RESPONSE IS NOT DISPENSABILITY: A CONTROLLED STUDY OF COUNTERFACTUAL DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting training evidence from a model's intervention response requires separating response magnitude from learning utility. We examine a low-response retention rule and show that probability saturation favors large-margin identities even at a fixed intervention-induced logit contrast. An exact binary decomposition and a multiclass saturation construction characterize this effect. We test its consequences in synthetic and rendered-digit tasks with known label-preserving interventions, matched class budgets, and shared coverage rules. In a linear factorial, increasing core strength changes rare-evidence retention from 48.29% to 1.15% under low-response selection, while a nuisance-logit ranking retains approximately half. Nonlinear experiments on synthetic and two colored-digit tasks separate retained composition from retrained subgroup accuracy. Across all 18 common training settings, applying the intervention during training on the same retained identities improves mean four-group worst accuracy over training without intervention. With validation-selected configurations, the gains are 20.00, 63.50, and 23.83 percentage points on synthetic data, colored digits, and colored MNIST. The results establish that response, retention, and intervention use are distinct decisions that require separate validation.

1 INTRODUCTION

Interventions provide a direct way to probe what a model uses. Given a label-preserving transformation T, the difference between predictions on x and Tx measures behavioral response without requiring a nuisance label for every example. This response can guide augmentation, invariance training, or data selection. Selection introduces a distinct inference: the score must identify which factual identities a new model should learn from.

Low probability response is not a direct measure of dispensability. A saturated prediction can change little in probability even when the intervention causes a substantial logit change. Conversely, a moderate-margin example can receive a large probability response under the same logit intervention. A retention rule based on low response therefore combines nuisance dependence with the model's confidence in the remaining evidence.

We define and test a low-response retention rule: retain factual identities with the smallest paired probability differences. This operational rule asks whether a small response is sufficient evidence for choosing the training subset. Our empirical setting supplies exact synthetic and rendered-digit interventions. First, an exact binary decomposition separates the midpoint of a factual–counterfactual logit pair from its intervention half-difference and characterizes their joint effect on probability response. Second, matched class and coverage quotas isolate the ranking from subset composition. Third, retraining evaluates whether the retained identities support subgroup learning and compares selection with using the same intervention during training. Figure 1 summarizes these decisions. Across 18 matched training settings, intervention use on the same identities improves the mean four-group minimum in every comparison.

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.

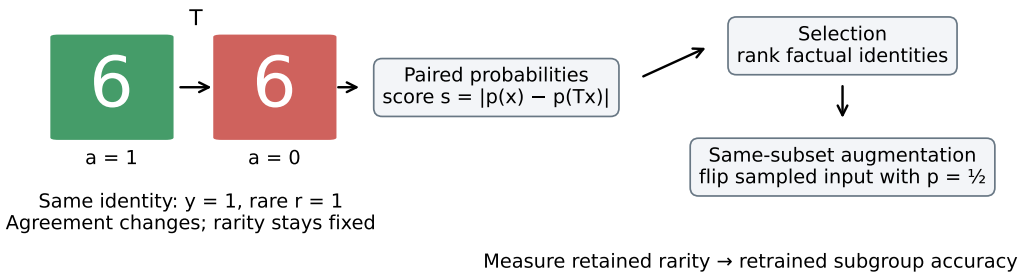


Figure 1: A label-preserving background intervention supports two different decisions. A response score ranks factual identities for selection; augmentation changes which version of each retained identity is shown during training. Evaluation measures retained composition and the subgroup accuracy of the retrained model.

2 Related work

Coverage and example importance have already been combined in data selection. Coverage-centric Coreset Selection preserves representation coverage at high pruning rates (Zheng et al., 2023), while D^2 Pruning integrates diversity and difficulty through graph message passing (Maharana et al., 2024). Our classwise KMeans quotas provide a common coverage control across score rankings, isolating the contribution of the counterfactual score.

Shortcut-oriented pruning and reweighting use different information regimes. Mulchandani & Kim (2025) prune difficult examples using early EL2N, classwise in their unidentifiable setting and with group information on Waterbirds. JTT identifies early training mistakes and upsamples them when fitting a new model, with group-labeled validation for tuning (Liu et al., 2021). We evaluate fixed-checkpoint, fixed-weight versions under a shared training protocol.

Pruning can also alter the balance of classification errors. DRoP studied classification bias and class pruning ratios (Vysogorets et al., 2025). Our comparison fixes class budgets and instead varies rankings within each class, measuring rare subtypes separately from label–nuisance groups. Influence functions traced predictions back through the training procedure (Koh & Liang, 2017); the sensitivity score studied here keeps parameters fixed. The deletion experiment directly tests a retraining outcome at the level of score bins.

Counterfactual selection uses several score definitions and directions. Zayed et al. (2023) use an ℓ_2 logit discrepancy to select high-discrepancy counterfactuals alongside random factual examples, whereas EL2N compares class probabilities with the training label (Paul et al., 2021). Our low-response rule ranks factual identities by their paired probability difference. These distinctions specify the decision examined here and motivate testing its output-scale dependence (Guo et al., 2017).

Data assessment also depends on the training protocol. Wang et al. (2026) showed that proxy settings can alter data-recipe rankings and studied assessment relative to recipe-specific tuning. Their analysis motivates our distinction between scorer checkpoints, common final-training settings, and equally budgeted validation selection. Our configuration rule selects exclusively by original-input validation loss under the training distribution.

Rare-example importance and shortcut-aware scoring provide complementary context. Feldman & Zhang (2020) connect memorization and influence to generalization on long-tail examples. Fazla & Hashemi (2026) study the dependence of sample scores on spurious features and propose a two-stage score separating core and spurious learning dynamics. Our controlled audit isolates a particular decision: how paired probability response ranks within-class rare predictive structure, and how the same intervention changes learning when applied to retained identities. The rare flag is distinct from nuisance–label disagreement, and evaluation reports both attributes jointly.

Group-robust learning supplies complementary training references. LfF trains biased and debiased predictors jointly using relative-difficulty weights (Nam et al., 2020); DFR retrains a frozen en-

coder's last layer on balanced reweighting data (Kirichenko et al., 2023). We include both under stated resource budgets. Waterbirds pairs CUB bird foregrounds with scene backgrounds and provides a natural-image benchmark for group shifts (Sagawa et al., 2020); CelebA provides annotated face attributes (Liu et al., 2015). Exact paired interventions in such images additionally require validated transformations.

3 METHOD

For binary labels $y_i \in \{0, 1\}$, let T be a known label-preserving nuisance intervention and p_θ the early classifier. We use absolute sensitivity $s_i = |p_\theta(1 \mid x_i) - p_\theta(1 \mid Tx_i)|$, signed loss benefit $b_i = \log p_\theta(y_i \mid x_i) - \log p_\theta(y_i \mid Tx_i)$, and binary EL2N (Paul et al., 2021) $d_i = \sqrt{2}(1 - p_\theta(y_i \mid x_i))$. Low-score selection retains the smallest values. Positive signed benefit means that the factual input has lower loss than its intervened counterpart under the fitted scorer. All selection rules act on training identities; augmentation determines which version of an identity is presented during retraining.

