When Does Data Value Reduce to Class Balance? A Coverage View of Per-Point Data Valuation

Anonymous authors
Paper under double-blind review

Abstract

Per-point data-valuation scores — Data Shapley, influence, leverage — are routinely used to *select* training data by taking the top k. **We show that fixed, mode-blind per-point rankings are provably suboptimal as selectors in an identifiable latent-mode regime, and characterize that regime exactly.** A point's value is not a scalar but a *context function* $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\})$; for a kernel learner it is the point's feature novelty after orthogonalizing against the selected set, so the interaction lives in the off-diagonal (redundant) Gram. In an equal-mode model we prove that any symmetric valuation, used as top-k, has expected excess risk *exactly* $\rho\,\delta(1, k)$ over greedy coverage at a binding budget — a fixed scalar cannot place one point per mode, so it covers only an occupancy fraction. **The decisive question is whether the learner's modes coincide with the labels available at selection.** *If yes* (standard classification on strong features), a one-line class-balance of the same scores recovers and beats coverage — the gap was just class imbalance. *If no*, the correction is coverage, not balancing: with *no* labels, label-free coverage wins by up to +0.20 accuracy across datasets (by several standard errors) while a raw top-k falls *below* random; with labels *coarser* than the modes, scalar balancing is insufficient and the best practical selector uses the coarse labels as strata and applies learner-geometry coverage within them. **The resulting diagnostic is simple: when the labels available at selection align with the learner's modes, balance within labels; otherwise add an explicit learner-geometry coverage or diversity correction.** The advantage is representation-relative — when the modes are not expressed in the learner geometry (spurious subgroups), coverage too fails, exactly as the theory predicts.

1 Introduction

Data valuation assigns each training point a scalar worth, and these scalars are widely used to *select* data: keep the top-k most valuable points to train cheaply, or to curate a fine-tuning set. Data Shapley (Ghorbani & Zou, 2019), leave-one-out and influence (Koh & Liang, 2017), datamodels (Ilyas et al., 2022), and leverage scores all produce such per-point numbers. We ask whether top-k-by-score is a good *selection* rule, and answer: precisely in the regime where data selection helps — a binding budget, a redundant representation, a non-interpolable task — it is provably suboptimal in a latent-mode model, and "improving" the per-point score can make it worse by concentrating the selected set (Figure 1 previews the full argument).

The reason is structural. The value of a point is not a number but a function of the *context* of other selected points: a point that is the only representative of a concept is valuable; once a near-duplicate is selected, its marginal value collapses. A scalar summary (Shapley averages this over contexts; influence evaluates it at full data) discards exactly the information a selector needs — which points are *jointly* non-redundant. We make this precise, prove a lower bound on the resulting selection gap, identify its source as the off-diagonal mass of the learner-geometry Gram, and verify it.

Contributions.

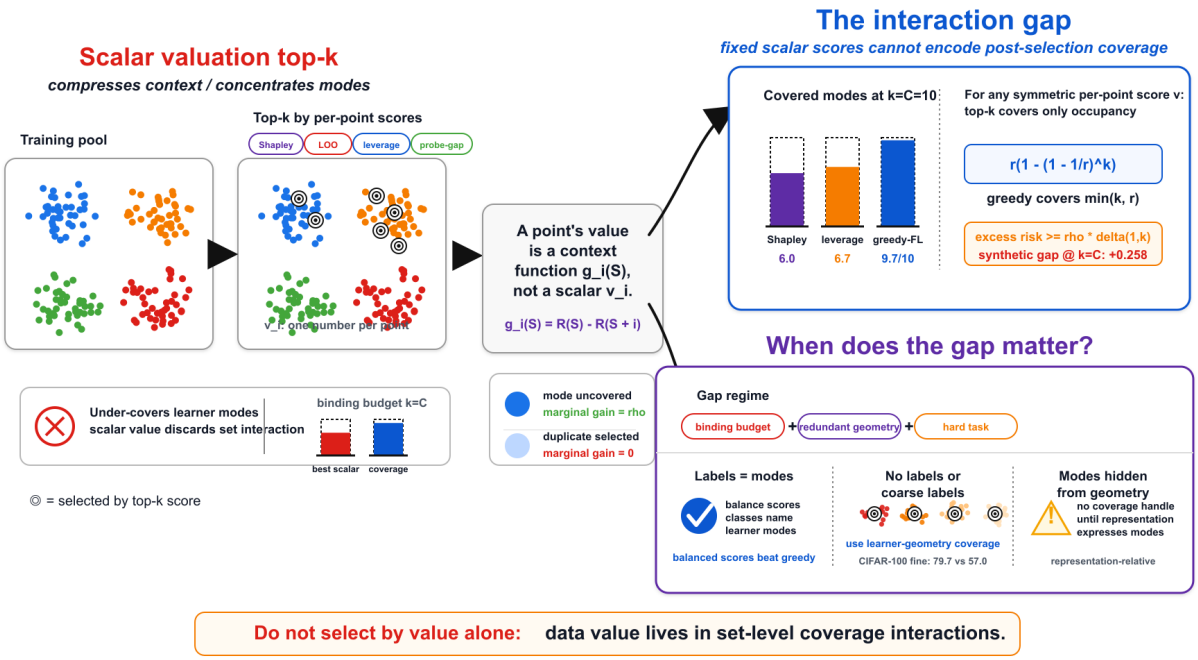

Figure 1: **Data value is a context function, not a scalar.** *Left:* per-point valuation methods collapse each example to a scalar and top-k selection concentrates, under-covering learner modes. *Center:* the true marginal is $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\})$: a point is valuable when its mode is uncovered and nearly worthless after a duplicate is selected. *Top right:* in the latent-mode model, any symmetric scalar top-k covers only the occupancy fraction, while greedy coverage covers $\min(k, r)$ modes and has excess-risk advantage $\rho\,\delta(1, k)$; empirically, at $k = C$ it covers 9.7/10 modes and gains $+0.258$ over the best scalar. *Bottom right:* the gap matters under a binding budget, redundant learner geometry, and a hard task; the remedy is scoped by the modes-vs-labels boundary: balance scores if labels name learner modes, use learner-geometry coverage when labels are missing or too coarse, and do not expect coverage to find modes absent from the representation.

- **Value is a context function (§2).** A closed form $g_i(S)$ for kernel learners; the interaction is the off-diagonal (redundant) Gram mass, and top-k by any within-mode-symmetric score under-covers the modes relative to greedy coverage.
- **An impossibility (§3).** In a latent-mode instance, a symmetric valuation selected top-k has expected excess risk *exactly* $\rho\,\delta_N(1, k)$ over greedy coverage (and $\geq$ for any within-mode-exchangeable score); LOO $\equiv 0$ and Shapley is mode-constant; a spiked-covariance analysis gives the soft-mode extension.
- **The modes-vs-labels boundary (§4).** Across synthetic and real ViT experiments we show that the relevant question is not whether a scalar score is accurate pointwise, but whether the groups available at selection name the learner's modes. When modes coincide with labels, a one-line class balance of the same scores recovers the missing coverage and can outperform label-free coverage. When labels are unavailable or coarser than the modes, scalar balancing is insufficient and an explicit learner-geometry coverage or diversity correction is needed. We also delimit the claim: coverage helps only when the target modes are expressed in the learner geometry, and it can fail when the representation hides the relevant subgroups.

We package this as a *diagnostic decision tree* (Algorithm 1): redundancy diagnostics flag whether a raw top-k is risky, and a label–mode alignment test routes each dataset to balancing or to coverage. The contribution is thus not only the negative result but an *operational boundary* a practitioner can evaluate before selecting.

Algorithm 1 Mode-aware selection: a diagnostic decision tree (balance-or-cover)

Input: pool features $\Phi \in \mathbb{R}^{N\times d}$, budget k, optional labels y with C classes
$K \leftarrow$ learner-geometry kernel of Φ; compute redundancy diagnostics (off-diagonal Gram mass, effective rank) ▷ is a raw top-k risky?
if labels unavailable **then**
  return learner-geometry coverage of K (greedy facility location; k-center as a cheap variant)
else
  cluster Φ into C groups; $a \leftarrow$ label–cluster alignment (e.g. NMI); check residual within-label multimodality
  if a high **and** little within-label multimodality **then** ▷ labels name the modes
    return class-balanced selection: $\approx\lfloor k/C\rfloor$ per class, ranked within class by the chosen score
  else ▷ labels coarser than / unaligned with the modes
    return learner-geometry coverage, optionally stratified by the coarse labels
  end if
end if

Effective rank and off-diagonal Gram mass are *redundancy* diagnostics: they indicate whether a raw top-k ranking is likely to concentrate in a redundant learner geometry, and should not be read literally as the number of semantic classes in high-class image datasets. The balance-vs-cover branch is instead set by whether the labels available at selection align with geometric clusters in the learner representation (and whether substantial within-label multimodality remains). The diagnostic thus separates three questions: whether the representation is redundant, whether the available labels name the learner's modes, and whether the labels are too coarse. It is a heuristic for choosing between two *safe* corrections, not an oracle: we do not tune its threshold on test accuracy, and the experiments report the balanced and coverage selectors *directly*, so our conclusions do not depend on the routing firing correctly. When the alignment statistic is ambiguous we recommend reporting *both* the balanced and the coverage selector, and avoiding a raw per-point top-k.

2 Value is a context function

A learner has risk $\mathrm{R}(S) = \mathbb{E}_{x\sim D}(f_S(x) - f^\star(x))^2$ for a selected set $S \subseteq P$, $|P| = N$. The **marginal value** of point i in context S is $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) \geq 0$ (monotone R). Scalar valuations collapse this function to a number: leave-one-out / influence $v_i^{\mathrm{LOO}} = g_i(P \setminus \{i\})$; Data Shapley $v_i^{\mathrm{Sh}} = \mathbb{E}_\pi g_i(S^\pi_{<i})$; leverage and probe-gap are geometry- or label-derived per-point scores. The **interaction** is $\mathrm{Int}_i = \max_S g_i(S) - \min_S g_i(S)$.

