TOWARDS THEORY-GUIDED AND GENERALIZABLE DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data selection aims to construct compact coresets that reduce training cost while preserving model performance. Most existing methods select data by preserving statistical, geometric, or optimization properties of the full training set, reflecting an empirical risk minimization (ERM)-oriented perspective. However, discrepancies between observed training data and unseen test data can arise from dataset collection, construction, and sampling effects, suggesting that preserving full-data ERM behavior alone may be insufficient for generalization. We therefore revisit data selection from a generalization perspective. Drawing on a domain generalization error bound, we identify four key factors associated with generalization and formulate corresponding principles for data selection: **preserving predictive utility**, **encouraging cross-subset invariance**, **maintaining semantic consistency**, and **promoting distributional coverage**. Guided by these principles, we propose **GenDS**, a theory-guided **Gen**eralizable **D**ata **S**election framework that operationalizes them through three complementary sample-level scores and a coverage-aware coreset selection strategy. Specifically, the scoring stage evaluates image-text alignment, subset invariance, and classification-structure consistency to capture predictive utility, cross-subset invariance, and semantic consistency, respectively, while the subsequent selection stage promotes distributional coverage and reduces redundancy. Extensive experiments on six standard and multi-domain image classification benchmarks demonstrate that GenDS consistently outperforms representative data selection methods across varying selection ratios, with particularly strong gains under limited data budgets.

1 INTRODUCTION

Large-scale training data have become increasingly important for deep neural networks (DNNs), but further expanding datasets incurs substantial computational and storage costs while yielding diminishing returns (Kaplan et al., 2020; Rosenfeld et al., 2020; Hoffmann et al., 2022). Moreover, real-world datasets often contain redundant or noisy samples, further reducing training efficiency (Yang et al., 2025). To improve data efficiency, **data selection** (Pooladzandi et al., 2022; Zheng et al., 2023; He et al., 2024; Yang et al., 2025; Cui et al., 2026) aims to identify a representative and informative subset, or coreset, that substantially reduces the training data while preserving the performance achievable with the full training set (Borsos et al., 2020; Killamsetty et al., 2021b).

Existing data selection methods employ diverse criteria, including sample-level importance scores (Toneva et al., 2019; Paul et al., 2021), feature-space representativeness (Sener & Savarese, 2018; Xia et al., 2023), and optimization-related signals (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a). Despite these differences, most methods share a common objective: preserving certain statistical, geometric, or optimization properties of the full training set. From a statistical learning perspective, most existing methods can be viewed in terms of **empirical risk minimization (ERM)**, where the coreset is expected to approximate the learning behavior induced by the full training set.

However, is preserving the full-data ERM solution sufficient for effective data selection? Not necessarily. ERM primarily focuses on fitting the observed training distribution, while in real-world scenarios, the test distribution may still differ due to dataset collection and sampling effects (Torralba & Efros, 2011; Koh et al., 2021). Such discrepancies may exist even within standard benchmarks constructed from the same underlying dataset (Rabanser et al., 2019; Yadav & Bottou, 2019). Con-

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.

sequently, faithfully approximating the full-data ERM solution may still be insufficient for strong generalization, motivating us to reconsider data selection from the perspective of generalization.

To make this generalization perspective concrete, we examine a domain generalization error bound (Albuquerque et al., 2020) to identify several key factors associated with the test risk. We interpret the four terms in the bound as motivating four principles for generalizable data selection: **(1) Preserve predictive utility**, **(2) Encourage cross-subset invariance**, **(3) Maintain semantic consistency**, and **(4) Promote distributional coverage**. Together, these principles clarify what a generalizable coreset should preserve beyond the full-data ERM solution and provide structured guidance for designing concrete data selection criteria in practice.

Guided by these four principles, we propose **GenDS**, a theory-guided **Gen**eralizable **D**ata **S**election framework that operationalizes the four principles through three complementary sample-level scores and a coverage-aware selection strategy. Specifically, the scoring stage captures **predictive utility**, **cross-subset invariance**, and **semantic consistency** through the Image-Text Alignment Score (S_{ITA}), Subset-Invariance Score (S_{SI}), and Classification-Structure Consistency Score (S_{CSC}), respectively. The subsequent selection stage integrates these scores while promoting **distributional coverage** over the training data, thereby yielding a representative and generalizable coreset.