Probability response combines the midpoint of a logit pair with its half-difference. For classifier logits $z_0(x), z_1(x)$, write the binary logit difference as $\ell(x) = z_1(x) - z_0(x)$ and define $t = [\ell(x) + \ell(Tx)]/2$ and $v = [\ell(x) - \ell(Tx)]/2$. With logistic function σ,

$$s(t, v) = |\sigma(t + v) - \sigma(t - v)| = \frac{\sinh |v|}{\cosh t + \cosh |v|}.$$

For fixed $|v| > 0$, this expression is even in t and strictly decreases with $|t|$: its derivative for $t > 0$ is $-\sinh |v| \sinh t/(\cosh t + \cosh |v|)^2 < 0$. Hence low probability response can favor large midpoint margins at the same intervention magnitude. This ordering also holds for any strictly increasing transformation of s. The linear construction gives these components a feature interpretation; in a nonlinear model, t and v are algebraic summaries of paired outputs. Retraining outcomes are measured separately. For multiclass softmax predictions, Appendix E gives a centered-logit construction in which probability response approaches zero at fixed intervention contrast as one common class margin grows.

For retention fraction f and training size n, all subsets contain $\lfloor fn \rfloor$ unique identities, with class budgets assigned by proportional largest remainders. Within each class we standardize stable features (standard-deviation floor 0.1) and run KMeans with ten initializations. The main comparison uses four clusters per class as a shared geometric coverage rule; clusters need not correspond to digit identities. A ten-seed study separately varies $K_c \in \{2, 4, 5, 8\}$ and measures rare composition and retrained accuracy. With class budget B_c and cluster count K_c, coverage reserves $\lfloor 0.5 B_c / K_c \rfloor$ examples per cluster, capped by its size, then fills the remaining budget by score. The selector receives training labels and stable representations; nuisance and subtype flags are reserved for evaluation.

The audit supplies a known stable representation to each coverage control. Numeric data use $(x + Tx)/2$, and rendered digits use the original grayscale pixels. Together with identical class and cluster quotas, these representations hold coverage resources fixed across rankings. Matched-subset augmentation receives the same known intervention, allowing its training use to be compared at a fixed retained-identity budget.

4 EXPERIMENTAL SETUP

The synthetic factorial varies nuisance-label agreement $\rho \in \{0.90, 0.99\}$ independently of rare-subtype probability $\pi \in \{0.02, 0.10\}$. Each condition has 2,400 training, 1,200 validation and 8,000 test examples. Labels $y \in \{0, 1\}$ are balanced in expectation. The nuisance label $a \in \{0, 1\}$ satisfies $\Pr(a = y \mid y) = \rho$ and is independent of rare status conditional on y. With sign $z = 2y - 1$ and independent rare flag r, coordinates zero and one have means $1.5z(1 - r)$ and $1.5zr$, respectively; coordinate two has mean $3r$. Independent Gaussian noise has standard deviation 0.65. Coordinate three equals $2(2a - 1)$ plus noise of standard deviation 0.2; coordinates four through seven are independent $\mathcal{N}(0, 0.65^2)$ noise. The intervention negates coordinate three. Test nuisance agreement and rare prevalence are both 0.5; validation follows the training distribution.

The colored-digits task uses the 1,797 grayscale 8×8 images provided by scikit-learn (scikit-learn developers, 2025). The binary label is one if the digit is at least five. Digits one and six define fixed

subtype identities, one in each class. We reserve 30% of identities for test and 20% of the remainder for validation, then reduce these subtypes to approximately 5% within each training and validation class. Test retains the original digit prevalence. For grayscale intensity g, the rendered image is $g + (1 - g)c$, where c is a red or green background of intensity 0.6. Its color agrees with the label with probability 0.95 in training and validation and 0.5 in test. Swapping backgrounds preserves the source grayscale image. The synthetic factorial isolates prevalence from identity; the image tasks test the same selection logic on fixed nonlinear subtype identities.

All local classifiers are MLPs with flattened inputs, 64 ReLU hidden units, and two logits. An early scorer trains for 30 updates; final models train for 300 updates with batch size 128, learning rate 0.03, momentum 0.9, and weight decay 0.001. Sampling is with replacement. Seeds are 11, 22, 33, 44 and 55, with paired final initialization across methods. Fixed final checkpoints are evaluated; the training grid separately evaluates six configurations per method.

Ten subset methods are evaluated at retention fractions 0.3, 0.5 and 0.7: random, low/high EL2N, low/high sensitivity, low signed benefit, and coverage with random, low EL2N, low sensitivity, or low signed benefit. Three references use all data: ERM, a JTT-style fixed error-set sampler with total error weight 20, and counterfactual augmentation with flip probability 0.5. Each final model processes 38,400 sampled examples. Equal final updates do not imply equal end-to-end compute; scorer, paired inference, clustering, selection and training costs are recorded separately.

The primary outcome is worst accuracy among the four (y, a) groups. We also record subgroup counts, subtype accuracy, the minimum over eight (y, a, r) subgroups, mean test probability response, and fractions of subtype or nuisance-minority training examples retained. Tables report five-seed means and sample standard deviations. Paired 95% Student-t intervals use the data-seed method-minus-control differences as the unit of analysis. These are descriptive intervals across exploratory comparisons; overlapping image-source populations limit their interpretation to the sampled training procedures. The ten fresh seeds below provide a separate precision check, with sample size fixed before their outcomes.

The CNN evaluation uses a custom colored-MNIST task with the same labels, fixed subtype identities, renderer, nuisance probabilities, and selection rules. Original MNIST training identities are split into a 50,000-example pool and 10,000-example validation set, stratified by digit. We sample 10,000 training identities before reducing digits one and six to approximately 5% per binary class; all 10,000 original test identities remain held out.

The CNN uses two padded 3×3 convolutions with 16 and 32 channels, ReLU activations, a 2×2 max pool after the first layer, and a 2×2 average pool after the second, followed by a two-logit linear classifier. In the fixed CNN suite, each final model trains for 2,000 SGD updates at learning rate 0.01, batch size 128, momentum 0.9 and weight decay 0.001. Each seed's scorer uses 100 updates. Seeds are 101, 202, 303, 404 and 505. Training runs on an RTX 5080 with PyTorch 2.7.1 and CUDA 12.8.

The deletion experiment uses five data seeds (611, 622, 633, 644, 655). It trains three paired optimizer seeds for each of ten class-matched bins under random, sensitivity, signed-benefit, and EL2N rankings, together with paired full-data references. Final MLP training uses 300 updates, and repeated fits are averaged within each data seed before inference.

The training grid holds each scorer and subset fixed while varying optimization over five data seeds (711, 722, 733, 744, 755). It evaluates the synthetic condition $\rho = 0.90, \pi = 0.02$, colored digits, and colored MNIST. This grid is separate from the fixed 2,000-update CNN suite. MLP learning rates are $0.01, 0.03, 0.1$ with 300 or 900 updates; CNN rates are $0.003, 0.01, 0.03$ with 1,000 or 3,000 updates. Each task compares five half-retention procedures. The matched augmentation procedure uses exactly the coverage-random identities and applies T to each sampled input with probability 0.5.

We report all six common configurations and a validation-selected result. Selection minimizes original-input validation cross-entropy independently for each method and seed, with ties ordered by learning rate and duration. Every method receives six candidates, and test outcomes are excluded from selection. Paired descriptive intervals use the five data seeds.

