WHAT DOES A DATA-SELECTION GAIN MEASURE?

Anonymous authors
Paper under double-blind review

ABSTRACT

A gain from selecting training data depends on the representation used to score examples and train the learner. It also depends on the candidate supply, the selected group mix and the labels used for guidance. We study what these comparisons measure using frozen feature maps and linear heads. Our central control redraws random examples at the same class–group counts as each selection. Across tasks from eleven benchmark sources, five registered comparisons show no clear worst-group benefit; secondary overall-accuracy residuals are negative in all 55 correlated source–implementation means. Two music comparisons are positive. Selecting real examples from unseen work families rather than generated variants of training works raises test accuracy by 2.88 percentage points. Random acquisition from the two supplies already gives a descriptive 2.60-point difference. On identical selected rows and random controls, retaining original descriptors alongside constructed features increases the selection advantage by 0.55 points relative to constructed features alone. This changes the learner, not the ranking rule. A larger reference gives a positive source-level estimate but an interval over repeated draws that spans zero; guidance shows no clear advantage over direct label use. Feature refresh fails its positive primary, and scorer–learner matching outside music shows no clear benefit. These findings do not provide a winning selector. They show why a selection gain should be reported with its candidate supply, scorer and learner, group-count control, and final accuracy. Code and numerical evidence are open-sourced at: anonymous repository.

1 INTRODUCTION

Data selection is often evaluated by training on the chosen examples and comparing the result with random acquisition at the same budget (Sener & Savarese, 2018; Paul et al., 2021). A positive difference is useful, but what is it evidence for? A subset can help one learner more than another because their representations differ. A rule can also change the group mix of the training data without finding better individual examples within each group. And choosing from a different candidate supply can help even when the choice within that supply is random.

Consider a schematic example. Two rules add the same number of bird images. One chooses more images with an uncommon background. Its final accuracy can differ even if the examples chosen within each class–background group are no better than random. To ask whether those particular images matter, redraw all acquired identities while keeping the chosen class–group counts. We call this the *same-count random* control. Now hold those row sets fixed and change the learner's features. Any change in the selected-over-random advantage belongs to that learner comparison, not to a new ranking rule. Neither step makes the ordinary-random comparison unnecessary; the comparisons answer different questions.

We ask: *what does an observed data-selection gain measure once the candidate supply, the scorer, the learner and the selected group counts are made explicit?* The representation question is central. Features determine which candidates a score prefers and which distinctions a learner can use. Those roles need not agree. We therefore evaluate the same selected sets under different feature maps, as well as different selections under the same learner. All studies use frozen maps and linear heads with small training and acquisition budgets. This setting lets us pair row-level comparisons without confounding them with feature updates during downstream training.

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.

The evidence proceeds from interpretation to possible improvements. First, a cross-source panel compares selected rows with ordinary random and same-count random acquisition. It shows no clear worst-group benefit, while the overall-accuracy residuals show a consistent secondary negative pattern. Next, a symbolic-music study gives two positive contrasts with different meanings: changing candidate supply and retaining the original features in the learner. Follow-on comparisons ask whether more reference labels, direct label use, refreshing a selection after a feature change, or matching the scorer to the learner improve the relevant quantity.

Contributions.

- We connect controls for supply, learner features and observed group counts to the interpretation of a selection gain (§§2–4). The controls are familiar; the contribution is the linked empirical comparison, not a new selector or causal decomposition.
- We distinguish the lack of a clear worst-group benefit from negative secondary overall residuals across eleven sources, and show why a positive music selection contrast must still be read beside random acquisition and absolute accuracy (§§3–4).
- We test proposed ways to improve selection value through guidance and representation changes. Their different outcomes delimit what follows from the earlier comparisons, rather than supplying a general prescription (§5).

2 Comparisons that distinguish different gains

We first separate the value of the selected examples from the value of their observed group mix. We then use the same logic to compare candidate supplies and feature maps.

Data roles and supervision. Each cross-source construction has labelled base training data, a labelled candidate pool, a small labelled reference and held-out test rows. A rule adds a fixed number of pool rows per class to the base data. Reference rows guide selection unless a study explicitly adds them to training. The downstream head-fitting procedure is the same for every acquired set within a learner. Group labels construct balanced references and the same-count control; they are not score inputs for the five reference-guided implementations. These are supervised comparisons, not label-free acquisition.

Comparison	What is held fixed	What the difference measures
Selected vs ordinary random	Learner, pool and per-class budget	Total selection difference at that budget
Selected vs same-count random	Also the selected class–group counts	Selected identities relative to redraws at those counts
Same rows, different learner	Selected rows and their random controls	How the learner's features change the selection advantage

A small accounting relation. Ordinary random acquisition fixes the budget. Same-count random acquisition also fixes the selected counts, but redraws *every* acquired identity. Draws may overlap the selected set; no selected row is deliberately held in place. Fix base data D, pool P, and a learner and evaluation pipeline. Let $V(A)$ be performance after adding subset A to D. Selector j chooses S_j with count vector n_j. Let U be ordinary random and $Q(n_j)$ be same-count random acquisition. Both are uniform over complete allowed sets. Then

$$\begin{aligned} T_j &= V(S_j) - \mathbb{E}V(U), & C_j &= \mathbb{E}V(Q(n_j)) - \mathbb{E}V(U), \\ R_j &= V(S_j) - \mathbb{E}V(Q(n_j)), & T_j &= C_j + R_j. \end{aligned} \tag{1}$$

Here T is the total difference, C the difference associated with observed composition, and R the residual after matching it. The estimates use the same selected, ordinary-random and mean same-count-random accuracies, so the identity holds exactly. It is accounting, not causal mediation: matching declared counts does not match difficulty or dependence within a group. In the cross-source studies the redraw samples without replacement within strata. In music it also obeys the same family and parent caps as selection. When every policy has a fixed quota, that quota already fixes the counts to be matched.

Scores and feature maps. The panel uses five implementations of four strategies. A denoised class-mean gap projects a candidate onto the reference-minus-training direction. Two logistic discriminators rank by reference-versus-pool probability, with either raw or standardized features. A gradient-alignment proxy uses the alignment between candidate and mean reference gradients. Kernel herding selects a set that covers the reference in an RBF geometry (Welling, 2009; Chen et al., 2010). The logistic variants are related to feature-space importance selection (Xie et al., 2023), not the original n-gram DSIR procedure; the gradient proxy is not full TracIn or TRAK (Pruthi et al., 2020; Park et al., 2023). Music uses a plain normalized mean-gap score instead of the denoised panel score. The later study outside music uses quota-stratified k-center selection and no reference. App. A gives the exact contracts.

Metrics and uncertainty. Worst-group accuracy is the minimum test accuracy over declared class–attribute groups; overall accuracy averages test rows. Music instead uses hierarchical accuracy, giving performances, work families and composers their stated weights. A residual is a difference in percentage points, not an accuracy level. For the fourteen constructions from eleven benchmark sources, encoder variants are averaged within source before sources receive equal weight. Source-t intervals and exhaustive source sign-flips describe variation across that fixed, purposive panel under their stated assumptions. Intervals over repeated randomization blocks ask a different question: how the result varies on the same panel. We retain both where available.

Relation to prior work. Group DRO and JTT change training objectives or example weights (Sagawa et al., 2020; Liu et al., 2021); balancing and last-layer retraining show the importance of group composition (Idrissi et al., 2022; Kirichenko et al., 2023; LaBonte et al., 2023). Our redraw evaluates an acquired subset rather than replacing these training methods. Model-aware selection connects data quality to downstream utility (Engstrom et al., 2024), and reference-guided methods use small target sets to steer selection (Xie et al., 2023; Xia et al., 2024). Selection comparisons can depend on budget, architecture and training variation (Munjal et al., 2022); data acquired by one model need not benefit a successor (Lowell et al., 2019). Work on proxy rankings, target-validation updates and compute-constrained selection already studies related transfer and resource questions (Wang et al., 2026; Jain et al., 2026; Yin & Rush, 2025). Our comparisons hold observed counts or selected identities fixed to ask which quantity changes. They do not establish a first transfer result or an end-to-end compute advantage.

3 SELECTION BEYOND THE CHOSEN GROUP MIX

Does selecting particular rows help beyond selecting their group counts? The count-control study compares each actual selection with ordinary random and eight complete redraws at its own counts. It uses fourteen constructions, eleven sources, seeds 45–52, four reference examples per class and a five-percent per-class acquisition budget. Five worst-group $R = 0$ tests, one per implementation, form the registered primary family.

The primary and secondary results answer different questions. All five worst-group residual point estimates are negative. Nominal intervals for the gap and unscaled-logistic implementations exclude zero, but the registered Holm-adjusted p-values range from 0.225 to 0.249. The total T and composition C intervals also span zero, so the count control does not explain away a clear total gain. Figure 1 separates the two endpoints, and Table 1 gives their exact summaries.

For overall accuracy, the residual is negative in all 55 source–implementation means. Equal-source means range from -2.04 to -1.56 points, with source sign-flip $p = 0.0009766$ for each implementation. These means share sources, selected sets and random draws: they are a correlated secondary pattern, not 55 independent confirmations. No registered secondary Holm family is added. This pattern concerns the examples chosen *within the observed counts*; it does not replace the worst-group primary or show that total acquisition harms accuracy.

What the control makes visible. The ordinary-random comparison and the residual differ because selection can change group mix. Acquired minority fractions, for example, differ between ordinary random and the scored rules (Table 29). The comparison does not establish that observed groups capture every useful property of a subset. It asks the narrower, reproducible question of whether

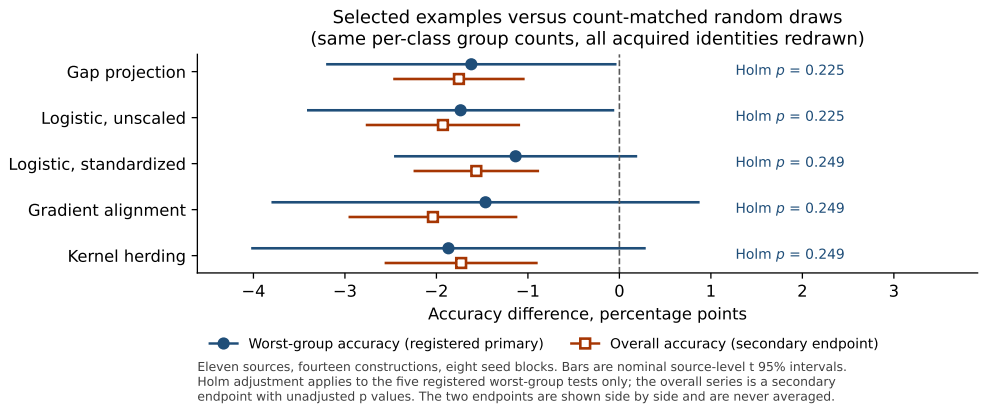

Figure 1: **Selected identities do not give a resolved worst-group advantage beyond observed counts.** Blue circles: registered worst-group residuals; orange squares: secondary overall residuals, in percentage points. Bars show nominal source-t 95% intervals. Table 1 gives exact values and registered verdicts.

Table 1: **Count-matched residual by selection rule.** Five implementations of four strategies; eleven sources ($G = 11$), fourteen constructions, eight seed blocks. Units: percentage points. Worst-group accuracy is the registered primary family: five sign-flip tests with one Holm adjustment. Overall accuracy is secondary; the endpoints are not averaged. Intervals are nominal source-level t 95% intervals, unadjusted for multiplicity.

	Worst-group accuracy (registered primary)			Overall accuracy (secondary)	
Implementation	R	95% CI	Holm p	R	95% CI
Gap projection	-1.62	[-3.21, -0.03]	0.225	-1.75	[-2.47, -1.04]
Logistic, unscaled	-1.74	[-3.41, -0.06]	0.225	-1.93	[-2.77, -1.09]
Logistic, standardized	-1.13	[-2.46, +0.19]	0.249	-1.56	[-2.25, -0.88]
Gradient alignment	-1.46	[-3.80, +0.88]	0.249	-2.04	[-2.96, -1.12]
Kernel herding	-1.87	[-4.02, +0.29]	0.249	-1.73	[-2.57, -0.89]

its identities outperform a complete redraw at those counts. Same-count random is therefore an evaluation control where group labels are available, not a deployable group-label-free selector.

Checks on the same evidence. A prospectively sealed sensitivity mask removes test rows whose feature vectors exactly match any same-cell train, pool or reference vector. Every group remains, and pooled worst-group R changes by less than 0.01 points. The check concerns exact feature equality, not all content or subject dependence. From the existing eight random draws, estimated conditional Monte Carlo standard errors are 0.075–0.082 points, compared with source-mean standard errors of 0.596–1.050 points. This estimates one noise component, not convergence. App. B retains every primary, secondary contrast, declared stratum and source-merging check.

The panel shows a consistent negative overall residual without a clear worst-group benefit. Neither endpoint alone says how the comparison would change with a different candidate supply or learner. We turn to those choices next.

4 Candidate supply and the learner change the comparison

The music study separates two questions that a single acquisition gain can hide. Does changing the supply improve the selected set? And, for the very same selected rows, does changing the learner alter the advantage over random acquisition?

One corpus, two supplies. The task classifies ten composers from 261 conservative work families, 979 performances and 3,868 symbolic-music segments in the retained MAESTRO corpus (Hawthorne

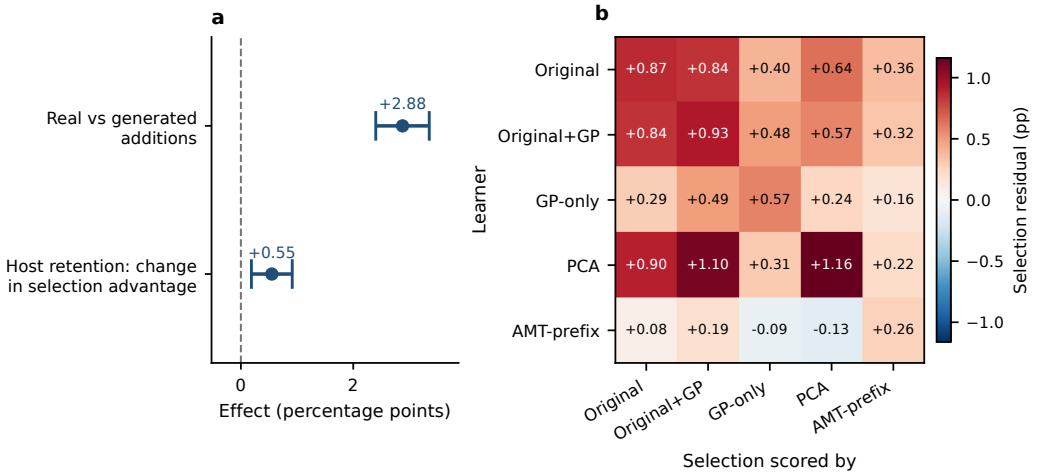

Figure 2: **Positive contrasts with different meanings.** (a) Supply: selected real versus generated additions under Original. Host retention: Original+GP versus GP-only selection advantage on identical mixed rows. Conditional 95% intervals use 128 paired randomizations. (b) Descriptive mixed-quota selected-minus-same-count-random residuals: learners in rows, scoring maps in columns. Original+GP minus GP-only in the Original column is P_2. AMT-prefix is a truncated, shared-checkpoint diagnostic. Units: percentage points.

et al., 2019). In each block, four families per composer supply training data, three supply a labelled reference, three supply real candidates and all remaining families supply test data. Related performances stay in one role. Every policy sees ten real candidate rows per class and ten generated rows. Each generated row is an infill of a training segment from one fixed Anticipatory Music Transformer checkpoint (Thickstun et al., 2024); both supplies receive the same whole-clip preprocessing.

