 $[-2.38, +1.30]$
Random fill (same pool)	61.30	294.25	$+4.94$ $[+2.17, +7.68]$	reference

D.3 Sensitivity to the Scoring Validation Set

The directional efficacy of RepairValue depends directly on the quality of its reference validation set. On CelebA (seeds 0–9, twenty-epoch frozen heads, with full clean validation retained for checkpointing), reducing the scoring set to 10% of clean examples minimally alters the RepairValue-minus-Loss advantage (shifting from 3.47 to 3.34 pp). In contrast, injecting a modest 5% label noise into the validation scoring set reverses this advantage to -1.39 pp. This demonstrates that first-order gradient guidance is robust to validation sample volume, but sensitive to validation label fidelity.

Table 16 pairs the saved retraining outcomes with recomputed scoring weights, gradients, and query sets. Appendix I lists the full sensitivity results and the Waterbirds loss-tail-fraction sweep.

D.4 Human Annotation Noise at a Smaller Budget

CIFAR-10N uses one fixed set of human-provided noisy labels and ten training seeds. Table 17 gives the 1% budget results, complementing the 2% results in the main text. The metric is WCA over semantic classes.

Table 16: Validation label noise on CelebA, seeds 0–9. Weight on noisy labels is their fraction of the total validation weight used for scoring. Gradient cosine similarity and query-set Jaccard similarity compare each condition with clean validation labels. The last column gives the recorded WGA difference between RV-Q and Loss.

Val. noise	Weight on noisy labels	Gradient cosine	Query Jaccard	Precision	RV-Q vs Loss (pp)
0%	0.0%	1.000	1.000	6.59%	+3.47
5%	49.4%	0.736	0.116	1.76%	−1.39
10%	51.3%	0.723	0.108	1.68%	−1.48
20%	54.1%	0.711	0.100	1.58%	−1.63

Table 17: CIFAR-10N, 1% budget. Ten training seeds share one human-noise realization. The metric is worst-class accuracy (WCA). Gains are relative to the uncorrected baseline, and paired differences compare each method with Loss. Brackets are 95% t intervals.

Method	Precision (%)	Corrected	ΔWCA (pp)	Paired vs Loss (pp)
Loss	99.88	499.4	$+0.32$ $[-3.25, +3.89]$	reference
NoiseScore	97.96	489.8	$+1.80$ $[-1.67, +5.27]$	$+1.47$ $[-0.37, +3.31]$
ClassPrior	98.24	491.2	$+2.61$ $[-1.28, +6.50]$	$+2.28$ $[-0.53, +5.10]$
Random	39.62	198.1	$+0.35$ $[-1.03, +1.73]$	$+0.02$ $[-3.02, +3.07]$

E Class-Balanced Queries and Influence Analysis

Two query sets can correct equally many labels while placing those corrections in different classes and on different examples. We examine class allocation with a simple quota rule, then derive and evaluate a first-order approximation to the effect of corrections on group loss.

E.1 Class-Balanced Loss

Query rule and protocol. Class-Balanced Loss (CB-Loss) first assigns half the budget to each observed class $\widetilde{Y} \in \{0, 1\}$, then selects the highest-loss examples within each class. An odd budget assigns the extra query to alternating classes across seeds. The rule uses the same probe as Loss, so the comparison tests the effect of adding class quotas.

We compare CB-Loss, Loss, and NoiseScore on cached Waterbirds, CelebA, and CivilComments features. The experiment covers 20% uniform and 40% minority-concentrated noise, 2% and 10% budgets, and seeds 0–9: 360 retraining runs across twelve conditions. Training uses the original AdamW schedules and clean-validation balanced-accuracy checkpoint selection. Table 18 gives the 24 paired comparisons.

Results on CelebA and CivilComments. Equal class quotas improve WGA over both Loss and NoiseScore in all eight conditions on CelebA and CivilComments. All sixteen paired comparisons pass BY correction at $q < 0.05$ across the 24 tests. At the 10% budget with 40% minority-concentrated noise on CelebA, CB-Loss exceeds Loss by 13.79 pp and NoiseScore by 2.45 pp. At the 10% budget with 20% uniform noise on CivilComments, the corresponding differences are 6.26 and 11.35 pp.

Results on Waterbirds. Waterbirds favors different allocations in different conditions. NoiseScore exceeds CB-Loss at the 2% budget with 20% uniform noise and at both budgets with 40% minority-concentrated noise. At the 10% budget with 20% uniform noise, CB-Loss exceeds NoiseScore by 6.18 pp ($p = 0.0004$), but both fall below the uncorrected model.

Under that condition's five-epoch training schedule, the uncorrected baseline achieves 60.33% WGA, while Loss reaches 56.73% (−3.60 pp), CB-Loss 54.26% (−6.07 pp), and NoiseScore 48.08% (−12.25 pp). Contextualizing retrained models against unverified baselines ensures that relative policy advantages are rigorously distinguished from net positive utility recovery.

Table 18: CB-Loss minus Loss or NoiseScore, in WGA percentage points across ten seeds. Brackets give pointwise 95% paired t intervals. Asterisks mark BY-adjusted $q < 0.05$ across the 24 comparisons. BH adjustment retains the same comparisons and additionally Waterbirds 20% uniform noise at 2% versus Loss.