Table 1: Worst-group accuracy (%, mean ± SD, five seeds). Selectors retain 50%; bottom references retain all data. All final models receive 300 updates.

Method	$\rho = .90, \pi = .02$	$\rho = .90, \pi = .10$	$\rho = .99, \pi = .02$	$\rho = .99, \pi = .10$	Digits
Random	63.1 ± 4.4	82.0 ± 3.1	28.5 ± 9.5	29.6 ± 13.0	24.0 ± 7.0
Low EL2N	14.8 ± 5.2	15.0 ± 6.5	8.4 ± 3.5	6.5 ± 2.8	0.0 ± 0.0
High EL2N	70.1 ± 8.1	86.4 ± 2.9	44.0 ± 4.9	49.4 ± 7.6	35.4 ± 11.1
Low sensitivity	50.8 ± 10.5	58.3 ± 5.7	32.2 ± 11.0	37.5 ± 9.5	27.0 ± 4.1
High sensitivity	71.9 ± 7.5	86.7 ± 4.2	39.4 ± 11.1	36.1 ± 10.4	19.1 ± 3.6
Coverage + sensitivity	61.4 ± 14.3	80.1 ± 5.0	29.3 ± 8.9	37.6 ± 10.3	28.4 ± 4.6
Coverage + random	66.5 ± 6.5	81.5 ± 4.4	27.4 ± 9.7	30.1 ± 11.9	24.0 ± 7.0
Coverage + signed	75.1 ± 15.4	87.6 ± 2.2	58.0 ± 9.0	75.2 ± 6.1	66.5 ± 4.3
Full ERM	67.0 ± 6.5	80.8 ± 5.9	36.3 ± 5.7	41.1 ± 9.4	27.0 ± 7.6
JTT-style (full)	81.9 ± 8.7	94.8 ± 2.0	61.0 ± 10.4	75.0 ± 3.7	57.5 ± 7.2
CF augment (full)	88.7 ± 1.8	96.3 ± 1.2	88.5 ± 1.8	96.1 ± 1.4	89.5 ± 1.8

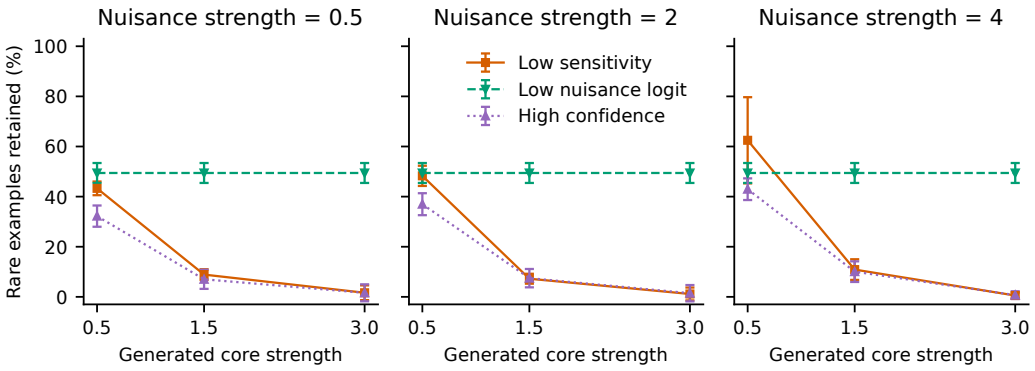


Figure 2: Probability and confidence ranking discard rare evidence as generated core strength increases. Each panel fixes generated nuisance strength. Curves show five-seed mean rare retention under classwise half-retention; bars are sample standard deviations. Nuisance-logit ranking retains roughly half. Both learned components may change after refitting.

5 Probability Response and Rare Evidence

Response ranking sharply redistributes rare predictive evidence. In the $\rho = 0.90, \pi = 0.02$ condition at 50% retention, low-response selection retains 8.6% of rare examples and reaches 50.8% worst-group accuracy, while high-response selection retains 91.4% and reaches 71.9%. Table 1 reports the corresponding worst-group accuracies; Table 7 reports these retention values. Their joint comparison motivates the matched-coverage and retraining analyses.

The linear factorial distinguishes probability ranking from nuisance-logit ranking. The 45 scorers cross core strengths $c \in \{0.5, 1.5, 3.0\}$ and nuisance strengths $u \in \{0.5, 2.0, 4.0\}$ over five seeds. Specifically, the generator replaces the first two coordinate means by $cz(1-r)$ and czr, and the nuisance mean by $u(2a-1)$; the rare-marker mean $3r$, noise distributions, $\rho = .95$ and $\pi = .10$ remain fixed. Define $t = [\ell(x) + \ell(Tx)]/2$ and $v = [\ell(x) - \ell(Tx)]/2$ for binary logit difference ℓ. Reconstruction of $|\sigma(t+v) - \sigma(t-v)|$ is an algebraic consistency check. A secondary analysis applies classwise half-retention using low probability sensitivity, low $|v|$ (the figure's low nuisance-logit rule), high $|t|$, or high $|t+v|$ (original confidence). At generated nuisance strength 2, increasing core strength from 0.5 to 3 changes mean rare retention from 48.29% to 1.15% under sensitivity, while low-$|v|$ ranking retains 49.43% at both endpoints. Original-confidence ranking changes from 36.97% to 1.46%. Figure 2 includes all nine factorial cells. The shared pattern under probability and confidence ranking links rare removal to margin-based selection in this construction. Both learned components may change across refitted factorial cells; the fixed-model identity above supplies the conditional relationship.

Fixed-model probability scaling. A temperature sweep holds the fitted linear scorer fixed and changes $s_\tau = |\sigma((t+v)/\tau) - \sigma((t-v)/\tau)|$ for $\tau \in \{0.25, 0.5, 1, 2, 4\}$. Positive scaling preserves

Table 2: Coverage-sensitivity minus coverage-random at 50% retention: paired worst-group accuracy differences in percentage points. Intervals are unadjusted descriptive 95% t intervals over five paired seeds.

Condition	Mean difference	Interval
$\rho = .90, \pi = .02$	-5.16	[-17.56, 7.24]
$\rho = .90, \pi = .10$	-1.34	[-11.73, 9.05]
$\rho = .99, \pi = .02$	1.90	[-10.29, 14.09]
$\rho = .99, \pi = .10$	7.44	[-9.73, 24.62]
Digits	4.45	[-7.48, 16.39]

argmax predictions. At generated core strength 1.5 and nuisance strength 2, increasing τ from 1 to 4 changes mean rare retention from 7.23% to 18.75% under classwise half-retention. Thus the selected evidence changes even when classifier decisions are fixed. This secondary analysis uses the same 45 fitted scorers; complete cell-wise results are in the numerical records. Its endpoint is retention, with nonlinear retraining effects requiring a separate experiment.

6 Coverage and subgroup accuracy

Coverage often raises rare retention without establishing a score-specific advantage. In the first synthetic condition it increases rare retention from 8.6% to 69.3%, but coverage-random achieves 66.5% worst-group accuracy compared with 61.4% for coverage-sensitivity. At 50% retention, all five condition-specific paired intervals for coverage-sensitivity minus coverage-random include zero. These intervals do not demonstrate equivalence; their widths also permit meaningful positive or negative effects. Table 2 reports these contrasts.