Each policy adds six rows per class: six real, six generated, or three of each. All arms cap acquisitions at two per family, two per performance and one generated child per ancestor segment. A normalized class-conditional training-to-reference mean-gap score chooses the highest-scoring allowed set. Four uniform draws over complete allowed subsets provide the same-count control. Every row set is evaluated under each learner map. Accuracy averages segments within performances, performances within families, families within composers, then the ten composers equally. The 128 blocks randomize this fixed corpus; they are not 128 datasets.

Five maps, two roles. *Original* contains 109 whole-clip descriptors (the feature host). *Original+GP* appends twelve coordinates from our train-only adaptation of the genetic-programming construction in Xue et al. (2026); *GP-only* keeps exactly those coordinates. *PCA* is a twelve-dimensional train-only projection of the original features. *AMT-prefix* is a fixed 768-dimensional map from the generator's checkpoint, mean-pooled over the first 1,023 tokens. Every scaler, construction and projection is fitted on training data alone. The same logistic-head settings are used under every learner. The no-acquisition arm uses the base rows only.

The two registered comparisons use Original to score. The supply primary P_1 also uses Original to learn and subtracts the six-generated selection from the six-real selection. The host-retention primary P_2 uses the identical mixed selection and four random controls: it subtracts the selected-over-random advantage under GP-only from that under Original+GP. It changes the learner's features, not the ranking rule.

Supply improves the selected set, but random supply differs too. The real-source selection is $+2.877$ points above the generated-source selection, with conditional 95% interval $[+2.401, +3.352]$ and Holm-adjusted $p = 3.0 \times 10^{-22}$. Under the same learner, random acquisition reaches 33.642% with real rows and 31.039% with generated rows: a descriptive 2.603-point source gap before ranking. The two differences are similar in size, but their comparison is not a causal share or an improvement in ranking. Unseen real works and infills of training works change novelty, dependence on training

Pre-registered contrast	Effect (pp)	95% CI (pp)	Holm p
Real vs generated additions	+2.88	[+2.40, +3.35]	< 0.001
Host retention: change in selection advantage	+0.55	[+0.19, +0.91]	0.003

Table 2: **Registered primaries** (percentage points of hierarchical accuracy, df=127). Two-sided Student-t 95% intervals; Holm p over these two tests at α=0.05. Inference is conditional on the fixed corpus, cache and procedure.

Learner representation	No-acq. (%)	Mixed-3/3 sel. (%)	Sel. gain (pp)	Sel.−rnd. (pp)
Original	31.00	33.44	+2.44	+0.87
Original+GP	31.66	34.04	+2.38	+0.84
GP-only	26.72	27.39	+0.68	+0.29
PCA	29.63	31.26	+1.63	+0.90
AMT-prefix	20.72	21.65	+0.92	+0.08

Table 3: **A larger acquisition gain need not give higher final accuracy.** No-acquisition and mixed real-3/generated-3 base-scored selected accuracy (%); gains over no-acquisition and matched random (percentage points). Rows are identical across learners. Levels are descriptive; only the two registered contrasts carry a test. AMT-prefix is a truncated, shared-checkpoint diagnostic map.

data and origin together. P_1 compares these supply policies, not real and synthetic data in general or generator fidelity.

Retaining features changes selection value under the learner. Host retention gives a $+0.551$-point larger residual under Original+GP than GP-only, with interval $[+0.188, +0.914]$ and Holm-adjusted $p = 0.0032$. Both primaries pass the two-test Holm correction. For the same Original-scored mixed selection, the residuals are $+0.865$ under Original, $+0.838$ under Original+GP and $+0.287$ under GP-only. The primary compares the last two, not GP against Original. The twelve-dimensional PCA learner has a larger descriptive residual than GP-only ($+0.90$ versus $+0.29$); dimension alone does not explain the ordering. Geometry, information, effective regularization, capacity and headroom change together, so the result isolates no single mechanism.

A larger gain need not give the best final model. Table 3 retains Original and PCA beside the tested learners and reports accuracy levels as well as gains. The largest residual does not belong to the highest-accuracy map. In the mixed-quota matrix, each learner's largest mean residual lies on the matching scorer column. This is one descriptive matrix, not five independent confirmations, and the pattern does not hold for every learner at the other quotas. App. C retains all 75 residual cells and the two primary summaries.

AMT-prefix reaches 20.724% without acquisition, 21.648% on the Original-scored mixed selection and 21.832% when it both scores and learns. These are diagnostic levels, not a competitive whole-clip encoder baseline: longer clips are truncated, the checkpoint is shared with the generator, and its pretraining overlap with this corpus is uncertified. More broadly, the intervals concern this corpus, cache and procedure. Composer-specific means do not all have the same sign. Metadata families are conservative groupings, not proven note-level identities, and generated labels are inherited from their parents. The 60 acquired rows use supplied composer labels, not new annotations.

Thus the positive findings are about two separate interventions. Supply changes which rows are available; retaining the feature host changes how the learner benefits from an unchanged selection. Neither establishes that updating the ranking or matching representations will itself help.

5 Can guidance or representation changes improve selection?

The preceding comparisons motivate specific follow-up questions rather than a single improvement recipe. We consider additional guidance, direct use of its labels, a selection refresh after a feature change, and scorer–learner matching outside music.

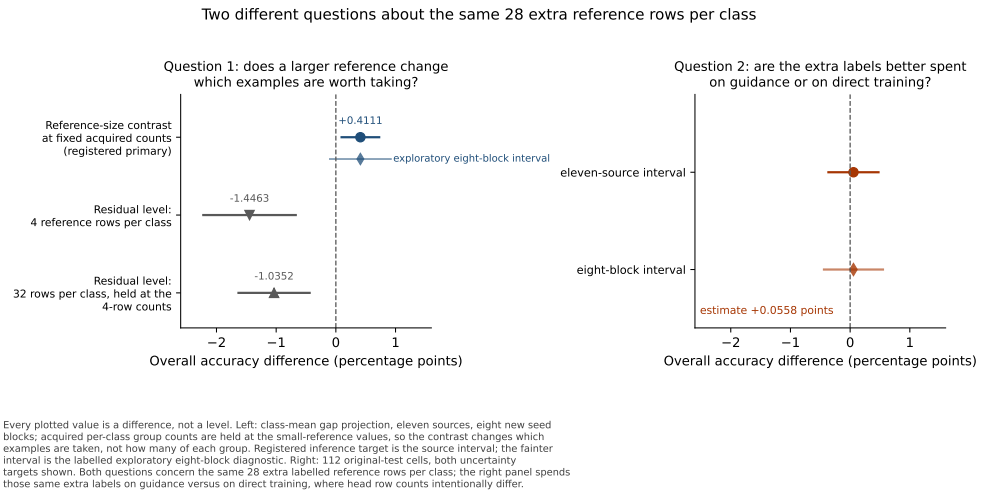

Figure 3: **More guidance and direct label use are different comparisons.** Left: reference-size contrast at fixed acquired group counts. The registered source interval is positive; the exploratory eight-block interval spans zero. Right: both guidance-versus-direct-training intervals span zero. Units: overall-accuracy percentage points. Source and block intervals describe different variation.

5.1 More reference labels at the same acquired counts

The reference-size study uses new blocks 53–60 on the same fourteen constructions and eleven sources. A four-row-per-class reference is nested within a balanced 32-row reference. Base training, candidate pool, test rows, budget and the smaller-reference selection's group counts are fixed within each pair. The extra 28 labels per class guide selection; they are not added to the head's training data.

Let S_4 be the smaller-reference selection with counts n_4, and $S_{32|n_4}$ take the larger-reference ranking at those same counts. Both share the same-count random baseline, so

$$\Delta_{\text{fixed}} = R_{32|n_4} - R_4 = V(S_{32|n_4}) - V(S_4). \tag{2}$$

This compares selected identities at fixed observed counts. Random-baseline Monte Carlo error cancels from the difference.

The registered gap/overall contrast is $+0.411$ points, with source-level 95% interval $[+0.096, +0.726]$ and sign-flip $p = 0.005859$. An exploratory interval over the same eight blocks is $[-0.099, +0.921]$, with nominal $p = 0.1015625$ and five-test Holm $p = 0.203125$. The first describes variation across the purposive source panel; the second describes repeated randomization on that fixed panel. Both are needed to read the result. The residual moves from -1.446 to -1.035 points: attenuation of a negative residual, not elimination. The unconstrained 32-reference total is $+0.827$ points with interval $[-1.960, +3.615]$, so improved fixed-count selection does not establish positive total value. App. D gives all implementations, levels, source means and sensitivities.

5.2 Use the extra labels to guide or to train?

The allocation study reuses the same 112 original-test cells, adding no source or seed. The guidance arm uses the 32-row reference to select candidates. The direct-use arm trains on the base data, the four-reference selection and the same 28 extra reference rows per class. Both arms have the same labels available, but the direct-use head has more training rows by design. This is a label-use comparison, not an equal-row-count effect.

For the gap rule, guidance minus direct use is $+0.0558$ points. The block 95% interval is $[-0.438, +0.550]$, the source interval is $[-0.364, +0.475]$, and the corresponding Holm-adjusted p-values are 0.75 and 0.87890625. The studies therefore support neither a general guidance-over-training recommendation nor a unique content, noise or shrinkage mechanism (App. E).

5.3 REFRESH THE SELECTION AFTER A FEATURE CHANGE?

An earlier music study asks whether changing features makes a previously chosen set stale. It selects again under Original+GP while holding the Original selection's class–origin counts fixed, then evaluates both sets under both learners. Its registered positive conjunction requires a fixed-count refresh gain D_1 above one point under Original+GP and a positive difference I between the learners' refresh gains.

That primary failed. The mean D_1 is -0.0647 points with conditional 95% interval $[-0.2434, +0.1139]$; I is -0.0536 with interval $[-0.2291, +0.1219]$. Only the D_1-specific registered bounded-null rule survives: its two-sided interval lies within the predeclared ± 1-point band for the mean effect on this fixed corpus and procedure. No interaction-equivalence rule was registered, and no successful co-optimizer follows. This differs from the positive host-retention comparison, which removes the original features from one learner instead of refreshing rows under a modest feature expansion. App. F retains the adaptive preparation history, levels, figure and all 65 endpoints.

5.4 MATCH THE SCORING MAP TO THE LEARNER OUTSIDE MUSIC?

The mixed music matrix suggests a further question: does matching scorer and learner help elsewhere? A separate registered study crosses two encoders on each of two reused sources, Waterbirds and CivilComments. Each encoder is projected to a 32-dimensional whitened PCA map fitted on 64 base training rows. Output dimension and head parameter count are matched, not information, geometry or effective regularization.

Selection uses quota-stratified greedy k-center (Gonzalez, 1985; Sener & Savarese, 2018) inside fixed class–attribute quotas. There is no reference set. This geometric selector differs from the reference-guided music score, so the study tests a related question in a different setting, not a replication of the music contrast. Every selected set is compared with four exact-quota random acquisitions. The single pre-specified contrast is the worst-group advantage of matched over crossed scorer–learner pairs, averaged across 24 blocks per source and then equally over sources. Within each paired difference the learner is unchanged, so its random baseline cancels exactly. Random draws describe each cell's residual; they do not drive the primary.

The contrast is $+0.265$ points with conditional 95% block-bootstrap interval $[-0.611, +1.114]$. The source means oppose each other, $+1.310$ on Waterbirds and -0.780 on CivilComments. These are descriptions of two fixed sources, not a between-dataset inference; no per-source interval is defined. The Waterbirds value comes almost entirely from its ResNet learner, while the CLIP-learner cells are descriptively close (Table 32).

Seven of the eight worst-group residual *cell means* are negative. Only Waterbirds with ResNet both scoring and learning is positive, at $+0.785$ points. Matching dimension does not itself make these selected sets useful relative to random. Distances are fitted from only 64 rows, and the smallest Waterbirds group supplies 16 acquisitions from 40 candidates. These are limits of the setting, not evidence of PCA degeneracy or an identified failure mechanism. Selection can also depend on budget (Hacohen et al., 2022), which this study does not vary.

The pairing has a source-specific provenance limit. Waterbirds alignment is conditional on a verified order-preserving chain from raw metadata through the extractor to packaged features, which lack self-authenticating row identifiers. CivilComments pairing is verified through its retained UID/text/extraction chain. App. H retains these contracts, every selected/random/residual cell and the conditional uncertainty recipe. This matching result does not extend the music learner comparison to the two outside-music sources.

6 WHAT TO REPORT, AND WHAT REMAINS OPEN

A data-selection gain is a comparison under a learner, not an intrinsic property of the rows. Count controls separate ordinary-random differences from residuals at observed counts, without resolving a worst-group benefit. Music supplies positive source and learner-feature comparisons; neither establishes that ranking caused the gain. More guidance attenuates a negative residual across sources,

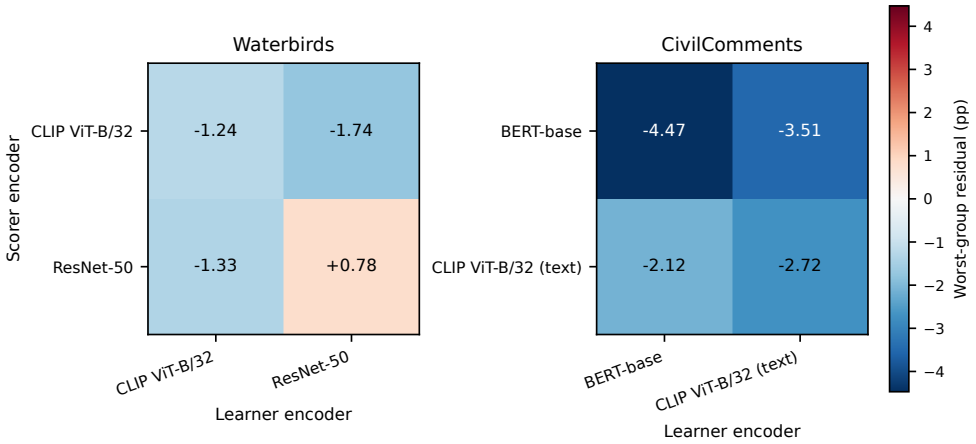

Figure 4: **Worst-group selection residual (percentage points), two-source representation crossing.** Cells show selected-minus-matched-random accuracy: learner encoders in columns, scorer encoders in rows, matched encoders on the diagonal. Both complete panels share a zero-centred scale. The equal-source primary is the diagonal-minus-crossed mean; the learner-specific random baseline cancels, so this mean is identical on residual and absolute scales.

but its repeated-block interval spans zero. Direct label use, selection refresh and outside-music matching do not establish their proposed benefits.

A compact reporting practice. Report candidate supply and budget, scoring and learner maps, group-count control and its supervision, and final accuracy alongside each gain. Ordinary random measures total selection value; same-count random asks about identities at the observed mix. For representation comparisons, state whether rows changed or only the learner did; for uncertainty, name sources or repeated draws as the unit.