Linear-algebra form. For a kernel/linear learner and *point* selection, the covered target is the feature second moment $\Sigma = \mathbb{E}_x[\varphi\varphi^\top]$: $\mathrm{R}(S) = \mathrm{tr}\big((I - P_S)\Sigma\big)$, with P_S the projector onto $\mathrm{span}\{\varphi_i : i \in S\}$. The marginal is the Σ-novelty of the point's feature after orthogonalizing against S (Gram–Schmidt / partial correlation):

$$g_i(S) = \frac{\langle \varphi_i^{\perp(S)}, \Sigma\, \varphi_i^{\perp(S)}\rangle}{\|\varphi_i^{\perp(S)}\|^2}, \qquad \varphi_i^{\perp(S)} = (I - P_S)\varphi_i. \tag{1}$$

What $\mathrm{R}(S)$ is, and when it is the supervised risk. As written, $\mathrm{R}(S) = \mathrm{tr}((I - P_S)\Sigma)$ is a *representation-coverage* objective: it measures the feature second-moment mass not reconstructed by the span of the selected points. It is not automatically the supervised risk for an arbitrary target. The connection requires a *mode-aligned target* assumption: the target variation is supported on learner-mode directions (equivalently, in the latent-mode instance the label is a function of mode identity and an uncovered mode incurs a positive cost). Under it, the representation-coverage risk is proportional to supervised excess risk — failing to span a target-carrying mode leaves that component unfit. In the latent-mode model of §3 this holds by construction, so $\mathrm{R}(S)$ *is* the mode classification risk; without it, $\mathrm{R}(S)$ should be read as a coverage *proxy*. The theorems characterize this proxy exactly; the experiments of §4 then test when optimizing it improves downstream supervised accuracy (it does, including for an SGD-trained classifier).

Theorem 1 (no-interaction geometry and redundancy)**.** *If the feature vectors are mutually orthogonal ($\varphi_i \perp \varphi_j$ for $i \neq j$), then $g_i(S)$ is independent of any context S not already containing i: selecting other points never changes the residual of φ_i, so there is no selection interaction. Conversely, when the learner geometry contains redundant directions, orthogonalizing against previously selected neighbours can change a point's marginal in two distinct ways.* (i) Rank collapse: *the residual may fall to zero once a redundant neighbour is selected, so g_i collapses to* 0 *— an all-or-nothing interaction.* (ii) Directional rotation: *among contexts where the residual stays nonzero, its direction may rotate, and if Σ is anisotropic on those directions the Rayleigh quotient in* (1) *changes. The hard latent-mode model below realizes mechanism (i) exactly, while the soft-mode and general-spectrum settings combine both. Both are carried by the off-diagonal / block mass of the learner-geometry Gram $G = \Phi\Phi^\top$.*

Proposition 1 (interaction has two sources)**.** *Let $\varphi_i^{\perp(S)}$ be the residual of φ_i after projecting away* $\mathrm{span}\{\varphi_j : j \in S\}$*. Interaction decomposes into two terms.* (i) Directional. *Over contexts S, S' with $\varphi_i^{\perp(S)} \neq 0$ and $\varphi_i^{\perp(S')} \neq 0$, write $w_i(S) = \varphi_i^{\perp(S)}/\|\varphi_i^{\perp(S)}\|$ and let $D_i = \sup \|w_i(S) - w_i(S')\|$ over such contexts be the diameter of the set of residual directions. Then*

$$\sup_{S,S' \colon \varphi_i^{\perp} \neq 0} |g_i(S) - g_i(S')| \;\leq\; \big(\lambda_{\max}(\Sigma) - \lambda_{\min}(\Sigma)\big)\, D_i.$$

D_i measures the amount of nonzero residual-direction variation induced by redundancy; directional interaction requires $D_i > 0$ and *anisotropy of Σ ($\lambda_{\max} > \lambda_{\min}$). Exact-duplicate or same-mode redundancy instead appears through the separate rank-collapse term below.* (ii) Rank collapse. *If some context S_0 has $\varphi_i^{\perp(S_0)} = 0$ (point i lies in the selected span), then $g_i(S_0) = 0$; this all-or-nothing term can be nonzero even when Σ is isotropic, and it is exactly the latent-mode mechanism (once a same-mode point is selected, another point's residual collapses).*

Proof. (i) $g_i(S) = R(w_i(S))$ with $R(w) = w^\top \Sigma w$. For any unit w, w', in the eigenbasis of Σ, $R(w) - R(w') = \sum_k (\lambda_k - c)(w_k^2 - w_k'^2)$ for any c; with $c = \frac{1}{2}(\lambda_{\max} + \lambda_{\min})$ and Cauchy–Schwarz, $|R(w) - R(w')| \leq (\lambda_{\max} - \lambda_{\min})\|w - w'\| \leq (\lambda_{\max} - \lambda_{\min})D_i$. (ii) Since φ_i lies in $\mathrm{span}(S_0)$, adding i does not change the selected span, so $\mathrm{R}(S_0 \cup \{i\}) = \mathrm{R}(S_0)$ and hence $g_i(S_0) = \mathrm{R}(S_0) - \mathrm{R}(S_0 \cup \{i\}) = 0$. □

This separates two mechanisms a single off-diagonal summary would conflate. *Directional* interaction is the classical marginal-screening-versus-forward-selection effect under collinearity: orthogonalization rotates the residual and anisotropy of Σ moves its Rayleigh quotient. *Rank-collapse* interaction is stronger: a duplicate or same-mode neighbour makes the residual vanish, collapsing the marginal to zero. The latent-mode theorem below uses the all-or-nothing mechanism; the soft-mode extension makes the collapse approximate.

Equivalently: scalar valuation is *marginal (univariate) screening*; the optimal context-aware selector is *forward selection / greedy coverage* (which orthogonalizes); their gap is the redundancy that orthogonalization removes. This is the classical failure of marginal screening under collinearity, here applied to data valuation. The off-diagonal Gram is a *contribution covariance*: the selection interaction lives in shared directions, not in per-point diagonals alone.

Terminology (modes vs. classes vs. clusters). A *mode* is a group of points the learner must each cover to generalize — a block of interchangeable features in the learner geometry (a block of the Gram). *Classes* are the labels available at selection; *clusters* are geometric groups in the representation. Our results turn on whether these coincide: when the modes *are* the classes, balancing labels suffices; when they are finer geometric clusters (or unlabeled), coverage is needed. Throughout, r is the number of modes, m the per-mode size, k the budget, and C the number of labels.

3 The interaction gap

Latent-mode instance. Let the pool be $P = \bigsqcup_{j=1}^{r} P_j$ with r modes of size $m = N/r$. A non-degenerate (block) learner geometry makes covering ≥ 1 point of mode j drive its residual to 0; an uncovered mode costs $\rho_j > 0$ (the task is non-interpolable across modes). Then $\mathrm{R}(S) = \sum_{j \colon S \cap P_j = \emptyset} \rho_j$.

Lemma 1 (all-or-nothing marginal)**.** *For $i \in P_j$, $g_i(S) = \rho_j \mathbf{1}[S \cap P_j = \emptyset] \in \{0, \rho_j\}$, so $\mathrm{Int}_i = \rho_j$. Any scalar $v_i \in [0, \rho_j]$ disagrees with the true marginal on an $\Omega(1)$ fraction of contexts.*

Lemma 2 (exchangeability collapses scores)**.** *In the* block-exchangeable *instance — where points within a mode are indistinguishable to the learner geometry and the valuation functional (the population / hard-block model, or finite samples with identical within-mode features) — any symmetric valuation is mode-constant: $v_i = c_j$ for $i \in P_j$. (Under merely* distributional *exchangeability the within-mode scores are only equal in distribution; the occupancy argument then uses only that the valuation induces no systematic within-mode preference, which uniform tie-breaking supplies.)*

The two canonical scalars, exactly. *LOO/influence:* at full data every mode is covered, and removing one point of a mode with $m \geq 2$ leaves it covered, so $v_i^{\mathrm{LOO}} = 0$ for *every* point — influence is identically zero in the redundant regime. *Shapley:* $v_i^{\mathrm{Sh}} = \rho_j \Pr[i \text{ first of its mode in } \pi] = \rho_j / m_j$, constant over points for equal modes. Both are Lemma-2 blind. (Numerically confirmed: LOO $\equiv 0$; Shapley $= \rho/m$ to three digits.)

Theorem 2 (interaction gap, exact for symmetric valuations)**.** *Take* equal *modes ($m_j \equiv m$, $N = rm$), equal costs ρ, and a budget $0 < k < r$, with ties in the score broken uniformly at random. Because the modes are statistically identical, a data-dependent* symmetric *valuation (the population / block-exchangeable versions of leave-one-out, Data Shapley, and leverage- or probe-gap-type scores) is constant not only within each mode (Lemma 2) but* across *modes, hence constant on the whole pool; its top-k set is then a uniformly random k-subset of P. (On finite noisy data these scores are constant only in distribution; Proposition 2 covers that case.) It therefore covers exactly the occupancy number of modes,*

$$\mathbb{E}[\#\mathit{covered}] = m_\star(k) = r\Big(1 - \tbinom{(r-1)m}{k} \Big/ \tbinom{rm}{k}\Big) \xrightarrow[N\to\infty]{} r\big(1 - (1 - \tfrac{1}{r})^k\big),$$

while greedy coverage S_{FL} covers $\min(k, r) = k$ modes and is $(1 - 1/e)$-optimal (Nemhauser et al., 1978). Hence, with $\mathrm{R}(S) = \rho\,(r - \#\mathit{covered})$,

$$\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) = \rho\,\delta_N(1, k), \qquad \delta_N(1, k) = k - m_\star(k) > 0.$$

As $N \to \infty$, $\delta_N(1, k) \to \delta(1, k) = k - r\big(1 - (1 - 1/r)^k\big)$; we use the finite-pool δ_N in statements and the closed-form limit δ in prose.

Proof. Equal modes are exchangeable, so a symmetric functional of the data assigns them equal values; with within-mode constancy (Lemma 2) the score is constant on P, and uniform tie-breaking makes S_v a uniform k-subset. A fixed mode is missed iff none of its m points is drawn, probability $\binom{(r-1)m}{k} \Big/ \binom{rm}{k}$; summing the complement over r modes gives $\mathbb{E}[\#\text{covered}] = m_\star(k)$. Greedy gains a fresh mode each step (a covered mode has zero marginal, Lemma 1), so it covers $\min(k, r) = k$. Subtract. (Full details, App. D.) □

Proposition 2 (general lower bound, equal modes)**.** *For* equal-size *modes, any* mode-constant *score — deterministic, or randomized with no systematic within-mode preference — used as top-k with uniform tie-breaking covers at most the occupancy, $\mathbb{E}[\#\mathit{covered}] \leq m_\star(k)$, so $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \geq \rho\,\delta_N(1, k)$, with equality iff the score is constant across modes (Theorem 2). A score whose mode-values* concentrate *the budget on a few modes covers strictly fewer (App. D, by a pooling / Schur-convexity argument that uses the equal mode sizes).*

Informally, a selector whose geometry is only *partially* aligned with the modes should interpolate between greedy coverage (fully aligned) and the mode-blind occupancy baseline; formalizing this interpolation for general representations is left open.

