FEATURE GEOMETRY FOR DATA-POOL SCREENING: UTILITY, STABILITY, AND LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting source-data pools for a target task can require training and validating a predictor for many combinations of pools. We propose a geometry-guided framework that shortlists combinations using unlabeled source and target features from frozen encoders, then trains predictors on source labels and selects one by target validation. Its core screening method applies target-weighted A-optimal design to pooled Gram matrices and the target second moment. Under a shared linear model, it favors combinations that reduce uncertainty about prediction coefficients, weighting feature directions by their energy in target data. A separate error decomposition distinguishes discarding good candidates from losing quality when validation chooses among those retained. On equal-sample-cost DomainNet combinations, an exploratory budget analysis with logistic regression finds lower mean test Brier from evaluating 50 A-opt candidates than 227 random candidates. In the original primary study, second-moment matching (MMD) matches A-opt on recall and final loss; it has higher mean recall at smaller budgets. Independent timing shows savings with cached features; adding measured feature-extraction costs preserves the logistic-regression saving but reverses it for ridge. These results support using target-aware geometry to reduce supervised data-pool evaluations, with savings depending on readout and feature preparation costs.

1 INTRODUCTION

Adapting a predictor to a target domain often starts with several source-data pools that differ in acquisition conditions, image style, or example coverage. Choosing a combination by training and validating a predictor on every candidate repeats much of the same work. We ask how geometry available before supervised fitting can reduce these evaluations while preserving a competitive final choice.

Frozen pretrained encoders provide a common comparison space. Their feature vectors can be reused across source combinations; CLIP and DINOv2 support this setting (Radford et al., 2021; Oquab et al., 2024). A classifier trained on these features is called a readout. Comparing feature geometry before fitting readouts provides a way to narrow the candidate set.

Source diversity alone does not identify target usefulness. Effective rank summarizes how evenly variation spans feature directions (Roy & Vetterli, 2007; Garrido et al., 2023). For intuition, if target feature energy lies mainly in shape-related directions, adding source coverage of background variation may matter less. Shape and background illustrate directions here; the screen does not identify these semantics.

We propose a geometry-guided framework for data-pool pre-screening, with Target A-opt as its core scoring method. We apply target-weighted experimental design (Yu et al., 2006; Hübotter et al., 2024) to candidate unions of source pools. Pooled Gram matrices record source directional coverage; the target second moment measures energy in those directions. Algorithm 1 gives the complete procedure: unlabeled features determine a shortlist, source labels train its readouts, and separate target validation chooses one. Geometry reduces fitting work; validation supplies label evidence for the final choice.

Under a shared Bayesian linear model, source coverage reduces uncertainty about the coefficients linking features to predictions. A-opt weights the remaining uncertainty by target energy, estimating

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.

risk beyond irreducible noise. This explains the score, not a classification guarantee. Separately, an error decomposition distinguishes good candidates lost during screening from quality lost when validation chooses among retained candidates. Retaining a good candidate does not ensure it will be selected, so we measure both retention and final loss.

Experiments on equal-sample-cost DomainNet combinations test this framework against random screening, target-blind diversity, and second-moment matching (MMD). With logistic regression, exploratory budget analyses show lower mean test Brier, a probability-prediction loss, for shorter A-opt lists than longer random lists. MMD remains competitive, and domain controls examine the value of block-level geometry. Independent timing finds savings with cached features; adding measured feature work preserves the logistic-regression saving but reverses it for ridge. The benefit is fewer supervised evaluations yielding a competitive final predictor, with actual savings depending on readout and feature costs.

2 Related Work

2.1 Target-aware experimental design

A-optimal design measures parameter uncertainty through a covariance trace (Fontaine et al., 2021). Transductive design instead weights prediction variance by target covariates (Yu et al., 2006). ITL and VTL distinguish where observations can be acquired from where predictions are needed, selecting observations to reduce target entropy or total variance (Hübotter et al., 2024). SIFT applies conditional uncertainty reduction to prompt-specific fine-tuning, accounting for redundancy among retrieved examples (Hübotter et al., 2025). Target A-opt follows the target-variance principle, but ranks fixed pool unions for subsequent supervised evaluation. Separate labeled validation then chooses among the retained combinations.

2.2 Moment matching and data selection

MMD compares distribution mean embeddings (Gretton et al., 2012); a homogeneous quadratic kernel compares second moments. Moment alignment also supports multi-source domain adaptation (Peng et al., 2019). SkMM sketches model gradients to control low-rank approximation bias, then matches moments to control variance during fine-tuning (Dong et al., 2024). Related high-dimensional analysis establishes covariance-matching optimality in specific synthetic-data augmentation regimes (Rezaei et al., 2026). Our MMD comparator matches uncentered frozen-feature moments, rather than inheriting those guarantees.

TSDS optimizes a sampling distribution using transport-based target alignment and density-aware regularization against near-duplicates (Liu et al., 2024). TAROT instead whitens loss-gradient features before transport-based selection (Feng et al., 2025). These methods construct training subsets or sampling weights. We hold candidate membership fixed and compare uncertainty weighting with direct moment matching as ways to reduce supervised evaluations. Appendix D details the different representations, selection units, and information requirements.

2.3 Geometry and computation budgets

RankMe assesses representations through effective rank without downstream labels (Garrido et al., 2023), and determinantal point processes favor complementary subsets (Kulesza & Taskar, 2012). They motivate target-blind comparators. Compute-constrained selection charges scoring against the same budget as training (Yin & Rush, 2025): a better subset need not justify its selection cost. We apply this accounting principle to evaluating pool combinations, measuring both final quality and the work needed to obtain it.

3 TARGET A-OPT DATA-POOL SCREENING

3.1 PROBLEM AND COMPLETE PROCEDURE

A pool is a fixed block of examples; a candidate $S \in \mathcal{C}$ is a feasible union of blocks. The shortlist contains combinations, not individual images. A fixed feature map h comprises the frozen encoder and shared label-independent preprocessing. Let $H_i \in \mathbb{R}^{n_i \times d}$ contain pool i's features, and let H_S, y_S stack the features and labels in S. All candidates share a readout fitting rule $\mathcal{A}$ and a validation loss L, fixed before selection.

Algorithm 1 uses the experimental setting $P_0^{-1} = I$, $\sigma^2 = 1$. The screen reads source features and an unlabeled target-selection set T_{sel}, but no labels or validation/test data. Only after freezing the shortlist does supervised evaluation read source labels and the separate target-validation set T_{val}; the records call these stages U0 and U2.

Algorithm 1 Target A-opt screening and validation

Require: Source pools $\{D_i\}$, candidate family $\mathcal{C}$, target sets $T_{\text{sel}}, T_{\text{val}}$.
Require: Fixed feature map h, fitting rule $\mathcal{A}$, loss L, shortlist size $1 \le s \le |\mathcal{C}|$.
Ensure: Shortlist $\mathcal{Q}_s$, selected combination $\widehat{S}$, and fitted readout $q_{\widehat{S}}$.

Form H_i and H_{sel} using h, without reading labels.
Cache $G_i = H_i^\top H_i$; estimate $\widehat{C}_T = H_{\text{sel}}^\top H_{\text{sel}} / |T_{\text{sel}}|$.
for each $S \in \mathcal{C}$ **do**
    $G_S \leftarrow \sum_{i \in S} G_i$; $f_A(S) \leftarrow \operatorname{tr}\left[\widehat{C}_T (I + G_S)^{-1}\right]$.
end for
Retain the first s candidates ordered by $(f_A(S), \mathrm{id}(S))$ as $\mathcal{Q}_s$.
for each $S \in \mathcal{Q}_s$ **do**
    Fit $q_S \leftarrow \mathcal{A}(H_S, y_S)$ using source labels.
    $v(S) \leftarrow |T_{\text{val}}|^{-1} \sum_{(x,y) \in T_{\text{val}}} L(q_S(h(x)), y)$.
end for
Choose $\widehat{S}$ minimizing $(v(S), \mathrm{id}(S))$ over $\mathcal{Q}_s$.
return $\mathcal{Q}_s, \widehat{S}, q_{\widehat{S}}$.

Both orderings are ascending, with combination identifiers breaking ties lexicographically. All candidates receive geometric scores, but only the s retained candidates require supervised fitting and validation. Summed Grams preserve source directional coverage; the target moment weights its relevance. Validation reuses frozen features and supplies label evidence for the final choice. Test data evaluate this frozen output and candidate retention, not any step of the algorithm.