To summarize, our main contributions are: (1) We revisit data selection from a generalization perspective. Motivated by a domain generalization error bound, we identify four key factors associated with generalization and formulate corresponding principles that provide guidance for data selection. (2) We propose GenDS, a theory-guided data selection framework that operationalizes these principles through three complementary sample-level scores targeting predictive utility, cross-subset invariance, and semantic consistency, together with coverage-aware coreset selection for distributional coverage. (3) Extensive experiments on six benchmarks demonstrate that GenDS consistently outperforms existing data selection methods across both standard and multi-domain settings.

2 Related Work

Data selection aims to identify a subset of original samples from the training set for learning. We focus on this real-sample selection setting, which differs from methods that synthesize or transform training data, such as dataset distillation (Wang et al., 2018; Cazenavette et al., 2022), condensation (Zhao et al., 2021; Zhao & Bilen, 2023), and quantization (Zhou et al., 2023a; Li et al., 2025). Within this setting, existing data selection methods can be broadly categorized into three types:

Score-based Data Selection. These methods assess sample importance using various sample-wise scoring criteria. EL2N and GraNd (Paul et al., 2021) score samples based on the expected ℓ_2-norm of the error vector and the gradient norm, respectively. Other commonly used metrics include Forgetting (Toneva et al., 2019), Memorization (Feldman & Zhang, 2020), and Uncertainty (Nguyen et al., 2022; He et al., 2024). Other approaches estimate sample importance through changes in the empirical risk, such as the moving-one-sample-out criterion (Tan et al., 2023).

Geometry-based Data Selection. Another line of research selects samples based on their geometric characteristics in feature space (Sorscher et al., 2022). These methods typically aim to improve the feature-space coverage or representativeness of the coreset. Herding (Welling, 2009) iteratively selects samples so that the mean representation of the selected subset approximates that of the full training data. Other methods leverage the distance to previously selected samples (Sener & Savarese, 2018) or the distance to the feature median (Xia et al., 2023) for sample selection.

Optimization-based Data Selection. These approaches construct coresets by optimizing subset selection objectives. Gradient matching methods (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a) select samples such that the gradients induced by the coreset approximate those of the full training set. Other approaches formulate data selection through bilevel optimization or validation-based objectives (Borsos et al., 2020; Killamsetty et al., 2021b; Zhou et al., 2022).

Existing methods mainly focus on sample-wise importance, feature-space representativeness, or optimization-related properties of the full training set. In contrast, we revisit data selection from a generalization perspective and use a domain generalization error bound (Albuquerque et al., 2020) to characterize what a generalizable coreset should preserve beyond the full-data ERM solution.

3 PRELIMINARY ANALYSIS

3.1 PROBLEM SETUP

Data selection aims to identify a representative subset, or coreset, that reduces the training data size while maximally preserving model performance (Sener & Savarese, 2018; Coleman et al., 2020; Xia et al., 2023; Yang et al., 2025). Let $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote the full training set of n samples, where $\mathbf{x}_i \in \mathcal{X}$ is the input and $y_i \in \mathcal{Y} = \{1, \ldots, C\}$ is the corresponding class label. Given a selection ratio $r \in (0, 1)$, the goal is to identify a coreset $\mathcal{D}'_{tr} \subset \mathcal{D}_{tr}$ with $|\mathcal{D}'_{tr}| = r \cdot n$ such that the hypothesis $h_{\mathcal{D}'_{tr}}$ trained on the selected data achieves strong generalization to unseen test data.

3.2 AN ERM VIEW OF DATA SELECTION

A natural way to view data selection is through empirical risk minimization (ERM) (Vapnik, 1991). Let P and Q denote the distributions of the observed training data and unseen test data, respectively. Under the i.i.d. assumption that training and test samples are drawn from the same distribution, ERM on the observed training set provides an empirical approximation to minimizing the expected risk on unseen data (Vapnik & Chervonenkis, 1971; Vapnik, 1999). From this perspective, data selection seeks a compact coreset that approximates the learning behavior of the full training set.