Dataset	Noise	Budget	CB-Loss − Loss (pp)	CB-Loss − NoiseScore (pp)
Waterbirds	Uniform (20%)	2%	$+2.53$ $[+0.39, +4.67]$	-1.64 $[-3.80, +0.51]$
Waterbirds	Uniform (20%)	10%	-2.48 $[-7.03, +2.08]$	$+6.18$ $[+3.61, +8.76]^*$
Waterbirds	Minority (40%)	2%	$+2.66$ $[-1.45, +6.78]$	-4.57 $[-6.31, -2.83]^*$
Waterbirds	Minority (40%)	10%	$+11.64$ $[+6.14, +17.13]^*$	-3.48 $[-8.36, +1.41]$
CelebA	Uniform (20%)	2%	$+5.09$ $[+3.67, +6.51]^*$	$+1.50$ $[+0.80, +2.21]^*$
CelebA	Uniform (20%)	10%	$+5.52$ $[+4.02, +7.01]^*$	$+2.88$ $[+1.95, +3.81]^*$
CelebA	Minority (40%)	2%	$+11.91$ $[+10.45, +13.36]^*$	$+1.63$ $[+0.59, +2.68]^*$
CelebA	Minority (40%)	10%	$+13.79$ $[+11.45, +16.13]^*$	$+2.45$ $[+1.26, +3.64]^*$
CivilComments	Uniform (20%)	2%	$+1.11$ $[+0.89, +1.33]^*$	$+3.72$ $[+3.58, +3.86]^*$
CivilComments	Uniform (20%)	10%	$+6.26$ $[+5.97, +6.55]^*$	$+11.35$ $[+10.92, +11.78]^*$
CivilComments	Minority (40%)	2%	$+0.52$ $[+0.31, +0.73]^*$	$+2.67$ $[+2.46, +2.87]^*$
CivilComments	Minority (40%)	10%	$+8.22$ $[+7.51, +8.93]^*$	$+7.91$ $[+7.44, +8.38]^*$

Table 19: Query allocation and group outcomes, averaged over ten seeds. Precision and initial-worst-group error recall are percentages. Corrections with $Y = 1$ count corrected label errors among examples whose true label is 1. Query classes refer to observed labels. U20 (20% uniform noise) and MH40 (40% minority-concentrated noise) denote the noise regimes.

Dataset	Condition	Method	WGA (%)	Prec. (%)	Recall (%)	Corrections ($Y = 1$)	Query classes
CelebA	U20, 2%	Loss	77.12	98.98	17.37	353.2	88.47% Class 1 (11.53% Class 0)
CelebA	U20, 2%	CB-Loss	82.21	92.98	66.65	1,405.6	Balanced 50:50
CelebA	U20, 2%	NoiseScore	80.71	38.89	58.41	1,266.0	100% Class 0
CelebA	U20, 10%	Loss	79.65	99.20	31.67	644.8	95.76% Class 1 (4.24% Class 0)
CelebA	U20, 10%	CB-Loss	85.17	61.74	92.07	1,947.1	Balanced 50:50
CelebA	U20, 10%	NoiseScore	82.28	10.29	78.77	1,674.1	100% Class 0
CivilComments	U20, 10%	Loss	28.33	92.53	0.01	1.2	$>$ 99.99% Class 1
CivilComments	U20, 10%	CB-Loss	34.59	57.59	41.69	2,559.6	Balanced 50:50
CivilComments	U20, 10%	NoiseScore	23.24	13.51	59.32	3,633.4	100% Class 0
CivilComments	MH40, 10%	Loss	16.53	71.37	0.26	13.4	$>$ 99.94% Class 1
CivilComments	MH40, 10%	CB-Loss	24.74	49.04	36.77	1,603.6	Balanced 50:50
CivilComments	MH40, 10%	NoiseScore	16.84	8.49	55.39	2,283.4	100% Class 0

Query allocation by observed class. The query rules place their budgets in markedly different classes (Table 19). NoiseScore selects from observed class 0 in the displayed conditions. Loss selects mostly from observed class 1: 88.47% and 95.76% of queries on CelebA, and more than 99.9% on CivilComments. CB-Loss reserves queries for both classes. Its quota fixes the queried class proportions; the errors found within each class determine the correction counts and flip directions.

Higher recall does not imply higher WGA. On CivilComments with 20% uniform noise and a 10% budget, NoiseScore corrects an average of 3,633.4 errors among examples with true label $Y = 1$. It recalls 59.32% of the initial worst group's errors and reaches 23.24% WGA, a gain of 13.89 pp over its uncorrected baseline. CB-Loss corrects fewer errors with $Y = 1$ (2,559.6) and has lower initial-group recall (41.69%), but reaches 34.59% WGA.

The lower-recall policy achieves the higher WGA, so initial-group recall alone cannot rank these query sets by repair gain. Their sample identities, total corrections, and observed-class allocations also differ. Separating these effects motivates examining the influence of individual corrections.

E.2 First-Order Influence on Group Loss

Regularized logistic regression gives a direct connection between the corrected labels and the fitted parameters. Its unique optimum lets us study that connection without variation from stochastic optimization.