Retention and learning provide complementary evidence. On digits, coverage-sensitivity retains 52.9% of subtype examples and has 48.3% subtype accuracy. Coverage-signed reaches higher four-group worst accuracy (66.5% versus 28.4%) with 32.7% subtype accuracy. Reporting both metrics distinguishes subset composition from the structure learned by the final model.

The CNN reproduces the separation between coverage and ranking. At half retention, coverage-sensitivity obtains $44.7 \pm 6.3\%$ four-group worst accuracy and $10.0 \pm 13.2\%$ eight-subgroup worst accuracy; coverage-random obtains $59.2 \pm 17.0\%$ and $45.3 \pm 29.7\%$, respectively. These image results concern the fixed digit subtypes. The complete thirteen-method CNN results, including both subgroup minima, appear in Table 12. These image experiments test nonlinear models; the synthetic factorial separately varies subtype prevalence.

Cluster sensitivity and fresh-seed precision. A separate frozen study uses ten fresh seeds (811 through 910, listed in the supplement), $K_c = 2, 4, 5, 8$, and the same fixed training settings. Table 3 reports every cluster count. Matching the number of clusters to five digit identities does not impose semantic assignments; the saved cluster compositions measure how rare examples distribute across them. At $K_c = 4$, the paired differences are -11.9 percentage points on synthetic data, $+1.7$ on digits and -15.6 on MNIST. Synthetic data change sign at $K_c = 8$ ($+3.0$ points), while MNIST remains negative at every tested count. Thus coverage granularity changes the empirical ordering, and low-response ranking supplies no consistent advantage across these settings. These ten-seed estimates are reported separately from the original five-seed comparisons.

7 Training protocol and intervention use

Figure 3 reports the complete matched-subset result behind the headline claim. Mean gains are positive in all 18 configurations. They range from 20.50 to 22.51 points on synthetic data, 51.86 to 68.54 points on colored digits, and 22.36 to 82.72 points on colored MNIST. The comparison isolates intervention use because each pair has identical retained identities and optimization settings.

Validation-selected comparisons test whether the same distinction survives method-specific optimization. Relative to coverage-random, coverage-sensitivity changes four-group worst accuracy by

Table 3: Cluster sensitivity with ten fresh seeds. Four-group accuracy differences, coverage-sensitivity minus coverage-random (percentage points), with paired descriptive 95% intervals. All configurations use half retention; synthetic $\rho = .90, \pi = .02$.

Clusters	Synthetic	Digits	MNIST
2	-10.2 [-15.0, -5.5]	2.0 [-9.6, 13.6]	-12.8 [-21.2, -4.3]
4	-11.9 [-19.9, -4.0]	1.7 [-6.2, 9.6]	-15.6 [-22.7, -8.6]
5	-6.2 [-16.0, 3.5]	4.0 [-3.4, 11.4]	-14.0 [-23.7, -4.2]
8	3.0 [1.6, 4.5]	-0.1 [-8.6, 8.3]	-9.8 [-18.8, -0.8]

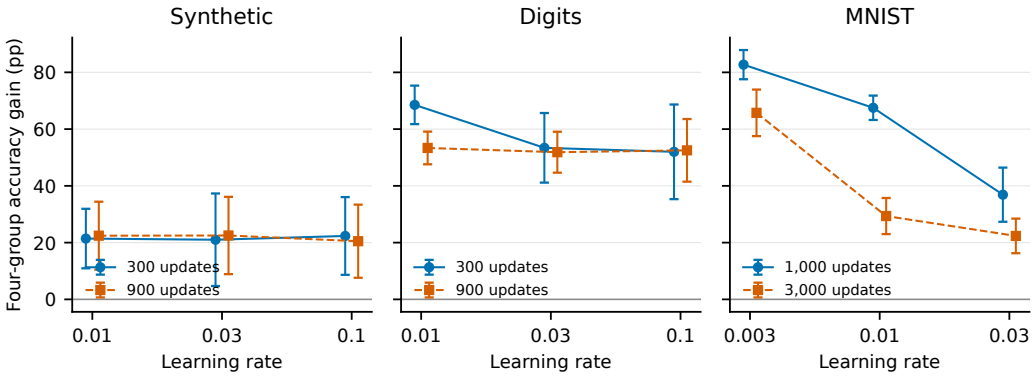


Figure 3: All 18 matched-subset comparisons. Each point is the paired four-group worst-accuracy difference between training with and without the intervention on the same coverage-random identities. Bars are descriptive 95% intervals over five data seeds. Identity budgets, update counts, learning rates, and durations are matched within every comparison.

-17.41, -5.64, and -13.66 points on synthetic data, digits, and MNIST. Applying the intervention to the coverage-random subset instead improves it by 20.00, 63.50, and 23.83 points. Table 4 reports both four- and eight-subgroup results.

The eight-subgroup gains under validation selection are 35.46, 60.75, and 38.02 points. Thus the matched-subset result extends to groups defined jointly by label, nuisance, and subtype status. The intervention improves what the retained model learns without changing which identities are retained.

Debiasing references. Table 5 places subgroup learning alongside LfF and DFR-Val using five fresh seeds from the cluster cohort. All references use the same CNN architecture. LfF trains two full-data networks; DFR freezes the full-data ERM encoder and uses group-labeled validation examples for head fitting and regularization selection. The augmentation row uses half the training identities. DFR-Val reaches 92.1% four-group accuracy versus ERM's 58.5%. The fixed-schedule LfF adaptation reaches zero worst-group accuracy in all five runs: predictions predominantly follow the opposite background label, exposing a failure of this optimization configuration. This result describes the specified SGD adaptation rather than a tuned estimate of LfF performance. These stated resource budgets provide context for the learned subgroup outcomes; the matched-subset experiment supplies the controlled intervention comparison.

8 Discussion

The study identifies three quantities that must be distinguished in intervention-based data selection: output response, retained evidence, and learned subgroup accuracy. The logit identity characterizes response geometry, matched coverage makes rankings comparable, and matched-subset augmentation measures the value of intervention during training. Together they replace a score-only interpretation with an evaluation of the model learned from the selected identities.

Table 4: Validation-selected accuracy differences versus coverage-random at half retention (percentage points). Brackets are paired descriptive 95% intervals over five data seeds. Each method selects among six configurations by validation label loss.

Task	Minimum	Coverage + sensitivity	Coverage-random + CF
Synthetic	Four-group	-17.41 $[-30.95, -3.87]$	20.00 $[7.08, 32.93]$
Synthetic	Eight-subgroup	-34.34 $[-58.82, -9.85]$	35.46 $[12.29, 58.62]$
Digits	Four-group	-5.64 $[-17.18, 5.89]$	63.50 $[48.20, 78.81]$
Digits	Eight-subgroup	-1.98 $[-5.67, 1.71]$	60.75 $[49.02, 72.48]$
MNIST	Four-group	-13.66 $[-28.89, 1.57]$	23.83 $[16.10, 31.55]$
MNIST	Eight-subgroup	-25.13 $[-52.12, 1.86]$	38.02 $[21.37, 54.68]$

Table 5: CNN reference comparisons on five fresh seeds. Accuracy in percent, mean $\pm$ SD. DFR-Val additionally uses target/background validation labels; LfF trains two networks. CF uses the known intervention on a half-retention coverage-random subset.