Remark 1 (unequal modes)**.** With *unequal* sizes the bound can reverse: a score favoring *small* (rare) modes can cover *more* than uniform. With $r = 10$ modes — nine singletons and one mode of 91 points — and $k = 9$, a score that prefers the singletons covers all nine, beating uniform random. Coverage then depends on whether the score's high values align with small modes (spreading, helpful) or large ones (concentrating, harmful); only equal modes give the unconditional bound.

Corollary 1 (improving a valuation need not improve selection)**.** *Whether a valuation helps or hurts coverage depends on how its high values are spread across modes. A score that* concentrates *— high values on many points of a few dense, typical modes — covers fewer than the uniform baseline and can fall* below *random at a binding budget; empirically the* estimated *Data Shapley and prototypicality do exactly this (their high scores cluster on typical points) and both dip to or below random near $k = C$ in §4. The equal-mode symmetric valuation instead ties random exactly (Theorem 2). Either way, a more accurate per-point number does not monotonically improve selection, because coverage is a set-level property the scalar does not control.*

Corollary 2 (phase boundary)**.** *The gap is positive iff all three hold: (i) the* budget binds *($k < r$), (ii) the* task is hard *($\rho > 0$, non-interpolable across modes), and (iii) the* learner geometry is redundant *(off-diagonal Gram > 0 / effective rank $\ll N$). Condition (iii) is geometry redundancy, not representation non-locality.*

All three claims are confirmed numerically to three digits (App. G, Table 11): uniform coverage equals $m_\star(k)$ (Theorem 2), a concentrating score stays below it (Proposition 2), and the unequal-mode reversal of Remark 1 occurs as predicted.

Soft modes / general spectra. Replace hard modes by a spiked covariance $\Sigma = \sum_j \pi_j u_j u_j^\top + \eta^2 P_\perp$ with within-mode noise η. From (1), $g_i(S) = \sigma_{j(i)}\mathbf{1}[\text{mode uncovered}] + O(\eta^2)$ (numerically: uncovered marginal $0.117 \approx \pi_j$, covered $0.004 = O(\eta^2)$), so $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \geq \rho\,\delta(1,k) - O(\eta^2 k) > 0$ for small η. The σ_j are the eigenvalues of Σ and η^2 the tail: the gap is large for a *peaked* spectrum (low effective rank, redundant) and vanishes as the spectrum *flattens* (high effective rank, near-orthogonal). General Σ reduces to the marginal-screening-vs-OMP gap of sparse approximation, governed by the Gram coherence.

Scope: what is and is not critiqued. The result is about a specific procedure — a *fixed, mode-blind, non-adaptive* per-point ranking used as top-k, with no balancing, grouping, or coverage constraint. Three escapes leave this regime, and the rest of the paper is precisely about them: (i) *adaptive* recomputation of the marginal after each pick is exactly greedy and covers $\min(k, r)$; (ii) a *group-aware* score (class quotas, subgroup balance) restores coverage when the groups are the modes (§4); and (iii) a score that already encodes mode-aware structure is, by definition, no longer mode-blind. What we claim is therefore sharp but bounded: $\delta_N(1,k)$ is the exact *price of non-adaptivity* — the regret of any fixed mode-blind score relative to adaptive coverage — so *within this regime* the remedy is structural (balance or cover), not a more accurate number. We do not claim that every conceivable scalar fails; we identify the regime in which a naive top-k provably under-covers.

4 Experiments

Findings at a glance.

- **Theory is exact.** A symmetric valuation's top-k covers $m_\star(k)$ to three digits; greedy covers all (Table 11).
- **Modes = labels** (4 datasets): class-balancing the *same* scores beats greedy coverage — the gap was just imbalance (Table 3). A real Data-Shapley method (KNN-Shapley) concentrates and dips below random too (Table 4).
- **No labels** (3 datasets): greedy coverage beats random by +0.08 to +0.20, and a raw top-k falls *below* random (Table 5); the gap persists for an SGD-trained MLP (Table 6).
- **Coarser labels** (CIFAR-100 coarse→fine): the correction is coverage/diversity, not scalar balancing — coarse-stratified facility location (0.419) beats label-free coverage (0.392) and scalar coarse balancing (0.24–0.27) (Table 7).
- **The diagnostic validates the balance branch** on five labeled datasets (pre-fixed threshold, Table 9); the no-label and coarse-label experiments separately instantiate the coverage and stratified-coverage branches.

Table 1: Test accuracy by selector vs. budget ($r = C = 10$). **Greedy-FL** (context-aware coverage) beats every per-point top-k at the binding budget; the gap is δ-shaped (peaks at $k = C$, $\to 0$ as $k \gg C$).

k	random	LOO	leverage	proto	marginal	shapley	greedy-FL	FL−best
5	0.419	0.249	0.444	0.417	0.293	0.390	**0.490**	+0.047
10	0.601	0.263	0.600	0.637	0.417	0.615	**0.895**	+0.258
15	0.690	0.267	0.741	0.719	0.500	0.678	**0.909**	+0.168
20	0.821	0.364	0.802	0.847	0.526	0.757	**0.928**	+0.081
30	0.821	0.343	0.819	0.907	0.612	0.837	0.908	+0.001
50	0.897	0.464	0.894	0.951	0.729	0.924	0.942	−0.010

Table 2: Redundancy drives the gap. As noise rises the block-contrast (off-diagonal mass) falls and the effective rank rises; the gap = greedy-FL − leverage-top-k at $k = C$ collapses.

noise	block-contrast	eff. rank	gap @ k=C
0.10	0.416	15.5	+0.349
0.18	0.231	22.8	+0.295
0.30	0.105	27.7	+0.035
0.50	0.040	29.4	−0.024
0.80	0.014	29.8	+0.003

- **Boundaries.** On spurious subgroups (Waterbirds) a generic representation hides the modes, so coverage fails too; and when quality is orthogonal to coverage, coverage over corrupted outliers collapses while quality-then-coverage wins (Table 8).

We instantiate the latent-mode regime synthetically (C=10 classes, 3 input-distant modes each, a non-degenerate embedding; learner = ridge probe). The full suite runs in seconds (numpy). Table 1 is the core result: every per-point scalar, selected top-k, is beaten by greedy coverage at the binding budget k=C, and the gap is δ-shaped in the budget; LOO is near-useless and Shapley stays far below the context-aware selector; Figure 2 summarizes the gap, its mechanism, and its cause.

Value is not a scalar. The marginal $g_i(S)$ over random contexts has spread/|mean| ≈ 1.5 and is bimodal (40% large gain, 57% ≈ 0) — all-or-nothing, as Lemma 1 predicts.

The mechanism, made visible: scalars concentrate. The failure is direct to see in the *classes covered* by the selected set. At the binding budget $k = C = 10$, greedy coverage selects points spanning 9.7 of the 10 classes, whereas every per-point top-k covers far fewer — leverage 6.7, prototypicality 6.5, *Data Shapley* 6.0. The reason is exactly the concentration in the proof: a per-point score is (near-)constant within a class, so the top-k fills up the highest-scoring class(es) before reaching others, leaving uncovered classes uncovered. The *more* a score concentrates on a small set of high-scoring classes, the fewer classes it covers — the counterintuitive corollary as a picture.

Redundancy causes the gap (causal). Sweeping the embedding noise to tune redundancy, the gap at $k = C$ tracks the off-diagonal / block structure monotonically (Table 2); as the effective rank rises $15 \to 30$ the gap collapses $+0.35 \to 0$. This is a controlled causal law, not a cross-task correlation.

Phase boundary. Crossing the three conditions of Corollary 2, the gap appears only under *binding budget* $\wedge$ *hard task* $\wedge$ *redundant geometry*, where it is largest (+0.278). Diffuse or loose-budget cells collapse to ≤ 0; the redundant easy/tight cell leaves only a smaller finite-probe residual. Representation locality is not a condition (that distinguishes this from the selection-geometry advantage it generalizes).

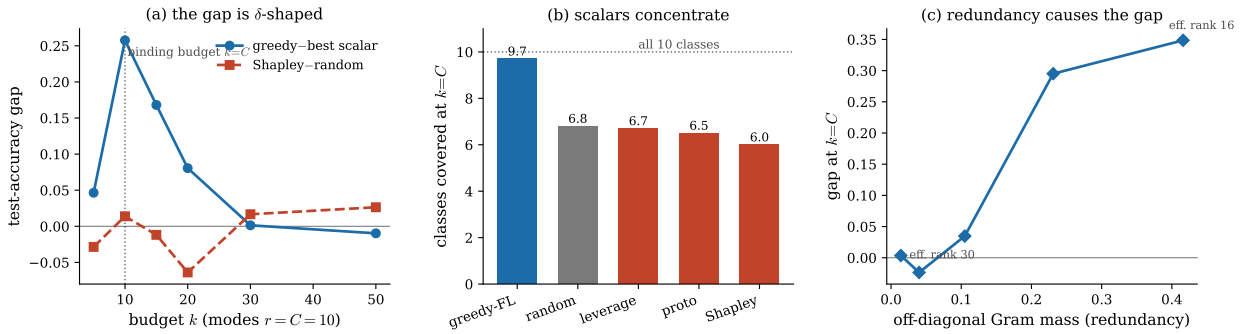

Figure 2: **The interaction gap, its mechanism, and its cause (synthetic, $r = C = 10$).** (a) The gap greedy−best-scalar is δ-shaped: it peaks at the binding budget $k = C$ (+0.258) and vanishes as $k \gg C$; top-k-by-Shapley even falls *below random* near the binding budget (red). (b) Mechanism: at $k = C$ greedy coverage spans 9.7/10 classes while every per-point scalar concentrates and covers far fewer (Shapley, the most mode-informative, covers fewest, 6.0). (c) Cause: sweeping embedding noise, the $k = C$ gap rises monotonically with the off-diagonal Gram mass (redundancy) and collapses to 0 as the effective rank rises $16 \to 30$ — a controlled causal law, not a cross-task correlation.

Table 3: Modes equal labels, using ViT features at budget $k = C$, with standard errors over five seeds. A raw per-point top-k loses to greedy facility location, but a class-balanced version of the same scores beats greedy facility location; in this regime the apparent interaction gap is class imbalance, fixed by balancing. The "best" entries are oracle summaries over a fixed, pre-specified selector set — not deployable model-selection procedures; the matched raw-versus-balanced variants are reported individually in Table 10 and show the same conclusion.