Objective and label corrections. Let x_i include a constant coordinate for the intercept, and regularize all parameters with $\lambda > 0$. The uncorrected objective is

$$\mathcal{F}_0(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\log(1 + e^{x_i^\top \theta}) - \tilde{y}_i x_i^\top \theta\right] + \frac{\lambda}{2}\|\theta\|_2^2. \tag{6}$$

Let θ_0 be its minimizer and $H_0 = \nabla^2 \mathcal{F}_0(\theta_0)$. Correcting the labels in Q adds a linear term:

$$\mathcal{F}_Q(\theta) = \mathcal{F}_0(\theta) + \delta g_Q^\top \theta, \qquad \delta g_Q = \frac{1}{N}\sum_{i \in Q}(\tilde{y}_i - y_i)x_i. \tag{7}$$

For $\theta_Q = \arg\min_\theta \mathcal{F}_Q(\theta)$ and clean group logistic loss L_g, define the first-order parameter change and predicted loss improvement as

$$v_Q = -H_0^{-1}\delta g_Q, \qquad I_g(Q) = -\nabla L_g(\theta_0)^\top v_Q = \nabla L_g(\theta_0)^\top H_0^{-1}\delta g_Q. \tag{8}$$

Approximation error. Suppose $\nabla^2 \mathcal{F}_0$ is M-Lipschitz and ∇L_g is β_g-Lipschitz on $[\theta_0, \theta_Q]$. With $\eta = \|\delta g_Q\|_2$,

$$\begin{aligned} \|\theta_Q - \theta_0 - v_Q\|_2 &\le \frac{M\eta^2}{2\lambda^3}, \\ |L_g(\theta_0) - L_g(\theta_Q) - I_g(Q)| &\le \left(\frac{\|\nabla L_g(\theta_0)\|_2 M}{2\lambda^3} + \frac{\beta_g}{2\lambda^2}\right)\eta^2. \end{aligned} \tag{9}$$

To see this, strong convexity gives $\|\theta_Q - \theta_0\|_2 \le \eta/\lambda$. Writing $d = \theta_Q - \theta_0$, stationarity implies $\delta g_Q + H_0 d + e = 0$, where $\|e\|_2 \le M\|d\|_2^2/2$. Thus $d - v_Q = -H_0^{-1}e$, which gives the first bound. Taylor expansion of L_g gives the second. □

Correction count and mean influence. Define the influence of correcting example i on group loss:

$$a_{i,g} = \nabla L_g(\theta_0)^\top H_0^{-1}\left[(\tilde{y}_i - y_i)x_i\right]. \tag{10}$$

Already-correct labels contribute zero. When $C(Q) > 0$, the prediction factors as

$$I_g(Q) = \frac{1}{N}\sum_{i \in Q,\ \text{erroneous}} a_{i,g} = \left(\frac{C(Q)}{N}\right)\text{mean}_{i \in Q,\ \text{erroneous}}\, a_{i,g}. \tag{11}$$

If $C(Q) = 0$, then $I_g(Q) = 0$. At a fixed budget, detection precision determines the first factor, the correction count. The second factor depends on which examples are corrected and how their features interact with the fitted model and group-loss gradient. This local loss approximation thus supplies information about the corrections beyond their count. WGA is evaluated separately below.

Numerical implementation. We utilize all cached examples and feature coordinates, scaling features by the root-mean-square norm of the training vectors. Optimization uses float64 preconditioned L-BFGS with convergence tolerance $\|\nabla \mathcal{F}_Q\|_\infty \leq 2 \times 10^{-7}$, where preconditioning rescales the parameter coordinates while preserving the convex objective. Analytical gradient shifts follow the exact label differences $(\tilde{y}_i - y_i)x_i$.

Regularization affects both approximation accuracy and classification. At $\lambda = 0.01$, group-loss prediction RMSE is below 0.0008 nats, while CelebA and CivilComments have zero WGA. Table 20 therefore emphasizes $\lambda = 10^{-4}$, where baseline WGA is 59.42% on Waterbirds, 52.04% on CelebA, and 6.05% on CivilComments.

Table 20: Prediction RMSE for logistic retraining with $\lambda = 10^{-4}$. Each row uses thirty models (ten seeds and three query rules), giving 120 group-loss evaluations. The two CE columns use test-group or validation-group loss gradients to predict test-group loss improvements (nats). WGA is evaluated at the linearized parameters (pp); the reference predicts zero WGA gain.

Dataset	Budget	CE: test gradients	CE: val. gradients	WGA: linearized	WGA: zero gain
Waterbirds	0.5%	0.00084	0.00279	0.114	1.057
Waterbirds	2%	0.00472	0.00682	0.487	5.246
Waterbirds	10%	0.05142	0.04943	3.942	3.603
CelebA	0.5%	0.00099	0.00192	0.028	3.122
CelebA	2%	0.00665	0.00544	0.502	6.775
CelebA	10%	0.01801	0.01295	0.729	8.286
CivilComments	0.5%	0.00031	0.00089	0.036	3.303
CivilComments	2%	0.00203	0.00271	0.329	8.806
CivilComments	10%	0.02206	0.02426	2.789	18.742

Prediction across budgets. The approximation is accurate at small budgets in the weak-regularization experiment, which contains 270 retrained models. At 0.5% and 2%, predicted group-loss improvements using test-group gradients have RMSE below 0.007 nats. Computing WGA at the approximate parameters $\theta_0 + v_Q$ gives RMSE below 0.503 pp across these settings.

At the 10% budget, higher-order parameter drift and boundary re-alignment become prominent: on Waterbirds, WGA prediction RMSE reaches 3.942 pp. Furthermore, loss reductions and zero-one accuracy shifts can decouple under non-linear margin geometry: on CivilComments at 10%, Loss improves WGA by 17.65 pp while worst-group cross-entropy increases by 0.136 nats. Consequently, first-order loss approximations and downstream classification accuracy reflect distinct aspects of the retrained model.

Table 21: Held-out RMSE under leave-one-seed-and-budget-out cross-validation (ninety predictions per dataset). Predictors successively add initial-group recall and validation-group influence to correction count. Group CE improvement is measured in nats and WGA gain in pp. The reference predicts no change.