3.2 WHY TARGET-WEIGHTED UNCERTAINTY GUIDES SCREENING

Uncertainty here concerns how feature changes affect predictions, not image clarity or disorder. We formalize it through scalar Bayesian regression with $\theta \sim N(0, P_0)$, $P_0 \succ 0$, and $y = h^\top \theta + \epsilon$, where independent noise satisfies $\epsilon \sim N(0, \sigma^2)$ and $\sigma^2 > 0$. Sources and target share θ and the noise model; target covariates reveal no further information about θ. With target second moment $C_T = \mathbb{E}_T[hh^\top]$, the posterior covariance and posterior-mean target risk conditional on the source design are

$$P_S = \left(P_0^{-1} + \sigma^{-2} G_S\right)^{-1}, \qquad R_T(S) = \sigma^2 + \operatorname{tr}(C_T P_S). \tag{1}$$

The risk is irreducible noise plus target-weighted parameter uncertainty, averaging the posterior and a new target observation. Appendix A gives the derivation.

The score $f_A(S) = \operatorname{tr}(\widehat{C}_T P_S)$ estimates the second term using unlabeled target features and is minimized with fixed hyperparameters. With an isotropic prior, it is the regularized target-variance criterion of transductive experimental design (Yu et al., 2006; Hübotter et al., 2024); we retain the recorded name Target A-opt. Adding a positive-semidefinite source Gram cannot increase this score.

For the experimental choice $P_0^{-1} = I$ and $\sigma^2 = 1$, write $G_S = U \operatorname{diag}(\lambda_j) U^\top$, including zero eigenvalues. Then

$$f_A(S) = \sum_j \frac{u_j^\top \widehat{C}_T u_j}{1 + \lambda_j}. \tag{2}$$

The reciprocal denominator is the remaining coefficient uncertainty along u_j; the numerator weights it by target feature energy. Source coverage reduces the former, while the latter is not known label importance. Combinations with identical spectra can thus receive different scores through their alignment with the target. The implementation evaluates this eigendecomposition form rather than explicitly forming the inverse in Algorithm 1.

The shared conditional model supplies the risk interpretation of A-opt. Its assumption is substantive: if all features equal 1, source pools labeled constantly $+1$ and -1 produce predictors whose risk ordering reverses between two target worlds with opposite constant labels. The feature geometry is unchanged. In classification, we therefore evaluate the scores using trained ridge-softmax and logistic-regression readouts rather than identify Equation (1) with their losses.

3.3 Moment matching and spectral comparators

Second-moment matching asks a different question: how close is the source's average directional energy to the target's? With $n_S = \sum_{i \in S} n_i$, it minimizes

$$f_M(S) = \left\| G_S / n_S - \widehat{C}_T \right\|_F^2. \tag{3}$$

This is empirical squared MMD with kernel $k(h, h') = (h^\top h')^2$, including diagonal terms. It matches average directional energy rather than weighting posterior uncertainty. Both criteria read the same features, testing whether inverse-covariance weighting adds value beyond direct moment matching.

Effective rank instead summarizes the evenness of singular values without target information (Roy & Vetterli, 2007). It measures spectral shape rather than sample evidence and is a comparator, not a step in Algorithm 1. Appendix A gives its definition, a numerical example, and why marginal spectra alone cannot replace the pooled Gram. These comparisons distinguish source diversity, target moment matching, and target-weighted uncertainty within the same selection procedure.

3.4 Separating retention from final selection

Screening can discard a good candidate; validation can also fail to choose the best retained one. Let $\ell(S)$ be target risk for a fixed trained predictor and $v(S)$ its validation loss. If $|v(S) - \ell(S)| \le \varepsilon$ throughout $\mathcal{Q}_s$, the validation minimizer satisfies

$$\ell(\widehat{S}) - \min_{S \in \mathcal{C}} \ell(S) \le \underbrace{\min_{S \in \mathcal{Q}_s} \ell(S) - \min_{S \in \mathcal{C}} \ell(S)}_{\text{shortlist omission}} + 2\varepsilon. \tag{4}$$

Omission compares the best retained risk with the best full-family risk. The validation-selection gap is the risk difference between the chosen and best retained candidates, not the validation loss itself. The deviation condition bounds this gap by 2ε. The argument needs no shared linear model (Appendix A). High recall need not yield a good final choice, nor does low final loss require retaining every top-ten candidate. We audit both gaps on finite test data alongside recall, without estimating a population guarantee.

3.5 Decision budgets and cost accounting

The full-search reference trains all candidates and chooses $S_{\text{full}} = \arg\min_{S \in \mathcal{C}} v(S)$. This is a deployable procedure. The test oracle, in contrast, is the candidate with lowest observed test loss and is used only for auditing. Let $b(S)$ be test Brier, the mean squared distance between predicted class probabilities and the one-hot label; lower is better. We report $b(\widehat{S}_s) - b(S_{\text{full}})$ and the fraction of repeated tasks exceeding a declared absolute tolerance. The gap can be negative because the validation and test rankings can differ.

With frozen features available, screening costs ranking plus fitting and validating retained candidates. Starting from images adds encoder initialization and feature extraction; target-aware scores also need target-selection features. Screening pays off only if its overhead, including extra feature work, is smaller than the evaluations it replaces. Candidate counts are not elapsed time: the matrix operations used in linear scoring can also comprise a substantial part of fitting ridge readouts.

4 Evaluation Protocol

The studies were conducted in stages. Synthetic tasks test the risk model; PACS and Office-Home (Li et al., 2017; Venkateswara et al., 2017) evaluate direct selection against a prespecified practical-effect threshold, followed by a post-hoc headroom audit. The subsequent DomainNet shortlist configuration was frozen before feature extraction and outcome evaluation. Resampling, projection/readout changes, budget reanalysis, and repartitioning are exploratory follow-ups. Independent timing separately measures the cost of executing the selection procedures. Appendix B records configurations and result locations.

From the official cleaned DomainNet data (Peng et al., 2019), we select 100 class names by a fixed hash rule. Each of six domains contributes disjoint sets of 1,536 target-selection images, 1,024 target-validation images, and 3,072 anchor-pool images from official training data, plus 2,048 official test images. The manifest contains 46,080 images. These roles separate information for screening, final selection, and independent evaluation. Class metadata define the benchmark vocabulary but are hidden from screening.

CLIP ViT-B/32 (Radford et al., 2021) serves as a frozen construction anchor. Within each domain, its first principal-component scores define low, middle, and high strata, each supplying a block of 512 images. For a given target, we exclude its own blocks and enumerate $\binom{15}{3} = 455$ combinations, each with 1,536 source images. ResNet-50 and DINOv2-B/14 provide evaluation features (He et al., 2016; Oquab et al., 2024). All rules share row normalization, a fixed Rademacher projection to 128 dimensions, and post-projection row normalization. Fixed sample count separates data quantity from combination quality; shared encoders and readout settings keep method comparisons from also changing the prediction model.

Target A-opt uses $P_0^{-1} = I$ and $\sigma^2 = 1$. Comparators are second-moment MMD, target energy $\mathrm{tr}(\widehat{C}_T G_S / n_S)$, Bayesian D-opt, merged effective rank, subspace DPP, domain balance, and 20 seeded random rankings. They score the same candidate family. Domain balance additionally reads source-domain identifiers. Kernel definitions and tie rules appear in Appendix B.

The screen produces nested shortlists of sizes $10, 25, 50, 100, 227$, with 227 designated primary. In the original protocol, U2 independently evaluates each method's primary shortlist using ridge regression with regularization 1, no intercept, and a fixed softmax temperature. Validation Brier selects the final combination, breaking ties lexicographically. Smaller lists reuse that method's nested evaluations; there is no cross-method cache. The test stage verifies upstream hashes before evaluating every combination to establish an exhaustive audit reference.

The follow-up adds four projection seeds and multinomial logistic regression with fixed regularization, preserving candidate members and data roles. Screen, validation, and test remain separate processes, but methods reuse each combination's fitted model for shortlist auditing. Its runtime measures a cached research audit, not separate deployments. Appendix B gives readout settings and provenance checks.

We measure how many test top-ten candidates survive screening and how far the validation choice lies from the test oracle S^*. Recall measures candidate retention, not classification accuracy; regret measures loss relative to the oracle:

$$\mathrm{Recall}_{10}(\mathcal{Q}_s) = \frac{|\mathcal{Q}_s \cap \mathrm{Top}_{10}(b)|}{10}, \qquad \mathrm{Regret}(\widehat{S}) = \frac{b(\widehat{S}) - b(S^*)}{|b(S^*)|}. \tag{5}$$

The original acceptance criteria require at least 50% removal, mean recall of at least 90%, and mean normalized regret at most 5%. Brier is primary. Follow-up negative log-likelihood (NLL) and classification-error audits use their corresponding validation metric to select candidates, with the metric fixed before test access.