Although existing data selection methods adopt diverse criteria, they can be interpreted through an ERM-oriented view, where the coreset is expected to approximate the learning behavior induced by the full training set. Specifically, these approaches focus on sample-wise importance, feature-space representativeness, or preserving optimization-related properties of the full training set.

However, preserving the full-data ERM solution alone may not be sufficient for effective data selection. ERM optimizes predictive performance with respect to the observed training distribution P, whereas the ultimate goal is to perform well on unseen data drawn from a test distribution Q. In practice, the classical i.i.d. assumption may not hold exactly, and discrepancies between P and Q can arise from dataset collection, construction, and sampling effects (Torralba & Efros, 2011; Koh et al., 2021). Importantly, such discrepancies do not necessarily require explicit domain changes: they may exist even between the training and test splits of standard benchmarks constructed from the same underlying dataset. For example, distributional discrepancies have been observed in MNIST (LeCun et al., 1998; Rabanser et al., 2019), potentially due to sampling biases introduced during dataset construction (Yadav & Bottou, 2019); we also observe similar discrepancies on ImageNet-1K (see Appendix A). Consequently, a coreset that closely reproduces the full-data ERM solution may still overlook properties that are important for generalization under distributional discrepancies. This motivates us to revisit data selection from a generalization perspective and ask what a generalizable coreset should preserve beyond the full-data ERM solution.

3.3 A GENERALIZATION VIEW OF DATA SELECTION

To characterize what a coreset should preserve beyond the full-data ERM solution, we draw on a domain generalization error bound and analyze data selection from a generalization perspective. We view the training set $\mathcal{D}_{tr}$ as M mutually disjoint subsets, $\mathcal{D}_{tr} = \bigcup_{i=1}^{M} \mathcal{D}_i$ with $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$ for $i \neq j$, where each subset $\mathcal{D}_i$ is associated with a distribution P_i. These subsets capture heterogeneous structures in the training data and need not correspond to predefined domains; they may be given by metadata or inferred from latent data structure (Muandet et al., 2013; Matsuura & Harada, 2020).

Following the generalization analysis of Albuquerque et al. (2020), we consider the convex hull of these subset distributions and examine how the expected risk on unseen test data can be related to several training-side and distributional factors. For a test distribution Q, we use the following generalization bound to guide our characterization of what a generalizable coreset should preserve.

Theorem 1 (Domain Generalization Error Bound (Albuquerque et al., 2020)). *Let* $\Lambda_{tr} = \left\{ \sum_{i=1}^{M} \pi_i P_i \,\middle|\, \pi \in \Delta_M \right\}$ *denote the convex hull of the training distributions, where* Δ_M *denotes the* $(M-1)$*-dimensional simplex. For any hypothesis* $h \in \mathcal{H}$*, the expected risk satisfies:*

$$\epsilon_Q(h) \leq \sum_{i=1}^{M} \pi_i^* \epsilon_{P_i}(h) + \frac{\rho}{2} + \lambda_{\mathcal{H},(P^*,Q)} + \frac{\gamma}{2}, \tag{1}$$

where $\pi^* \in \arg\min_{\pi \in \Delta_M} d_{\mathcal{H}}\left(\sum_{i=1}^{M} \pi_i P_i, Q\right)$ *denotes mixture weights whose corresponding training mixture is closest to the test distribution* Q*, and* $P^* = \sum_{i=1}^{M} \pi_i^* P_i$ *denotes a closest training mixture within* Λ_{tr}*. Here,* $\rho = \max_{1 \le i < j \le M} d_{\mathcal{H}}(P_i, P_j)$ *denotes the maximum pairwise discrepancy among the component distributions and upper-bounds the diameter of* Λ_{tr}*, while* $\lambda_{\mathcal{H},(P^*,Q)} = \min_{h \in \mathcal{H}} [\epsilon_{P^*}(h) + \epsilon_Q(h)]$ *denotes the ideal joint error. Finally,* $\gamma = d_{\mathcal{H}}(P^*, Q)$ *quantifies the discrepancy between the test distribution* Q *and its closest approximation within* Λ_{tr}*.*