Scope of the evidence. The source panel is fixed and purposive; encoders are nested within sources, and some tests are fixed. Counts do not capture all coverage or dependence. Cross-source results use one acquisition budget and two reference sizes, not scaling curves. Music uses one corpus, candidate cache and checkpoint; the outside-music study reuses sources and changes both selector and domain. Frozen maps and linear heads do not establish transfer to end-to-end representation learning.

Before crediting a rule with better example identities, state what was fixed and show the random comparison under its learner. Useful supply, features and ranking are different findings.

AI USE STATEMENT

We used generative AI for experiment design, implementation, data cleaning and reformatting, and simulated review. The Anticipatory Music Transformer generated music infills. We did not use AI for conceptual and mathematical formulations, hypotheses, result analysis and interpretation, and translation; proofs and qualitative or thematic analysis were not applicable. AI also assisted literature research, code, artifacts, and writing. We reviewed all AI-assisted work through source checks, code tests and independent reconstruction of saved numerical evidence. We take responsibility for the final text, methods, claims, citations and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–H document data processing, methods and inference. The code package provides implementations, saved evidence and reproduction commands. Its coverage ledger distinguishes prediction and aggregate replay from omitted raw-data-to-training steps.

REFERENCES

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision – ECCV 2018*, pp. 472–489. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01270-0_28.

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. ACM, 2019. doi: 10.1145/3308560.3317593.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 886–893. IEEE, 2005. doi: 10.1109/CVPR.2005.177.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128. ACM, 2019. doi: 10.1145/3287560.3287572.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-aware dataset selection with datamodels. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 12491–12526, 2024. URL https://proceedings.mlr.press/v235/engstrom24a.html.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *2013 IEEE International Conference on Computer Vision*, pp. 1657–1664. IEEE, 2013. doi: 10.1109/ICCV.2013.208.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning (ICML)*, 2022.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations (ICLR)*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, May 1994. doi: 10.1109/34.291440.

Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *PMLR*, pp. 336–351, 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Train on validation (ToV): Fast data selection with applications to fine-tuning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1c3d419b754cb4de0a67a453cb28d959-Paper-Conference.pdf.

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.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, April 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 11552–11579, 2023. doi: 10.52202/075280-0509. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/265bee74aee86df77e8e36d25e786ab5-Abstract-Conference.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. IEEE, 2017. doi: 10.1109/ICCV.2017.591.

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.

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. IEEE, 2015. doi: 10.1109/ICCV.2015.425.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. Practical obstacles to deploying active learning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 21–30, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1003. URL https://aclanthology.org/D19-1003/.

Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 223–232, June 2022. doi: 10.1109/CVPR52688.2022.00032.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.

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

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

John Thickstun, David Hall, Chris Donahue, and Percy Liang. Anticipatory music transformer. *Transactions on Machine Learning Research (TMLR)*, 2024.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394. IEEE, 2017. doi: 10.1109/CVPR.2017.572.

Jiachen (Tianhao) 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/file/c7d04ce5e71efdfc7a4a78f7161238a9-Paper-Conference.pdf.

Max Welling. Herding dynamical weights to learn. In *International Conference on Machine Learning (ICML)*, 2009.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chenhao Xue, Weitao Hu, Joyraj Chakraborty, Zhijin Guo, Kang Li, Tianyu Shi, Martin Reed, and Nikolaos Thomos. Constructing composite features for interpretable music-tagging. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13032–13036, 2026. doi: 10.1109/ICASSP55912.2026.11462876. URL https://arxiv.org/abs/2603.28644.

Junjie Oscar Yin and Alexander Rush. Compute-constrained data selection. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5acb720a361eecb34ee62d356859d246-Paper-Conference.pdf.

A SOURCE AND SELECTION CONTRACTS

Navigation. The appendices follow the main argument. Record names map to count control (M179v2), reference size (M183), label allocation (M184) and representation refresh (M185–M186). The source-by-representation music study and the outside-music crossing have separate protocols.

A.1 SOURCES AND ATTRIBUTES

Waterbirds (Sagawa et al., 2020) predicts bird type with background as the binary attribute. CelebA (Liu et al., 2015) predicts blond-hair status with the supplied binary sex attribute. The bias-in-bios (De-Arteaga et al., 2019) construction is nurse versus surgeon with the supplied binary gender code; MultiNLI (Williams et al., 2018) has three inference labels. Its derived negation-word grouping follows Sagawa et al. (2020). For CivilComments (Borkan et al., 2019), the frozen preprocessing uses `toxicity>0` as the binary label and the upstream `identity_any>0` flag as the attribute. That flag is not recomputed as the union of the eight separately retained identity columns. The two encoders share each source's label contract. These frozen preprocessing definitions, not an assumed alternative benchmark threshold, define the tasks.

The six added constructions and exact group counts are in Tables 4–5. Every binary source uses $g = 2y + a$; MultiNLI uses the same formula with three class labels. References are balanced by class and attribute. The original five sources label the less frequent class-group in the current pool as minority (ties select the lower group ID); the added-source loader takes the complement of the class's majority attribute in the frozen source split (majority ties select attribute 0). No definition uses observed test accuracy. The current controls preserve both groups' counts, so their pairing does not depend on which of the two is called minority.

The M179v2 seed block is 45–52. M183 uses 53–60 with a new reference/randomization namespace. The later seeds do not imply new datasets or globally unused raw examples. Exact group counts and row identities are retained for each cell. The base training set has 100 rows per class, and class pool budgets use $\min(\max(5, \mathrm{round}(.05n_c)), n_c - 1)$; the floor/cap are inactive in the frozen cells. The inherited pool cap is 3,000 per class. Test sets are paired across all arms within a cell.

A.2 SCORE DEFINITIONS

All score functions are class-conditional and access the labelled base training set, reference and candidate features, but not the candidate group labels. Larger scores are preferred.

Gap projection. For feature coordinate d, let $g_d = \bar{x}_{\mathrm{ref},d} - \bar{x}_{\mathrm{train},d}$ and $e_d = \sqrt{\widehat{\mathrm{Var}}(x_d)(1/n_{\mathrm{train}} + 1/n_{\mathrm{ref}}) + 10^{-12}}$, where the variance is computed on pooled class training/reference rows with the implemented population variance convention. The denoised gap is $\widetilde{g}_d = \mathrm{sign}(g_d)\max(0, |g_d| - 1.5e_d)$. The score is $(x - \bar{x}_{\mathrm{train}})^\top \widetilde{g}/\|\widetilde{g}\|$; a norm below 10^{-12} gives a zero primary score followed by the declared tie rule. Changing reference size can change both estimated direction and shrinkage; the experiment does not separate those mechanisms.

Logistic reference/pool discrimination. A logistic classifier distinguishes class-reference rows from class-candidate rows with $C = 1$. One implementation uses unscaled features and a 400-iteration cap; the other fits a `StandardScaler` on these discriminator inputs and uses a 1,000-iteration cap. The preference score is the predicted reference probability. Convergence warnings stop preparation or execution. These are specified feature-discriminator variants, not the original n-gram DSIR procedure or its importance-resampling functional.

Gradient alignment. A logistic surrogate is fitted on the standardized biased base training data with an 800-iteration cap. For a labelled candidate, its score is the cosine between its loss gradient with respect to the surrogate's weight matrix and the mean reference gradient, stabilized by adding 10^{-9} to the product of their norms in the denominator. The implementation uses the residual $(p - \mathrm{onehot}(y))$ outer product with standardized features; the bias gradient and feature-extractor updates are not included. This is a gradient-alignment proxy, not full TracIn/TRAK.

Kernel herding. Within each class, the RBF bandwidth is the square root of the median squared pairwise distance among the first $\min(n_c, 1024)$ candidate rows in frozen pool order. The greedy ordering maximizes mean kernel similarity to the reference minus accumulated similarity to previously selected points, divided by the next prefix size. The recorded first-step utility supplies its distinctness diagnostic; the final preference score encodes the total greedy order. This order is subsequently subjected to the same class/group selection controls as the other strategies.

Ordering and learning. Scores are rounded to 12 decimal places before a deterministic tie rule: squared distance from the class training centroid descending, then a SHA-256 row key ascending. M183 sorts selected pool indices into canonical order before head fitting, so equal subsets cannot differ merely by row ordering. The downstream head is independent of the surrogate used by the gradient score. Numerical threads are pinned to one; original and new outcome grids retain their own immutable code/registration hashes.

A.3 Reading the uncertainty

The empirical unit for panel inference is the source. Seeds and encoders are nested repetitions, not extra source units. Source-t and sign-flip calculations express their model assumptions; the sign-flip floor at 11 sources is $2/2048$, not the observed Holm-adjusted p. In M179v2 the five-test family is worst-group $R = 0$. In M183 the only primary is gap/overall Δ_{fixed}. Secondary nominal intervals and tests are not silently converted into additional primaries. The practical one-point band is a descriptive reference, not a retrospectively selected significance threshold or a uniform equivalence guarantee.

B Count-control study: complete results and checks

B.1 Design and decision rules

The count-control study (M179v2) uses fourteen constructions from eleven sources: bias-in-bios, MultiNLI, CelebA, Waterbirds, CivilComments, MNIST–USPS (LeCun et al., 1998; Hull, 1994), CIFAR-100 (Krizhevsky, 2009), OfficeHome (Venkateswara et al., 2017), TerraIncognita (Beery et al., 2018), PACS (Li et al., 2017) and VLCS (Fang et al., 2013). CelebA, Waterbirds and CivilComments each contribute two encoders. Seeds are 45–52, reference size is four per class and acquisition is five percent of the class pool. All $14 \times 8 \times 51 = 5,712$ arm records are present.

The local registration, `M179v2_PREREG.md`, includes amendments A19–A24 made before any outcome was fitted or inspected. They clarified inherited splits, made bottom-k secondary rather than an R gate, made the R decision symmetric, withdrew a defective wild-t helper, distinguished five implementations from four strategies, separated zero nulls from practical margins, and added integrity checks.

bias-in-bios, MultiNLI, CelebA and CivilComments use inherited seed-dependent, group-stratified test carve-outs of 300 rows per group, disjoint in row index from same-cell train, pool and reference rows. Waterbirds uses its full official test split; the six added constructions use their full fixed tests. The head is training-fitted standardization followed by logistic regression with $C = 1$ and convergence enforced. App. A defines labels, scores, budgets and ordering.

B.2 Additional source contracts

Tables 4 and 5 give the derived binary tasks and frozen group counts; these are not the original full multi-class tasks. The added-source loader preserves each class's frozen majority/minority proportions in train and pool up to rounding, and draws a two-per-attribute reference. For OfficeHome, the positive class is membership in this frozen original-label-ID set:

$$\left\{\begin{matrix} 3, 5, 6, 9, 10, 13, 14, 19, 21, 22, 24, 25, 28, 30, 31, 35, 37, \\ 39, 40, 47, 48, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 63 \end{matrix}\right\}.$$

No semantic category name is inferred from the IDs. MNIST–USPS uses HOG4, a histogram-of-oriented-gradients descriptor (Dalal & Triggs, 2005); CIFAR-100 and PACS use CLIP ViT-B/32; OfficeHome and TerraIncognita use ResNet-50 penultimate features with a fixed Gaussian projection from 2048 to 512 dimensions; VLCS uses public ResNet-50 features.

Table 4: The six added binary construction tasks, not their original full multi-class benchmark tasks. Groups are class crossed with the declared binary attribute. The first named attribute category is coded 1.

source	binary task	attribute 1 versus 0
MNIST–USPS	Digits 5–9 versus 0–4	USPS versus MNIST
CIFAR-100	Vehicles-1 versus people (cached coarse label)	Bicycle/bus or baby/boy versus other fine classes within the superclass
OfficeHome	Membership in the frozen `HOME_LABELS` set	Product/Real World versus Art/Clipart
TerraIncognita	Bobcat/cat/coyote/dog versus bird/opossum/rabbit/raccoon/squirrel	Locations 100/43 versus 38/46
PACS	Dog/elephant/giraffe/horse versus guitar/house; person excluded	Photo/art painting versus cartoon/sketch
VLCS	Person versus bird/car/chair/dog	Caltech/Pascal versus LabelMe/SUN

Table 5: Frozen source sizes and cell budgets for the added constructions. Each has 100 training examples per class and a four-per-class reference. Test counts are ordered by group $2y+a = 0, 1, 2, 3$. Budgets are class 0/class 1.

source	full rows	dimensions	cell pool	test group counts	budgets
MNIST–USPS	18476	441	6000	1250/1143/1250/975	150/150
CIFAR-100	6000	512	2800	450/300/450/300	70/70
OfficeHome	15588	512	6000	824/1136/873/1063	150/150
TerraIncognita	23115	512	6000	2582/1334/1158/703	150/150
PACS	8545	512	3930	277/236/1149/473	46/150
VLCS	9976	2048	4790	799/563/581/550	131/108

Table 6: Complete-set WG ranking heterogeneity and Monte Carlo resolution. I^2 is the registered descriptive Q-based statistic for this eleven-source panel, not the historical ten-family statistic. Conditional MC SE and source-mean SE are in percentage points. The MC diagnostic was computed after review from the existing eight draws, without new fits.

implementation	I^2 (%)	conditional MC SE	source-mean SE	first-four / last-four R
Gap	83.7	0.0752	0.7122	$-1.555/-1.682$
Logistic, unscaled	87.0	0.0789	0.7536	$-1.743/-1.728$
Logistic, standardized	75.6	0.0816	0.5964	$-1.079/-1.189$
Gradient alignment	83.1	0.0823	1.0502	$-1.536/-1.391$
Herding	89.2	0.0783	0.9677	$-1.921/-1.816$

B.3 Complete outcomes

Each implementation has one actual subset and eight complete independent redraws at its class–group counts. One ordinary-random acquisition is shared across implementations. Both majority and minority identities are redrawn. The mean of eight *accuracies*, not an ensemble prediction, enters C and R. Source means average encoders within seed, then seeds within source. Bottom-k, denoted A in the complete tables, changes composition as well as identity and does not gate R.

All five worst-group $R = 0$ source sign-flip tests have Holm-corrected $p = 0.2246$–0.2490; no interval lies within the practical ± 0.01 accuracy margin. Nominal source-t intervals and adjusted sign-flip decisions need not agree. Every total and composition interval also spans zero. Overall residuals are negative in all 55 correlated source–implementation means, with nominal sign-flip $p = 2/2^{11} = 0.0009766$ for each implementation; no secondary Holm family was registered. The saved prediction transitions have source-balanced fixed rates 0.0694–0.0715 and broken rates 0.0868–0.0918. Their difference reproduces overall accuracy, not necessarily minimum-group differences. These are retrospective categories, not acquisition features or an identified mechanism.