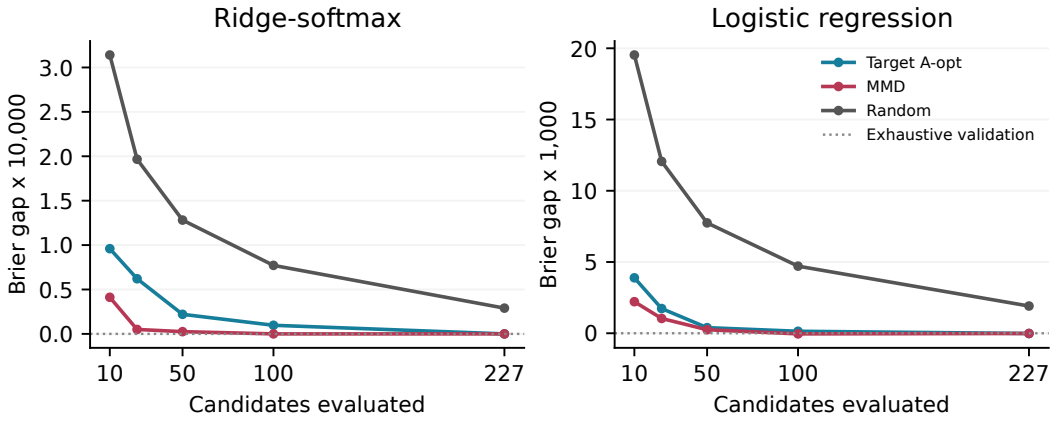


Figure 1: Mean final test Brier minus the exhaustive-validation choice is plotted against the number of candidates evaluated. The two panels use different vertical scales. The zero line is a validation-selected model, not the test oracle. Curves are post-hoc summaries of the four new projection seeds; the horizontal axis counts fits, not measured runtime.

We average encoder results within each dataset–target-domain unit and random rankings within each task. PACS/Office-Home contain eight such units and DomainNet six. Follow-up seeds are also averaged within domains. Original bootstrap intervals resample target units; the test-sample analysis fixes trained models and resamples examples within each target. Seeds, encoders, and overlapping combinations are repeated measurements, not additional independent datasets.

5 Empirical Findings

We first ask whether screening preserves final quality and saves computation on DomainNet. Retention then separates screening omissions from validation choices, before domain controls and sensitivity tests examine when the geometry is useful. The earlier shared-model and direct-selection studies explain the score's motivation and the distinction between improving predictions and reducing evaluation work.

5.1 Final choices across evaluation budgets

Figure 1 tests whether a smaller evaluation budget still yields a useful final predictor, reanalyzing the four new projection seeds. The reference fits every candidate and selects by the same validation metric. Target-aware lists have lower mean Brier than equally sized random lists at small budgets and approach the exhaustive-validation reference before all candidates are evaluated.

For logistic regression, evaluating 50 A-opt or MMD candidates gives lower mean Brier than evaluating 227 random candidates. More random evaluations do not produce a better mean final choice in this comparison: which candidates receive evaluation matters. Both comparisons favor screening in five of six domain averages. Paired mean differences are -0.001518 for A-opt and -0.001664 for MMD; descriptive domain-bootstrap ranges are $[-0.003517, 0.000023]$ and $[-0.003720, -0.000119]$. Shared source pools prevent interpreting these as independent-dataset confidence statements.

Task-level exceedance rates check whether favorable averages hide larger losses on individual tasks. At size 50, each geometric method exceeds an absolute Brier gap of 0.005 to exhaustive validation in one of 48 repeated encoder–seed–domain tasks, or 2.08%. Random screening at size 227 exceeds it in 15.52% after the prescribed averaging. Appendix Table 5 reports final losses. The tolerance is part of an exploratory grid, not a new success gate; these post-hoc curves compare budgets but do not validate an automatic budget-selection rule.

5.2 Measured decision and feature costs

Fewer fits save time only if the avoided work exceeds screening and any extra feature costs. Table 1 tests this condition with independent scoring, fitting, and validation for each method, without sharing fitted models. Unlike the preceding four-seed audit, this profile fixes one projection and uses three executions per encoder–target–readout task. Quality and time are paired within these runs; Appendix Figure 4 gives all measured budgets.

With projected features in memory, A-opt and MMD at size 50 take about one tenth of exhaustive logistic-regression time, with slightly higher Brier than exhaustive validation and lower Brier than random at size 227. Adding measured image-feature work leaves their accounted cost at about 56% of exhaustive validation. Ridge gives nearly identical losses across these choices, but its cheap fits save less than the extra target-selection features cost. The cost reversal identifies when screening overhead matters, despite fewer evaluations. Image totals add measured components, excluding pool construction, downloads, cache serialization, and test auditing. Disk loading and projection appear separately in Appendix Table 7.

Table 1: Matched final Brier and mean accounted time in seconds are reported for the two readouts. Resident time includes scoring, fitting, and validation from in-memory projected features. Images adds the separately measured feature work for the fixed pool inventory; it is an additive profile, not a complete pipeline replay. All means use the same timed choices.

Readout	Screen	Fits	Brier	Resident (s)	Images (s)
Ridge	A-opt	50	0.986655	0.53	39.59
Ridge	MMD	50	0.986643	0.21	39.27
Ridge	Random	227	0.986653	0.75	35.30
Ridge	Full validation	455	0.986639	1.48	36.03
Logistic	A-opt	50	0.863170	4.49	43.54
Logistic	MMD	50	0.863110	4.31	43.37
Logistic	Random	227	0.864241	21.23	55.77
Logistic	Full validation	455	0.862992	42.83	77.38

5.3 Candidate retention and validation choices

To distinguish screening omissions from validation choices, Table 2 returns to the original DomainNet study. A-opt retains nearly all test top-ten candidates while discarding approximately half the combinations. Every primary A-opt and MMD shortlist contains a test Brier oracle; remaining regret comes from validation choosing another member. In terms of Equation (4), no empirical oracle omission remains in these lists, but validation selection still affects the final loss.

MMD matches A-opt's primary recall and regret and leads validation to the same final combination in every encoder–target task. At smaller budgets, MMD has higher mean recall (Appendix Figure 2). Full-ranking correlation measures agreement across all candidates, whereas recall depends on which good candidates cross the shortlist boundary. A-opt's higher correlation therefore need not give better retention. Both outperform the tested target-blind rules here, but the weaker target-energy comparator shows that using target information alone is insufficient: how the score uses it matters.

5.4 Candidate construction and source-domain controls

Does a good shortlist require distinguishing blocks within source domains, or is domain composition enough? We compare mean matching and domain-level moment matching, with projection and readout parameters fixed, on the original blocks and a label- and feature-independent hash partition of the same images. Domain-level matching pools all blocks within each domain before scoring mixtures; it cannot distinguish candidates with identical domain composition. Appendix Table 6 reports this exploratory challenge, separating randomized tie handling from the frozen lexicographic protocol.

Repartitioning preserves the broad Brier advantage over random screening. At size 50, domain-level matching and MMD give similar mean final Brier: domain-level moments suffice for a competitive

Table 2: The original DomainNet study retains $s = 227$ of 455 combinations. Means average encoders within six target domains. Recall concerns the test Brier top ten; regret is relative to oracle Brier. Spearman measures agreement between the complete screening and test-loss rankings.

Screen	Recall (%)	Regret (%)	Spearman
Target A-opt	98.333	0.001311	0.7443
Second-moment MMD	98.333	0.001311	0.6902
Bayesian D-opt	82.500	0.002658	0.4996
Merged effective rank	81.667	0.002909	0.4951
Random (20 repeats)	48.958	0.003899	0.0016
Target energy	46.667	0.009347	-0.0399
DPP subspace	27.500	0.009067	-0.0633
Domain balance	25.000	0.012083	0.0681

global shortlist in this setting. Mean matching is competitive on the original blocks but less effective after repartitioning. These comparisons ask whether coarse information suffices overall, not whether blocks within a domain are interchangeable.

The within-composition comparison addresses the latter question. We group candidates containing the same number of blocks from each source domain and retain the floor of half of each non-singleton group. On the original blocks with logistic regression, A-opt and MMD retain the empirical Brier optimum in about 77% and 76% of groups, against 45% for random selection. Repartitioning lowers these rates to about 56% and 55%; both fall below random for classification error. These are within-group optimum-retention rates, not global top-ten recall. Useful block information can coexist with competitive domain-level matching, but its benefit depends on construction and metric. Neither comparison assigns a causal share of the overall benefit to domain composition.

5.5 Projection and readout sensitivity

The original ridge-softmax losses differ little: every candidate lies within the 5% regret gate, so it does not distinguish screens. Two checks address this result. Four new projections test dependence on one coordinate compression; logistic regression tests retention when candidate losses are more separated. Source membership and sample cost stay fixed. Appendix Table 3 reports these exploratory new-seed averages, with the original seed as an implementation control.