Target	Dataset	Count	Count + recall	Count + recall + influence	Zero change
Initial Worst Group CE Imp.	Waterbirds	0.12157	0.09987	**0.04172**	0.11838
Initial Worst Group CE Imp.	CelebA	0.12554	0.05787	**0.00743**	0.10959
Initial Worst Group CE Imp.	CivilComments	0.07480	0.07483	**0.01496**	0.06480
WGA Gain (pp)	Waterbirds	8.518	10.331	6.279	**3.725**
WGA Gain (pp)	CelebA	6.608	2.833	**1.413**	6.437
WGA Gain (pp)	CivilComments	8.145	9.154	**2.069**	12.107

Predictive value beyond counts and recall. We next ask whether influence improves prediction beyond correction counts and group recall. Leave-one-seed-and-budget-out cross-validation gives ninety held-out predictions per dataset. Three nested predictors use correction count; count plus initial-group recall; and both quantities plus influence from validation-group loss gradients. This offline analysis uses group annotations and verified correction information.

Influence reduces group-loss RMSE relative to count and recall by 58% on Waterbirds, 87% on CelebA, and 80% on CivilComments. This improvement on all three datasets shows that the selected examples carry predictive information beyond their number and initial-group coverage.

For downstream WGA gain, adding first-order influence reduces prediction RMSE over count-and-recall predictors by 4.05 pp (95% CI $[3.70, 4.43]$) on Waterbirds, 1.42 pp ($[1.28, 1.53]$) on CelebA, and 7.08 pp ($[6.86, 7.42]$) on CivilComments. This highlights that individual instance influence provides actionable signal for downstream robustness beyond raw sample counts, though higher-order optimization curvature sets a natural upper bound on local linear extrapolations.

F Class-Wise Ranking on CivilComments

NoiseScore combines two choices: ranking within observed classes and combining four signals. We separate them to examine how each changes the queried examples and the retrained model.

Factorial design. We cross two ranking scopes (global or within observed class) with two signals (Loss or the four-component NoiseScore). The four combinations share probes, initializations, and minibatch randomization. With cached MiniLM features, 20% uniform noise, 2% and 5% budgets, and seeds 0–9, the design gives 80 retrained heads.

AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 4096, thirty epochs, and patience eight. Checkpoints maximize clean-validation balanced accuracy. Group annotations serve solely for diagnostic evaluation. We report paired Student's t confidence intervals to characterize the internal variance of the factorial comparisons.

Class noise rates and query precision. For positive-class prevalence π and symmetric flip probability η,

$$\Pr(Z = 1 \mid \widetilde{Y} = 0) = \frac{\eta\pi}{(1-\eta)(1-\pi) + \eta\pi}, \qquad \Pr(Z = 1 \mid \widetilde{Y} = 1) = \frac{\eta(1-\pi)}{\eta(1-\pi) + (1-\eta)\pi}. \tag{12}$$

With a clean positive prevalence of 11.34%, the observed classes have very different noise rates: 3.10% in class 0 and 66.13% in class 1 on average. Query precision also depends on where the score ranks those errors. For a single continuous score, a common within-class percentile approximately preserves observed-class proportions, whereas an equal class quota assigns the same number of queries to each class.

Discrete disagreement ranks. If $d > 0$ of a class's $n > 1$ examples disagree with the probe prediction, the positive disagreement entries have normalized average rank

$$\mathrm{nr}(1) = \frac{n - (d+1)/2}{n-1}. \tag{13}$$

The tied entries occupy zero-based positions $n-d, \ldots, n-1$; averaging and dividing by $n-1$ gives the expression. Disagreement occurs in 0.65% of observed class 0 and 96.55% of class 1, giving positive ranks of 0.997 and 0.517. Consequently, the same disagreement indicator contributes more to NoiseScore in class 0. The factorial experiment below measures the combined score's effect on retraining.

Effects of normalization and score composition. Class-wise Loss ranking improves WGA while finding fewer errors. At a 2% budget, replacing global ranks with within-class ranks reduces precision from 97.31% to 48.74% and increases WGA by 0.31 pp (95% CI $[0.16, 0.46]$). At 5%, precision falls from 95.63% to 41.90%, while WGA rises by 0.75 pp ($[0.60, 0.90]$).

Adding the other NoiseScore components reduces WGA under both ranking scopes. The two composite scores also differ sharply: at 5%, the global composite has higher precision (28.28% versus 18.65%) but lower WGA (2.60% versus 19.92%) than the class-wise composite. Figure 7 connects these outcomes to query allocation and detection.

Table 22: CivilComments normalization–composition ablation: 20% uniform noise, ten seeds, shared probes and paired 30-epoch head retraining. Class 1 is the observed toxic label; all proportions and WGA are percentages. Differences from global Loss are pp with exploratory paired 95% t intervals. This is a separately rerun experiment block.

Budget	Ranking	Signal	Class 1 queries	Precision	WGA	vs. Loss (pp)
2%	Global	Loss	100.00	97.31	18.55	reference
2%	Class	Loss	26.79	48.74	18.87	$+0.31$ $[+0.16, +0.46]$
2%	Global	Composite	81.69	21.60	8.11	-10.45 $[-10.64, -10.26]$
2%	Class	Composite	0.00	27.12	16.56	-2.00 $[-2.15, -1.85]$
5%	Global	Loss	100.00	95.63	24.34	reference
5%	Class	Loss	26.79	41.90	25.09	$+0.75$ $[+0.60, +0.90]$
5%	Global	Composite	91.05	28.28	2.60	-21.73 $[-21.99, -21.48]$
5%	Class	Composite	0.00	18.65	19.92	-4.41 $[-4.56, -4.27]$