	CIFAR-10	Oxford Pets	CIFAR-100	Flowers-102
greedy − best *raw* scalar	$+0.196 \pm 0.042$	$+0.166 \pm 0.021$	$+0.119 \pm 0.012$	$+0.223 \pm 0.006$
greedy − best *balanced* scalar	-0.092 ± 0.020	-0.004 ± 0.015	-0.081 ± 0.012	-0.171 ± 0.006

4.1 The modes-vs-labels boundary (real ViT-B/16 features)

We use frozen pretrained ViT-B/16 features (learner = ridge probe; PCA-64; 3–5 seeds).

Modes = labels: balancing suffices, and beats coverage. A *raw* per-point top-k fails badly at the binding budget $k = C$: it concentrates and under-covers classes, losing +0.12 to +0.22 to greedy coverage (Table 3, "FL−best raw"). But pretrained ViT collapses each class to roughly one cluster, so here the *modes are the labels*. Class-balancing the very same scores (equal quota per class, ranked within) then *recovers and beats* greedy facility location at every budget: greedy minus the best balanced score is *negative* everywhere (Table 3, "FL−best bal."), e.g. −0.09/−0.00/−0.08/−0.17 on CIFAR-10/Pets/CIFAR-100/Flowers at $k = C$; balanced *prototypicality* is typically the single best selector. So in standard classification the selection interaction is exactly class coverage, and the right — and best — fix is a one-line class balance, not set-level coverage.

A mainstream valuation method concentrates too. To check the failure is not specific to leverage or prototypicality, we add *exact* KNN-Shapley (Jia et al., 2019) — a real Data-Shapley method — on the same frozen features (Table 4). Used as a raw top-k it behaves identically: on CIFAR-10 it covers 5.3 of 10 classes and scores 0.43, *below* random (0.54); on CIFAR-100 it trails random. Class-balancing the very same Shapley values restores full coverage and lifts accuracy (0.68, 0.36). Per-point data valuation thus concentrates exactly as predicted — the critique is about the per-point ranking *procedure*, not one score.

No labels to balance: label-free coverage wins across datasets. When selection has *no* labels (the common curation setting), class-balancing is unavailable and the only label-free options are random, a

Table 4: **A real Data-Shapley method (KNN-Shapley) concentrates like the other scalars** (frozen ViT, $k = C$, 3 seeds; accuracy / classes covered). Values are computed on a held-out validation split *drawn from the training pool*, disjoint from the final test set. Raw top-k trails random and covers few classes; balancing the same values recovers; greedy coverage is strongest. The mechanism is the per-point ranking, not the particular score.

selector ($k = C$)	CIFAR-10 (acc / cov /10)	CIFAR-100 (acc / cov /100)
random	0.541 / 6.3	0.265 / 64.0
KNN-Shapley, raw top-k	0.433 / 5.3	0.245 / 60.3
leverage, raw top-k	0.352 / 4.7	0.190 / 55.3
KNN-Shapley, class-balanced	0.682 / 10.0	0.363 / 100.0
greedy facility location	**0.779** / 9.0	**0.391** / 76.7

Table 5: No labels to balance (frozen ViT features, $k = C$, mean ± se over 5 seeds). With no labels, label-free greedy coverage beats random on every dataset by many standard errors, while a *raw* per-point top-k (leverage / global typicality) concentrates and falls *below* random on both accuracy and classes covered on all three datasets — the concentration mechanism on real data.

$k = C$	random	leverage	typicality	k-center	greedy-FL
test accuracy					
CIFAR-100	$0.274_{\pm.009}$	$0.175_{\pm.007}$	$0.224_{\pm.006}$	$0.247_{\pm.004}$	$\mathbf{0.405}_{\pm.004}$
Food-101	$0.175_{\pm.004}$	$0.129_{\pm.005}$	$0.063_{\pm.002}$	$0.204_{\pm.008}$	$\mathbf{0.254}_{\pm.006}$
Flowers-102	$0.490_{\pm.016}$	$0.375_{\pm.006}$	$0.224_{\pm.014}$	$0.642_{\pm.011}$	$\mathbf{0.686}_{\pm.006}$
classes covered (/C)					
CIFAR-100	$64.6_{\pm1.6}$	$58.2_{\pm1.4}$	$51.8_{\pm0.8}$	$65.6_{\pm0.8}$	$\mathbf{77.8}_{\pm1.4}$
Food-101	$65.8_{\pm0.5}$	$53.2_{\pm1.5}$	$50.6_{\pm1.3}$	$62.2_{\pm1.7}$	$\mathbf{71.0}_{\pm0.9}$
Flowers-102	$61.2_{\pm2.4}$	$49.8_{\pm0.8}$	$36.8_{\pm1.3}$	$76.6_{\pm1.2}$	$\mathbf{78.6}_{\pm0.7}$

per-point scalar top-k, or set coverage. The modes are then the many fine classes — geometric clusters in ViT space — exactly the regime the theory targets. Across three many-class datasets at the binding budget $k = C$ (Table 5, 5 seeds), greedy facility location covers far more classes and wins on accuracy by +0.08 to +0.20 over random (by many standard errors), while a *raw per-point top-k* (leverage or global typicality) concentrates, covers *fewer* classes than random, and falls *below* random accuracy on all three datasets — the counterintuitive corollary, now on real data and across datasets. The gap is δ-shaped (largest at $k = C$, shrinking by $4C$).

Beyond the ridge probe. A natural worry is that the gap is an artifact of the closed-form linear probe. It is not: replacing the probe with a small *SGD-trained* MLP (Adam, cross-entropy, dropout) reproduces the coverage gap almost exactly — within 0.02 accuracy of the ridge probe on both datasets — and keeps the raw leverage top-k below random (Table 6). The interaction is a property of the selected *data*; a trained nonlinear classifier benefits from coverage just as the linear probe does.

Modes ≠ labels: an explicit coverage or diversity correction is needed. The picture changes when the labels available at selection are coarser than the learner's modes. On CIFAR-100 we let selectors use only the 20 coarse superclasses, but *evaluate* on the 100 fine classes (Table 7, Figure 3). Balancing the coarse labels with a *scalar* cannot resolve the fine modes inside a superclass: at the binding budget $k = C_{\text{fine}}$ it reaches only 0.266 (coarse-balanced random) or 0.240 (coarse-balanced prototypicality) fine accuracy, below even label-free greedy coverage (0.392). But the coarse labels *are* useful when paired with an explicit within-coarse coverage step: coarse-balanced facility location — equal quota per superclass, chosen by facility location *within* each — reaches 0.419, above label-free coverage and second only to the unavailable fine-label oracle (0.474). So the correction that matters is coverage/diversity, not scalar balancing: when labels are coarser than the modes, use them as strata for coverage rather than as targets for a per-point ranking.

Table 6: **Beyond the ridge probe: the gap is a property of the data, not the estimator.** Label-free selection at $k = C$, scored with the closed-form ridge probe vs. an SGD-trained MLP (5 seeds, ± se on the MLP). The coverage gap (greedy-FL − random) is nearly identical under both learners, and a raw leverage top-k stays below random under both.

dataset	learner	random	leverage	k-center	greedy-FL	gap (FL−rand)
CIFAR-100	ridge	0.274	0.175	0.247	**0.405**	+0.131
	MLP	$0.267_{\pm.007}$	$0.186_{\pm.010}$	$0.200_{\pm.008}$	$\mathbf{0.397}_{\pm.010}$	+0.130
Flowers-102	ridge	0.490	0.375	0.642	**0.685**	+0.195
	MLP	$0.473_{\pm.016}$	$0.357_{\pm.006}$	$0.567_{\pm.015}$	$\mathbf{0.653}_{\pm.007}$	+0.180

Table 7: Modes not equal labels: CIFAR-100 selected with only the 20 coarse superclasses, evaluated on the 100 fine classes at $k = C_{\text{fine}}$ (3 seeds). Balancing the coarse labels with a *scalar* is insufficient (it cannot resolve fine modes inside a superclass), but pairing the coarse quotas with an explicit *within-coarse coverage* step (facility location) is the best practical selector — above label-free greedy coverage — and trails only the unavailable fine-label oracle. The correction that matters is coverage/diversity, not scalar balancing. Not all diversity objectives are equally effective, though: within-coarse k-center improves fine-class coverage (73.3) but selects boundary or atypical points and scores poorly (0.251), whereas facility location better represents the dense fine modes and gives the best practical accuracy.

selector ($k = C_{\text{fine}}$)	fine accuracy	fine classes (/100)
coarse-balanced random (scalar)	0.266	68.3
coarse-balanced prototypicality (scalar)	0.240	57.0
coarse-balanced k-center (coverage)	0.251	73.3
coarse-balanced facility location (coverage)	**0.419**	**83.0**
label-free greedy coverage	0.392	79.7
oracle (fine-balanced)	0.474	100.0

Boundary: coverage needs the modes to live in the representation. The positive coverage results share a precondition — the modes are geometric clusters *expressed* in the learner representation. When they are not, coverage fails. On Waterbirds (Sagawa et al., 2020) the modes are four spurious subgroups (bird × background), but a frozen generic ViT organizes points by the visually dominant background rather than the subgroup, so facility location spends its budget on the two majority subgroups (covering only ∼2 of the 4 even at $k = 100$) and the coverage selectors place atypical points rather than representative minority ones. At the tight budget $k = 20$, worst-group accuracy is then *worst* for the coverage methods (0.21–0.23), below random (0.36) and label-balancing (0.43), and is recovered only by an oracle that balances the true subgroups (0.67). This is exactly what the theory predicts: the gap is governed by the *learner-geometry* Gram (§2), so coverage helps only when the modes *are* that geometry's structure. A generic representation that does not express the subgroups offers no coverage handle on them; discovering representation-hidden modes (and worst-group robustness as the objective) is outside our scope.

The quality axis, controlled. When point *quality* varies independently of coverage, pure coverage can be actively harmful — the trade-off we flag for imitation learning (§6). We exhibit it cleanly: inject 30% corrupted examples (feature outliers with random labels) into the CIFAR-100 pool and select $k = C$ (Table 8). Pure coverage is *catastrophic* — k-center and greedy facility location fill 97–99% of the budget with the corrupted outliers (they maximize spread) and collapse to chance. A quality score (typicality) avoids the outliers but concentrates. The *decoupled* recipe — restrict to the high-quality half, then cover within it — selects no corrupted points and wins by a wide margin, instantiating the quality-controlled-pool-then-coverage rule of §6.