A-opt retains 98.33% of the Brier top ten on average under each readout, with MMD slightly higher. The mean relative best-to-worst Brier span increases from 0.248% under ridge-softmax to 14.65% under logistic regression. High recall thus persists in a setting with more separation between candidates, rather than appearing only when their Brier losses are almost equal. The two target-aware scores also retain most NLL- and error-ranked candidates on average, with MMD higher in these comparisons.

At the primary shortlist size, both methods retain a Brier test oracle in every new-seed task and have the same mean selected regret. Metric-specific outcomes vary: one A-opt task has only 40% error-ranked recall, and validation misses a retained test-error oracle. Appendix C gives the task details and full comparator table.

5.6 Target weighting under the shared risk model

The earlier synthetic study asks whether target weighting helps when the shared risk model holds. It varies feature dimension, candidate count, source budget, target sample size, and target rank, using 20 seeds for 216 settings. A-opt reduces modeled risk by 4.23% relative to the strongest target-blind effective-rank comparator and by 14.85% relative to MMD. Descriptive 95% intervals are $[3.86, 4.62]\%$ and $[13.99, 15.70]\%$, resampling settings after averaging seeds. Here the score is evaluated on modeled linear risk; classification tests its usefulness for shortlisting. A-opt's advantage here and MMD's competitive classification results answer different questions, with no contradiction. These comparisons alone do not identify why their relative performance differs.

A separate source-conditional-shift sweep changes the regression mechanism. Mean normalized prediction regret increases from 0.211 without shift to 3.143 at the largest tested shift. This measures realized prediction, not posterior risk: covering target directions and transferring labels across conditional mechanisms are different requirements.

5.7 Direct selection and available improvement headroom

PACS and Office-Home test whether the score can directly choose a better source combination. At block budget $K = 3$, A-opt improves mean within-unit relative Brier by 0.351% over DPP and 0.273% over MMD. Both fall below the prespecified 2% practical-effect threshold. The paired difference interval crosses zero against DPP and is negative against MMD; detailed values are in Appendix C. These comparisons hold block count, rather than source sample count, fixed.

The post-hoc exhaustive audit measures headroom: the loss reduction available from the test-best candidate over the reference. Mean headroom over DPP is only 0.781% (Appendix Figure 3), limiting average gains without changing the failed prespecified test. This motivates distinguishing better predictions from less evaluation work. These studies provide context, not independent equal-cost screening replications.

6 Discussion

6.1 What makes screening useful?

Screening should be evaluated within the selection procedure. Final loss determines whether fewer evaluations leave an acceptable choice; omission and validation diagnostics explain quality losses. Timing establishes whether work is saved. The readout-dependent cost reversal is central: a useful shortlist need not be economical when fitting is already cheap.

MMD and domain matching show why detailed geometry needs simple matching baselines. Good global choices need not distinguish individual blocks, although block information can help within fixed domain mixtures. Operating points should reflect the target metric, acceptable loss gap, and measured cost together, not recall or full-ranking correlation alone.

6.2 Limitations

The equal-sample-cost evidence uses two constructions on the same DomainNet images. Target domains share source pools; added seeds and readouts measure within-benchmark sensitivity, not independent-data replication. Differences from PACS/Office-Home include dataset, projection, sample budget, and construction, preventing causal attribution to candidate heterogeneity alone.

The derivation concerns shared linear regression; classification evaluates the score empirically. Test-example resampling conditions on fitted models and validation choices, excluding uncertainty from training or selection. Absolute quality also depends on the candidate family, as the supplementary adapter study illustrates. Image costs add measured components, excluding pool construction and cache serialization, with features profiled once on one hardware setup. Independent equal-cost datasets, controlled validation budgets, and checks for encoder-pretraining overlap remain important.

7 Conclusion

We evaluated target-aware geometry as a way to reduce supervised data-pool evaluations while retaining a competitive final choice. The analysis explains the design score under shared linear risk and separates candidate omission from the quality gap due to validation selection. On equal-sample-cost DomainNet combinations, evaluating 50 A-opt or MMD candidates gives lower mean test Brier than evaluating 227 random candidates. Simple moment matching remains competitive. Fewer evaluations save time with cached features; adding measured feature costs preserves the logistic-regression saving but reverses it for ridge. Screening is therefore best assessed through the final loss and the actual work required to obtain that choice, with candidate retention explaining rather than replacing those outcomes.

AI USE STATEMENT

Generative AI tools assisted conceptual reformulation, mathematical claim and proof drafting, experiment design, software implementation, interpretation of results, literature search, manuscript writing, figure scripting, and Chinese translation. Numerical findings in this draft are taken from recorded experiment artifacts rather than generated observations. AI assistance is not independent validation of those artifacts or of the mathematical claims. Final verification and responsibility for a submitted version remain with the human authors.

REPRODUCIBILITY STATEMENT

Proof details appear in Appendix A; preprocessing, access rules, candidate construction, metrics, and result paths appear in Section 4 and Appendix B. The code repository is https://anonymous.4open.science/r/collapsemodel-1DB6. Outcome tables and manifests support aggregate reproduction; large feature caches are identified by hashes rather than duplicated in Git. Exact hardware-independent bitwise feature reproduction is not claimed.

ETHICS STATEMENT

The study uses existing benchmark images and pretrained encoders, with no new participant recruitment. Public availability does not remove image licensing, privacy, or representational-bias concerns. We do not claim that a diverse feature spectrum or a high shortlist recall makes a selected pool socially representative, fair, or private. The candidate taxonomy, pretraining overlap uncertainty, and restriction to visual classification limit downstream deployment claims. Dataset and model access remain subject to their original terms.

REFERENCES

Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007. doi: 10.1088/1751-8113/40/28/S18.

Yijun Dong, Hoang Phan, Xiang Pan, and Qi Lei. Sketchy moment matching: Toward fast and provable data selection for finetuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4c79c359b3c5f077c0b955f93cb0f53e-Abstract-Conference.html.

Lan Feng, Fan Nie, Yuejiang Liu, and Alexandre Alahi. TAROT: Targeted data selection via optimal transport. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16837–16852, 2025. URL https://proceedings.mlr.press/v267/feng25l.html.

Xavier Fontaine, Pierre Perrault, Michal Valko, and Vianney Perchet. Online A-optimal design and active linear regression. In *International Conference on Machine Learning*, pp. 3374–3383, 2021. URL https://dblp.org/rec/conf/icml/FontainePVP21.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pp. 10929–10974, 2023. URL https://dblp.org/rec/conf/icml/GarridoBNL23.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://dblp.org/rec/journals/jmlr/GrettonBRSS12.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://dblp.org/rec/conf/cvpr/HeZRS16.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/e17fe6fe9990fffb637b42c98c005515-Abstract-Conference.html.

Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. Efficiently learning at test-time: Active fine-tuning of LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ba942323c447c9bbb9d4b638eadefab9-Abstract-Conference.html.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044. URL https://dblp.org/rec/journals/ftml/KuleszaT12.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017. URL https://dblp.org/rec/conf/iccv/LiYSH17.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: Data selection for task-specific model finetuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13848b5893119ff772b69812c95914fa-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://dblp.org/rec/journals/tmlr/OquabDMVSKFHMEA24.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019. URL https://dblp.org/rec/conf/iccv/PengBXHSW19.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://dblp.org/rec/conf/icml/RadfordKHRGASAM21.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/97c5b2707228e7e3fb67e4ecc2e0e607-Abstract-Conference.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://dblp.org/rec/conf/eusipco/RoyV07.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017. URL https://dblp.org/rec/conf/cvpr/VenkateswaraECP17.

Junjie Oscar Yin and Alexander M. Rush. Compute-constrained data selection. In *International Conference on Learning Representations*, 2025. URL https://dblp.org/rec/conf/iclr/YinR25.

Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 1081–1088, 2006. doi: 10.1145/1143844.1143980. URL https://dblp.org/rec/conf/icml/YuBT06.

A Proofs and Scope of the Mathematical Claims

The singular value decomposition $H = U\Sigma V^\top$ separates feature directions from their strengths. For fixed nonnegative weights w_i, stacking the scaled source matrices gives

$$G_S = H_S^\top H_S = \sum_{i \in S} w_i G_i. \tag{6}$$

The eigenvalues of this sum are the squared singular values of the pooled features. It preserves the strength and cross-pool alignment of directions, provided preprocessing coordinates remain fixed. The experiments use unit weights.