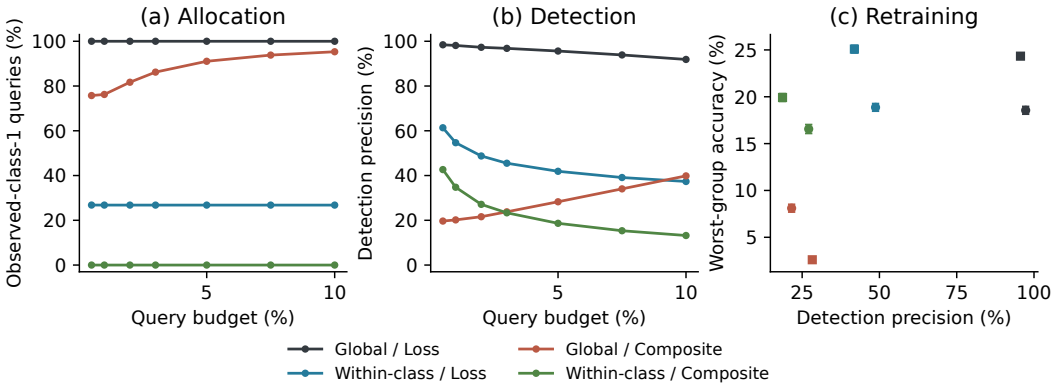

Figure 7: **Normalization and score composition.** Panels (a) and (b) show mean query allocation and precision across ten seeds at several budgets. Retraining is performed only at 2% and 5%. Panel (c) shows WGA with pointwise 95% seed t intervals (circles: 2%; squares: 5%).

Group outcomes. Higher error recall again coexists with lower group accuracy. At 5%, the class-wise composite recalls more errors in the toxic/Black group than class-wise Loss (51.83% versus 44.02%), but reaches lower accuracy on that group (25.49% versus 31.22%). The non-Black toxic group has the lowest final accuracy in all ten seeds under both methods.

The loss changes follow the same ordering. Global composite scoring raises mean test cross-entropy on the two toxic groups by 0.067 and 0.099 nats relative to baseline; class-wise Loss lowers it by 0.077 and 0.097 nats. The probe's mean validation balanced accuracy is 54.90%.

These experiments identify effects of ranking scope and score composition on the fixed MiniLM features. The query sets change sample identities along with group allocation, so a causal effect of group membership alone is not identified. The separate question of why validation-based method selection fails across budgets on CivilComments remains open.

Reproduction. All forty rerun Loss/NoiseScore WGAs match the saved results. Thirty-nine query sets match exactly; one differs by a single replacement (Jaccard similarity 0.99985). The supplement includes per-seed records, input hashes, and scripts that reconstruct queries and predictions.

G Two-World Construction and Proof

The construction makes every queried error useful for its own prediction while hiding which block determines the worst-group accuracy. The proof has three steps: realize this prediction rule with a common logistic objective, show that class-label feedback cannot distinguish the group assignments, and match the resulting lower bound with balanced querying.

G.1 A Common Classification and Retraining Problem

Fix integers $B \geq 2$ and $M \geq \max\{B+1, B(B-1)\}$. The training set has $N = 2B+3$ atoms: two candidate blocks A_0, A_1, each of size B, a positive anchor a_+, and two negative anchors $a_{-,0}, a_{-,1}$. Every candidate has observed label zero and true label one. The three anchors are correctly labeled. Each atom has its own standard-basis feature in $\mathbb{R}^N$. Clean validation observations are any fixed collection of these atoms with their true class labels, shared by both worlds and without group labels.

For a query set Q, let y_i^Q be its corrected training labels. The common retraining algorithm returns the unique minimizer of

$$L_Q(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\log(1+e^{\theta_i}) - y_i^Q\theta_i\right] + \frac{\lambda}{2}\|\theta\|_2^2, \qquad \lambda > 0, \tag{14}$$

with no intercept and prediction $\mathbf{1}\{\theta_i > 0\}$. Strong convexity gives a unique solution. Each coordinate derivative is $(\sigma(\theta_i) - y_i^Q)/N + \lambda\theta_i$, strictly increasing and equal to $(1/2 - y_i^Q)/N$ at zero. The optimum is negative for label zero and positive for label one. Thus, exactly the corrected candidates switch from incorrect to correct predictions. Querying a clean anchor leaves the model unchanged.

The shared test distribution assigns mass $1/(2B+M+2)$ to each candidate and negative anchor and mass $M/(2B+M+2)$ to the positive anchor. In world w, the four evaluation groups are

$$G_{w,1} = A_w, \quad G_{w,2} = A_{1-w} \cup \{a_+\}, \quad G_{w,3} = \{a_{-,0}\}, \quad G_{w,4} = \{a_{-,1}\}. \tag{15}$$

The four nonempty groups have masses proportional to $(B, M+B, 1, 1)$ in either world. Swapping the positive candidate blocks changes their group assignments while preserving the (X, Y) distribution and group proportions. These groups can be encoded as intersections of the class label with a hidden binary attribute. All observable inputs to acquisition, including training dynamics and scores computed from class labels, remain identical.

G.2 Utility, Oracle Optimum, and Indistinguishability

The uncorrected WGA is zero because every candidate in A_w is misclassified. For any feasible query, including queries to anchors, let $k_j = |Q \cap A_j|$. Direct evaluation of the retrained predictions

gives

$$U_w(Q) = \min\left\{\frac{k_w}{B}, \frac{M + k_{1-w}}{M + B}\right\}, \qquad k_0 + k_1 \le |Q| \le B. \tag{16}$$

Querying all of A_w gives $M/(M+B)$. If $k_w \le B-1$, utility is at most $(B-1)/B \le M/(M+B)$ by $M \ge B(B-1)$. If $k_w = B$, the budget forces $k_{1-w} = 0$. Thus the group-aware optimum is exactly $U_w^\star = M/(M+B)$ in both worlds.