The diagnostic is not post-hoc. Algorithm 1 routes between balance and coverage without using test accuracy. The redundancy statistics (effective rank, off-diagonal Gram mass) diagnose whether a raw top-k ranking is risky; the branch is then made by label–cluster alignment. With a threshold fixed *in advance*, the

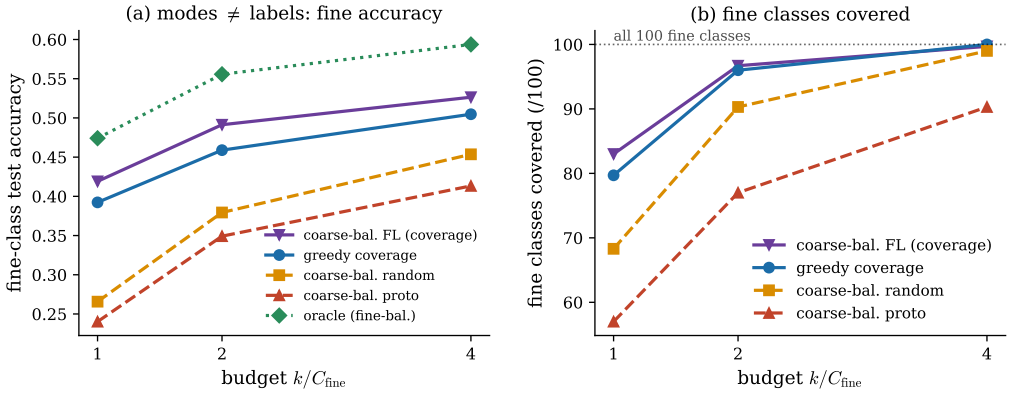

Figure 3: **Modes not equal labels: coverage, not scalar balancing, is the needed correction (CIFAR-100 coarse-to-fine).** Selectors use only the 20 coarse superclasses, evaluation is on the 100 fine classes. (a) Scalar coarse balancing (orange/red) is insufficient; *coarse-balanced facility location* (purple) — coverage within each superclass — is the best practical selector, above label-free greedy coverage (blue) and second only to the unavailable fine-label oracle (green). (b) The advantage is coverage of finer modes: the coverage selectors span the most fine classes at every budget, decisively at the binding budget $k = C_{\text{fine}}$.

Table 8: **Quality and coverage are orthogonal: a controlled demonstration** (CIFAR-100, 30% corrupted feature-outliers, $k = C$, 5 seeds). Pure coverage selects almost only corrupted points and collapses to chance; quality-then-coverage avoids them and is best — the decoupled recipe.

selector ($k = C$)	test accuracy	% selected corrupted
random	$0.156_{\pm.009}$	34%
quality (typicality) top-k	$0.226_{\pm.004}$	0%
k-center (coverage)	$0.011_{\pm.002}$	99%
greedy facility location (coverage)	$0.011_{\pm.003}$	97%
quality-filter + coverage	$\mathbf{0.387}_{\pm.007}$	0%

labeled standard-classification datasets have high alignment (NMI $\geq 0.54 > \tau = 0.45$, Table 9) and route to BALANCE, matching the best selector family (Table 10). In contrast the no-label setting routes directly to coverage, and the coarse-label setting has unresolved within-label fine modes and routes to coverage (Tables 5, 7). The Waterbirds experiment illustrates the remaining failure mode: when the relevant subgroups are not expressed in the learner representation, neither label balancing nor geometry coverage can be expected to recover them. The labeled-dataset rows here validate the BALANCE branch; the no-label and coarse-label experiments (Tables 5, 7) separately instantiate the COVER branch. So the routing is predictive, not a post-hoc rationalization.

Reproducibility. All selectors are the few-line numpy routines of §4 (RBF facility location, k-center, ridge leverage, class-balanced top-k); the synthetic suite runs in seconds, and the real-data results use frozen ImageNet ViT-B/16 features over 3–5 seeds. Full hyperparameters — kernel bandwidth, the facility-location objective, prototypicality, ridge λ, the MLP architecture and optimizer, and each dataset's pool/budget construction — are in Appendix F; code and exact configurations are released with the paper.

5 Related work

Per-point data valuation. A large literature assigns each training point a scalar worth: Data Shapley (Ghorbani & Zou, 2019) and its efficient or reweighted variants — KNN-Shapley (Jia et al., 2019), Beta-Shapley (Kwon & Zou, 2022), Data Banzhaf (Wang & Jia, 2023), the learning-algorithm-agnostic LAVA (Just et al., 2023), and learned valuation (Yoon et al., 2020) — alongside gradient / influence at-

Table 9: **Diagnostic validation** (frozen ViT, 3 seeds). Effective rank and mean off-diagonal Gram mass measure redundancy (whether a raw top-k is risky); NMI/ARI of k-means(C) clusters with the labels measure whether the labels name the geometric modes; *wLMM* (within-label multimodality, the average per-class effective rank) measures residual within-label structure, lower meaning the labels are more nearly unimodal. With a *pre-fixed* alignment threshold $\tau = 0.45$, all five labeled datasets have high alignment and modest wLMM and route to BALANCE, matching the empirically best selector family (Table 10); the no-label and coarse-label settings instead route to coverage (text). The effective rank ($\sim$25 for all, the intrinsic dimension) is a redundancy signal, not the class count.

dataset	C	eff. rank	off-diag	NMI(C)	ARI(C)	wLMM	diagnostic branch
CIFAR-10	10	23.6	0.38	0.749	0.681	15.9	BALANCE (labels = modes)
Oxford Pets	37	26.4	0.37	0.841	0.670	12.3	BALANCE
CIFAR-100	100	25.6	0.38	0.638	0.288	10.3	BALANCE
Flowers-102	102	26.9	0.39	0.864	0.616	5.9	BALANCE
Food-101	101	24.8	0.37	0.544	0.187	10.6	BALANCE

tribution: influence functions (Koh & Liang, 2017), TracIn (Pruthi et al., 2020), datamodels (Ilyas et al., 2022), TRAK (Park et al., 2023), and memorization estimates (Feldman & Zhang, 2020). All return per-point numbers. Our claim is not that every conceivable scalar score fails, but that using a fixed, mode-blind, non-adaptive per-point ranking as a top-k selector is provably suboptimal in the redundant latent-mode regime. Our real-image experiments instantiate the critique with geometry-based scores and KNN-Shapley; gradient-based and datamodel-style valuations are covered by the theory when used as fixed mode-blind rankings, while a full large-scale empirical evaluation of those families is left to future work. This complements two recent observations specific to Data Shapley: for data-*selection* tasks it can perform no better than random absent a structured utility function (Wang et al., 2024), and as a filtering layer for *batch* active learning it must be paired with an explicit diversity / coverage criterion (Ghorbani et al., 2022). Our contribution is the exact occupancy gap and the modes-vs-labels boundary that determine when balancing suffices and when explicit coverage is needed.

Coresets and data pruning. Set-level selection instead optimizes a subset: submodular coresets (Wei et al., 2015; Mirzasoleiman et al., 2020; Killamsetty et al., 2021), facility-location summarization (Lin & Bilmes, 2011), k-center / core-set active learning (Sener & Savarese, 2018), and leverage / Nyström sampling (Drineas & Mahoney, 2005). Difficulty- or dynamics-based pruning scores — forgetting events (Toneva et al., 2019), EL2N/GraNd (Paul et al., 2021), proxy selection (Coleman et al., 2020) — are again per-point, while data-pruning at scale shows that retaining a *coverage* of easy and hard examples is what matters (Sorscher et al., 2022). We supply the theory for *why* an explicit set-level coverage or diversity correction is needed in the mode-mismatched regime, the exact scalar gap δ, and the modes-vs-labels boundary that says when the cheap class-balance proxy already suffices.

Data selection for language models. Curating pretraining and fine-tuning mixtures is now central: importance resampling (Xie et al., 2023a) and domain reweighting (Xie et al., 2023b) select or balance *groups*. Our boundary predicts when balancing the available groups is enough (group = mode) versus when label-free coverage in the model's own representation is required.

Marginal screening vs. forward selection. Our gap is the classical failure of univariate screening under collinearity — marginal / sure-independence screening (Fan & Lv, 2008) versus orthogonal matching pursuit and forward selection (Tropp, 2004). We instantiate it for data valuation, giving its exact form (δ, occupancy), the three-condition phase boundary, and the counterintuitive "more accurate score $\Rightarrow$ worse selection."

Subgroup robustness. The Waterbirds boundary connects to spurious-correlation robustness (Sagawa et al., 2020; Liu et al., 2021): there the modes are subgroups a generic representation does not express, so neither balancing the labels nor coverage in that geometry recovers them — precisely our learner-geometry condition.

6 Limitations and conclusion

The results are for non-adaptive valuations and a fixed (probe) learner; the latent- and soft-mode analyses are clean instances of a general coherence-governed gap whose matching constants we leave open, and the real experiments use frozen ViT features with linear and small-MLP learners (the gap persists for an SGD-trained MLP, §4) rather than end-to-end backbone fine-tuning, which together with LLM/LoRA validation is a natural next step. A second axis lies *outside* our model. The theory treats a point's value as coverage of the learner's modes, but when point *quality* varies independently of coverage, a quality scalar can dominate. This is sharpest in imitation learning or embodied-control data, where demonstrations of the same behavior can differ substantially in execution quality. Pure geometric coverage may deliberately select trajectory outliers that are unique in feature space but low-quality or near-failure in execution — a set well spread over the geometry but poor for policy learning. The controlled corruption experiment in Table 8 is our formal evidence for this orthogonal quality axis: pure coverage selects almost only corrupted outliers and collapses, while quality-filter-then-coverage avoids them and performs best. This suggests a *decoupled* recipe for such domains — first restrict to a quality-controlled pool, then apply geometric coverage within it — and formalizing the quality–coverage trade-off for imitation learning, reinforcement learning, and embodied data is an important extension. The message for the regimes we model is sharp but scoped. A point's value is a context function, not a number. In the regimes where selection matters most — a binding budget, a redundant learner geometry, and non-interpolable modes — a raw fixed per-point top-k can concentrate and under-cover, and should not be used without a balancing or coverage correction. Which correction is appropriate depends on one question: *do the labels available at selection name the learner's modes?* If yes, class-balance the scores; this recovers the missing coverage and can outperform generic set-level coverage. If no — because labels are absent, coarser than the modes, or not aligned with the learner geometry — add an explicit learner-geometry coverage or diversity correction. **Finally, if the target modes are not expressed in the representation, coverage in that geometry has no handle on them; representation learning or subgroup discovery is then the missing step.**

Broader Impact Statement

This work studies how training subsets are chosen, and subset choice shapes who and what a model represents. Two implications deserve care. First, our Waterbirds result shows that geometric coverage inherits the blind spots of the representation it covers: subgroups the encoder does not express will not be recovered by coverage (nor by label balancing), so practitioners using any selector studied here should audit subgroup representation in the selected set directly rather than assume diversity objectives protect minorities. Second, selection criteria can be dual-use: the same diagnostics that improve budgeted training could be used to systematically exclude data sources. On the positive side, replacing opaque per-point scores with an explicit, auditable balance-or-cover rule makes the composition of the selected set easier to inspect and contest.