Effective rank normalizes the singular values into a probability vector and exponentiates its entropy (Roy & Vetterli, 2007):

$$r_{\text{eff}}(H) = \exp\left(-\sum_j p_j \log p_j\right), \qquad p_j = \frac{\sigma_j(H)}{\sum_k \sigma_k(H)}. \tag{7}$$

For a nonzero matrix, $0 \log 0 = 0$ and $1 \le r_{\text{eff}}(H) \le \operatorname{rank}(H)$, with the upper endpoint attained by equal positive singular values. Two values in ratio $9 : 1$ give about 1.38, versus two for equal strengths. Gram eigenvalues must be square-rooted before normalization; normalizing them directly defines another score. Repeating the entire matrix leaves effective rank unchanged: it measures spectral shape, not sample evidence. It is a target-blind comparator, not a step required by the target-aware scores.

Marginal spectra and principal angles alone need not determine this pooled statistic: the same strengths can be assigned differently to shared and unshared directions. The following counterexample and local formula identify the missing directional information.

Proposition 1 (Directional assignment matters)**.** *There exist row-normalized pairs (A, B_1) and (A, B_2) with identical marginal singular spectra and principal angles between their row spaces, but different effective ranks after concatenation.*

Proof. Take A with four rows $e_1^\top$ and one row $e_2^\top$. Let B_1 have four rows $e_1^\top$ and one $e_3^\top$, and let B_2 have one row $e_1^\top$ and four $e_3^\top$. All three nonzero singular spectra are $(2, 1)$; both row-space pairs have principal angles $(0, \pi/2)$. Their merged Gram spectra are $(8, 1, 1)$ and $(5, 4, 1)$, giving effective ranks approximately 2.6260 and 2.8495. The shared direction receives different energy despite the identical marginal summaries. □

Let u, v be unit directions with $u^\top v = c \in [0, 1]$. In the orthonormal basis u, z, where $v = cu + \sqrt{1-c^2}z$, the operator $a^2 uu^\top + b^2 vv^\top$ is

$$\begin{pmatrix} a^2 + b^2c^2 & b^2c\sqrt{1-c^2} \\ b^2c\sqrt{1-c^2} & b^2(1-c^2) \end{pmatrix}. \tag{8}$$

Its trace is $a^2 + b^2$ and determinant $a^2b^2(1-c^2)$, so its eigenvalues are

$$\lambda_\pm = \frac{a^2 + b^2 \pm \sqrt{(a^2-b^2)^2 + 4a^2b^2c^2}}{2}. \tag{9}$$

At $c = 1$, the second branch is zero. These pairs give a full spectrum when the associated blocks are mutually orthogonal and invariant for both source Grams; unpaired directions contribute their own eigenvalues. This compatibility condition matters because principal vectors need not be singular vectors.

For fixed H_S and Gaussian likelihood, completing the square in the prior times likelihood gives posterior precision $P_0^{-1} + \sigma^{-2} H_S^\top H_S$ and mean $\mu_S = P_S \sigma^{-2} H_S^\top y_S$. For an independent target observation $y = h^\top \theta + \epsilon$, conditional on the training data and h,

$$\mathbb{E}[(y - h^\top \mu_S)^2 \mid H_S, y_S, h] = \sigma^2 + h^\top P_S h. \tag{10}$$

Integrating h gives Equation (1). We assume the target covariate distribution does not reveal additional information about θ beyond this model. If a source has a different regression function, or

labels are multiclass with softmax readout, this derivation no longer identifies the evaluated loss. The proof does not bound this misspecification error. Replacing C_T by a finite-sample estimate adds another discrepancy not removed by the identity.

If $G \succeq \widetilde{G} \succeq 0$, inverse order implies $(P_0^{-1} + \sigma^{-2}G)^{-1} \preceq (P_0^{-1} + \sigma^{-2}\widetilde{G})^{-1}$. Taking a trace against a PSD target moment preserves this order. More generally, suppose the sum of omitted sketch tails satisfies $0 \preceq G - \widetilde{G} \preceq \delta I$. Define $B = P_0^{-1} + \sigma^{-2}\widetilde{G}$. Then

$$\operatorname{tr}\Big(\widehat{C}_T(B + \sigma^{-2}\delta I)^{-1}\Big) \le f_A(G) \le \operatorname{tr}(\widehat{C}_T B^{-1}). \tag{11}$$

One admissible δ is the sum of valid spectral-norm bounds on each PSD tail. A candidate's proxy ordering is certified when its upper bound is below every competing lower bound. Thus interval width, as well as coverage, determines whether a sketch can resolve the next choice. The archived sweep evaluates this type of proxy comparison with its documented byte accounting.

Let S_Q minimize ℓ over $\mathcal{Q}_s$. Uniform deviation on that shortlist gives

$$\ell(\widehat{S}) \le v(\widehat{S}) + \varepsilon \le v(S_Q) + \varepsilon \le \ell(S_Q) + 2\varepsilon. \tag{12}$$

Subtracting $\min_{S\in\mathcal{C}} \ell(S)$ proves Equation (4). No probability is asserted by this algebra. For population generalization, one must additionally justify the deviation event using the actual dependence and selection scheme. Exhaustive finite-test scores alone do not supply that justification.

The feature map $\phi(h) = hh^\top$, equipped with Frobenius inner product, satisfies $\langle \phi(h), \phi(h')\rangle_F = (h^\top h')^2$. The squared distance between source and target empirical feature means is therefore Equation (3). This is the biased empirical squared MMD (a V-statistic), not the diagonal-excluded unbiased estimator. It depends on uncentered second moments and cannot distinguish all distributions sharing those moments. We use it as a computationally simple same-information comparator, not a complete discrepancy measure.

B Implementation and Audit Details

The original DomainNet study fits a ridge readout and transforms its scores into class probabilities. Its weights and Brier loss are

$$\begin{aligned} W_S &= (I + G_S)^{-1} \sum_{i\in S} H_i^\top Y_i, \qquad p(x) = \operatorname{softmax}(h(x)^\top W_S), \\ b(S) &= \frac{1}{n_T} \sum_t \|p(x_t) - e_{y_t}\|_2^2 . \end{aligned} \tag{13}$$

Y_i is the one-hot source-label matrix over the fixed 100-class vocabulary. There is no intercept or temperature tuning. Linear algebra uses float64 after loading float32 features. For effective rank obtained through a Gram eigendecomposition, negative roundoff eigenvalues are clipped to zero, square roots form singular values, and only $\sigma_j > \sigma_{\max}\sqrt{8\epsilon_{64}\max(n_S, d, 1)}$ are retained. The zero matrix returns rank zero in code. This tolerance concerns numerical rank; the entropy definition itself is unchanged. The shared projection uses seed 20260911 and $d = 128$; no outcome-dependent dimension choice is used within the frozen DomainNet protocol.

Bayesian D-opt maximizes $\log\det(I + G_S)$. Subspace DPP takes each block's top 16 Gram eigenvectors U_i, defines $K_{ij} = \|U_i^\top U_j\|_F^2/16$ with $K_{ii} = 1$, numerically projects the matrix to a PSD correlation matrix, and maximizes $\log\det(K_{S,S} + 10^{-8}I)$. Its kernel describes block subspaces, not individual-image features. Domain balance maximizes the number of represented domains minus 10^{-3} times the sum of squared domain block counts. Deterministic scores break ties lexicographically by combination identifier. These choices are fixed comparators; poor performance of this DPP construction is not a result about all determinantal selection methods.

DomainNet records image content hashes, sample identifiers, official split membership, role assignments, and class vocabulary. Archive and split checksums are validated before feature extraction. Disjoint role membership and duplicate-content checks address overlap within the materialized benchmark; they do not rule out semantic near-duplicates or pretraining contamination. All screen

methods use the same source blocks, projection, target-selection rows, and candidate family. Feature extraction ignores labels, while benchmark materialization uses taxonomy metadata and stores labels for later authorized stages.

Screen, validation, and audit manifests bind configuration and upstream artifact hashes. Stage-specific loaders enforce label and split permissions. In the original protocol, validation evaluates each method or random repeat separately. The test audit covers all 6×455 combinations per encoder after choices are frozen; its exhaustive oracle serves as a research reference. The follow-up caching policy is described below.

The CUDA reconstruction retains the original sample identifiers, candidate membership, preprocessing, and hyperparameters. All twelve Brier top-ten sets agree with the original study. Per-example loss means reproduce the reconstructed aggregate audit within 1.78×10^{-15}. For each target domain, 2,000 paired bootstrap draws resample its 2,048 test examples, sharing weights across combinations, encoders, and four losses. Fitted models and validation choices stay fixed; combination identifiers resolve ties. These draws measure sensitivity to test-example sampling. The original Brier top ten overlaps with NLL by 90.83%, raw squared loss by 30.83%, and classification error by 29.17% on average.