Table 7: **Complete contrasts, worst-group accuracy, scope `primary_g11`.** Percentage points. T is actual minus ordinary random, C is count-matched random minus ordinary random, R is actual minus count-matched random, and $T = C + R$ holds exactly because the same mean matched accuracy enters all three. A is the bottom-ranked anti diagnostic, which also changes composition and therefore does not test identity alone. Sign-flip p values are nominal; the Holm column is populated only for the one registered worst-group R family.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	11	+1.978	[-5.030, +8.987]	0.5439453125	–
T	Logistic, unscaled	11	+2.407	[-3.523, +8.337]	0.38671875	–
T	Logistic, standardized	11	+1.415	[-2.751, +5.581]	0.470703125	–
T	Gradient alignment	11	+0.940	[-4.961, +6.841]	0.7080078125	–
T	Kernel herding	11	+2.388	[-3.552, +8.328]	0.4130859375	–
C	Gap projection	11	+3.597	[-2.543, +9.736]	0.26171875	–
C	Logistic, unscaled	11	+4.142	[-0.896, +9.181]	0.189453125	–
C	Logistic, standardized	11	+2.549	[-0.840, +5.939]	0.18359375	–
C	Gradient alignment	11	+2.403	[-3.800, +8.606]	0.396484375	–
C	Kernel herding	11	+4.256	[-0.378, +8.891]	0.0185546875	–
R	Gap projection	11	-1.619	[-3.205, -0.032]	0.052734375	0.225
R	Logistic, unscaled	11	-1.735	[-3.415, -0.056]	0.044921875	0.225
R	Logistic, standardized	11	-1.134	[-2.463, +0.194]	0.087890625	0.249
R	Gradient alignment	11	-1.463	[-3.803, +0.877]	0.193359375	0.249
R	Kernel herding	11	-1.868	[-4.025, +0.288]	0.0830078125	0.249
A	Gap projection	11	-5.891	[-9.362, -2.420]	0.0009765625	–
A	Logistic, unscaled	11	-5.240	[-9.319, -1.162]	0.0009765625	–
A	Logistic, standardized	11	-4.120	[-6.700, -1.540]	0.0029296875	–
A	Gradient alignment	11	-8.601	[-13.329, -3.872]	0.0009765625	–
A	Kernel herding	11	-5.861	[-9.611, -2.111]	0.0009765625	–

Table 8: **Complete contrasts, overall accuracy, scope `secondary_ov`.** Secondary percentage-point comparisons; definitions are as in Table 7. Intervals and sign-flip p-values are nominal, without a secondary Holm family.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	11	-0.119	[-2.723, +2.484]	0.9375	–
T	Logistic, unscaled	11	-0.180	[-2.715, +2.354]	0.8994140625	–
T	Logistic, standardized	11	-0.408	[-1.979, +1.162]	0.57421875	–
T	Gradient alignment	11	-1.172	[-3.425, +1.080]	0.271484375	–
T	Kernel herding	11	-0.089	[-2.329, +2.151]	0.9248046875	–
C	Gap projection	11	+1.634	[-0.683, +3.951]	0.16015625	–
C	Logistic, unscaled	11	+1.749	[-0.464, +3.961]	0.05078125	–
C	Logistic, standardized	11	+1.156	[-0.165, +2.476]	0.0361328125	–
C	Gradient alignment	11	+0.866	[-1.381, +3.113]	0.435546875	–
C	Kernel herding	11	+1.641	[-0.300, +3.582]	0.0361328125	–
R	Gap projection	11	-1.754	[-2.472, -1.036]	0.0009765625	–
R	Logistic, unscaled	11	-1.929	[-2.772, -1.085]	0.0009765625	–
R	Logistic, standardized	11	-1.564	[-2.250, -0.878]	0.0009765625	–
R	Gradient alignment	11	-2.038	[-2.961, -1.115]	0.0009765625	–
R	Kernel herding	11	-1.730	[-2.566, -0.893]	0.0009765625	–
A	Gap projection	11	-3.168	[-4.802, -1.533]	0.0009765625	–
A	Logistic, unscaled	11	-2.780	[-4.552, -1.008]	0.0009765625	–
A	Logistic, standardized	11	-2.291	[-3.530, -1.052]	0.0029296875	–
A	Gradient alignment	11	-3.988	[-5.805, -2.170]	0.0009765625	–
A	Kernel herding	11	-3.098	[-4.707, -1.489]	0.0009765625	–

Table 9: **Complete contrasts, worst-group accuracy, scope `stratum_S`.** The seven-source stratum was declared before outcomes. Percentage points and definitions as in Table 7; all intervals and tests here are secondary and nominal.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	7	+5.777	[-4.775, +16.330]	0.328125	–
T	Logistic, unscaled	7	+5.818	[-2.946, +14.582]	0.1875	–
T	Logistic, standardized	7	+3.968	[-1.943, +9.879]	0.1875	–
T	Gradient alignment	7	+3.671	[-5.584, +12.926]	0.4375	–
T	Kernel herding	7	+5.788	[-3.013, +14.589]	0.1875	–
C	Gap projection	7	+6.606	[-3.032, +16.244]	0.140625	–
C	Logistic, unscaled	7	+6.634	[-1.296, +14.564]	0.171875	–
C	Logistic, standardized	7	+4.487	[-0.547, +9.521]	0.09375	–
C	Gradient alignment	7	+4.836	[-5.233, +14.904]	0.28125	–
C	Kernel herding	7	+6.738	[-0.391, +13.866]	0.015625	–
R	Gap projection	7	-0.828	[-2.957, +1.301]	0.4375	–
R	Logistic, unscaled	7	-0.816	[-2.973, +1.341]	0.390625	–
R	Logistic, standardized	7	-0.519	[-2.289, +1.251]	0.4375	–
R	Gradient alignment	7	-1.165	[-4.077, +1.748]	0.375	–
R	Kernel herding	7	-0.950	[-4.131, +2.232]	0.484375	–
A	Gap projection	7	-6.803	[-12.521, -1.085]	0.015625	–
A	Logistic, unscaled	7	-6.529	[-13.314, +0.257]	0.015625	–
A	Logistic, standardized	7	-4.679	[-8.941, -0.418]	0.03125	–
A	Gradient alignment	7	-9.605	[-17.086, -2.125]	0.015625	–
A	Kernel herding	7	-7.231	[-13.302, -1.160]	0.015625	–

Table 10: **Complete contrasts, overall accuracy, scope `stratum_S`.** Secondary nominal comparisons for the declared seven-source stratum. Percentage points and definitions as in Table 7.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	7	+1.171	[-2.804, +5.146]	0.625	–
T	Logistic, unscaled	7	+1.074	[-2.784, +4.932]	0.609375	–
T	Logistic, standardized	7	+0.483	[-1.739, +2.704]	0.65625	–
T	Gradient alignment	7	+0.015	[-3.161, +3.192]	0.984375	–
T	Kernel herding	7	+1.037	[-2.358, +4.432]	0.578125	–
C	Gap projection	7	+2.681	[-1.049, +6.410]	0.109375	–
C	Logistic, unscaled	7	+2.781	[-0.753, +6.316]	0.015625	–
C	Logistic, standardized	7	+1.843	[-0.202, +3.888]	0.015625	–
C	Gradient alignment	7	+1.919	[-1.534, +5.372]	0.15625	–
C	Kernel herding	7	+2.636	[-0.384, +5.657]	0.015625	–
R	Gap projection	7	-1.510	[-2.243, -0.777]	0.015625	–
R	Logistic, unscaled	7	-1.707	[-2.684, -0.730]	0.015625	–
R	Logistic, standardized	7	-1.360	[-2.163, -0.558]	0.015625	–
R	Gradient alignment	7	-1.904	[-2.787, -1.020]	0.015625	–
R	Kernel herding	7	-1.599	[-2.707, -0.491]	0.015625	–
A	Gap projection	7	-3.213	[-5.484, -0.942]	0.015625	–
A	Logistic, unscaled	7	-3.020	[-5.770, -0.271]	0.015625	–
A	Logistic, standardized	7	-2.302	[-4.240, -0.364]	0.03125	–
A	Gradient alignment	7	-4.063	[-6.774, -1.353]	0.015625	–
A	Kernel herding	7	-3.110	[-5.425, -0.794]	0.015625	–

B.4 Sensitivity and dependence

Before outcomes, 112 test masks were sealed against frozen inputs. A mask excludes a test row whose canonical feature vector equals a same-cell training, pool or reference vector, regardless of the selected arm. Across repeated test-row evaluations, $792/318,456$ rows are excluded in 35 cells; OfficeHome contributes 747 across eight seeds. Every group remains, with at least 236 rows. Pooled worst-group R changes by at most 0.000094, or 0.0094 points, preserving signs and verdicts. The saved `endpoint_comparison.csv` contains paired endpoint changes. Exact feature equality

Table 11: **Complete contrasts, worst-group accuracy, scope `stratum_F`.** The four-source stratum was declared before outcomes. Percentage points and definitions as in Table 7; all intervals and tests here are secondary and nominal.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	4	-4.671	[-9.644, +0.303]	0.125	–
T	Logistic, unscaled	4	-3.562	[-7.280, +0.157]	0.125	–
T	Logistic, standardized	4	-3.053	[-6.488, +0.383]	0.125	–
T	Gradient alignment	4	-3.840	[-8.624, +0.945]	0.125	–
T	Kernel herding	4	-3.562	[-7.210, +0.085]	0.125	–
C	Gap projection	4	-1.669	[-4.457, +1.118]	0.125	–
C	Logistic, unscaled	4	-0.217	[-1.450, +1.015]	0.875	–
C	Logistic, standardized	4	-0.841	[-2.797, +1.116]	0.5	–
C	Gradient alignment	4	-1.854	[-6.679, +2.971]	0.375	–
C	Kernel herding	4	-0.086	[-0.957, +0.786]	1.000	–
R	Gap projection	4	-3.002	[-6.181, +0.178]	0.125	–
R	Logistic, unscaled	4	-3.344	[-6.726, +0.037]	0.125	–
R	Logistic, standardized	4	-2.212	[-5.094, +0.671]	0.125	–
R	Gradient alignment	4	-1.986	[-9.108, +5.137]	0.375	–
R	Kernel herding	4	-3.477	[-7.125, +0.172]	0.125	–
A	Gap projection	4	-4.294	[-8.563, -0.025]	0.125	–
A	Logistic, unscaled	4	-2.985	[-6.375, +0.404]	0.125	–
A	Logistic, standardized	4	-3.141	[-6.585, +0.303]	0.125	–
A	Gradient alignment	4	-6.843	[-15.237, +1.551]	0.125	–
A	Kernel herding	4	-3.465	[-7.243, +0.314]	0.125	–

Table 12: **Complete contrasts, overall accuracy, scope `stratum_F`.** Secondary nominal comparisons for the declared four-source stratum. Percentage points and definitions as in Table 7.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	4	-2.378	[-4.854, +0.099]	0.125	–
T	Logistic, unscaled	4	-2.375	[-4.911, +0.160]	0.125	–
T	Logistic, standardized	4	-1.968	[-3.969, +0.033]	0.125	–
T	Gradient alignment	4	-3.250	[-6.746, +0.245]	0.125	–
T	Kernel herding	4	-2.059	[-4.300, +0.181]	0.125	–
C	Gap projection	4	-0.197	[-0.710, +0.317]	0.375	–
C	Logistic, unscaled	4	-0.059	[-0.582, +0.464]	0.625	–
C	Logistic, standardized	4	-0.048	[-0.470, +0.375]	0.625	–
C	Gradient alignment	4	-0.977	[-3.391, +1.437]	0.375	–
C	Kernel herding	4	-0.101	[-0.680, +0.477]	0.625	–
R	Gap projection	4	-2.181	[-4.526, +0.164]	0.125	–
R	Logistic, unscaled	4	-2.316	[-4.935, +0.302]	0.125	–
R	Logistic, standardized	4	-1.920	[-3.999, +0.158]	0.125	–
R	Gradient alignment	4	-2.273	[-5.591, +1.045]	0.125	–
R	Kernel herding	4	-1.958	[-4.312, +0.395]	0.125	–
A	Gap projection	4	-3.089	[-7.491, +1.313]	0.125	–
A	Logistic, unscaled	4	-2.359	[-5.968, +1.250]	0.125	–
A	Logistic, standardized	4	-2.272	[-4.817, +0.274]	0.125	–
A	Gradient alignment	4	-3.855	[-8.119, +0.408]	0.125	–
A	Kernel herding	4	-3.077	[-7.159, +1.005]	0.125	–

does not establish raw-image identity or cover near duplicates, subject dependence or differently encoded content.

Two secondary strata were defined from earlier random-dose results, not selected from these outcomes. S contains bias-in-bios, CelebA, CivilComments, MultiNLI, Waterbirds, MNIST–USPS and PACS; F contains CIFAR-100, OfficeHome, TerraIncognita and VLCS. Their complete T, C, R, A results appear in Tables 9–12; every worst-group R interval spans zero. The sign-flip floors, not observed p-values, are $2/2^7 = 0.015625$ and $2/2^4 = 0.125$. The mandatory PACS/VLCS-merged sensitivity

Table 13: **Complete contrasts, worst-group accuracy, scope `pacs_vlcs_merged`.** The mandatory source-merging sensitivity has ten units and does not replace the primary. Percentage points and definitions as in Table 7; tests are nominal.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	10	+2.515	[-5.170, +10.200]	0.4921875	–
T	Logistic, unscaled	10	+2.874	[-3.643, +9.390]	0.3515625	–
T	Logistic, standardized	10	+1.744	[-2.829, +6.317]	0.41796875	–
T	Gradient alignment	10	+1.428	[-4.968, +7.824]	0.603515625	–
T	Kernel herding	10	+2.876	[-3.642, +9.394]	0.37890625	–
C	Gap projection	10	+4.066	[-2.701, +10.833]	0.25390625	–
C	Logistic, unscaled	10	+4.515	[-1.063, +10.092]	0.203125	–
C	Logistic, standardized	10	+2.907	[-0.770, +6.584]	0.169921875	–
C	Gradient alignment	10	+2.703	[-4.216, +9.623]	0.39453125	–
C	Kernel herding	10	+4.701	[-0.376, +9.777]	0.009765625	–
R	Gap projection	10	-1.551	[-3.267, +0.166]	0.07421875	–
R	Logistic, unscaled	10	-1.641	[-3.380, +0.098]	0.060546875	–
R	Logistic, standardized	10	-1.163	[-2.647, +0.321]	0.107421875	–
R	Gradient alignment	10	-1.275	[-3.646, +1.096]	0.25	–
R	Kernel herding	10	-1.825	[-4.198, +0.549]	0.1171875	–
A	Gap projection	10	-6.139	[-9.886, -2.392]	0.001953125	–
A	Logistic, unscaled	10	-5.492	[-9.947, -1.037]	0.001953125	–
A	Logistic, standardized	10	-4.386	[-7.153, -1.619]	0.001953125	–
A	Gradient alignment	10	-8.780	[-13.627, -3.933]	0.001953125	–
A	Kernel herding	10	-6.113	[-10.192, -2.035]	0.001953125	–

Table 14: **Complete contrasts, overall accuracy, scope `pacs_vlcs_merged`.** Secondary nominal source-merging comparisons. Percentage points and definitions as in Table 7; no Holm family is added.

Contrast	Implementation	G	estimate	95% CI	sign-flip p	Holm p
T	Gap projection	10	+0.057	[-2.773, +2.887]	0.97265625	–
T	Logistic, unscaled	10	-0.029	[-2.797, +2.739]	0.986328125	–
T	Logistic, standardized	10	-0.321	[-2.027, +1.385]	0.6796875	–
T	Gradient alignment	10	-1.014	[-3.347, +1.318]	0.357421875	–
T	Kernel herding	10	+0.058	[-2.386, +2.502]	0.984375	–
C	Gap projection	10	+1.822	[-0.733, +4.377]	0.150390625	–
C	Logistic, unscaled	10	+1.943	[-0.491, +4.376]	0.0234375	–
C	Logistic, standardized	10	+1.288	[-0.154, +2.731]	0.0234375	–
C	Gradient alignment	10	+1.005	[-1.487, +3.497]	0.359375	–
C	Kernel herding	10	+1.823	[-0.304, +3.951]	0.0234375	–
R	Gap projection	10	-1.765	[-2.429, -1.101]	0.001953125	–
R	Logistic, unscaled	10	-1.972	[-2.805, -1.138]	0.001953125	–
R	Logistic, standardized	10	-1.609	[-2.305, -0.914]	0.001953125	–
R	Gradient alignment	10	-2.019	[-2.764, -1.275]	0.001953125	–
R	Kernel herding	10	-1.765	[-2.628, -0.902]	0.001953125	–
A	Gap projection	10	-3.354	[-5.091, -1.617]	0.001953125	–
A	Logistic, unscaled	10	-2.957	[-4.874, -1.040]	0.001953125	–
A	Logistic, standardized	10	-2.395	[-3.688, -1.102]	0.001953125	–
A	Gradient alignment	10	-4.160	[-6.038, -2.281]	0.001953125	–
A	Kernel herding	10	-3.236	[-4.945, -1.527]	0.001953125	–

has ten units: worst-group R remains negative, with zero-spanning source intervals and sign-flip $p = 0.0605$–0.2500. It upgrades no verdict.