Identical verification transcripts. Couple a sequential policy's internal random bits across worlds. It starts from identical observations and, after any identical transcript, makes the same next query. That atom has the same true label in both worlds, so the response is also identical. Induction gives the same complete transcript, stopping decision, and final-query distribution. The argument covers clean-anchor queries and intermediate group-blind model fitting.

Revealing every error indicator, or even all true training class labels as side information, leaves the group assignment strictly unidentifiable. With the correction budget retained, the same allocation problem remains.

G.3 Lower Bound and Attainment

For every feasible Q, Equation (16) implies

$$U_0(Q) + U_1(Q) \le \frac{k_0 + k_1}{B} \le 1. \tag{17}$$

Since a group-blind policy has the same distribution over Q in both worlds,

$$\mathbb{E}[\mathrm{Reg}_0(Q_\pi)] + \mathbb{E}[\mathrm{Reg}_1(Q_\pi)] = 2\frac{M}{M+B} - \mathbb{E}[U_0(Q_\pi) + U_1(Q_\pi)] \tag{18}$$

$$\ge \frac{M-B}{M+B}. \tag{19}$$

At least one world's expected regret is therefore at least $(M-B)/(2(M+B))$, regardless of the policy's randomization or use of verification feedback.

For even B, query $B/2$ candidates from each block. For odd B, allocate $\lfloor B/2 \rfloor$ and $\lceil B/2 \rceil$ queries to the two blocks, swapping the allocations with probability $1/2$. In either case the second term in Equation (16) is at least $M/(M+B) \ge \lceil B/2 \rceil / B$ under the stated conditions (with $B = 2$ checked directly). The active block therefore determines WGA, and each world's expected utility is exactly $1/2$. The resulting regret is $M/(M+B) - 1/2 = (M-B)/(2(M+B))$, proving equality in Equation (3).

Detection precision and regret. Every size-B candidate-only query has 100% precision. Querying all of A_0 gives utilities $(M/(M+B), 0)$, while querying all of A_1 reverses them. Balanced querying attains expected utility $1/2$ in each world. The regret arises from uncertainty about which allocation maximizes group performance, even after all error locations are known.

Structural stability of the minimax bound. The lower bound is structurally robust to continuous feature perturbations. For each feasible corrected label vector, the unique regularized logistic optimum exhibits strictly positive prediction margins. By strong convexity of the objective, the parameter minimizer and its classification margins vary continuously with respect to the input vectors. Because the space of query choices is finite, the intersection of positive perturbation radii defines an open neighborhood around the canonical atom basis within which all prediction signs, utilities, and the minimax value remain invariant.

Numerical verification. We numerically validate the construction across $B \in \{2, 3, 4\}$ at minimal M and scaled $M = 10B^2$ via `verify_information_boundary.py`. The verification suite enumerates all feasible query allocations, fits the corresponding logistic models, and validates empirical WGA against Equation (16). Finite linear programming independently confirms the randomized minimax value, corroborating the optimality of the balanced allocation strategy.

H METHOD-SELECTION ANALYSIS

We compare what happens when the same candidate methods are selected by detection precision, validation gain, or historical WGA gain. A condition fixes the dataset, noise setting, and query budget. The extended sweep supplies twelve non-oracle candidates; the separately evaluated CB-Loss is outside this set.

H.1 VALIDATION-BASED SELECTION

Validation gain measures how much a candidate's retrained model improves on clean validation data, defined as the difference between post-retraining and pre-retraining class-balanced validation accuracy: $\Delta\mathrm{Acc}_{\mathrm{val}} = \mathrm{Acc}_{\mathrm{val}}(Q) - \mathrm{Acc}_{\mathrm{val}}(\emptyset)$. These metrics are computed on the same clean validation examples utilized for checkpoint selection. The analysis evaluates twelve non-oracle candidate policies across ten seeds, two noise regimes, and five budgets from the extended benchmark sweep.

The three selection criteria use historical detection precision, validation gain, or WGA gain. All rely on previous candidate evaluations. Precision uses verified error counts, validation gain uses clean class labels, and the WGA reference additionally uses group evaluations. Test-set balanced accuracy is excluded from selection.

Selection folds. Let $A_{m,b,s}$ be a historical score for method m, budget b, and seed s, within a fixed dataset and noise setting. For target (b, s), within-budget selection maximizes $\frac{1}{9}\sum_{t\neq s} A_{m,b,t}$. Cross-budget selection maximizes

$$\frac{1}{4}\sum_{b'\neq b}\frac{1}{9}\sum_{t\neq s} A_{m,b',t}. \tag{20}$$

Both protocols strictly exclude the evaluation seed from every source history. Methods with identical validation scores receive equal probability, and evaluation averages the held-out WGA across ties.

We average the ten seed folds and then the ten noise–budget conditions with equal weights. Dataset-macro results weight the three datasets equally. Since conditions and folds reuse runs, uncertainty is estimated by resampling seeds together across conditions.

Statistical estimation. We estimate confidence distributions via 10,000 multinomial seed-cluster bootstrap resamples independently per dataset. Resampling clusters seeds jointly across all methods, budgets, and noise settings, preserving cross-condition correlation structures. Percentile intervals characterize sampling variability across experimental folds.