Seeds 20260917, 20260918, 20260919, and 20260920 generate new 128-dimensional Rademacher projections with the same pre/post row normalization. Seed 20260911 is an implementation control, excluded from new-seed means. Logistic regression uses L2 regularization, $C = 1$, no intercept, LBFGS, tolerance 10^{-6}, and at most 1,000 iterations. A convergence warning aborts a task; no such warning occurred, and the largest iteration count was 24. The class vocabulary is fixed, with zero probability assigned to source-absent classes.

All methods retain their screening rules, candidate combinations, and five shortlist sizes. Separate processes screen, fit and validate, then test frozen models. The audit checks source, configuration, lineage, model, and selection hashes before reading test features. The follow-up fits every combination once per readout and reuses these fits across methods. This cached exhaustive evaluation supports counterfactual comparisons of shortlists. Brier, NLL, and error use their own validation losses for final selection. Original-seed ridge reproduces 5,460 reconstructed test combinations with maximum loss difference 1.78×10^{-15} and all 1,620 Brier validation choices exactly.

The two encoders and five seeds produce 54,600 fits across both readouts. New-seed summaries first average random repeats within a task, then encoders and seeds within each target domain, and finally the six domains. Each readout contributes 48 repeated encoder–seed–domain measurements. They remain six domain units sharing the same source pools.

The archive retains historical experiment IDs to identify results without confusing them with the paper's scientific questions. The following paths are relative to the anonymous code repository:

- Shared-model and shift studies (E1): `results/target_conditioned/e1_synthetic/expanded_v3_controlled_rng/`.
- PACS/Office-Home comparison (E2): `results/target_conditioned/e2_dataset_selection/method_v1/`.
- Post-hoc headroom (E2a): `results/target_conditioned/e2_dataset_selection/oracle_headroom_v1/`.
- Equal-cost shortlist (E2b): `results/target_conditioned/e2b_fixed_cost_shortlist/domainnet_v1/`.
- Risk sketch sweep (E5): `results/target_conditioned/e5_sketch_certificate/expanded_v3_controlled_rng/`.
- Reconstruction and test resampling (R1/R2): `results/target_conditioned/revision_20260916/`.
- Projection and readout follow-up (R3/R4): `results/target_conditioned/revision_20260917/`.

The DomainNet data manifest begins `d99e2a4c637cadc5`, candidate-set ID `b093ac0926b6a196`, and summary ID `ad71c79b7d2ad81d`; full SHA-256 identifiers are in the artifacts. The original configuration is

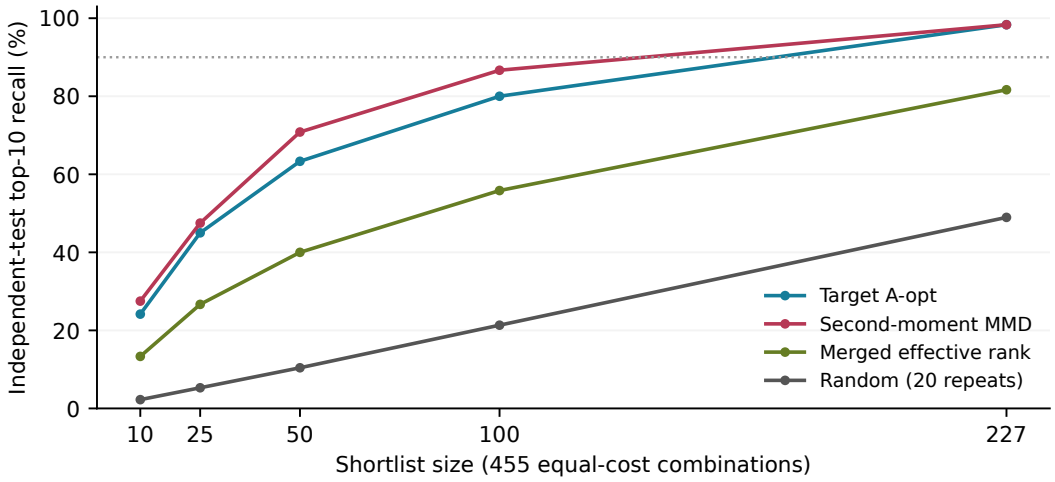


Figure 2: Mean test Brier top-ten recall is shown across shortlist sizes in the original DomainNet study. The dotted line marks the prespecified recall threshold. All curves use the same candidate combinations and feature access appropriate to each rule.

`code/configs/target_conditioned_e2b/domainnet_v1.json`, implemented in `scripts/run_target_conditioned_e2b_shortlist.py`. The follow-up uses `code/configs/target_conditioned_e2b/readout_projection_v1.json` and `scripts/run_e2b_readout_projection.py`. Its summary tables are in `revision_results/e2b_readout_projection_v1/summary/` under the R3/R4 directory. The manuscript's `build_assets.py` regenerates tables, figures, numeric macros, and the English abstract from recorded sources, logging input hashes in `generated/asset_manifest.json`.

The three original feature matrices are not duplicated in Git. Metadata record their exact file and checkpoint hashes, shapes, preprocessing, extraction code revision, and runtime. Re-extraction requires access to the original public data and model weights; bitwise identity across different hardware is not assumed. The repository contains manifests and sufficient outcome tables to reproduce reported aggregate results without those matrices.

C Supplementary Results

At $K = 3$ on PACS/Office-Home, mean Brier is 0.826769 for A-opt, 0.829162 for DPP, and 0.828719 for MMD. The paired A-opt-minus-baseline intervals are $[-0.007449, 0.000731]$ and $[-0.003392, -0.000712]$, respectively. The post-hoc audit covers all 220 combinations per task. Oracle headroom over DPP has descriptive 95% interval $[0.162, 1.617]\%$; only PACS photo exceeds the original 2% threshold at the target-unit level. The oracle improves on the post-hoc test-best envelope of parent methods by 0.087% on average. This envelope is an outcome-based reference, not a selectable baseline.

Table 4 gives the complete Brier recall comparison at the primary size. A-opt and MMD have the same mean relative selected regret: 0.001006% for ridge and 0.103% for logistic regression. All their new-seed primary Brier shortlists contain an empirical oracle.

The lowest A-opt classification-error recall occurs for DINOv2, Quickdraw, logistic regression, and projection seed 20260918. The test-error oracle is 85.9863%, compared with 87.6465% for the validation choice. Its oracle remains in the shortlist despite the 40% top-ten recall. Thus retaining the best member and retaining the best ten answer different questions, and both are distinct from the absolute quality of the available models.

The four-encoder, ten-construction-seed spectral sweep uses six overlap settings and four budgets, with 960 configurations per sketch rank. At ranks $10, 20, 50$, relative exact merged-effective-rank

Table 3: The exploratory DomainNet follow-up evaluates shortlists of size $s = 227$. Test top-ten recall (%) is averaged over two encoders and four new projection seeds within each target domain, then over six domains. Ridge denotes ridge-softmax. NLL and error use matching validation losses for final selection.

Readout	Metric	A-opt	MMD	Random
Ridge	Brier	98.33	99.38	50.12
Ridge	NLL	98.54	99.38	49.86
Ridge	Error	96.25	97.50	48.63
Logistic	Brier	98.33	98.96	48.97
Logistic	NLL	98.54	99.79	48.61
Logistic	Error	97.29	97.92	49.04

Table 4: Brier top-ten recall (%) is reported for all follow-up scoring rules at $s = 227$. Averages use the same domain-level procedure as Table 3; original-seed controls are excluded.

Screen	Ridge	Logistic
Target A-opt	98.33	98.33
Second-moment MMD	99.38	98.96
Bayesian D-opt	82.29	86.46
Merged effective rank	80.42	85.42
Random (20 repeats)	50.12	48.97
Target energy	41.04	33.96
DPP subspace	28.75	35.42
Domain balance	26.46	31.67

gains over rank-only are respectively $2.09, 3.81, 4.94\%$ for the Gram sketch and $4.90, 5.48, 5.68\%$ for DPP. These gains are in the geometry objective, not classification accuracy. Its entropy-continuity bounds follow the type of bound in Audenaert (2007); all 2,880 selected-prefix intervals cover, but none of 7,200 evaluated greedy steps is certified. The archived source is `results/two_stage_controlled_rank_sweep_corrected_v3/`. These observations do not imply that every low-rank sketch must fail.

The separate E5 study compares sketch payloads to packed symmetric float64 Grams, not inflated dense storage. Among full-grid strategies whose mean byte ratio is at most 25%, the best selected-set match to Full-Gram greedy is 15.42%, achieved by adaptive rank cap 2 with mean byte ratio 14.91%. Its pairwise certificate rate is 3.91% and fully certified selection rate zero. Adaptive fallback reaches full selected-set agreement but costs approximately 160% of packed-Gram bytes on average when previous transmissions are counted. This fails the joint decision-and-communication requirement. The reference is Full-Gram greedy, not the exhaustive global minimizer.