Monte Carlo resolution. The exploratory diagnostic uses only the existing eight redraws. If s_c^2 is their accuracy variance in cell c and $w_c = 1/(11n_s)$, where source s has n_s construction–seed cells, the estimated conditional Monte Carlo variance is $\sum_c w_c^2 s_c^2/8$. It assumes independent redraws conditional on the cell and excludes source, reference, train and test sampling uncertainty. Table 6

gives the standard errors, both four-draw halves and panel-specific I^2. Both halves remain negative; this is not a convergence claim or a new primary test.

Evidence checks. Two single-threaded CPU workers completed the grid. Independent reconstruction checked every stored accuracy, primary source summary, grid key, checksum, input binding, paired count and prediction label, including the exact $T = C + R$ relation. This verifies the saved evidence, not independence of source effects.

C CANDIDATE SUPPLY AND LEARNER FEATURES IN MUSIC

C.1 CORPUS, ROLES AND PRE-OUTCOME CHOICES

§4 defines the ten-composer corpus and role assignment. Non-test performances are capped at two per family; every eligible segment of a retained performance is kept. Related performances never straddle roles. The 128 seeded blocks use stage `fuller_b_v1` with the inherited pure seed function and its `M185|` prefix. The new stage token supplies distinct streams; no earlier lifecycle or validation assignment was rerun.

Two choices changed before outcomes. The block count rose from the planning default of 32 to 128 using training-only and synthetic runtimes and outcome-blind assumed-SD precision scenarios, not measured power or an observed effect. The fifth representation was included because its encoder assets were verified within the available wall-clock before freezing all maps together. It was not added after a four-map result. This is an adaptive follow-up on a studied corpus, not new-corpus confirmation.

C.2 CANDIDATES AND COMPLETE-SET CONTROLS

Each class offers ten real rows and ten generated rows from capped training parents, with one cached child per ancestor chosen by a frozen random rule. Rosters first take two distinct rows or ancestors from every source family, then fill to ten without replacement. All policies share the roster. Each adds six rows per class under real/generated quotas 6/0, 3/3 or 0/6, with caps of two per family, two per performance and one generated row per ancestor. All quotas were feasible in all 1,280 class–block combinations without relaxing a cap.

For each quota, four independent uniform draws over complete allowed subsets are retained, including overlaps or coincidences. Selection maximizes the plain normalized class-conditional mean-gap score in the scoring map, using exact float64 sums and lexicographic clip-identifier ties without score rounding. Matching class and source counts does not match all latent properties of the rows.

C.3 FEATURE FITTING AND LEARNERS

The 109 Original descriptors include pitch range and entropy, key-relative pitch-class histograms, melodic intervals, onset and duration distributions, polyphony, instrumentation and velocity. They are a project-specific whole-clip bank, not jSymbolic. Codes M0–M4 denote Original, Original+GP, GP-only, PCA and AMT-prefix. Original has training-weighted scaling and clipping. Original+GP appends the twelve coordinates used alone by GP-only. PCA is a training-weighted twelve-dimensional projection of the host with training-only output scaling and clipping. The encoder scaler is fitted and saved during preparation, then only loaded at evaluation.

The GP recipe is fixed: population 400, ten generations, four outputs for each of three training LDA axes. Duplicate constructed coordinates and constant base columns use the existing unit-scale convention. Exactly constant constructed training features, degenerate PCA directions and nonfinite coordinates are invalid, without replacement or search. No program was changed in response to an outcome. AMT-prefix mean-pools the first 1,023 tokens; 33.95% of real-context clips and 37.34% of infills in the returned bank are truncated. It shares the generator checkpoint, whose pretraining overlap is uncertified.

All heads use logistic regression with $C = 100$, total sample weight 1, training/acquisition masses $2/3$ and $1/3$, tolerance 10^{-6} and a 2,000-iteration cap. The no-acquisition arm uses only base rows. Failed fits are not silently retried; the largest observed iteration count was 300. Per block there

are 221–280 training rows (mean 257.12), 166–217 reference rows (mean 194.99), exactly 200 candidates and 60 acquisitions, and 1,967–2,645 test rows (mean 2,353.66). These are the resources per randomization, not the corpus totals.

C.4 PRIMARIES AND COMPLETE RESIDUAL GRID

Hierarchical accuracy and the two paired primaries are defined in §4. Inference uses approximate Student-t intervals with 127 degrees of freedom, conditional on this corpus, cache and procedure. Tests are two-sided with Holm correction over exactly P_1, P_2 at $\alpha = 0.05$; intervals are per-contrast, not simultaneous. No inferential Original+GP-versus-Original contrast was registered.

Each block has 15 selected sets, 12 random sets and one no-acquisition set, crossed with five learners: 140 cells per block and 17,920 total. All other quotas, crossings, levels and random-source comparisons are secondary. Table 15 preserves all 75 residual cells. The machine-readable study tables retain the selected and random absolute levels for every cell.

Table 15: **Full learner×scorer selection residual crossing, all source quotas (percentage points).** Each row is the selection *scored by* the named map minus the mean of its four count-matched random draws (paired per block, averaged). Every cell is retained. Descriptive: no test, no starred cell and no best-cell selection. Random controls are count-matched on class and source only.

Quota	Learner	Selection scored by	Residual (pp)
Real 6 / Gen 0	Original	Original	+0.47
	Original	Original+GP	+0.37
	Original	GP-only	+0.12
	Original	PCA	+0.09
	Original	AMT-prefix	+0.21
	Original+GP	Original	+0.47
	Original+GP	Original+GP	+0.44
	Original+GP	GP-only	+0.06
	Original+GP	PCA	+0.24
	Original+GP	AMT-prefix	+0.21
	GP-only	Original	+0.24
	GP-only	Original+GP	+0.32
	GP-only	GP-only	+0.44
	GP-only	PCA	+0.12
	GP-only	AMT-prefix	+0.04
	PCA	Original	+0.41
	PCA	Original+GP	+0.41
	PCA	GP-only	+0.06
	PCA	PCA	+0.41
	PCA	AMT-prefix	+0.00
	AMT-prefix	Original	+0.07
	AMT-prefix	Original+GP	+0.05
	AMT-prefix	GP-only	-0.04
	AMT-prefix	PCA	-0.04
	AMT-prefix	AMT-prefix	+0.13
Real 3 / Gen 3	Original	Original	+0.87
	Original	Original+GP	+0.84
	Original	GP-only	+0.40
	Original	PCA	+0.64
	Original	AMT-prefix	+0.36
	Original+GP	Original	+0.84
	Original+GP	Original+GP	+0.93
	Original+GP	GP-only	+0.48
	Original+GP	PCA	+0.57
	Original+GP	AMT-prefix	+0.32
	GP-only	Original	+0.29
	GP-only	Original+GP	+0.49
	GP-only	GP-only	+0.57
	GP-only	PCA	+0.24
	GP-only	AMT-prefix	+0.16

Quota	Learner	Selection scored by	Residual (pp)
	PCA	Original	+0.90
	PCA	Original+GP	+1.10
	PCA	GP-only	+0.31
	PCA	PCA	+1.16
	PCA	AMT-prefix	+0.22
	AMT-prefix	Original	+0.08
	AMT-prefix	Original+GP	+0.19
	AMT-prefix	GP-only	-0.09
	AMT-prefix	PCA	-0.13
	AMT-prefix	AMT-prefix	+0.26
Real 0 / Gen 6	Original	Original	+0.19
	Original	Original+GP	+0.16
	Original	GP-only	-0.00
	Original	PCA	+0.16
	Original	AMT-prefix	-0.09
	Original+GP	Original	+0.27
	Original+GP	Original+GP	+0.21
	Original+GP	GP-only	+0.12
	Original+GP	PCA	+0.14
	Original+GP	AMT-prefix	-0.09
	GP-only	Original	+0.08
	GP-only	Original+GP	+0.15
	GP-only	GP-only	+0.51
	GP-only	PCA	+0.07
	GP-only	AMT-prefix	+0.06
	PCA	Original	+0.48
	PCA	Original+GP	+0.42
	PCA	GP-only	+0.04
	PCA	PCA	+0.58
	PCA	AMT-prefix	+0.12
	AMT-prefix	Original	-0.06
	AMT-prefix	Original+GP	-0.09
	AMT-prefix	GP-only	-0.07
	AMT-prefix	PCA	-0.03
	AMT-prefix	AMT-prefix	+0.10

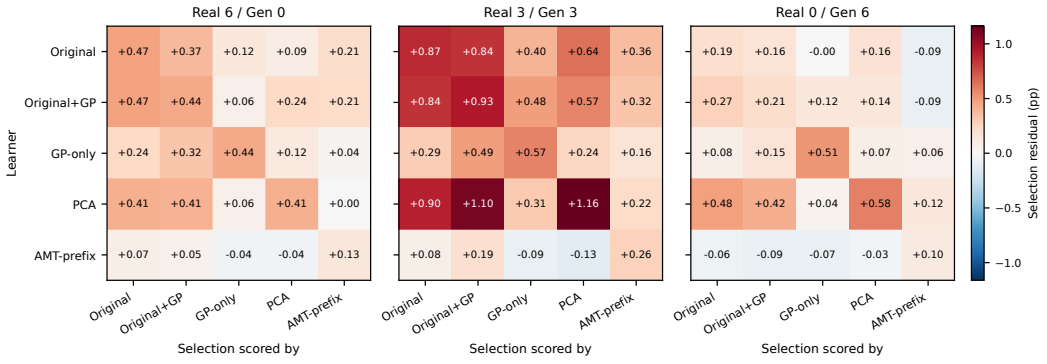

Figure 5: **All three source quotas, with every scorer–learner cell.** Real-only, mixed and generated-only panels share a scale. Rows are learners and columns are scorers; a residual subtracts the mean of four same-count random draws, paired per block. These are descriptive percentage-point differences. Exact displayed values are in Table 15.

C.5 Class spread and evidence checks

A post-outcome breakdown of saved predictions by composer reproduces both primaries within 10^{-12} points. For supply, nine class means are positive and one is -1.72 points. For host retention, eight

are positive and two are -0.67 and -0.17. Leaving out one class gives point estimates of $+2.459$ to $+3.388$ for supply and $+0.296$ to $+0.686$ for retention. This descriptive reweighting supplies no new test, interval or independent replication.

The run completed 384 construction-axis fits, 128 LDA fits, 384 descriptor-map scalers, 128 encoder scalers, 3,584 sealed sets and 17,920 heads, with zero failures and input reuses. Peak stage working set was 668,090,368 bytes. Receipts total 3,979.328 CPU seconds and 4,643.118 controller wall seconds; GPU encoder extraction took 2,682 seconds and job-to-merge time was 3,232 seconds. These costs exclude the generator, cache creation and earlier research.

An independent checker importing no producer modules reconstructed every cell and the registered inference, with maximum per-cell discrepancy 4.44×10^{-16}. Frozen-input, result and parent reconstruction digest prefixes are respectively `5a83d9946658`, `5ee56ec726df` and `230634e60b0a`. Per-block contrasts, levels, gains and residuals accompany the exporter manifest and its hashes.

D Reference-size study: complete results and checks

D.1 Design and primary comparison

M183 was specified after the M179v2 secondary overall-accuracy pattern and before any M183 outcome fit. The registration names gap/overall fixed-composition moderation as its sole primary; it does not retroactively change M179v2's five worst-group ranking tests. The practical ± 0.01 accuracy band is descriptive. Evidence for a positive effect requires both a positive lower source-t95% bound and a two-sided source sign-flip $p \leq .05$; evidence for a negative effect requires the corresponding negative conditions.

Fourteen constructions cover the same eleven source units, with new split-seed block 53–60 and a new reference/matching namespace. The designated reference frame is disjoint in row indices from the base training, candidate pool and test data. Within each class-group, one permutation independent of reference size supplies two rows to the $m = 4$ reference and 16 to $m = 32$. Both are balanced, sampled without replacement, and exactly nested. At least 16 eligible reference rows per group are required. These are new draws within existing datasets, not newly collected or globally unseen test data.

D.2 Arms and fixed-count identification

For each cell there is one ordinary-random acquisition shared by every implementation and reference size. For each of five implementations there are actual4, actual32 and yoked32 selections, eight matched4 draws at actual4's counts and eight matched32 draws at actual32's counts. This gives $1 + 5(3 + 8 + 8) = 96$ arms and $14 \times 8 \times 96 = 10{,}752$ records.

The larger-reference yoked selection takes its highest-ranked rows within each stratum at the smaller reference's exact acquired count. The smaller-reference global top-k set equals its own highest-ranked set under those stratum counts. Thus the fixed-quota comparison does not give only one conceptual arm a different group-allocation target. It remains a group-informed analysis intervention, not a group-label-free deployed selector.

The base training set, candidate pool, test rows, per-class acquisition budgets and head specification are fixed within the pair. Acquired identities differ and are canonically ordered before fitting. Extra reference examples affect selection but are not directly added to the head's training set. The larger reference requires 28 additional labelled examples per class—56 per binary construction and 84 for three-class MultiNLI—besides the group information used in reference construction.

The fixed-count residuals share the same matched4 outcomes, so $\Delta_{\text{fixed}} = R_{32|n_4} - R_4 = V(S_{32|n_4}) - V(S_4)$. Matched-baseline Monte Carlo error cancels from that primary contrast. The own-composition R_{32} uses its own matched32 baseline and is a different, secondary quantity. All secondary estimates retain nominal intervals/tests; no implementation or endpoint is selected after outcomes to replace the primary.