References

C. Coleman, C. Yeh, S. Mussmann, B. Mirzasoleiman, P. Bailis, P. Liang, J. Leskovec, M. Zaharia. Selection via proxy: efficient data selection for deep learning. *ICLR*, 2020.

P. Drineas, M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 2005.

J. Fan, J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 2008.

V. Feldman, C. Zhang. What neural networks memorize and why: discovering the long tail via influence estimation. *NeurIPS*, 2020.

A. Ghorbani, J. Zou. Data Shapley: equitable valuation of data for machine learning. *ICML*, 2019.

A. Ilyas, S. Park, L. Engstrom, G. Leclerc, A. Madry. Datamodels: predicting predictions from training data. *ICML*, 2022.

A. Ghorbani, J. Zou, A. Esteva. Data Shapley valuation for efficient batch active learning. *56th Asilomar Conference on Signals, Systems, and Computers*, pp. 1456–1462, IEEE, 2022.

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gürel, B. Li, C. Zhang, C. J. Spanos, D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 2019.

H. A. Just, F. Kang, J. T. Wang, Y. Zeng, M. Ko, M. Jin, R. Jia. LAVA: data valuation without pre-specified learning algorithms. *ICLR*, 2023.

K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, A. De, R. Iyer. GRAD-MATCH: gradient matching based data subset selection for efficient deep model training. *ICML*, 2021.

P. W. Koh, P. Liang. Understanding black-box predictions via influence functions. *ICML*, 2017.

Y. Kwon, J. Zou. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. *AISTATS*, 2022.

H. Lin, J. Bilmes. A class of submodular functions for document summarization. *ACL*, 2011.

E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, C. Finn. Just train twice: improving group robustness without training group information. *ICML*, 2021.

A. W. Marshall, I. Olkin, B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

B. Mirzasoleiman, J. Bilmes, J. Leskovec. Coresets for data-efficient training of machine learning models (CRAIG). *ICML*, 2020.

G. L. Nemhauser, L. A. Wolsey, M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 1978.

S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, A. Madry. TRAK: attributing model behavior at scale. *ICML*, 2023.

M. Paul, S. Ganguli, G. K. Dziugaite. Deep learning on a data diet: finding important examples early in training. *NeurIPS*, 2021.

G. Pruthi, F. Liu, S. Kale, M. Sundararajan. Estimating training data influence by tracing gradient descent. *NeurIPS*, 2020.

S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. *ICLR*, 2020.

O. Sener, S. Savarese. Active learning for convolutional neural networks: a core-set approach. *ICLR*, 2018.

B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, A. S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *NeurIPS*, 2022.

M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, G. J. Gordon. An empirical study of example forgetting during deep neural network learning. *ICLR*, 2019.

J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 2004.

J. T. Wang, R. Jia. Data Banzhaf: a robust data valuation framework for machine learning. *AISTATS*, 2023.

J. T. Wang, T. Yang, J. Zou, Y. Kwon, R. Jia. Rethinking data Shapley for data selection tasks: misleads and merits. *ICML*, 2024.

K. Wei, R. Iyer, J. Bilmes. Submodularity in data subset selection and active learning. *ICML*, 2015.

S. M. Xie, S. Santurkar, T. Ma, P. Liang. Data selection for language models via importance resampling. *NeurIPS*, 2023.

S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, A. W. Yu. DoReMi: optimizing data mixtures speeds up language model pretraining. *NeurIPS*, 2023.

J. Yoon, S. Arik, T. Pfister. Data valuation using reinforcement learning. *ICML*, 2020.

Appendix

The appendix gives complete proofs of all results in the main text. Appendix A derives the closed form (1) and proves the no-interaction geometry theorem (Theorem 1). Appendix B proves the latent-mode Lemmas 1–2; Appendix C computes leave-one-out and Data Shapley exactly; Appendix D proves the interaction-gap Theorem 2 and its corollaries; and Appendix E gives the soft-mode (spiked-covariance) extension. Appendix F gives experimental details (selectors, learners, features, and per-dataset configurations), and Appendix G records the numerical verification.

A The coverage risk and the marginal-value closed form

Setup. For a kernel/linear learner and *point* selection, the quantity the selected set must reconstruct is the feature second moment $\Sigma = \mathbb{E}_x[\varphi\varphi^\top] \succeq 0$, where $\varphi = \varphi(x)$ is the feature map. Writing P_S for the orthogonal projector onto $\mathrm{span}\{\varphi_i : i \in S\}$, the (population) risk of the best learner supported on S is the uncovered second-moment mass

$$\mathrm{R}(S) = \mathrm{tr}\big((I - P_S)\,\Sigma\big) = \sum_\ell \sigma_\ell \,\big\|(I - P_S)e_\ell\big\|_\Sigma^2, \tag{2}$$

which is nonincreasing in S (adding a column can only enlarge the range of P_S), so $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) \geq 0$.

Derivation of the closed form (1). Let $w = \varphi_i^{\perp(S)}/\|\varphi_i^{\perp(S)}\|$ with $\varphi_i^{\perp(S)} = (I - P_S)\varphi_i$ the component of φ_i orthogonal to the current span (assume $\varphi_i^{\perp(S)} \neq 0$; otherwise i is already in the span and $g_i(S) = 0$). Adding φ_i extends the projector by exactly the rank-one term along w:

$$P_{S\cup\{i\}} = P_S + ww^\top.$$

Substituting into (2),

$$g_i(S) = \mathrm{tr}\big((P_{S\cup\{i\}} - P_S)\Sigma\big) = \mathrm{tr}\big(ww^\top\Sigma\big) = w^\top\Sigma\, w = \frac{\langle \varphi_i^{\perp(S)},\, \Sigma\,\varphi_i^{\perp(S)}\rangle}{\|\varphi_i^{\perp(S)}\|^2},$$

which is (1): the marginal value is the Σ-Rayleigh quotient of the point's feature *after orthogonalizing against the selected set* (a Gram–Schmidt / partial-correlation residual). □

Theorem 1 (no-interaction geometry, restated). *If $\varphi_i \perp \varphi_j$ for all $i \neq j$, then for every context $S \not\ni i$ the marginal $g_i(S)$ equals the constant $\langle \varphi_i, \Sigma\varphi_i\rangle/\|\varphi_i\|^2$, independent of S. Conversely, with redundant features the marginal becomes context-dependent through two distinct mechanisms:* (i) directional rotation *of the residual when its direction changes and Σ is anisotropic on the shared directions, and* (ii) rank collapse *when a context drives the residual to zero, both controlled by the off-diagonal mass of $G = \Phi\Phi^\top$.*

Proof. If the features are mutually orthogonal then for any $S \not\ni i$ we have $P_S\varphi_i = 0$, hence $\varphi_i^{\perp(S)} = \varphi_i$ and (1) gives the stated constant for every such S; the interaction $\mathrm{Int}_i = 0$. Conversely, suppose φ_i has a nonzero component along $\mathrm{span}\{\varphi_j : j \in S\}$ for some S. *(i)* If the residual stays nonzero, its unit direction $w_i(S) = \varphi_i^{\perp(S)}/\|\varphi_i^{\perp(S)}\|$ varies with S, and since $g_i(S) = w_i(S)^\top\Sigma\, w_i(S)$ is a Rayleigh quotient, an anisotropic Σ makes it vary as $w_i(S)$ rotates; Proposition 1(i) bounds this by $(\lambda_{\max} - \lambda_{\min})$ times the residual-direction diameter. *(ii)* If some S_0 makes $\varphi_i^{\perp(S_0)} = 0$, then $g_i(S_0) = 0$ regardless of Σ (Proposition 1(ii)). The latent-mode instance of Appendix B realises mechanism (ii) in the extreme, where the range is $\mathrm{Int}_i = \rho_{j(i)} = \Omega(1)$. □

This is the classical statement that *marginal (univariate) screening differs from forward selection exactly by the redundancy that orthogonalization removes*; here applied to data valuation. The off-diagonal Gram is the *contribution covariance* referred to in the main text.

B The latent-mode instance: Lemmas 1–2

Recall the instance: $P = \bigsqcup_{j=1}^{r} P_j$ with r modes of size $m = N/r$, a block learner geometry in which covering ≥ 1 point of mode j drives that mode's residual to 0, and an uncovered mode costs $\rho_j > 0$. Equation (2) then reduces to

$$\mathrm{R}(S) = \sum_{j:\, S \cap P_j = \emptyset} \rho_j. \tag{3}$$

Lemma 1 (all-or-nothing marginal, restated). *For $i \in P_j$, $g_i(S) = \rho_j \, \mathbf{1}[S \cap P_j = \emptyset] \in \{0, \rho_j\}$, so $\mathrm{Int}_i = \rho_j$. Any scalar $v_i \in [0, \rho_j]$ disagrees with the true marginal on an $\Omega(1)$ fraction of contexts.*

Proof. Take $i \in P_j$ and any context S with $i \notin S$. If $S \cap P_j = \emptyset$, mode j is uncovered in S but covered in $S \cup \{i\}$, so by (3) $g_i(S) = \mathrm{R}(S) - \mathrm{R}(S \cup \{i\}) = \rho_j$. If $S \cap P_j \neq \emptyset$, mode j is already covered and remains so, the term ρ_j is absent from both risks, and no other mode's coverage changes by adding a mode-j point, so $g_i(S) = 0$. Hence $g_i(S) = \rho_j \mathbf{1}[S \cap P_j = \emptyset]$ and $\mathrm{Int}_i = \rho_j - 0 = \rho_j$. Finally, a fixed scalar v_i predicts a single value, whereas the true marginal takes both 0 and ρ_j on positive-probability context families (e.g. random S of size k with $k < r$ has $\Pr[S \cap P_j = \emptyset] = \Omega(1)$ and $\Pr[S \cap P_j \neq \emptyset] = \Omega(1)$); so $|v_i - g_i(S)| \geq \rho_j/2$ on at least one of the two families, an $\Omega(1)$ fraction. □