Comparison with fixed methods. Table 23 places the selection gains alongside fixed Loss, fixed NoiseScore, and the historical WGA reference. On Waterbirds, validation and WGA select AUTO-D3M-Q in all 100 original folds under both protocols. Their point estimates coincide, although within-budget bootstrap choices can differ. On CelebA, validation selection assigns AUTO-D3M-Q total fold weights of 78/100 within budget and 88/100 across budgets; the remaining weight goes to NoiseScore or GatedCoverage.

Because historical WGA selection is evaluated out-of-sample on held-out seeds, it represents a realistic empirical reference rather than an artificial theoretical ceiling.

Candidate-pool robustness. To verify that validation selection does not depend on specific prominent acquisition algorithms, we remove AUTO-D3M-Q and TracIn-CP and repeat selection among the remaining ten candidate methods. Validation-based selection consistently outperforms precision-based selection across both image datasets under the restricted pool (Table 24), demonstrating that validation recovery reliably identifies superior policies even when the candidate space is constrained.

CivilComments. The CivilComments results explain why a within-budget gain need not transfer across budgets. The within-budget advantage over precision selection is 0.295 pp, with approximately 85% of the condition-summed advantage coming from the 40% minority-concentrated noise

Table 23: Held-out WGA differences between validation-based selection, fixed methods, and historical WGA-based selection. Values are percentage points with pointwise 95% seed-cluster bootstrap intervals.

Dataset	Budget	Val. − Loss	Val. − NoiseScore	WGA − Val.
Waterbirds	Within	$+11.17$ $[+9.45, +12.94]$	$+7.11$ $[+6.20, +8.05]$	$+0.00$ $[-0.23, +0.00]$
Waterbirds	Cross	$+11.17$ $[+9.45, +12.94]$	$+7.11$ $[+6.20, +8.05]$	$+0.00$ $[+0.00, +0.00]$
CelebA	Within	$+4.88$ $[+4.13, +5.36]$	$+0.69$ $[+0.08, +0.87]$	-0.09 $[-0.14, +0.53]$
CelebA	Cross	$+4.48$ $[+3.79, +5.12]$	$+0.29$ $[+0.08, +0.67]$	-0.06 $[-0.26, +0.24]$
CivilComments	Within	$+0.28$ $[+0.23, +0.33]$	$+1.60$ $[+1.53, +1.68]$	$+0.01$ $[-0.03, +0.03]$
CivilComments	Cross	-0.04 $[-0.08, +0.01]$	$+1.27$ $[+1.18, +1.40]$	$+0.01$ $[-0.16, +0.06]$

Table 24: Exploratory candidate-set sensitivity on the image benchmarks: remove both TracIn-CP and AUTO-D3M-Q, then repeat selection over the remaining ten methods in every fold and bootstrap draw. The same seed-cluster weights as the full-candidate analysis are used. Entries are WGA differences in pp with 95% intervals.

Dataset	Budget	Val. − Precision	Val. − NoiseScore
Waterbirds	Within	$+3.47$ $[+2.37, +5.68]$	-0.69 $[-1.55, +0.89]$
Waterbirds	Cross	$+3.48$ $[+2.30, +4.80]$	-0.42 $[-1.51, +0.23]$
CelebA	Within	$+5.15$ $[+3.82, +5.53]$	-0.05 $[-1.16, +0.00]$
CelebA	Cross	$+5.64$ $[+5.17, +6.06]$	$+0.00$ $[-0.03, +0.00]$

condition at the 10% budget (2.520 pp). Across budgets, eight condition gaps are zero and two are negative, giving a mean of -0.030 pp. Validation selection exceeds fixed NoiseScore by 1.273 pp, but precision selection remains at least as effective on average.

Operational scope and transfer properties. The transfer mechanism characterized here identifies policy preferences across query scales within a given dataset, representation, and candidate pool. In operational pipelines where exploratory trials are routinely conducted across sample sizes or hyperparameter sweeps, historical validation signals provide actionable guidance without requiring additional data-collection overhead.

Reproduction. The supplementary repository provides complete selection records across the 600 evaluation folds, candidate trajectory logs, and 10,000 bootstrap replicates. The reproduction runner `build_validation_selection_assets.py` reconstructs all summary tables and candidate sensitivity analyses directly from raw outputs.

H.2 Historical WGA-Based Selection

Historical WGA selection measures the benefit of choosing methods using previous group-performance results. We compare it with precision selection on both frozen-feature sweeps. The original sweep excludes ten method–condition combinations without complete ten-seed pairs.

Criteria and estimands. Let $D_{m,c,s}$ and $U_{m,c,s}$ denote detection precision and WGA gain for method m, condition c, and seed s. For each held-out seed, select

$$m^{D}_{c,-s} \in \arg\max_{m} \frac{1}{9} \sum_{t \neq s} D_{m,c,t}, \qquad m^{U}_{c,-s} \in \arg\max_{m} \frac{1}{9} \sum_{t \neq s} U_{m,c,t}. \tag{21}$$

The held-out selection gap is $L^{\mathrm{CV}}_{c} = \frac{1}{10} \sum_{s} (\mathbb{E}[U_{m^{U}_{c,-s},c,s}] - \mathbb{E}[U_{m^{D}_{c,-s},c,s}])$, where expectations average uniformly across exact ties. Cross-validated WGA selection uses group outcomes exclusively from held-out source seeds, while in-sample selection serves as an empirical comparison ceiling evaluated across all seeds.

Statistical estimation. We apply the multinomial seed-cluster bootstrap procedure detailed in Appendix H.1 (10,000 resamples), capturing sampling variability across experimental seeds under fixed dataset splits and candidate sets.