An earlier, different experiment trains source-supervised bottleneck adapters on 21 sources and audits seven targets with four encoders and five adapter seeds. Target-train validation uses three stratified five-fold partitions, with held-out target test accessed separately after selection freezing. Frozen identity reaches mean accuracy 0.83654, an untrained random adapter 0.77144, and the validation-oracle source-supervised adapter 0.76951. The last trails identity by 0.06703, with target-cluster 95% interval $[-0.09771, -0.04400]$. Small shortlist regret relative to this adapter family is therefore not downstream benefit. This is not a DomainNet experiment or evidence against all adaptation methods. Results reside in `results/two_stage_repeated_cv_utility_corrected_v3/`.

C.1 Decision-budget reanalysis

The post-hoc budget audit uses the four new projection seeds and both readouts without selecting a favorable seed. For each task and endpoint, exhaustive validation chooses its combination using validation losses only. Its test loss is then the paired reference for the frozen shortlist choices. In

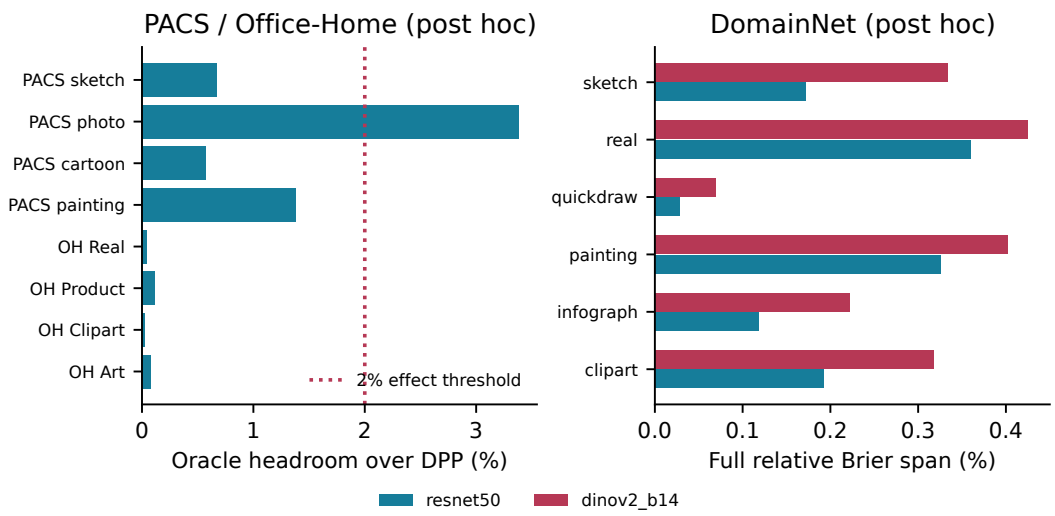


Figure 3: The panels report post-hoc diagnostics for different candidate families. The left panel shows oracle Brier improvement over DPP, averaging encoders within each PACS/Office-Home target. The right panel shows relative best-to-worst Brier spans for the original DomainNet tasks. The original evaluation thresholds are unchanged.

contrast, the empirical test oracle remains a lower audit reference. All five original budgets are retained.

Table 5: Final logistic-regression Brier and the rate of exceeding the exhaustive-validation reference by more than 0.005 are reported. Means average random repetitions before encoders and seeds within each domain, then the six domains. The tolerance is exploratory, not a new success gate.

Screen	Fits	Test Brier	Gap > 0.005 (%)
Random	227	0.866030	15.52
Target A-opt	50	0.864511	2.08
MMD	50	0.864365	2.08
Exhaustive validation	455	0.864118	0.00

We report absolute-gap sensitivity grids of $0.001, 0.005, 0.01$ for Brier, $0.01, 0.05, 0.1$ for NLL, and $0.5, 1, 2$ percentage points for error. Exceedance indicators are computed before averaging random repetitions; thresholding an averaged loss would answer a different question. Descriptive paired bootstrap ranges use 2,000 resamples of the six domain averages. They do not account for dependence induced by shared source pools, validation resampling, or independent datasets. All endpoint curves, paired domain differences, and threshold rates are in `results/target_conditioned/revision_20260922/budget_audit/`. The script `scripts/analyze_e2b_decision_budget.py` verifies recorded loss hashes and constructs the exhaustive reference from validation files.

C.2 Loss resolution and test-sample stability

The post-hoc loss-span diagnostic in Appendix Figure 3 puts the original regret criterion in context. Every candidate's relative Brier regret lies below the 5% threshold, so that threshold cannot distinguish successful screens in this experiment. Recall provides the more informative comparison, but it raises another question: how sensitive is the test top ten to the sampled test examples?

A CUDA reconstruction first reproduces all original Brier top-ten sets. We then apply paired test-example bootstrap resampling while fixing fitted models and validation choices. Although resampled and original top-ten sets overlap by only 85.15% on average, A-opt and MMD retain 97.42%

and 97.85% of the resampled top tens. Broad shortlists are less sensitive than exact rank membership to these test-sample changes.

Changing the metric has a larger effect on which combinations count as good: the original Brier and classification-error top tens overlap by only 29.17% on average. A screen should therefore be evaluated against the intended deployment metric, even when its average recall is high for several metrics. The resampling protocol and additional metric comparisons appear in Appendix B.

C.3 Candidate construction and domain controls

The fixed challenge uses projection seed 20260917, both encoders, and both readouts. Each domain's original 1,536 source images are sorted by SHA-256 of a fixed seed, domain identifier, and sample identifier, then partitioned into three blocks of 512. No labels, feature values, or outcomes determine membership. Target-selection, validation, and test sets do not change. The original construction is rerun as a control: all 65,520 recorded validation/test loss values agree exactly with the corresponding earlier seed. Repartitioning does not reuse CLIP stratum labels, but still conditions on the originally selected source images; it is neither an independent dataset nor a new test sample.

Mean matching minimizes $\|\bar{h}_S - \bar{h}_T\|_2^2$, the linear-kernel MMD score. Domain matching additionally uses source-domain identifiers: it replaces each block moment by its full source-domain moment, weights these moments by the candidate's domain sample proportions, and minimizes squared Frobenius distance to the target moment. The original challenge breaks ties lexicographically. Because domain matching creates many structural ties, an additional post-hoc audit randomizes the order within each tied domain-composition stratum using 20 fixed seeds. It freezes every new validation choice before reading test losses and does not refit or change classifier parameters. The original tie protocol remains available in the released curves.

Table 6: The construction challenge uses logistic Brier. Final loss is validation-selected at $s = 50$. The last columns measure retention of one empirical optimum when half of each non-singleton domain-composition stratum is kept, not final selection accuracy. Means average strata and encoders within targets, then six targets. Domain matching uses 20 randomized tie orders for final loss; its within-stratum entry is the exact uniform-tie expectation.

	Final Brier ($s = 50$)		Within-composition (%)	
Screen	Original	Hash	Original	Hash
Target A-opt	0.863170	0.869699	77.22	55.83
MMD	0.863110	0.869257	76.11	55.00
Mean matching	0.863713	0.873038	75.28	57.78
Domain moment (random ties)	0.863667	0.869356	45.68	45.68
Random	0.870013	0.874244	45.04	45.31

There are 20 strata of nine candidates and ten of 27 per target; five singleton strata are excluded. Within a stratum, every candidate uses the same domain multiplicities and total sample count. Random repeats are averaged first, followed by strata, encoders, and target domains. A domain-only score is constant inside a stratum, giving expected optimum retention $(20 \cdot 4/9 + 10 \cdot 13/27)/30 = 45.68\%$ under uniform ties. Lexicographic within-stratum results are recorded but not interpreted as domain discrimination.

After hash repartitioning, logistic classification-error optimum retention is 41.94% for A-opt and 35.28% for MMD, versus 45.75% for random selection. Mean Brier omission remains lower than random for both scores, but the advantage is smaller than on the original blocks. These metric-dependent observations describe conditional ranking performance; they do not establish a population guarantee or isolate the causal contribution of heterogeneity from the properties of this particular partition.