Method	Training identities	Four-group	Eight-subgroup
ERM	All	58.5 ± 3.8	35.5 ± 18.8
LfF	All	0.0 ± 0.0	0.0 ± 0.0
DFR-Val	All	92.1 ± 1.7	83.9 ± 6.4
Coverage-random + CF	Half	94.7 ± 0.8	91.0 ± 2.5

The theoretical result applies to binary prediction under any label-preserving intervention, while the empirical study uses interventions and stable representations known from the data generator. Synthetic data vary subtype prevalence independently of nuisance agreement; the image experiments test fixed digit subtypes with nonlinear models. Natural-image transfer requires paired transformations that preserve foreground identity and target information. Foreground-mask recomposition can support this evaluation on Waterbirds-style images; source identities, mask validity and background-swap controls must be checked. The current evidence concerns the known generators, while Appendix E characterizes multiclass response geometry.

9 Conclusion

A known nuisance intervention supports distinct decisions about which identities to retain and how to train on them. Probability-response geometry can favor confident evidence independently of intervention magnitude, while subgroup evaluation shows whether selected evidence is learned. Across every matched training setting, applying the intervention during training improves mean worst-group accuracy on the same retained identities. Response-based selection should therefore be evaluated through retained composition and retrained subgroup accuracy under matched identity and optimization budgets.

AI use statement

In this work, we used a generative AI system for research and manuscript preparation, including analysis, code, figures, grammar, and wording. We reviewed the AI-assisted work and take responsibility for the final content.

Ethics statement

The experiments use synthetic data and existing handwritten-digit datasets. The study concerns rare predictive patterns in controlled classification tasks; its rare-subtype labels do not represent demographic groups. Any application to people would require separate validation of the intervention and subgroup definitions. Dataset access and attribution are documented in the supplementary materials.

Reproducibility statement

The supplementary code and data records contain executable configurations, split identities, selection scores, predictions, checkpoints, and result tables. Sections 3 and 4 specify the selection and evaluation protocol. The archive README documents data acquisition, execution, metric recomputation, and checksum verification. Source images are obtained through their original dataset loaders.

References

Arda Fazla and Abolfazl Hashemi. Mitigating spurious correlations with memorization-guided dataset de-biasing. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 1540–1573. PMLR, 2026. URL https://proceedings.mlr.press/v337/fazla26a.html.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/1e14bfe2714193e7af5abc64ecbd6b46-Abstract.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2204.02937.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/koh17a.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 *International Conference on Machine Learning*, 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity & difficulty in data pruning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f646a4c970799f789e031b63161d018d-Abstract-Conference.html.

Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/427048354ac2db22d43149c51346bafd-Abstract-Conference.html.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2107.07075.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

scikit-learn developers. load_digits: scikit-learn 1.6.1 documentation. https://scikit-learn.org/1.6/modules/generated/sklearn.datasets.load_digits.html, 2025. Accessed September 12, 2026.

Artem Vysogorets, Kartik Ahuja, and Julia Kempe. DRoP: Distributionally robust data pruning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/110efb00ba39be543b75d097a0aae531-Abstract-Conference.html.

Jiachen T. Wang, Tong Wu, Kaifeng Lyu, James Y. Zou, Dawn Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c7d04ce5e71efdfc7a4a78f7161238a9-Abstract-Conference.html.

Abdelrahman Zayed, Prasanna Parthasarathi, Goncalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. Deep learning on a healthy data diet: Finding important examples for fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2211.11109.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.15809.

A BATCH DELETION AND RETRAINING

The deletion study connects rank bins to retraining outcomes. Using 600 training examples ($n_{\text{train}} = 600, \rho = 0.95, \pi = 0.10$), each score partitions training examples into ten within-class bins. This smaller training pool makes repeated deletion retraining tractable; its intermediate nuisance agreement defines a separate stress condition. We remove one bin at a time, retrain for 300 updates, and compare with a paired full-data fit. Across seeds, Spearman correlations between mean removed score and change in test response range from -0.04 to 0.59 for absolute sensitivity and from -0.94 to -0.15 for signed benefit. The latter associates higher-score removal with larger reductions in test response.

Matched random deletions isolate score-bin effects from removing the same number of class-matched examples. After averaging three optimizer repeats, correlations between mean removed score and the change in test response relative to random deletion range from -0.09 to 0.42 for absolute sensitivity and from -0.66 to -0.36 for signed benefit. Complete per-bin intervals and seed-level correlations are provided in the supplement.

Predictive consequences. Figure 4 reports changes in overall, four-group, eight-subgroup, and rare accuracy relative to matched random deletion, recomputed from the same 600 deletion fits. Three optimizer repeats are averaged within each data seed. Extreme signed-benefit and EL2N bins account for the largest accuracy changes: across interior bins, their mean overall-accuracy contrasts range from -0.15 to 1.62 points and -0.15 to 1.64 points, respectively, compared with minima of -34.41 and -31.96 when endpoint bins are included. Sensitivity-ranked deletion ranges from -2.23 to 1.07 points across all bins. These companion outcomes establish why a reduction in intervention response must be interpreted together with predictive accuracy. Per-bin rare and nuisance-minority composition and endpoint summaries are supplied in the analysis records.

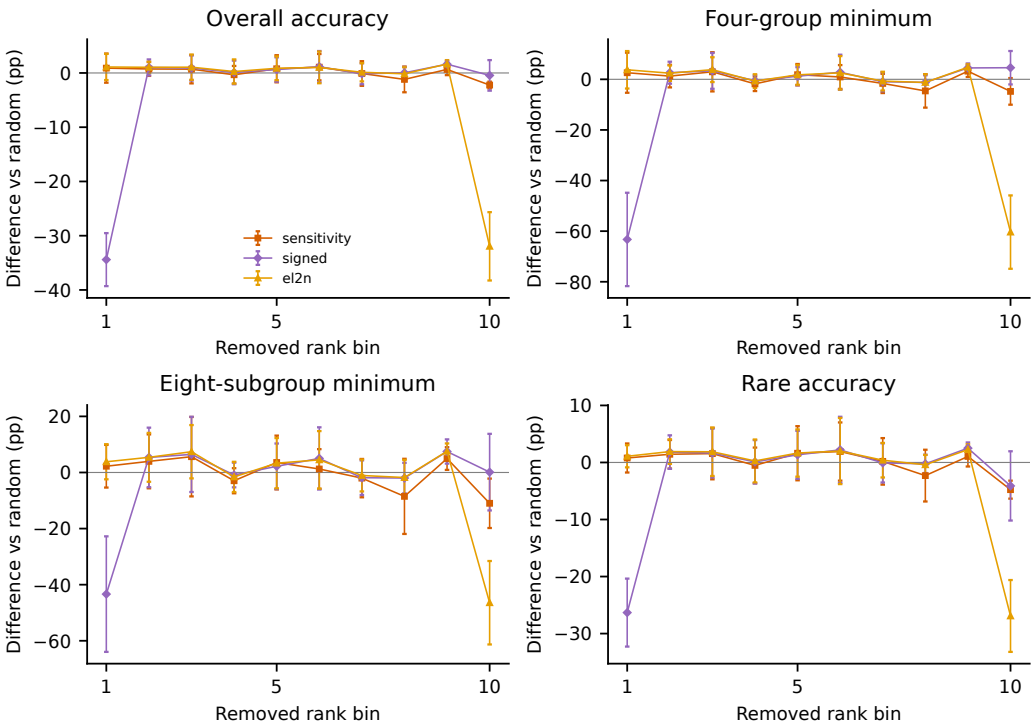


Figure 4: Predictive effects of removing score bins, relative to class-matched random removal. Points average three paired optimizer repeats within each of five data seeds; bars are descriptive 95% intervals across seeds. Axes report percentage points and use panel-specific scales.