Table 16: M183 fixed-composition moderation, in accuracy percentage points. Only gap/overall is the registered primary; all other rows are secondary/descriptive. Intervals are source-t95%, $G = 11$, and the printed sign-flip values are nominal.

implementation	metric	moderation	95% CI	sign-flip p
Gap	overall (primary)	$+0.411$	$[+0.096, +0.726]$	0.005859
Gap	worst-group	$+1.174$	$[+0.030, +2.318]$	0.048828
Logistic, unscaled	overall	$+0.663$	$[+0.299, +1.027]$	0.001953
Logistic, unscaled	worst-group	$+1.503$	$[+0.522, +2.483]$	0.005859
Logistic, standardized	overall	$+0.759$	$[+0.288, +1.230]$	0.003906
Logistic, standardized	worst-group	$+1.406$	$[+0.164, +2.648]$	0.015625
Gradient alignment	overall	$+0.468$	$[-0.034, +0.971]$	0.060547
Gradient alignment	worst-group	$+1.534$	$[+0.166, +2.902]$	0.020508
Herding	overall	$+1.064$	$[+0.534, +1.595]$	0.000977
Herding	worst-group	$+1.782$	$[+0.696, +2.869]$	0.002930

Table 17: Gap levels and comparisons in M183. Accuracy percentage points, eleven sources. Only the overall fixed-count moderation is primary; these level and own-composition estimates are required secondary evidence. $R_{32,\mathrm{own}}$ and $R_{32|n_4}$ use different count-matched baselines and must not be conflated.

metric	quantity	estimate	95% CI	sign-flip p
overall	R_4	-1.446	$[-2.221, -0.672]$	0.000977
overall	$R_{32\|n_4}$	-1.035	$[-1.631, -0.439]$	0.000977
overall	$R_{32,\mathrm{own}}$	-0.905	$[-1.367, -0.444]$	0.000977
overall	T_4	$+0.010$	$[-2.565, +2.585]$	0.990234
overall	C_4	$+1.456$	$[-0.894, +3.806]$	0.212891
overall	T_{32}	$+0.827$	$[-1.960, +3.615]$	0.558594
overall	C_{32}	$+1.733$	$[-1.098, +4.563]$	0.215820
overall	Δ_{own}	$+0.541$	$[+0.098, +0.984]$	0.002930
worst-group	R_4	-0.630	$[-2.628, +1.369]$	0.486328
worst-group	$R_{32\|n_4}$	$+0.544$	$[-1.820, +2.909]$	0.617188
worst-group	$R_{32,\mathrm{own}}$	$+0.290$	$[-1.846, +2.427]$	0.788086
worst-group	T_4	$+3.862$	$[-4.394, +12.119]$	0.399414
worst-group	C_4	$+4.492$	$[-2.881, +11.865]$	0.222656
worst-group	T_{32}	$+6.091$	$[-2.953, +15.136]$	0.202148
worst-group	C_{32}	$+5.801$	$[-2.663, +14.265]$	0.196289
worst-group	Δ_{own}	$+0.920$	$[-0.058, +1.898]$	0.061523

Table 18: M183 gap/overall source means in percentage points. These are source-specific point estimates, not extra independent tests per seed or encoder.

source	R_4	$R_{32\|n_4}$	Δ_{fixed}	T_4	T_{32}
bias-in-bios	-2.832	-2.530	$+0.302$	$+1.823$	$+3.344$
CelebA	-0.469	-0.381	$+0.089$	$+5.443$	$+5.672$
CIFAR-100	-0.219	-0.002	$+0.217$	-0.267	-0.092
CivilComments	-1.723	-1.311	$+0.411$	-2.156	-1.661
MNIST–USPS	-1.480	-0.884	$+0.595$	-3.551	-2.888
MultiNLI	-1.076	-0.653	$+0.424$	$+0.028$	$+0.694$
OfficeHome	-3.632	-2.025	$+1.607$	-3.029	-1.328
PACS	-0.326	-0.232	$+0.094$	-0.258	-0.129
TerraIncognita	-2.696	-2.308	$+0.387$	-4.825	-3.975
VLCS	-0.552	-0.792	-0.241	-1.038	-0.852
Waterbirds	-0.906	-0.269	$+0.636$	$+7.939$	$+10.316$

Table 19: M183 gap/worst-group source means in percentage points. These are source-specific point estimates, not extra independent tests per seed or encoder.

source	R_4	$R_{32\|n_4}$	Δ_{fixed}	T_4	T_{32}
bias-in-bios	-5.755	-6.380	-0.625	$+6.083$	$+9.417$
CelebA	$+3.568$	$+3.776$	$+0.208$	$+35.000$	$+35.854$
CIFAR-100	-0.556	-0.042	$+0.514$	-0.597	-0.167
CivilComments	-2.435	-1.352	$+1.083$	-2.250	-1.500
MNIST–USPS	-0.186	$+2.020$	$+2.206$	-6.312	-4.158
MultiNLI	-0.495	$+2.089$	$+2.583$	$+3.208$	$+5.750$
OfficeHome	-5.044	-1.829	$+3.215$	-5.171	-1.998
PACS	-1.185	-0.799	$+0.386$	-0.779	-0.247
TerraIncognita	-0.058	-1.356	-1.298	-5.343	-5.784
VLCS	$+2.654$	$+2.999$	$+0.345$	$+1.346$	$+2.443$
Waterbirds	$+2.567$	$+6.861$	$+4.294$	$+17.302$	$+27.395$

Table 20: M183 primary sensitivity checks. Each is gap/overall fixed-count moderation, in percentage points. The original eleven-source row is primary; neither mask exclusion nor source merging replaces or upgrades it.

test endpoint	G	estimate	95% CI	sign-flip p
Original	11	$+0.411$	$[+0.096, +0.726]$	0.005859
Original, PACS/VLCS merged	10	$+0.460$	$[+0.132, +0.787]$	0.003906
Duplicate-excluded	11	$+0.409$	$[+0.096, +0.722]$	0.005859
Duplicate-excluded, PACS/VLCS merged	10	$+0.457$	$[+0.132, +0.783]$	0.003906

Table 21: **Exploratory eight-block reference-size diagnostic, complete family.** Percentage points. Block p-values have their own five-test Holm adjustment. These diagnostics do not replace the registered gap/overall source-level primary.

Endpoint	Implementation	estimate	block 95% CI; p; Holm	source 95% CI; p
Overall	Gap projection	$+0.4111$	$[-0.0993, +0.9215]$; .1015625; .203125	$[+0.0963, +0.7259]$; .005859375
Overall	Logistic, unscaled	$+0.6630$	$[+0.4340, +0.8920]$; .0078125; .0390625	$[+0.2986, +1.0274]$; .001953125
Overall	Logistic, standardized	$+0.7586$	$[+0.5239, +0.9933]$; .0078125; .0390625	$[+0.2877, +1.2295]$; .00390625
Overall	Gradient alignment	$+0.4684$	$[-0.2632, +1.2001]$; .1875; .203125	$[-0.0338, +0.9707]$; .060546875
Overall	Kernel herding	$+1.0644$	$[+0.6991, +1.4297]$; .0078125; .0390625	$[+0.5335, +1.5953]$; .0009765625
Worst-group	Gap projection	$+1.1738$	$[-0.1578, +2.5054]$; .0859375; .0859375	$[+0.0300, +2.3175]$; .048828125
Worst-group	Logistic, unscaled	$+1.5026$	$[+0.7903, +2.2149]$; .0078125; .0390625	$[+0.5224, +2.4828]$; .005859375
Worst-group	Logistic, standardized	$+1.4060$	$[+0.9154, +1.8966]$; .0078125; .0390625	$[+0.1636, +2.6484]$; .015625
Worst-group	Gradient alignment	$+1.5344$	$[-0.1540, +3.2228]$; .015625; .046875	$[+0.1665, +2.9024]$; .0205078125
Worst-group	Kernel herding	$+1.7824$	$[+0.8018, +2.7630]$; .015625; .046875	$[+0.6961, +2.8687]$; .0029296875

D.3 Integrity, sensitivity and scope

The outcome-blind freeze binds the code, registration, inherited helpers, numerical environment, both input hashes per cell, all selected-set hashes, exact reference/selection identities and test masks. A separate process reproduced every cell and selection before the immutable execution lock. All 112 cells passed independent nested-reference and acquisition-budget checks. The minimum retained test-group count is 236; the minimum stored distinct diagnostic-value count is 930. For herding, that diagnostic is first-step kernel utility, not the total greedy rank order; cutoff-tie sizes were not stored, and we do not claim zero cutoff ties.

The common sensitivity mask excludes a test feature vector if it equals any same-cell base-training, candidate-pool or larger-reference vector. It is applied to every arm at both reference sizes. All groups remain. Using the same predictions, the primary becomes $+0.409$ points $[0.096, 0.722]$, sign-flip $p = 0.005859$. The mandatory PACS/VLCS-merged analysis gives $+0.460$ points $[0.132, 0.787]$, $G = 10$, $p = 0.003906$. These checks do not replace the frozen primary or certify absence of all raw-content, near-duplicate or subject-level dependence.

An independent checker reconstructed both accuracies for all 10,752 records, verified source-row membership, reference nesting, every fixed/matched quota, the complete grid and the primary contrast. The original and independent source summaries and verdict agree. The numerical environment was Python 3.12.10, NumPy 2.4.6, pandas 3.0.3 and scikit-learn 1.9.0, with one numerical thread per worker. Runtime is reported for reproducibility, not as a throughput comparison: two workers completed the grid and official analysis in approximately 637 seconds.

Interpretation. The new R_4 is negative, so partial attenuation of a cost is supported; $R_{32|n_4}$ remains negative. The positive primary does not imply positive total value versus ordinary random acquisition. The overall T_{32} estimate is positive but its interval spans zero. The secondary worst-group moderation is positive at a nominal threshold, not a confirmatory worst-group primary.

The intervention changes reference information and, through the estimator, may change direction, shrinkage and other score properties. It does not uniquely identify stochastic reference noise, establish an optimal reference size, or prove generality to adapted representations or other budgets. The source panel is purposive and heterogeneous. Source-level uncertainty is reported under the stated model assumptions rather than interpreted as a universal law.

E Direct-use allocation study

E.1 Question and comparison

M184 asks whether the same 28 additional labelled reference rows per class are better used to guide acquisition or added directly to the classifier's training set. It reuses the exact 112 original-test cells from M183, with no new sources or seeds. Let $V_{\mathrm{train}}(A)$ denote performance when the downstream head is trained on row set A. For implementation j, the allocation contrast is

$$Q_j = V_{\mathrm{train}}(D \cup S^{\mathrm{actual}}_{32,j}) - V_{\mathrm{train}}(D \cup S^{\mathrm{actual}}_{4,j} \cup (\mathcal{R}_{32} \setminus \mathcal{R}_4)).$$

Both arms have access to the same labels. The direct-use head contains more training rows by design, while the guidance arm uses the extra rows only to form the reference-guided ranking. Thus Q_j is an operational allocation comparison, not a fixed-composition or equal-row-count effect.

E.2 Registered results

The study retains fourteen registered directions: allocation contrasts for five implementations on overall and worst-group accuracy, plus four gap-specific mechanism contrasts, assessed with the specified block/source intersection and Holm rules. For the gap/overall primary, guidance minus direct use is $+0.0558$ percentage points, with block 95% interval $[-0.4381, +0.5497]$, source 95% interval $[-0.3636, +0.4752]$, block $p = 0.75$, and source $p = 0.87890625$. The declared one-point band is descriptive and is not an equivalence margin.

E.3 Mechanism checks and descriptive levels

A count-factor-only hybrid changes the reference-count factor while retaining the smaller reference's content, pooled variance, and group quotas. It changes the selected rows in every cell, with mean turnover 24.044% and range 5%–62%. Its gap/overall effect is -0.0259 points. The remaining configuration contrast is $+0.4370$ points. These two quantities telescope to the M183 fixed-count contrast, but neither registered mechanism direction clears its gate. They do not identify a unique content, noise, or shrinkage mechanism. Content and pooled variance change together in the remaining configuration; neither miscalibration of the smaller-reference estimator nor the necessity of the extra labels is established.

E.4 Execution and verification

The run fitted 672 new heads over the 112 cells, with no prediction reuse and no execution failure. A separate reconstruction recovered all paired effects, source and block summaries, multiplicity adjustments, and verdicts from saved artifacts without refitting. Runtime and artifact hashes are retained in the local study records; they are reproducibility information rather than evidence for the allocation claim.

Table 22: **M184 registered allocation and mechanism contrasts.** Percentage points. Each block/source p-value is Holm adjusted within its declared family.

Contrast	Implementation	Endpoint	estimate	block 95% CI; Holm p	source 95% CI; Holm p
Allocation	Gap projection	overall	$+0.0558$	$[-0.438, +0.550]$; .75	$[-0.364, +0.475]$; .87890625
Allocation	Logistic, unscaled	overall	$+0.1493$	$[-0.197, +0.495]$; .703125	$[-0.372, +0.671]$; 1.000
Allocation	Logistic, standardized	overall	$+0.0636$	$[-0.210, +0.337]$; .703125	$[-0.230, +0.357]$; 1.000
Allocation	Gradient alignment	overall	-0.4314	$[-1.084, +0.222]$; .46875	$[-1.464, +0.601]$; 1.000
Allocation	Kernel herding	overall	$+0.5779$	$[+0.187, +0.969]$; .09375	$[+0.230, +0.926]$; .00390625
Allocation	Gap projection	worst-group	$+0.3817$	$[-0.542, +1.305]$; 1.000	$[-1.094, +1.857]$; 1.000
Allocation	Logistic, unscaled	worst-group	$+1.1675$	$[+0.407, +1.928]$; .0390625	$[-0.576, +2.911]$; .71875
Allocation	Logistic, standardized	worst-group	$+0.2034$	$[-0.752, +1.158]$; 1.000	$[-0.774, +1.180]$; 1.000
Allocation	Gradient alignment	worst-group	-0.0903	$[-1.822, +1.641]$; 1.000	$[-1.780, +1.599]$; 1.000
Allocation	Kernel herding	worst-group	$+0.7149$	$[-0.227, +1.657]$; .4375	$[-0.002, +1.432]$; .2490234375
Count factor	Gap projection	overall	-0.0259	$[-0.290, +0.238]$; 1.000	$[-0.182, +0.131]$; 1.000
Count factor	Gap projection	worst-group	$+0.1319$	$[-0.463, +0.727]$; 1.000	$[-0.498, +0.762]$; 1.000
Remaining configuration	Gap projection	overall	$+0.4370$	$[+0.118, +0.756]$; .125	$[+0.163, +0.711]$; .01171875
Remaining configuration	Gap projection	worst-group	$+1.0419$	$[+0.116, +1.968]$; .125	$[+0.163, +1.921]$; .041015625

Table 23: **Allocation primary: descriptive source and block values.** Percentage points. Constructions are averaged within a source first, then the eleven sources are averaged equally; the block column averages the same quantity within each of the eight seed blocks. These are descriptive views of the one primary, not a new per-source or per-block hypothesis family, and the negative block is retained.

Source	Q overall, points	Seed block	Q overall, points
bios	-0.5104	53	-0.1428
celeba	-0.0729	54	-1.1326
cifar100	+0.0167	55	+0.0859
civilcomments	-0.1823	56	+0.9449
mnist_usps	-0.4953	57	+0.2809
multinli	+0.2361	58	-0.0387
officehome	+0.3369	59	+0.0204
pacs	+0.1230	60	+0.4284
terraincognita	-0.2791		
vlcs	-0.3008		
waterbirds	+1.7421		

Table 24: **Absolute arm levels.** Equal-source and equal-block averages of the original test metrics, in percent. Descriptive only: these levels are not a new common-population accuracy claim and no arm is compared here outside its registered contrast.

Arm	Overall %	Worst-group %
c2st_m32_actual	78.778	71.226
direct_c2st	78.714	71.023
direct_dratio	78.939	71.796
direct_gap	79.099	72.322
direct_gradalign	78.533	70.967
direct_herding	79.014	72.229
dratio_m32_actual	79.088	72.964
gap_count32	78.312	70.606
gap_m32_actual	79.155	72.703
gap_m32_yoked	78.749	71.648
gap_m4_actual	78.337	70.474
gradalign_m32_actual	78.101	70.877
herding_m32_actual	79.592	72.944
random	78.328	66.612

F REPRESENTATION-REFRESH STUDY

F.1 ADAPTIVE ORIGIN AND FIXED DESIGN

The preparation study M185 asked whether the fixed feature recipe changed enough selections to proceed. Four of eight blocks passed, below the required six; it stopped before validation-map fitting, downstream heads or test utility. M186 then evaluated the unchanged recipe on the previously frozen, unexecuted validation blocks, following an authorized, outcome-free decision. Both the stop and that adaptive decision remain part of the study's provenance.