Table 25: WGA gaps in the extended and original frozen-feature experiments. In-sample estimates use the same ten seeds for selection and evaluation. Held-out estimates use leave-one-seed-out evaluation with uniform tie breaking. Alphabetical tie breaking uses method identifiers for both criteria. All values are pp; brackets are 95% seed-cluster bootstrap CIs.

Experiments	Dataset	In-sample	Held-out [95% CI]	Alphabetical
Extended	Waterbirds	11.35	11.33 [9.26, 13.30]	11.32
Extended	CelebA	9.21	9.11 [8.34, 9.92]	9.12
Extended	CivilComments	0.31	0.31 [0.23, 0.36]	0.31
Extended	Equal-dataset mean	6.96	6.91 [6.17, 7.63]	6.92
Original	Waterbirds	5.59	5.14 [3.02, 6.54]	5.11
Original	CelebA	8.44	8.32 [7.51, 9.12]	8.32
Original	CivilComments	1.29	1.28 [1.24, 1.33]	1.28
Original	Equal-dataset mean	5.11	4.92 [4.15, 5.46]	4.90

Table 26 compares fixed Loss and NoiseScore with precision selection. In the extended sweep, the dataset-macro gap is 6.91 pp with uniform tie averaging and 6.92 pp with alphabetical tie-breaking. Choosing the best or worst held-out WGA among precision-tied methods gives bounds of 6.77–7.06 pp. The original-sweep mean gap is 4.92 pp.

Table 26: Fixed baselines and tie breaking in the extended experiments (pp). Positive differences favor the named baseline over precision-based selection. The last column gives the selection-gap range obtained by taking the best or worst held-out WGA among methods tied on precision.

Dataset	Loss − precision	NoiseScore − precision	Tie gap range
Waterbirds	0.15 [−0.26, 0.82]	4.22 [2.62, 5.98]	[10.88, 11.76]
CelebA	4.32 [3.74, 4.92]	8.51 [7.66, 9.34]	[9.11, 9.11]
CivilComments	0.01 [−0.03, 0.06]	−1.30 [−1.42, −1.20]	[0.31, 0.31]

The historical WGA-minus-precision gap exceeds 1 pp in 21 of 30 extended conditions: all ten Waterbirds conditions, all ten CelebA conditions, and one CivilComments condition (Figure 8).

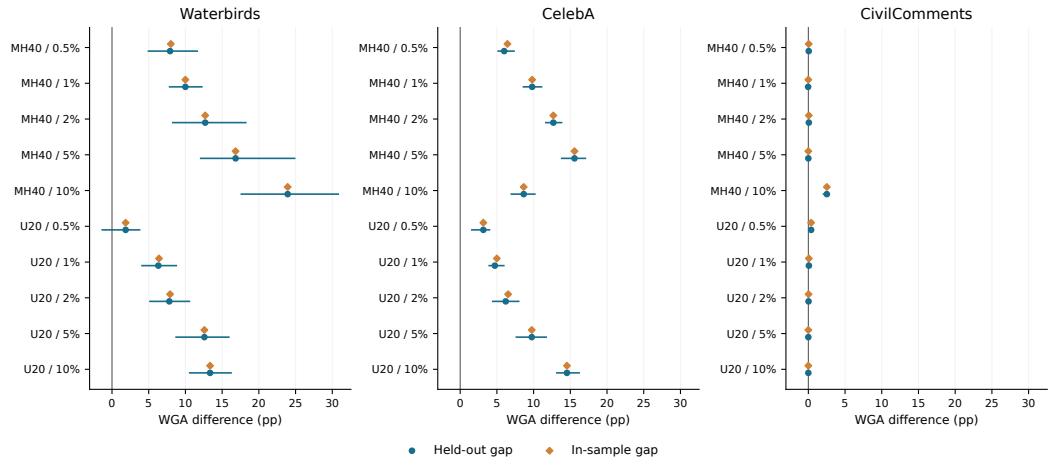

Figure 8: Historical WGA-based minus precision-based selection in thirty extended conditions. Blue points show held-out estimates with 95% seed-cluster bootstrap intervals; orange diamonds select and evaluate on the same runs. U20 and MH40 denote 20% uniform noise and 40% minority-concentrated noise, respectively.

I Supplementary Results and Reproduction

The results archive includes analysis inputs, tables, figures, and reconstruction scripts. Its `RESULTS_INDEX.md` maps every experiment to its files; `MANIFEST.json` lists checksums. The main directories in `supplementary_results.zip` are:

- `data/s2_*.csv`, `data/matrix_audit.csv`, and `data/supplement_*.csv`: frozen-feature benchmark sweeps, paired comparisons, and multiplicity-adjusted statistics.
- `data/selection_audit/` and `data/validation_selection/`: policy selection inputs, cross-validation folds, bootstrap replicates, candidate sensitivity sweeps, and analysis scripts.
- `data/prospective_selection_all20/`, `data/failure_*.csv`, `data/random_fill_control/`, and `data/cifar10n.csv`: end-to-end results, group outcomes, controls, and human-noise experiments.
- `data/civilcomments_mechanism/` and `output/repair_bridge_20260918/`: factorial, class-allocation, and influence experiments, with protocols and reproduction checks.
- `verify_information_boundary.py` and `data/information_boundary_check.json`: computational checks of the theoretical construction.
- `tables/` and `figures/`: full result displays, including validation sensitivity and model references. The index identifies their rebuilding scripts.

The scripts rebuild analyses and displays from saved results. Model weights and raw query and prediction arrays are outside this archive. Training-level reproduction uses the experiment code, datasets, and larger archives described in the supplementary README.