B PAIRED TRAINING INTERACTIONS

The synthetic task in this appendix has $\rho = .90, \pi = .02$. For each data seed, let $\Delta_{s,h}$ be coverage-sensitivity minus coverage-random accuracy at setting h. The interaction between settings h and h' is $\Delta_{s,h'} - \Delta_{s,h}$; intervals are computed over the five paired seed values. The numerical records contain all 15 setting pairs per task at both subgroup resolutions. Figure 5 shows the seed-level four-group contrasts across the six configurations.

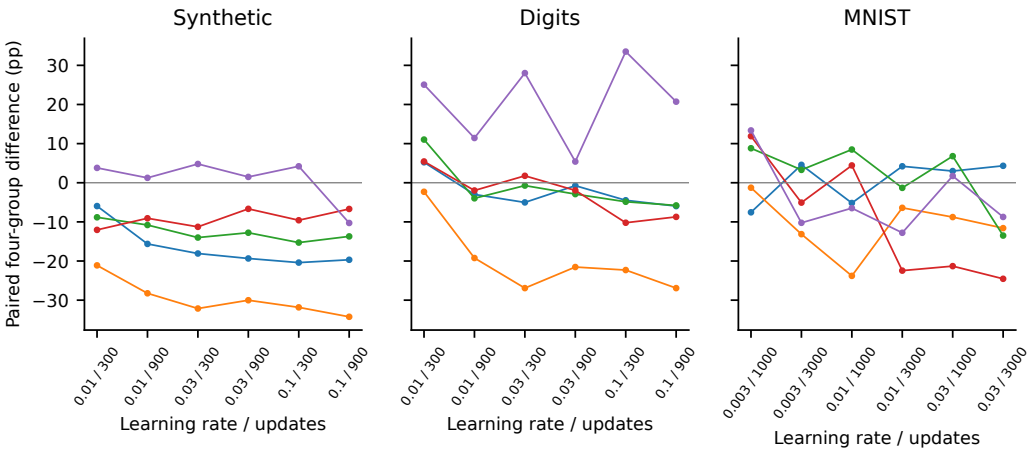


Figure 5: Each line tracks one data seed across the six settings. Vertical values are coverage-sensitivity minus coverage-random four-group accuracy. Repeated settings within a seed are paired observations.

Table 6: Paired training-duration interactions: change in the coverage-sensitivity versus coverage-random contrast when duration increases from 300 to 900 updates (MLPs) or 1,000 to 3,000 updates (CNN). Entries are percentage points with descriptive 95% intervals over five paired data seeds. These are secondary exploratory comparisons.

Task	Learning rate	Four-group interaction	Eight-subgroup interaction
Synthetic	0.01	$-3.67\ [-9.76, 2.42]$	$-8.45\ [-18.80, 1.91]$
Synthetic	0.03	$0.67\ [-3.12, 4.47]$	$0.87\ [-5.74, 7.48]$
Synthetic	0.1	$-2.34\ [-11.11, 6.43]$	$-4.54\ [-19.08, 10.01]$
Digits	0.01	$-12.24\ [-17.47, -7.01]$	$-2.83\ [-9.90, 4.24]$
Digits	0.03	$-3.77\ [-17.74, 10.20]$	$-2.69\ [-11.67, 6.29]$
Digits	0.1	$-3.67\ [-10.56, 3.22]$	$-5.10\ [-7.12, -3.09]$
MNIST	0.003	$-9.17\ [-26.10, 7.76]$	$-0.83\ [-2.30, 0.64]$
MNIST	0.01	$-3.24\ [-24.71, 18.23]$	$-17.68\ [-38.99, 3.62]$
MNIST	0.03	$-7.10\ [-17.67, 3.47]$	$-14.83\ [-37.42, 7.77]$

C COMPLETE HALF-RETENTION RESULTS

Table 7: $\rho = .90, \pi = .02$: half-retention selectors and full-data references. Percent, mean $\pm$ sample SD over five seeds. Rare retention measures retained/original rare training count.

Method	Worst-group acc.	Rare acc.	Rare retention
Random	63.1 ± 4.4	68.2 ± 5.4	49.2 ± 4.5
Low sensitivity	50.8 ± 10.5	58.4 ± 6.9	8.6 ± 4.4
High sensitivity	71.9 ± 7.5	75.4 ± 6.9	91.4 ± 4.4
Coverage + sensitivity	61.4 ± 14.3	70.8 ± 14.3	69.3 ± 42.6
Coverage + random	66.5 ± 6.5	73.6 ± 5.8	78.7 ± 29.2
Coverage + EL2N	34.9 ± 18.4	65.3 ± 17.2	63.1 ± 50.5
Coverage + signed	75.1 ± 15.4	75.8 ± 16.1	66.7 ± 45.7
Low signed benefit	62.1 ± 5.9	62.6 ± 6.0	13.4 ± 8.4
Low EL2N	14.8 ± 5.2	49.9 ± 2.2	5.7 ± 2.8
High EL2N	70.1 ± 8.1	75.2 ± 6.7	94.3 ± 2.8
Full ERM	67.0 ± 6.5	70.9 ± 4.7	100.0 ± 0.0
JTT-style (full)	81.9 ± 8.7	86.2 ± 5.3	100.0 ± 0.0
CF augment (full)	88.7 ± 1.8	85.2 ± 2.0	100.0 ± 0.0

Table 8: $\rho = .90, \pi = .10$: half-retention selectors and full-data references. Percent, mean $\pm$ sample SD over five seeds. Rare retention measures retained/original rare training count.

Method	Worst-group acc.	Rare acc.	Rare retention
Random	82.0 ± 3.1	87.7 ± 1.9	51.5 ± 3.9
Low sensitivity	58.3 ± 5.7	63.8 ± 9.9	11.4 ± 8.2
High sensitivity	86.7 ± 4.2	90.7 ± 2.0	88.6 ± 8.2
Coverage + sensitivity	80.1 ± 5.0	88.7 ± 2.7	60.1 ± 3.1
Coverage + random	81.5 ± 4.4	87.9 ± 1.7	60.0 ± 2.9
Coverage + EL2N	14.4 ± 12.3	53.7 ± 4.5	59.6 ± 2.8
Coverage + signed	87.6 ± 2.2	89.6 ± 1.0	59.7 ± 2.9
Low signed benefit	68.9 ± 6.2	73.7 ± 4.8	18.9 ± 8.9
Low EL2N	15.0 ± 6.5	50.5 ± 2.3	11.0 ± 7.3
High EL2N	86.4 ± 2.9	90.5 ± 1.8	88.9 ± 7.3
Full ERM	80.8 ± 5.9	86.0 ± 3.2	100.0 ± 0.0
JTT-style (full)	94.8 ± 2.0	96.0 ± 1.5	100.0 ± 0.0
CF augment (full)	96.3 ± 1.2	95.7 ± 0.7	100.0 ± 0.0

Table 9: $\rho = .99, \pi = .02$: half-retention selectors and full-data references. Percent, mean $\pm$ sample SD over five seeds. Rare retention measures retained/original rare training count.