Lemma 2 (exchangeability collapses scores, restated). *Every valuation that is a symmetric functional of the training distribution is constant on each mode: $v_i = c_j$ for all $i \in P_j$.*

Proof. Points within a mode are distributionally identical (exchangeable): there is a measure-preserving relabelling of the pool that swaps any $i, i' \in P_j$ and fixes all other points and the distribution D. Leave-one-out, Data Shapley, leverage and probe-gap are all symmetric functionals — their definitions are invariant to the names of the points and depend only on the multiset of features/labels. Applying the swap leaves the functional unchanged, so $v_i = v_{i'}$. Hence v is constant on P_j. □

C The two canonical scalars, computed exactly

Leave-one-out / influence. By definition $v_i^{\mathrm{LOO}} = g_i(P \setminus \{i\}) = \mathrm{R}(P \setminus \{i\}) - \mathrm{R}(P)$. At full data every mode is covered, so $\mathrm{R}(P) = 0$. For $i \in P_j$ with $m \geq 2$, the set $P \setminus \{i\}$ still contains the other $m - 1 \geq 1$ points of mode j, so mode j stays covered and $\mathrm{R}(P \setminus \{i\}) = 0$. Therefore $v_i^{\mathrm{LOO}} = 0$ for *every* point: in the redundant regime each point is individually removable and influence sees no value, leaving top-k an arbitrary tie-break.

Data Shapley. For a random permutation π and $i \in P_j$, the marginal $g_i(S^{\pi}_{<i})$ equals ρ_j if no point of mode j precedes i in π (mode j still uncovered when i arrives) and 0 otherwise (Lemma 1). Hence

$$v_i^{\mathrm{Sh}} = \mathbb{E}_{\pi}\, g_i(S^{\pi}_{<i}) = \rho_j \Pr_{\pi}[\, i \text{ is the first point of mode } j \,] = \frac{\rho_j}{m_j},$$

because the m_j points of mode j are equally likely to be the first of their mode in a uniform random order. For equal modes v^{Sh} is constant over *all* points; for unequal modes it is mode-constant (up-weighting rare modes) but still mode-measurable, hence subject to Lemma 2 and the gap below.

D Proof of the interaction gap

Throughout, $\mathrm{R}(S) = \rho\,(r - N_{\mathrm{cov}}(S))$ with $N_{\mathrm{cov}}(S) = \#\{j : S \cap P_j \neq \emptyset\}$ (Eq. (3)), so comparing selectors reduces to comparing expected covered modes. Greedy facility location gains ρ from a fresh mode and 0 from a covered one (Lemma 1), so it covers a new mode each step: $N_{\mathrm{cov}}(S_{\mathrm{FL}}) = \min(k, r) = k$ for $k \leq r$, exactly optimal (a fortiori $(1 - 1/e)$-optimal by submodularity (Nemhauser et al., 1978)). We use a uniform tie-breaking rule throughout (Theorem 2's hypothesis).

Equality for symmetric valuations (Theorem 2). With *equal* modes the whole pool is exchangeable: any relabelling permuting points within *and across* modes preserves the data distribution. A data-dependent symmetric valuation is invariant under such relabellings, so it assigns one common value to all $N = rm$ points. Its top-k set, ties broken uniformly, is therefore a uniformly random k-subset of P. A fixed mode j is uncovered iff none of its m points is drawn, whence

$$\Pr[\text{mode } j \text{ missed}] = \frac{\binom{N-m}{k}}{\binom{N}{k}} = \frac{\binom{(r-1)m}{k}}{\binom{rm}{k}}, \qquad \mathbb{E}[N_{\text{cov}}] = r\Big(1 - \frac{\binom{(r-1)m}{k}}{\binom{rm}{k}}\Big) = m_\star(k),$$

exact for finite N; as $N \to \infty$ with $m/N = 1/r$, $\binom{(r-1)m}{k}/\binom{rm}{k} \to (1-1/r)^k$, recovering $m_\star(k) \to r(1-(1-1/r)^k)$. Subtracting from greedy gives $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) = \rho\,(k - m_\star(k)) = \rho\,\delta_N(1,k)$, strictly positive for $0 < k < r$ since $k > m_\star(k)$ (each successive uniform draw covers a fresh mode with probability < 1 once a mode is hit). □

Upper bound for arbitrary scores (Proposition 2, equal modes of size m). A general within-mode-exchangeable score is mode-constant, $v_i = c_j$ on P_j (Lemma 2), but the c_j may differ. (Equal sizes are used in the pooling step below; Remark 1 shows the bound can fail for unequal modes.) Given the (tie-break-random) per-mode counts $s_j = |S_v \cap P_j|$ with $\sum_j s_j = k$, within-mode exchangeability makes the chosen s_j points uniform inside P_j, and mode j is covered iff $s_j \geq 1$; thus $N_{\text{cov}} = \sum_j \mathbf{1}[s_j \geq 1]$ is a function of the count vector $s = (s_1, \ldots, s_r)$ alone. We show the uniform score (whose s is multivariate hypergeometric, the Theorem 2 case) maximizes $\mathbb{E}[N_{\text{cov}}]$.

Lemma 3 (pooling does not decrease coverage)**.** *Fix all per-mode counts except two modes a, b with $s_a + s_b = t$.* Pooling *the pair — redrawing the t points uniformly without replacement from the $2m$ points of $P_a \cup P_b$ — gives the pair expected coverage $2 - 2\binom{m}{t}/\binom{2m}{t}$. When the original allocation is* concentrated *($s_a \in \{0, t\}$, so the pair covers 1), pooling does not decrease coverage:*

$$2 - 2\,\frac{\binom{m}{t}}{\binom{2m}{t}} \;\geq\; 1 \qquad (t \geq 1),$$

with strict inequality for $t \geq 2$.

Proof. After pooling, a given mode of the pair is missed iff all t points fall in the other mode's m slots, probability $\binom{m}{t}/\binom{2m}{t}$; summing the complement over the two modes gives the stated pair coverage. The inequality is $\binom{m}{t}/\binom{2m}{t} \leq \frac{1}{2}$, i.e. the t draws are at least as likely to touch both halves as to fall in one; for $t \geq 2$ the event "all in one half" has probability $< \frac{1}{2}$ strictly (e.g. $\binom{m}{2}/\binom{2m}{2} = \frac{m-1}{2(2m-1)} < \frac{1}{2}$). □

From pairwise to global. By Lemma 2 a within-mode-exchangeable score is mode-constant; with uniform tie-breaking its top-k takes whole value-tiers and a uniform subset of the single active tier, so within a tier the counts are already exchangeable and every *split* pair (one mode fully taken, the other untaken) is concentrated — exactly the case of Lemma 3. Repeatedly pooling split pairs weakly increases $\mathbb{E}[N_{\text{cov}}]$ and converges to the fully exchangeable law, the uniform k-subset, with $\mathbb{E}[N_{\text{cov}}] = m_\star(k)$ (Theorem 2). Hence $\mathbb{E}[N_{\text{cov}}(S_v)] \leq m_\star(k)$, with equality iff the mode-values are equal. This is the majorization statement that $\sum_j \mathbf{1}[s_j = 0]$ is Schur-convex in the counts and the uniform allocation is majorized by every tiered one (Marshall et al., 2011). Subtracting from greedy gives $\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \geq \rho\,\delta_N(1,k)$. □

Corollary (mode signal can help or hurt). For equal modes, $\mathbb{E}[N_{\text{cov}}]$ is maximised at zero mode signal and decreases as the signal concentrates the budget within a few modes, so a concentrating score can fall *below* uniform random at a binding budget. With *unequal* sizes the effect can go either way (Remark 1): a score whose high values spread budget across *rare* modes can cover *more* than uniform, whereas one that concentrates within a few large modes covers fewer. The unconditional lower bound is therefore specific to equal-size modes; in general, whether mode signal helps or hurts coverage depends on whether it spreads the budget across modes or concentrates it.

Corollary (phase boundary, restated). A positive scalar-selection gap $\rho\,\delta_N(1,k) > 0$ requires all three of: (i) $k < r$ (budget binds; else $\delta_N = 0$), (ii) $\rho > 0$ (task non-interpolable across modes), and (iii) a redundant learner geometry, i.e. $r \ll N$ with off-diagonal Gram mass, so that there exist modes to under-cover. If any fails, the scalar top-k is optimal in the model. Condition (iii) is redundancy of the *learner* geometry, not non-locality of the representation.

E Soft modes: the spiked-covariance extension

Replace hard modes by r orthonormal directions $u_1, \dots, u_r$ with masses $\sigma_1 \geq \cdots \geq \sigma_r$ and within-mode noise η: a point of mode j is $\varphi_i = u_j + \eta\,\xi_i$ with $\xi_i \perp \mathrm{span}\{u_\ell\}$ i.i.d. isotropic, so

$$\Sigma = \sum_{j=1}^{r} \pi_j\, u_j u_j^\top + \eta^2 P_\perp, \qquad \pi_j \propto \sigma_j,$$

with $P_\perp$ the projector onto the orthogonal complement of $\mathrm{span}\{u_\ell\}$.