The study imported 128 validation role files without resampling. It used the same 261 metadata families, 979 performances and 3,868 segments as the later music study, with four training, three reference and three real-pool families per composer and all remaining families for test. Non-test roles cap performances at two per family. Every class offers ten real and ten generated candidates and acquires six labelled rows. Caps are two per family, two per performance and one generated child per capped training parent. Ordinary and matched random acquisitions are uniform over complete allowed sets, not independent row draws.

F.2 MAPS, SELECTIONS AND UTILITY

M0 uses 109 whole-clip descriptors with training-weighted scaling and clipping. M1 appends twelve GP outputs, four for each of three LDA axes; LDA, GP and scalers use training rows only. Duplicate constructed columns occur in 76 blocks and constructed training rank ranges from 5 to 12. The twelve columns are one fitted map, not twelve established independent directions.

The plain class-conditional score projects a candidate onto the unit-normalized reference-minus-training mean. It differs from the panel's denoised gap, which subtracts 1.5 pooled standard errors coordinate-wise. The recorded `score_spearman` measures correlation between pool score vectors, not fitness or generalization. The stale selection S_0 uses M0; fixed-count refresh $S_{1|q_0}$ uses M1 at S_0's class–origin counts; operational refresh S_1 uses M1 at its own counts. Four complete redraws are retained per quota, with identical quotas sharing them. Every row set is evaluated under both maps. Logistic heads use $C = 100$ and training/acquisition weight masses $2/3, 1/3$. Utility averages segments within performances, performances within families, families within composers and the ten composers equally.

F.3 FAILED POSITIVE PRIMARY AND THE SURVIVING BOUND

D_1 is the M1-learner benefit of $S_{1|q_0}$ over S_0; D_0 is the same row change under M0 and $I = D_1 - D_0$. The registered intersection-union claim $D_1 > 0.01$ (one percentage point) and $I > 0$, both one-sided, failed with $p_{\text{joint}} = 1$. Approximate Student-t intervals use 127 degrees of freedom over randomizations conditional on this corpus, cache and procedure. The prospective precision gate, requiring both primary 95% interval half-widths to be at most one point, passed. The interval half-widths for $D_1, I, D_0, D_{\text{OP},M1}$ are respectively $0.1787, 0.1755, 0.1736, 0.2140$ points.

Only D_1's registered bounded-null rule applies: its interval lies strictly inside $[-1, +1]$ points, recorded as `BOUNDED_REFRESH`. The band was a prospective research threshold, not a validated deployment loss. It gives no equivalence decision for I, D_0, operational quotas or individual works. Table 25 shows the four levels behind the main contrasts. All exact levels, intervals and recorded inference fields appear once in Tables 26–28.

Holding stale rows fixed gives M1 minus M0 $= +0.5071$ points. Fixed-count refresh then gives $D_1 = -0.0647$, for a descriptive joint change $J = +0.4423$. Because the primary failed, the stronger $HN/HU/J$ family was not evaluated; its raw values in the complete record are not confirmatory results. Count-matched residual means are positive here. This is not a replication of the panel's overall residual: corpus, score, reference allocation, caps, head weighting, groups and metric all differ.

Table 25: **Which map chooses the rows, and which map learns from them.** Hierarchical accuracy (%). D_1 is the second column's refresh difference, D_0 the first column's, and $I = D_1 - D_0$. These are the frozen S0 and S1q0 utility means; exact levels, intervals and recorded decisions appear in the complete endpoint tables.

Rows chosen on	base learner M_0	constructed learner M_1
the base map (S0, stale)	33.513	34.020
the constructed map (S1q0, refreshed at the fixed quota)	33.501	33.955

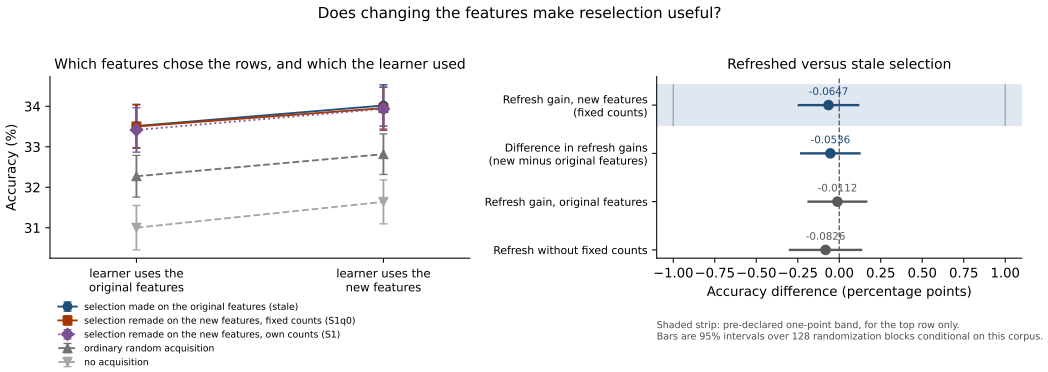

Figure 6: **Changing the learner is not the same as refreshing the rows.** Left: stale and fixed-count-refreshed sets under both learners, with no acquisition and ordinary random for scale. Right: D_1, I, D_0 and operational refresh. The ± 1-point band applies to D_1 alone. Intervals use approximate Student-t inference over 128 fixed-corpus randomizations (127 degrees of freedom).

F.4 ALL 65 ENDPOINTS

The following tables retain the canonical producer order, roles, intervals, p-values and validity flags. They add no primary or multiplicity family. R denotes selected minus matched random, C matched minus ordinary random and T selected minus ordinary random. Q0 is the stale quota, FIXED denotes its refreshed selection, and OP the operational quota. MAPDIFF is M1 minus M0 for the named arm; U endpoints are absolute utility levels. N is no acquisition, U is ordinary random, QBAR is the mean of the four named redraws. S1q0 and S1 remain distinct arms.

F.5 EXECUTION AND PROVENANCE

All 128 blocks completed, with 3,328 logical cells from 3,254 physical head fits and 74 exact-input reuses, zero failures, 384 GP-axis fits, 128 LDA fits and 256 scalers. Maps, scores, selections and redraws were sealed before test access. Fixed-count turnover averaged 8.37240%, ranging from 0 to 18.33333%; zero-turnover and reused cells were retained. Sixteen blocks had the pinned constant-base-column scaler warning. The permitted unit-scale handling left finite arrays, without ad hoc recovery.

Independent reconstruction checked saved-coefficient predictions, hierarchical utilities, all 65 endpoints and inference, roles, seals, quotas and resource counts without refitting, returning `full_contract_pass=true` with full-contract scope. Before execution, a hard-linked fixture changed two unused M185 role files. Their bytes were restored to the original pins, the 270-file tree digest matched, and 151 links were detached. No result used the damaged records.

Metadata families do not prove note-level identity, generated labels are inherited and generator-pretraining disjointness is uncertified. Matched rendering removes the identified processing mismatch but changes the task from untreated MIDI features. Reported CPU, wall time and memory cover instrumented stages, not cache creation, earlier engineering or an end-to-end speedup.

Table 26: **All 65 frozen endpoints, part 1 of 3.** Canonical producer order; nothing is sorted by effect size and no arm is selected. Values are percentage points. Uncertainty is over 128 randomization blocks conditional on this corpus, not over new works or independent datasets.

Endpoint	Role	mean 95% CI	SD	one-sided p	interval type	valid
D1	primary component	-0.0647 [-0.2434, +0.1139]	1.0216	1.000	student_t_two_sided	yes
D0	secondary	-0.0112 [-0.1847, +0.1624]	0.9924	0.550497	student_t_two_sided	yes
I	primary component	-0.0536 [-0.2291, +0.1219]	1.0034	0.726599	student_t_two_sided	yes
J	joint system	+0.4423 [+0.1968, +0.6879]	1.4041	0.000257	student_t_two_sided	yes
HN	joint system	+2.3149 [+1.9104, +2.7195]	2.3132	2.82e-21	student_t_two_sided	yes
HU	joint system	+1.1380 [+0.7401, +1.5358]	2.2747	4.79e-08	student_t_two_sided	yes
R_Q0_M0	secondary	+0.6181 [+0.3049, +0.9313]	1.7906	7.6e-05	student_t_two_sided	yes
R_FIXED_Q0_M0	secondary	+0.6069 [+0.2817, +0.9322]	1.8595	0.000164	student_t_two_sided	yes
T_U_M0	secondary	+1.2397 [+0.8096, +1.6697]	2.4588	3.9e-08	student_t_two_sided	yes
C_Q0_M0	secondary	+0.6216 [+0.2770, +0.9661]	1.9700	0.000253	student_t_two_sided	yes
D_OP_M0	secondary	-0.0968 [-0.2915, +0.0980]	1.1135	0.836333	student_t_two_sided	yes
F_OP_M0	secondary	-0.0976 [-0.3570, +0.1618]	1.4830	0.771023	student_t_two_sided	yes
C_OP_M0	secondary	+0.0008 [-0.1966, +0.1982]	1.1286	0.496762	student_t_two_sided	yes
R_Q0_M1	secondary	+0.5404 [+0.2449, +0.8359]	1.6893	0.000213	student_t_two_sided	yes
R_FIXED_Q0_M1	secondary	+0.4757 [+0.1674, +0.7839]	1.7624	0.001378	student_t_two_sided	yes
T_U_M1	secondary	+1.2027 [+0.7919, +1.6135]	2.3488	2.57e-08	student_t_two_sided	yes
C_Q0_M1	secondary	+0.6623 [+0.3302, +0.9945]	1.8990	6.54e-05	student_t_two_sided	yes
D_OP_M1	secondary	-0.0826 [-0.2966, +0.1314]	1.2236	0.776732	student_t_two_sided	yes
F_OP_M1	secondary	-0.1383 [-0.4158, +0.1393]	1.5870	0.836927	student_t_two_sided	yes
C_OP_M1	secondary	+0.0557 [-0.1401, +0.2515]	1.1194	0.287301	student_t_two_sided	yes

Table 27: **All 65 frozen endpoints, part 2 of 3.** Continuation in canonical producer order. Roles, intervals, and inference-validity flags are copied from the frozen analysis.

Endpoint	Role	mean 95% CI	SD	one-sided p	interval type	valid
MAPDIFF_N	secondary	+0.6372 [+0.4071, +0.8673]	1.3155	1.1e-07	student_t_two_sided	yes
MAPDIFF_U	secondary	+0.5440 [+0.3401, +0.7480]	1.1659	2.72e-07	student_t_two_sided	yes
MAPDIFF_S0	secondary	+0.5071 [+0.3055, +0.7086]	1.1525	1.02e-06	student_t_two_sided	yes
MAPDIFF_S1q0	secondary	+0.4535 [+0.2559, +0.6511]	1.1300	6.43e-06	student_t_two_sided	yes
MAPDIFF_S1	secondary	+0.5213 [+0.3409, +0.7016]	1.0313	3.64e-08	student_t_two_sided	yes
MAPDIFF_Q0_0	secondary	+0.4489 [+0.2470, +0.6508]	1.1541	1.13e-05	student_t_two_sided	yes
MAPDIFF_Q0_1	secondary	+0.6170 [+0.4181, +0.8160]	1.1375	4.96e-09	student_t_two_sided	yes
MAPDIFF_Q0_2	secondary	+0.6196 [+0.4155, +0.8236]	1.1668	9.26e-09	student_t_two_sided	yes
MAPDIFF_Q0_3	secondary	+0.6536 [+0.4622, +0.8450]	1.0942	2.27e-10	student_t_two_sided	yes
MAPDIFF_Q1_0	secondary	+0.6507 [+0.4566, +0.8448]	1.1097	4.24e-10	student_t_two_sided	yes
MAPDIFF_Q1_1	secondary	+0.4174 [+0.2046, +0.6301]	1.2163	8.27e-05	student_t_two_sided	yes
MAPDIFF_Q1_2	secondary	+0.6349 [+0.4645, +0.8053]	0.9743	9.5e-12	student_t_two_sided	yes
MAPDIFF_Q1_3	secondary	+0.8556 [+0.6501, +1.0611]	1.1748	9.14e-14	student_t_two_sided	yes
MAPDIFF_QBAR0	secondary	+0.5848 [+0.4405, +0.7291]	0.8251	3.03e-13	student_t_two_sided	yes
MAPDIFF_QBAR1	secondary	+0.6396 [+0.5086, +0.7707]	0.7494	3.51e-17	student_t_two_sided	yes
U_M0_N	secondary	+31.0027 [+30.4538, +31.5515]	3.1380	5.28e-129	student_t_two_sided	yes
U_M0_U	secondary	+32.2729 [+31.7576, +32.7881]	2.9457	1.18e-134	student_t_two_sided	yes
U_M0_S0	secondary	+33.5125 [+32.9763, +34.0487]	3.0655	1.55e-134	student_t_two_sided	yes
U_M0_S1q0	secondary	+33.5014 [+32.9675, +34.0352]	3.0525	9.44e-135	student_t_two_sided	yes
U_M0_S1	secondary	+33.4157 [+32.8674, +33.9641]	3.1351	3.76e-133	student_t_two_sided	yes
U_M0_Q0_0	secondary	+32.8031 [+32.2628, +33.3433]	3.0888	5.94e-133	student_t_two_sided	yes
U_M0_Q0_1	secondary	+32.9324 [+32.4289, +33.4359]	2.8789	5.11e-137	student_t_two_sided	yes

Table 28: **All 65 frozen endpoints, part 3 of 3.** Final continuation in canonical producer order. The three parts together reproduce every endpoint without creating an additional primary or multiplicity family.