Method	Worst-group acc.	Rare acc.	Rare retention
Random	28.5 ± 9.5	53.2 ± 2.3	49.2 ± 4.5
Low sensitivity	32.2 ± 11.0	52.8 ± 3.9	19.8 ± 11.2
High sensitivity	39.4 ± 11.1	55.9 ± 3.0	80.2 ± 11.2
Coverage + sensitivity	29.3 ± 8.9	55.1 ± 4.9	73.8 ± 39.5
Coverage + random	27.4 ± 9.7	54.5 ± 3.3	78.7 ± 29.2
Coverage + EL2N	4.4 ± 3.9	51.4 ± 4.9	62.2 ± 51.8
Coverage + signed	58.0 ± 9.0	59.5 ± 10.1	68.4 ± 43.8
Low signed benefit	62.9 ± 10.0	55.5 ± 14.2	12.9 ± 12.6
Low EL2N	8.4 ± 3.5	49.5 ± 1.0	6.4 ± 5.5
High EL2N	44.0 ± 4.9	58.6 ± 3.7	93.6 ± 5.5
Full ERM	36.3 ± 5.7	55.3 ± 2.2	100.0 ± 0.0
JTT-style (full)	61.0 ± 10.4	67.0 ± 7.6	100.0 ± 0.0
CF augment (full)	88.5 ± 1.8	85.2 ± 2.2	100.0 ± 0.0

Table 10: $\rho = .99, \pi = .10$: half-retention selectors and full-data references. Percent, mean $\pm$ sample SD over five seeds. Rare retention measures retained/original rare training count.

Method	Worst-group acc.	Rare acc.	Rare retention
Random	29.6 ± 13.0	58.2 ± 4.6	51.5 ± 3.9
Low sensitivity	37.5 ± 9.5	55.7 ± 4.9	20.0 ± 9.6
High sensitivity	36.1 ± 10.4	63.8 ± 4.3	80.0 ± 9.6
Coverage + sensitivity	37.6 ± 10.3	61.4 ± 8.1	60.8 ± 2.7
Coverage + random	30.1 ± 11.9	59.1 ± 4.4	60.0 ± 2.9
Coverage + EL2N	3.6 ± 1.5	49.6 ± 0.9	59.6 ± 2.8
Coverage + signed	75.2 ± 6.1	79.4 ± 5.1	60.6 ± 2.6
Low signed benefit	69.0 ± 12.5	68.8 ± 12.9	17.5 ± 12.7
Low EL2N	6.5 ± 2.8	49.5 ± 1.0	14.6 ± 10.6
High EL2N	49.4 ± 7.6	66.6 ± 4.2	85.5 ± 10.6
Full ERM	41.1 ± 9.4	60.9 ± 3.9	100.0 ± 0.0
JTT-style (full)	75.0 ± 3.7	81.6 ± 2.1	100.0 ± 0.0
CF augment (full)	96.1 ± 1.4	95.7 ± 0.8	100.0 ± 0.0

Table 11: Digits: half-retention selectors and full-data references. Percent, mean $\pm$ sample SD over five seeds. Rare retention measures retained/original rare training count.

Method	Worst-group acc.	Rare acc.	Rare retention
Random	24.0 ± 7.0	50.5 ± 7.0	47.6 ± 6.1
Low sensitivity	27.0 ± 4.1	46.6 ± 9.3	53.3 ± 9.3
High sensitivity	19.1 ± 3.6	45.9 ± 8.0	46.7 ± 9.3
Coverage + sensitivity	28.4 ± 4.6	48.3 ± 7.2	52.9 ± 8.8
Coverage + random	24.0 ± 7.0	50.5 ± 7.0	47.6 ± 6.1
Coverage + EL2N	0.0 ± 0.0	48.8 ± 7.0	8.1 ± 3.2
Coverage + signed	66.5 ± 4.3	32.7 ± 9.3	10.5 ± 4.0
Low signed benefit	66.1 ± 5.2	34.3 ± 7.9	11.0 ± 4.9
Low EL2N	0.0 ± 0.0	48.8 ± 7.0	8.6 ± 2.7
High EL2N	35.4 ± 11.1	53.6 ± 8.2	91.0 ± 3.1
Full ERM	27.0 ± 7.6	50.3 ± 6.1	100.0 ± 0.0
JTT-style (full)	57.5 ± 7.2	55.6 ± 7.0	100.0 ± 0.0
CF augment (full)	89.5 ± 1.8	80.9 ± 4.7	100.0 ± 0.0

Table 12: Complete colored-MNIST results for the fixed 2,000-update CNN suite. Percent, mean ± SD over five seeds. Selection rows retain half the training identities; the final three references retain all.

Method	Four-group	Eight-subgroup	Rare acc.	Rare retention
Random	58.8 ± 16.8	42.8 ± 28.2	74.3 ± 12.8	49.2 ± 3.6
Low sensitivity	33.9 ± 4.7	1.2 ± 2.1	52.9 ± 15.8	25.4 ± 10.1
High sensitivity	54.7 ± 8.5	27.2 ± 17.0	68.3 ± 10.5	74.6 ± 10.1
Coverage + sensitivity	44.7 ± 6.3	10.0 ± 13.2	59.3 ± 14.5	26.5 ± 8.7
Coverage + random	59.2 ± 17.0	45.3 ± 29.7	75.5 ± 13.4	49.2 ± 3.6
Coverage + EL2N	0.0 ± 0.0	0.0 ± 0.0	49.2 ± 0.7	8.6 ± 4.0
Coverage + signed	70.2 ± 6.1	26.7 ± 10.1	59.4 ± 5.4	8.8 ± 2.8
Low signed benefit	60.2 ± 6.3	14.3 ± 8.8	45.8 ± 5.5	7.4 ± 2.4
Low EL2N	0.0 ± 0.0	0.0 ± 0.0	49.2 ± 0.7	5.8 ± 3.3
High EL2N	69.2 ± 8.4	61.7 ± 17.4	84.8 ± 7.2	94.2 ± 3.3
Full ERM	59.1 ± 6.3	39.3 ± 15.8	70.2 ± 8.0	100.0 ± 0.0
JTT-style (full)	88.8 ± 2.4	80.0 ± 4.6	92.4 ± 1.2	100.0 ± 0.0
CF augment (full)	93.6 ± 1.7	89.1 ± 4.2	94.4 ± 1.4	100.0 ± 0.0

D BUDGET SENSITIVITY

Figure 6 compares the three budgets across both image sources and the four synthetic conditions.