Soft Lemma 1. *$g_i(S) = \sigma_{j(i)}\,\mathbf{1}[\text{mode } j(i) \text{ uncovered by } S] + O(\eta^2)$.*

Proof. Uncovered mode. If no selected point shares mode $j = j(i)$, then $u_j \notin \mathrm{span}\{\varphi_\ell : \ell \in S\}$ up to $O(\eta)$, so $\varphi_i^{\perp(S)} = \varphi_i + O(\eta)$. Using $\Sigma\varphi_i = \pi_j u_j (u_j^\top \varphi_i) + \eta^2 P_\perp \varphi_i = \pi_j u_j + \eta^3 \xi_i$ and $\|\varphi_i\|^2 = 1 + \eta^2 \|\xi_i\|^2$,

$$g_i(S) = \frac{\langle \varphi_i, \Sigma\varphi_i \rangle}{\|\varphi_i\|^2} + O(\eta^2) = \frac{\pi_j + O(\eta^4)}{1 + O(\eta^2)} + O(\eta^2) = \sigma_j + O(\eta^2),$$

identifying $\sigma_j \equiv \pi_j$. (The apparent first-order correction from $\varphi_i^{\perp(S)} = \varphi_i + O(\eta)$ lies in the noise subspace orthogonal to the mode directions; since Σ has no first-order cross term between the mode subspace and the noise subspace, its contribution to the Rayleigh quotient is $O(\eta^2)$.) *Covered mode.* If some selected point shares mode j, then $u_j \in \mathrm{span}$ up to $O(\eta)$, so $\varphi_i^{\perp(S)} = \eta\,\xi_i^\perp + O(\eta^2)$ is pure noise of norm $O(\eta)$. Since $\xi_i \in \mathrm{range}(P_\perp)$, $\Sigma\varphi_i^{\perp(S)} = \eta^2 \varphi_i^{\perp(S)} + O(\eta^3)$, whence the Rayleigh quotient is $g_i(S) = \eta^2 + O(\eta^3) = O(\eta^2)$. □

As $\eta \to 0$ this recovers the $\{0, \sigma_j\}$ dichotomy of Lemma 1, and the interaction $\mathrm{range}_S\, g_i \approx \sigma_{j(i)}$ stays $\Omega(1)$.

Soft theorem. Within-mode exchangeability (Lemma 2) is unchanged, so the occupancy argument of Appendix D applies to the leading term, and the $O(\eta^2)$ covered-mode residuals contribute at most $O(\eta^2 k)$ to either risk:

$$\mathbb{E}[\mathrm{R}(S_v)] - \mathrm{R}(S_{\mathrm{FL}}) \;\geq\; \rho\,\delta(1,k) - O(\eta^2 k) \;>\; 0 \quad \text{for } \eta \text{ small and } 0 < k < r.$$

Spectral reading. The σ_j are the eigenvalues of Σ, r is the effective rank, and η^2 is the tail. The gap is large for a *peaked* spectrum (low effective rank $\Rightarrow$ redundant, large off-diagonal Gram) and vanishes as the spectrum *flattens* (high effective rank $\Rightarrow$ near-orthogonal points, diagonal Gram) — the controlled monotone law of the noise sweep (Table 2). For general Σ the problem reduces to the classical marginal-screening-versus-orthogonal-matching-pursuit gap of sparse approximation, governed by the Gram coherence $\mu = \max_{i \neq j} |G_{ij}|$; the spiked model is the clean computable instance.

F Experimental details

Selectors. The kernel is RBF with the median-heuristic bandwidth $\sigma^2 = \mathrm{median}_{i \neq j} \|\varphi_i - \varphi_j\|^2$. *Greedy facility location* maximizes $\sum_j \max_{i \in S} K_{ji}$ by the standard greedy step; *ridge leverage* is $\tau_i = [K(K + \lambda I)^{-1}]_{ii}$ with $\lambda = 0.1$; *k-center* is farthest-first from the medoid. *Prototypicality* is $-\|\varphi_i - \bar{\varphi}_{c(i)}\|$ (distance to the class mean) and *global typicality* is $-\|\varphi_i - \bar{\varphi}\|$. Class-balanced selectors take $\lfloor k/C \rfloor$ points per class, ranking within a class by the scalar and filling any shortfall at random.

KNN-Shapley. We compute exact KNN-Shapley (Jia et al., 2019) on the frozen ViT feature space after PCA projection, with 10 nearest neighbours and Euclidean distance. The valuation set is a held-out validation split drawn from the training pool (1500 examples per seed), disjoint from both the selected training subset and the final test set; the scores are computed once per seed before selection. The same scores are used two ways — raw top-k, and class-balanced (labels only impose equal per-class quotas, with points ranked within class by their KNN-Shapley value) — so the comparison isolates the *selection rule*: the per-point values are unchanged while the raw top-k procedure is replaced by a coverage-preserving quota.

Learners. The probe is one-hot ridge regression, $W = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y$ with $\lambda = 0.1$, predicting $\arg\max$. The MLP (Table 6) is $64 \to 256 \to C$ with ReLU and dropout 0.3, trained by Adam (learning rate 10^{-3}, weight decay 10^{-4}) for 300 steps with early stopping on a 10% held-out split of the selected set.

Features and data. Images are resized to 224^2, normalized with ImageNet statistics, and embedded by a frozen timm ViT-B/16 (`vit_base_patch16_224`); the embeddings are PCA-projected to 64 dimensions (fit on the pool). Each seed draws a fresh pool of $N \leq 5000$ training points and as many test points; budgets are $k \in \{C, 2C, 4C\}$, binding at $k = C$. CIFAR-10/100, Oxford Pets, Flowers-102, and Food-101 use their standard splits; Waterbirds uses the WILDS subgroup metadata. We report means over 3–5 seeds with standard errors. The synthetic suite uses $C = 10$ classes with 3 input-distant modes each, a non-degenerate embedding, and the same ridge probe.

Facility location vs. the theory's greedy coverage, and scalability. The theory's greedy coverage adds, at each step, a point of an uncovered mode; RBF facility location is its empirical instantiation, with the *recovered* modes depending on the kernel bandwidth, the PCA dimension, and feature normalization (the median heuristic and PCA-64 above worked across all datasets without tuning; the synthetic redundancy sweep, Table 2, isolates the bandwidth/effective-rank dependence). Facility location needs the $N \times N$ Gram, i.e. $O(N^2)$ time and memory, which is the relevant cost at the binding budgets we study (k in the hundreds, pools of thousands). For larger pools, standard approximations include lazy (accelerated) greedy, k-NN graph sparsification, approximate-neighbor kernels, and stochastic-greedy or streaming submodular maximization; these trade exactness for scalability, and evaluating them at larger scale is orthogonal to the balance-vs-cover boundary studied here.

Table 10: **Full per-selector test accuracy at** $k = C$ (modes = labels; mean ± se, 5 seeds), complementing Table 3, which reports only the gap to the *best* selector. Class-*balanced* prototypicality is the single best selector on every dataset; raw per-point scalars (leverage, prototypicality, LOO, Data Shapley) lose to their balanced versions and often to random; greedy coverage is strong but is itself beaten by balanced prototypicality when modes = labels. Data Shapley (a TMC estimate) is omitted on the 100-class datasets for cost.

selector ($k = C$)	CIFAR-10	Oxford Pets	CIFAR-100	Flowers-102
raw scalar / random				
random	$0.554_{\pm.017}$	$0.453_{\pm.026}$	$0.246_{\pm.010}$	$0.479_{\pm.005}$
leverage	$0.373_{\pm.037}$	$0.413_{\pm.015}$	$0.193_{\pm.010}$	$0.364_{\pm.013}$
prototypicality	$0.353_{\pm.021}$	$0.187_{\pm.010}$	$0.231_{\pm.007}$	$0.100_{\pm.007}$
LOO / influence	$0.419_{\pm.060}$	$0.235_{\pm.029}$	$0.275_{\pm.005}$	$0.459_{\pm.006}$
Data Shapley	$0.563_{\pm.032}$	$0.505_{\pm.017}$	—	—
class-balanced (same scores, equal per-class quota)				
bal-random	$0.672_{\pm.012}$	$0.606_{\pm.008}$	$0.348_{\pm.007}$	$0.762_{\pm.013}$
bal-leverage	$0.620_{\pm.019}$	$0.592_{\pm.014}$	$0.276_{\pm.005}$	$0.746_{\pm.004}$
bal-prototypicality	$\mathbf{0.851}_{\pm.007}$	$\mathbf{0.675}_{\pm.010}$	$\mathbf{0.475}_{\pm.005}$	$\mathbf{0.852}_{\pm.005}$
bal-LOO	$0.579_{\pm.031}$	$0.626_{\pm.036}$	$0.365_{\pm.012}$	$0.761_{\pm.009}$
bal-Shapley	$0.781_{\pm.015}$	$0.638_{\pm.014}$	—	—
coverage				
k-center	$0.527_{\pm.022}$	$0.552_{\pm.014}$	$0.249_{\pm.006}$	$0.649_{\pm.007}$
greedy facility location	$0.759_{\pm.020}$	$0.671_{\pm.008}$	$0.394_{\pm.008}$	$0.681_{\pm.002}$

G Numerical verification

All analytic claims are reproduced by the self-contained synthetic suite (`experiments.py`, numpy, ∼1.5s): (i) the marginal $g_i(S)$ over random contexts has spread/|mean| ≈ 1.5 and is bimodal (40% large, 57% ≈ 0), matching Lemma 1; (ii) leave-one-out is $\equiv 0$ to machine precision and Data Shapley equals ρ/m to three digits, matching Appendix C; (iii) at the binding budget $k = C$ greedy coverage scores 0.895 versus best per-point top-k 0.637 (gap +0.258), and *estimated* Data Shapley top-k dips below random near $k = C$, matching Theorem 2 and the concentration corollary; (iv) the soft-mode marginals are $0.117 \approx \pi_j$ (uncovered) and $0.004 = O(\eta^2)$ (covered), matching Appendix E; (v) sweeping embedding noise, the $k = C$ gap tracks the off-diagonal block mass monotonically and collapses $+0.35 \to 0$ as the effective rank rises $15 \to 30$ (Table 2), the controlled causal law.

Direct validation of the occupancy theorem. Independently of the ML suite, we check the latent-mode combinatorics directly (`verify_theory.py`, 2×10^4 trials, Table 11): for $r = 10$ equal modes of size $m = 5$, a uniform k-subset — the top-k of any symmetric valuation (Theorem 2) — covers exactly $m_\star(k)$ to three digits (max error 0.007 across k); a concentrating mode-constant score stays at or below $m_\star(k)$ (Proposition 2); greedy covers $\min(k, r) = k$; and with *unequal* sizes (nine singletons and one mode of 91 points) a rare-favouring score covers 9.0 modes versus the uniform 1.8, exactly the reversal of Remark 1.

Table 11: Direct validation of the occupancy theorem ($r = 10$ modes, equal size $m = 5$ unless noted; 2×10^4 trials). Uniform sampling matches the closed form $m_\star(k)$ to three digits (Thm 2); a concentrating score stays $\leq m_\star$ (Prop 2); unequal modes reverse the bound (Rem 1).

k	$m_\star(k)$ (theory)	uniform (sim)	concentrating (sim)	greedy
2	1.918	1.918	1.000	2
5	4.234	4.235	1.000	5
8	5.985	5.984	2.000	8
10	6.894	6.899	2.000	10
unequal sizes 9×1+91, $k = 9$:		uniform 1.807,	rare-favouring 9.000 (Rem 1)	