Endpoint	Role	mean 95% CI	SD	one-sided p	interval type	valid
U_M0_Q0_2	secondary	+32.8983 [+32.3874, +33.4092]	2.9210	3.64e-136	student_t_two_sided	yes
U_M0_Q0_3	secondary	+32.9440 [+32.4199, +33.4680]	2.9963	7.54e-135	student_t_two_sided	yes
U_M0_Q1_0	secondary	+32.7091 [+32.1581, +33.2602]	3.1507	1.04e-131	student_t_two_sided	yes
U_M0_Q1_1	secondary	+33.0192 [+32.4578, +33.5806]	3.2099	3.29e-131	student_t_two_sided	yes
U_M0_Q1_2	secondary	+32.8764 [+32.3435, +33.4092]	3.0464	7.89e-134	student_t_two_sided	yes
U_M0_Q1_3	secondary	+32.9763 [+32.4324, +33.5201]	3.1095	7.1e-133	student_t_two_sided	yes
U_M0_QBAR0	secondary	+32.8944 [+32.4216, +33.3673]	2.7035	2.13e-140	student_t_two_sided	yes
U_M0_QBAR1	secondary	+32.8952 [+32.3993, +33.3912]	2.8357	8.75e-138	student_t_two_sided	yes
U_M1_N	secondary	+31.6399 [+31.0975, +32.1823]	3.1010	9.17e-131	student_t_two_sided	yes
U_M1_U	secondary	+32.8169 [+32.3168, +33.3170]	2.8593	3.37e-137	student_t_two_sided	yes
U_M1_S0	secondary	+34.0196 [+33.5094, +34.5298]	2.9169	4.46e-138	student_t_two_sided	yes
U_M1_S1q0	secondary	+33.9548 [+33.4367, +34.4730]	2.9624	3.98e-137	student_t_two_sided	yes
U_M1_S1	secondary	+33.9370 [+33.4032, +34.4708]	3.0519	1.82e-135	student_t_two_sided	yes
U_M1_Q0_0	secondary	+33.2519 [+32.7156, +33.7883]	3.0665	4.31e-134	student_t_two_sided	yes
U_M1_Q0_1	secondary	+33.5494 [+33.0558, +34.0430]	2.8221	3.99e-139	student_t_two_sided	yes
U_M1_Q0_2	secondary	+33.5179 [+33.0225, +34.0132]	2.8322	7.04e-139	student_t_two_sided	yes
U_M1_Q0_3	secondary	+33.5975 [+33.0722, +34.1228]	3.0034	8.54e-136	student_t_two_sided	yes
U_M1_Q1_0	secondary	+33.3598 [+32.8279, +33.8917]	3.0411	1.01e-134	student_t_two_sided	yes
U_M1_Q1_1	secondary	+33.4366 [+32.8754, +33.9977]	3.2082	6.35e-132	student_t_two_sided	yes
U_M1_Q1_2	secondary	+33.5112 [+32.9685, +34.0540]	3.1030	7.21e-134	student_t_two_sided	yes
U_M1_Q1_3	secondary	+33.8319 [+33.3066, +34.3571]	3.0029	3.49e-136	student_t_two_sided	yes
U_M1_QBAR0	secondary	+33.4792 [+33.0172, +33.9411]	2.6411	1.22e-142	student_t_two_sided	yes
U_M1_QBAR1	secondary	+33.5349 [+33.0406, +34.0291]	2.8260	5.01e-139	student_t_two_sided	yes

G ACQUIRED COUNTS AND ADDITIONAL UNCERTAINTY SUMMARIES

G.1 ACTUAL ACQUIRED COMPOSITION

Table 29 reports the count-control study's acquired minority fractions relative to ordinary random. Table 30 gives the reference-size study's gap-selection counts by source and class. Averages aid interpretation; the frozen cell records retain exact integer vectors and quota checks.

Table 29: M179v2 mean acquired minority fractions (%) after averaging encoders within source/seed and seeds within source. This is a readable composition summary, not a replacement for the exact per-cell count vectors in the frozen feasibility file.

source	random	gap	unscaled	standardized	gradient	herding
bias-in-bios	6.2	26.8	39.1	23.9	26.5	31.7
CelebA	26.3	48.2	31.4	27.4	19.6	35.9
CIFAR-100	38.7	59.0	57.1	46.9	55.6	47.2
CivilComments	37.6	40.7	40.2	38.0	36.0	39.6
MNIST–USPS	46.2	90.6	51.1	41.8	64.0	49.8
MultiNLI	12.8	14.2	14.7	15.1	14.6	15.1
OfficeHome	43.8	39.8	41.2	45.0	47.2	46.1
PACS	33.4	68.6	42.1	30.5	37.8	46.5
TerraIncognita	37.2	57.1	46.8	48.0	33.9	48.4
VLCS	45.9	40.8	45.4	47.7	36.7	45.9
Waterbirds	4.6	38.0	33.1	14.2	27.2	23.8

Table 30: M183 gap acquired counts, averaged within each source/class. Each pair is attribute0/attribute1. The yoked32 arm uses the actual4 counts exactly in every cell; the unconstrained actual32 counts need not coincide. Fractional values arise only from averaging the frozen integer counts over seeds/encoders.

source	class code	budget	actual4 (also yoked32)	actual32
bias-in-bios	0	150.0	39.5/110.5	57.1/92.9
bias-in-bios	1	150.0	101.5/48.5	71.6/78.4
CelebA	0	150.0	80.9/69.1	76.8/73.2
CelebA	1	150.0	85.4/64.6	66.9/83.1
CIFAR-100	0	70.0	37.0/33.0	13.8/56.2
CIFAR-100	1	70.0	31.4/38.6	23.0/47.0
CivilComments	0	150.0	84.0/66.0	85.6/64.4
CivilComments	1	150.0	55.1/94.9	60.0/90.0
MNIST–USPS	0	150.0	2.1/147.9	22.2/127.8
MNIST–USPS	1	150.0	17.8/132.2	8.2/141.8
MultiNLI	0	150.0	141.0/9.0	133.9/16.1
MultiNLI	1	150.0	141.0/9.0	137.1/12.9
MultiNLI	2	150.0	106.5/43.5	95.2/54.8
OfficeHome	0	150.0	57.4/92.6	63.5/86.5
OfficeHome	1	150.0	59.1/90.9	64.9/85.1
PACS	0	46.0	24.1/21.9	15.4/30.6
PACS	1	150.0	25.4/124.6	1.6/148.4
TerraIncognita	0	150.0	41.2/108.8	9.8/140.2
TerraIncognita	1	150.0	79.0/71.0	63.8/86.2
VLCS	0	131.0	63.1/67.9	59.5/71.5
VLCS	1	108.0	62.1/45.9	56.9/51.1
Waterbirds	0	150.0	86.9/63.1	47.1/102.9
Waterbirds	1	51.0	15.2/35.8	29.5/21.5

G.2 ADDITIONAL REFERENCE-SIZE UNCERTAINTY SUMMARIES

These exploratory diagnostics use saved results without new fits. They do not replace the registered source-level primary. First, averaging OfficeHome, TerraIncognita, PACS and VLCS into one

unit, then weighting that unit equally with seven other sources, gives gap/overall $+0.392$ points $[+0.238, +0.546]$, nominal sign-flip $p = 0.0078125$. This changes the target weighting, not just the original estimate's standard error. Sharing a fixed encoder does not by itself establish stochastic dependence.

Second, equal weighting of each source's observed class–groups, rather than empirical test-row frequencies, gives $+0.390$ points $[+0.098, +0.683]$, nominal $p = 0.001953$. This changes the evaluation metric; it is not the registered overall-accuracy endpoint.

Third, each of the eight common seed blocks averages encoders within source and then the same eleven sources equally. The mean remains $+0.411$ points, with block standard error 0.216 and $t(7)$ interval $[-0.099, +0.921]$. This measures randomization variation conditional on the panel, not variation across sources. Eight blocks and the t approximation limit its precision. Table 21 retains the complete implementation family and its separate block-wise Holm adjustments.

H Representation crossing outside music

H.1 Sources and alignment

This study reuses two cached sources. Waterbirds predicts landbird versus waterbird with the land/water background attribute and retains the official train/validation/test split, with 11,788 rows. CivilComments is the retained 60,000-row WILDS subsample (Koh et al., 2021) with the frozen `toxicity>0` label and upstream `identity_any>0` attribute. Official split labels survive proportional sampling over split–group strata. Both use $g = 2y + a$. Validation is unused; no tuning or model selection occurs. Retained official test rows were inaccessible to utility computation before clearance. Neither source is a new or untouched corpus.

CivilComments pairing is verified through the retained UID/text/extraction chain. Waterbirds pairing is conditional on the verified order-preserving provenance chain from raw metadata through the extractor to packaged features: its feature archives do not carry self-authenticating row IDs. Waterbirds results remain conditional on accepting that chain.

H.2 Blocks, quotas and maps

Each source has 24 paired randomized blocks, IDs 0–23, with seed root 20260923. Each block draws 16 training rows without replacement from each of four groups, giving 64 base rows. After their removal, a deterministic permutation retains at most 512 pool rows per group. Every arm acquires exactly 16 per group, 64 total, and uses four exact-quota random controls. Class and group labels define these quotas. Waterbirds pools are $[512, 168, 40, 512]$, with acquisition fractions $[0.031, 0.095, 0.400, 0.031]$. Taking 16 of 40 rows in the smallest group can compress scorer differences; this is a setting limitation, not an observed explanation.

Waterbirds uses CLIP (Radford et al., 2021) ViT-B/32 image features (512 dimensions) and ResNet-50 (He et al., 2016) (2048); CivilComments uses BERT-base-uncased (Devlin et al., 2019) (768) and CLIP ViT-B/32 text features (512). Each encoder gets its own 32-dimensional whitened PCA fitted only on the 64 base rows, using `sklearn.decomposition.PCA`, `svd_solver=full`, `whiten=True`, without extra scaling. This matches output dimension and head parameter count, not capacity, information, geometry, PCA estimation quality or effective regularization. Budget and geometric sensitivity are not varied.

H.3 Selector and learner

Within each class–attribute stratum, greedy k-center repeatedly chooses the candidate farthest from its nearest base or already-acquired center in Euclidean whitened-PCA space. Squared nearest-center scores are rounded to 12 decimals, with SHA-256 row-key ties. This uses the published geometric primitive (Gonzalez, 1985; Sener & Savarese, 2018) inside four fixed quotas, not unrestricted Core-Set active learning. No reference or reference-guided score is used.

Source	Worst-group difference (pp)	95% conditional interval (pp)
Waterbirds	+1.310	—
CivilComments	-0.780	—
Pooled (equal sources)	+0.265	[-0.611, +1.114]

Table 31: **Same-map versus crossed selection, in worst-group accuracy points.** The pooled estimate weights the two sources equally; its 95% interval comes from paired-block resampling within each fixed source. Other rows show source means, for which no interval is defined (—).

The learner is logistic regression on its encoder's own PCA map, with 64 base plus 64 acquired rows, $C = 1$, `lbfgs`, tolerance 10^{-6}, 2,000 maximum iterations, `random_state=0`, no class weighting and no extra scaler. Four execution stages fitted 576 heads in 36.332 seconds.

H.4 Primary and uncertainty

For scorer S and learner L, write U_{SL} for the worst-group residual against the mean of four random controls. The sole pre-specified quantity is $\frac{1}{2}((U_{AA} - U_{BA}) + (U_{BB} - U_{AB}))$, averaged over 24 blocks per source and then equally over sources. Within each difference the learner is the same, so the random terms cancel. The primary compares selected sets; random controls enter the descriptive cell residuals.

The interval uses 10,000 paired-block bootstrap draws within each fixed source, seed 2026092301, source order (Waterbirds, CivilComments), ascending block order and one sequential `numpy.random.default_rng` stream. It uses percentile endpoints at linear quantiles 0.025 and 0.975. The analysis record and exported machine-readable tables retain the 24 ordered values per source. This is conditional randomization uncertainty, not dataset-population, independent-subject, new-corpus or feature-extractor uncertainty. Overlapping blocks are not independent subjects. The recorded 0.5-point band provides scale for reading the estimate.

The contrast is $+0.264988$ points with 95% interval $[-0.611126, +1.113999]$. Waterbirds is $+1.310111$, CivilComments -0.780134; no source-level interval was defined.

H.5 Every cell and its accuracy level

Table 32 retains all cells of both sources for worst-group, macro-group and overall accuracy in the original order. A residual subtracts the learner-specific matched-random level from selected accuracy. Figure 4 shows all worst-group residuals on one zero-centred scale. Seven of eight worst-group cell means are negative; these are descriptive cells, not eight tests.

Selected worst-group levels span 54.90–83.19%, and matched random spans 58.63–82.41%. Cross-learner level differences also reflect the learner; they cannot be assigned to selection. The result establishes neither a general matching advantage nor a penalty. Because the selector family differs from music, it is a related comparison rather than a direct replication. The study ran once: no source, block, budget, dimension, map, head, endpoint or margin changed after test access.

H.6 Saved-evidence replay

The verified record is `analysis/results.json` (SHA-256 `47b3591e...c52f97`); a separate implementation rebuilt it before completion (`independent_done.json`, `d99409db...c7c308`). A later independent check read only the 48 saved prediction archives, imported no study code and fitted no model. It rebuilt all 576 cells and group metrics, complete absolute/random/residual matrices, 48 ordered contrasts, source and pooled means, and the exact seeded interval. Across 1,230 arrays and scalars, the maximum absolute discrepancy was 4.44×10^{-16}. A clean copy outside the project tree reproduced the same replay receipt. This verifies saved numerical evidence, not upstream feature extraction or training regeneration.

Reproduction coverage across studies. The numerical supplement also replays all 17,920 music source-by-representation cell utilities, both primaries, 140 absolute levels and 135 acquisition gains;

Source	Metric	Scorer	Learner	Selected (%)	Random (%)	Residual (pp)
Waterbirds	Worst-group	CLIP ViT-B/32	CLIP ViT-B/32	72.95	74.19	-1.239
		CLIP ViT-B/32	ResNet-50	80.67	82.41	-1.739
		ResNet-50	CLIP ViT-B/32	72.85	74.19	-1.335
		ResNet-50	ResNet-50	83.19	82.41	+0.785
	Macro-group	CLIP ViT-B/32	CLIP ViT-B/32	81.08	81.18	-0.105
		CLIP ViT-B/32	ResNet-50	85.78	86.50	-0.722
		ResNet-50	CLIP ViT-B/32	80.59	81.18	-0.594
		ResNet-50	ResNet-50	86.65	86.50	+0.149
	Overall	CLIP ViT-B/32	CLIP ViT-B/32	79.81	79.99	-0.177
		CLIP ViT-B/32	ResNet-50	84.47	85.98	-1.507
		ResNet-50	CLIP ViT-B/32	78.99	79.99	-1.003
		ResNet-50	ResNet-50	86.71	85.98	+0.725
CivilComments	Worst-group	BERT-base	BERT-base	54.90	59.37	-4.470
		BERT-base	CLIP ViT-B/32 (text)	55.12	58.63	-3.505
		CLIP ViT-B/32 (text)	BERT-base	57.25	59.37	-2.121
		CLIP ViT-B/32 (text)	CLIP ViT-B/32 (text)	55.91	58.63	-2.716
	Macro-group	BERT-base	BERT-base	65.12	64.82	+0.303
		BERT-base	CLIP ViT-B/32 (text)	63.99	64.49	-0.503
		CLIP ViT-B/32 (text)	BERT-base	64.33	64.82	-0.493
		CLIP ViT-B/32 (text)	CLIP ViT-B/32 (text)	64.26	64.49	-0.232
	Overall	BERT-base	BERT-base	60.30	64.10	-3.799
		BERT-base	CLIP ViT-B/32 (text)	60.74	64.99	-4.246
		CLIP ViT-B/32 (text)	BERT-base	62.18	64.10	-1.918
		CLIP ViT-B/32 (text)	CLIP ViT-B/32 (text)	61.82	64.99	-3.168

Table 32: **Full selected / matched-random / residual matrices, both corpora and all three metrics.** Absolute selected accuracy and the matched-random baseline in accuracy %; the selected-minus-matched-random residual in percentage points, paired with its absolute levels. Every 2×2 scorer$\times$learner cell is shown and none is omitted. These per-cell summaries are descriptive. Random controls are quota-stratified count-matched draws.

the earlier refresh study's 3,328 cells and 65 endpoints; and count-control, reference-size and allocation inference from saved arm or contrast records. Upstream feature generation, model fitting and core prediction-level reproduction are not included. The coverage ledger identifies remaining sensitivity and arm-level omissions. Frozen digests and reconstruction receipts accompany the evidence; local protocols and historical records are retained separately. Local registration means hash-frozen local commitments, not an externally time-stamped service. The code package contains no raw music or model checkpoints and grants no licence to redistribute upstream assets.