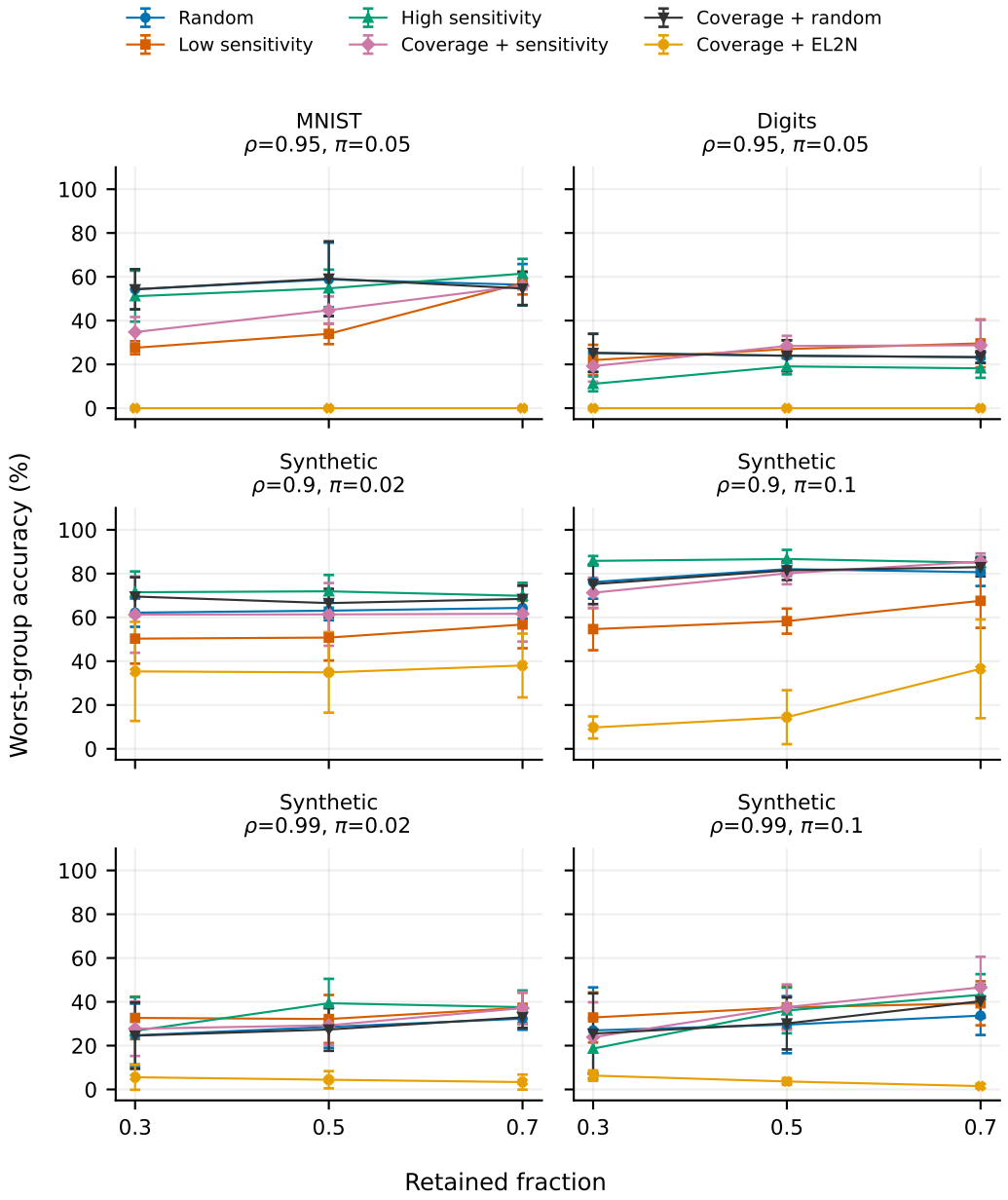


Figure 6: Ranking effects vary across retention budgets. Points are five-seed means; bars are sample standard deviations, not confidence intervals. The six conditions and six displayed controls are shown. Complete thirteen-method results are in the artifact tables.

E MULTICLASS RESPONSE GEOMETRY

For C classes, center logits with $P = I - \mathbf{1}\mathbf{1}^\top/C$ and define $m = P(z(x) + z(Tx))/2$, $v = P(z(x) - z(Tx))/2$. Softmax is invariant to common shifts, giving the exact representation $p = \mathrm{softmax}(m+v)$ and $q = \mathrm{softmax}(m-v)$. Use total variation $R(m, v) = \frac{1}{2}\|p-q\|_1$, which reduces to the absolute probability difference for binary predictions.

If v is fixed and $m_k - m_j \to \infty$ for every $j \neq k$, both distributions converge to the same point mass at class k, so $R(m, v) \to 0$. An exact three-class instance takes $m_M = (2M/3, -M/3, -M/3)$ and $v = (0, a, -a)$ with $a > 0$:

$$R(m_M, v) = \frac{2\sinh a}{e^M + 2\cosh a} \longrightarrow 0 \quad \text{as } M \to \infty, \qquad \|v\|_2 = \sqrt{2}a.$$

The denominators of both softmax distributions coincide, and their last two probabilities interchange, which gives the formula directly. At $a = 1$, response changes from 0.575210 at $M = 0$ to 0.000107 at $M = 10$. Increasing midpoint norm alone does not suffice: $M \to -\infty$ in the same construction instead gives $R \to \tanh a$. Thus the multiclass statement concerns dominance of a common class. The binary scalar monotonicity result and the multiclass dominance result have distinct conditions. Numerical checks cover 400 random centered-logit pairs; these validate the algebra, while the fitted classification experiments remain binary.

F CLUSTER AND REFERENCE PROTOCOLS

The ten cluster-study seeds are 811, 822, 833, 844, 855, 866, 877, 888, 899 and 910. Each task has 90 final fits: two ranking rules at four cluster counts, plus coverage-random augmentation at $K_c = 4$. Synthetic data use $\rho = .90, \pi = .02$; digit tasks use training $\rho = .95$ and rare prevalence approximately $.05$. Scorers retain the original 30-update MLP and 100-update CNN protocols. Final training is 300 MLP or 2,000 CNN updates. All settings are retained, and inference uses paired data-seed differences. At $K_c = 5$, rare examples occupy an average of 1.85 clusters per target class on digits and 3.55 on MNIST (averaged over ten seeds and two classes). Every rare-occupied image cluster also contains common examples at every tested cluster count. Thus equal digit and cluster counts do not isolate rare identities. Saved assignments and subtype labels reproduce these composition checks. In the original five-seed digits study at half retention, random and coverage-random select identical identities: the shared random ranking already satisfies the coverage floors. Their identical predictions give the same $24.0\% \pm 7.0$ four-group accuracy.

LfF uses the first five seeds, two independently initialized CNNs, and 2,000 SGD updates per network at learning rate $.01$, momentum $.9$ and weight decay $.001$. Its released loss implementation uses detached qp_y^q times cross-entropy with $q = .7$. Per-example CE moving averages use coefficient $.7$, initialize on first observation, and are normalized by the maximum stored EMA within each target class. The debiased CE is weighted by the biased normalized EMA divided by the sum of both normalized EMAs plus 10^{-8}. Shuffled batches contain at most 128 examples. Only the debiased network is evaluated. This adapts the loss mechanics to the study's CNN protocol.

DFR-Val uses the same full-data ERM encoder with all encoder parameters frozen. Stratified validation halves provide head-fitting and regularization-selection sets. Standardization is fit on the fitting half. L1 logistic regression uses liblinear and $C \in \{.01, .03, .07, .1, .3, .7, 1\}$; a common group-balanced fitting subset is used across candidates. Selection maximizes four-group accuracy on the other half, with smaller-C tie breaking. The final head averages ten fits on balanced resamples of all validation examples, with standardization refit there. Each group contributes its smallest available group count. Target/background validation labels construct these groups. Test identities remain untouched. Saved tuning predictions verify every selected C.