The complete curves and domain-conditional audits are in `results/target_conditioned/revision_20260922/challenge_summary/`; randomized domain ties are in `results/target_conditioned/revision_20260922/domain_tie_audit/`. The raw three-stage records bind candidate membership, source features, configuration, upstream choices, and

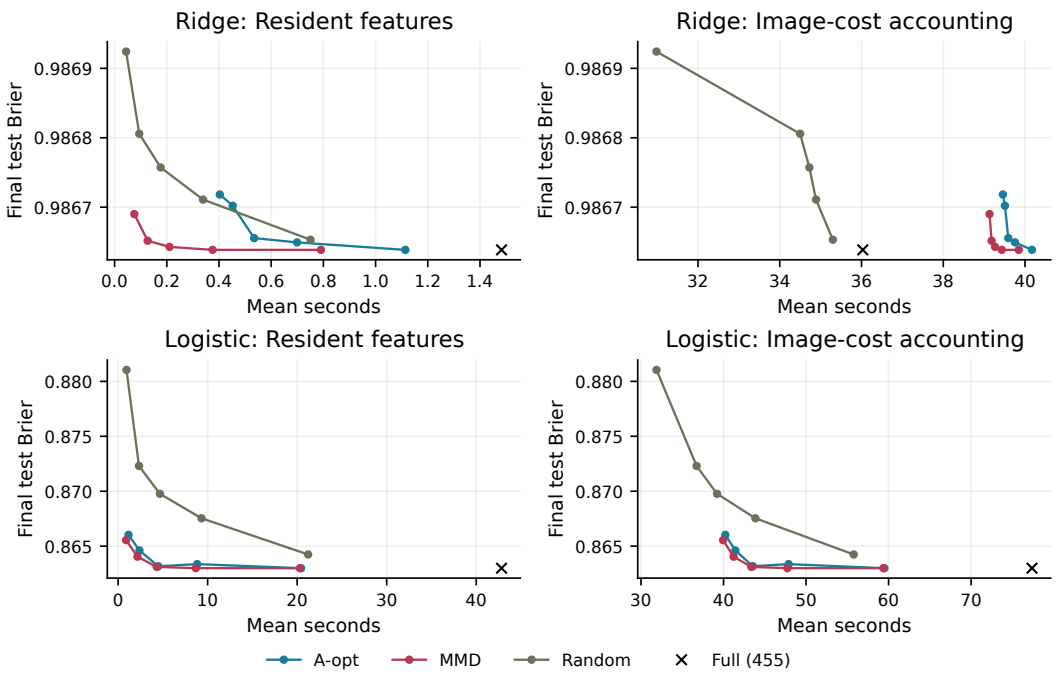


Figure 4: Final test Brier is plotted against matched decision time and additive image-cost accounting. Each curve has budgets $10, 25, 50, 100, 227$; crosses mark exhaustive validation over 455 candidates. Rows have different vertical scales. Means use two encoders, six targets, and three independent executions per method. Raw-image costs exclude pool construction, downloads, cache serialization, and test auditing.

numerical payloads by hashes. Local summaries verify all numerical hashes and upstream seals; frozen model hashes were checked on the server before test access.

C.4 Matched quality and computation costs

The timing study uses the original candidate construction and projection seed 20260917. Each encoder–target–readout task runs A-opt, MMD, random screening, and exhaustive validation independently, with three repetitions and a seeded method order. BLAS uses one thread. Shortlist checkpoints measure cumulative elapsed time after a single ranking; no fitted model is shared across methods. Ridge lazily caches each used block's float64 Gram and label right-hand side, including in exhaustive search. Preliminary float32-Gram ridge timings are retained in the archive but excluded from all tables.

Every timed validation loss and choice is checked against the frozen scientific audit. Quality in Figure 4 uses the actual choices from these same three runs, with Brier, NLL, and error each using its matching validation metric. Thus the random quality coordinates here average three timed rankings, not the 20 rankings in the main scientific audit. We first average repetitions within each task, then encoders within targets and the six targets equally; timing medians are retained as diagnostics.

The resident scope excludes loading stage bundles, checking their integrity, and projecting cached features. Table 7 accounts for these operations separately for two six-target encoder batches. Target-aware methods load both selection and validation bundles; random and exhaustive methods need only validation bundles. Shared preparation is charged once per encoder batch, not converted into a fictitious single-task time.

Raw-image profiling measures model initialization, image reads and hash checks, decoding, preprocessing, inference, CPU feature transfer, and normalization/projection. Each source block is timed separately; random screening is charged only for the union of blocks in its actual prefix, while geometric scoring and exhaustive validation use all 15 eligible source blocks. Each method also needs

Table 7: Disk-cache batch costs in seconds include input loading, integrity checks, projection, and all twelve encoder–target decisions. They are totals for two six-target batches, whereas Table 1 reports per-task means. The choices and losses do not change.

Readout	A-opt (50)	MMD (50)	Random (227)	Full (455)
Ridge	9.35	5.46	10.40	19.18
Logistic	56.77	54.63	255.89	515.12

target-validation features; only target-aware scores need the target-selection sample. No test images are processed. The fixed candidate inventory is supplied to this measurement, so CLIP-based pool construction is outside its scope.

The machine uses an Intel Xeon Platinum 8352V CPU and an NVIDIA vGPU-32GB. Feature extraction uses CUDA 12.4, PyTorch 2.5.1, float32 inference without TF32, batch size 64, and four data-loading workers. The readout stack uses NumPy 1.26.4, SciPy 1.16.1, and scikit-learn 1.7.1. Feature components are measured in one pass, with operating-system disk caches left unchanged. Block-level loader startup is included. The totals therefore describe this measured implementation and reuse scenario, not a hardware-independent speedup.

All 32 quality/time curve entries, three loss endpoints, paired differences, and 384 task-level budget entries are in `results/target_conditioned/revision_20260922/cost_summary/`. Standalone timing runs and per-block feature profiles accompany them. The manuscript verifier independently reconstructs the matched quality coordinates, random block unions, and all additive totals from those records.

D Connections to Recent Data-Selection Methods

Recent selection methods differ in what they select, which representation describes relevance, and what information selection can access. These distinctions locate the present framework and determine how a future comparison should be constructed. The methods discussed here provide literature context; our measured comparators remain those in Section 4.

Transductive active learning separates an accessible sampling region from a prediction target region (Hübotter et al., 2024). ITL reduces target entropy, whereas VTL reduces the sum of target predictive variances. Their Gaussian-process formulation explains how an observation can be useful for a target without being drawn from that target region. The variance objective is the closest connection to Equation (1): source geometry matters through the uncertainty remaining at target covariates. We evaluate complete, fixed pool combinations and then validate retained predictors, rather than sequentially acquiring individual observations.

SIFT makes redundancy important to target-specific selection (Hübotter et al., 2025). It selects fine-tuning sequences by the uncertainty remaining about a prompt after conditioning on previously selected sequences. Consequently, an example's relevance depends on what has already been selected, not just its distance to the prompt. This provides a useful perspective on pooled Gram matrices, whose combined coverage also depends on cross-pool alignment. SIFT's experiments concern prompt-specific language-model fine-tuning; here the encoder remains frozen and the supervised stage compares readouts trained on different pool unions.

SkMM addresses the bias–variance tradeoff of selecting data for high-dimensional fine-tuning (Dong et al., 2024). Gradient sketching first identifies a low-dimensional parameter subspace with controlled approximation bias; moment matching then controls variance within that subspace. Its gradient representation comes from model derivatives, so it should not be conflated with label-dependent loss-gradient selection. SkMM matches selected data to the original dataset, whereas our target moment is estimated from a separate target-selection sample. Both use matrix geometry, but our fixed feature projection does not perform SkMM's gradient-space construction.

TSDS balances distribution alignment and diversity by optimizing sampling probabilities with an optimal-transport objective (Liu et al., 2024). Its density-aware regularizer reduces concentration on near-duplicate candidates, and its metric space can use semantic embeddings or model-specific

representations. TAROT builds transport costs from projected, whitened, normalized loss-gradient features (Feng et al., 2025). Whitening reduces the dominance of correlated gradient directions before matching the selected and target distributions. Unlike our unlabeled screen, its supervised loss-gradient construction includes both inputs and labels. These are relevant transport-based comparison methods, but their sample-level outputs must be reconciled with our fixed-block candidate family. A frozen-embedding OT score would be a useful controlled adaptation, not a reproduction of TAROT's loss-gradient method.

The covariance-matching analysis of Rezaei et al. (2026) studies a predictor trained on real data augmented with synthetic data under high-dimensional linear regression. It characterizes test error and establishes covariance-matching optimality under specified conditions. This gives matching a risk-based motivation beyond simple geometric resemblance. The distinction between centered covariance and uncentered second moments matters here: Equation (3) compares the latter on frozen features. Our classification shortlists address a different training and selection setting from the synthetic-augmentation analysis.

A comparison with these methods would need to hold the candidate family, sample cost, target-data roles, and final validation rule fixed. Extra label access would require a separately identified information setting. Feature or gradient construction, transport optimization, screening, and supervised evaluation would all enter the cost account. This follows the selection-plus-training budget principle of Yin & Rush (2025) and makes the comparison about the final predictor obtained for the available computation, rather than the screening score alone.