Although Theorem 1 is stated using the training mixture P^* closest to Q, the same form applies to any $\pi \in \Delta_M$. For $P_\pi = \sum_{i=1}^{M} \pi_i P_i$ and $\gamma_\pi = d_{\mathcal{H}}(P_\pi, Q)$, the bound can be written with π, P_π, γ_π, and $\lambda_{\mathcal{H},(P_\pi,Q)}$ replacing their counterparts associated with P^*. Since several terms depend on the unavailable test distribution Q, we use the generalization bound as design guidance rather than a directly optimizable objective, and translate its terms into the following principles for data selection:

(1) Preserve predictive utility ($\sum_i \pi_i \epsilon_{P_i}(h)$). This term is the expected risk over the training mixture P_π, suggesting that a coreset should preserve reliable task-relevant supervision.

(2) Encourage cross-subset invariance ($\rho/2$). This term characterizes discrepancies among the observed subset distributions, motivating the selection of samples that rely less on subset-specific variations and more on shared predictive patterns that remain stable across subsets.

(3) Maintain semantic consistency ($\lambda_{\mathcal{H},(P_\pi,Q)}$). The ideal joint error reflects whether a common hypothesis can perform well across the training mixture and unseen data. Since Q is unavailable, we use consistency of predictive structure across the observed subsets as a training-side characterization.

(4) Promote distributional coverage ($\gamma_\pi/2$). This term measures the discrepancy between a training mixture within Λ_{tr} and the unseen distribution Q. Since the mixture weights most relevant to Q are unknown, the coreset should maintain broad coverage of the observed subset distributions.

4 Generalization-Theory-Guided Data Selection

4.1 Overview

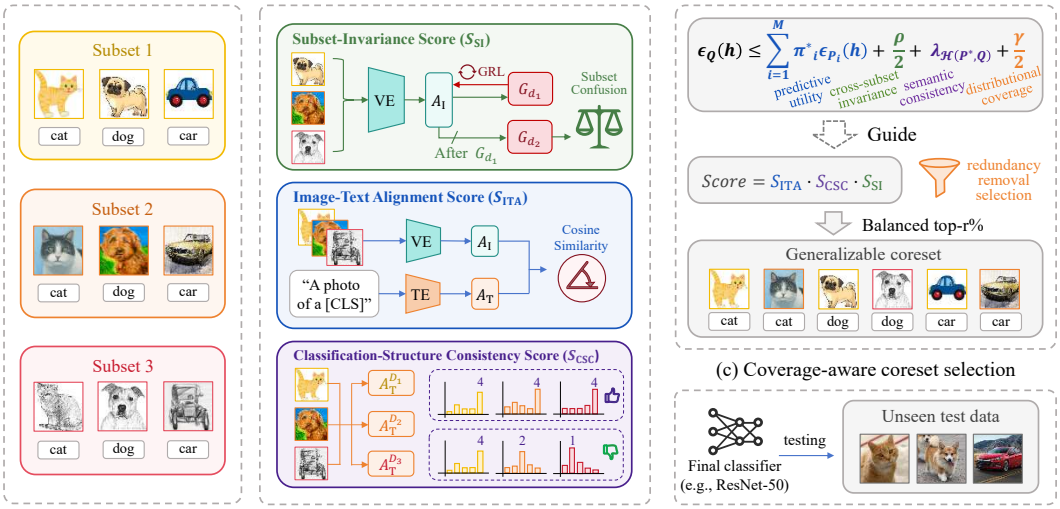
Figure 1 presents the overall pipeline of our **Gen**eralizable **D**ata **S**election (**GenDS**) framework. Guided by the four principles established in Section 3.3, GenDS consists of two stages: **Sample-level Multi-dimensional Scoring** and **Coverage-aware Coreset Selection**. Before scoring, the training data are partitioned into multiple subsets using either available annotations or latent subset discovery. The scoring stage evaluates each sample in terms of predictive utility, cross-subset invariance, and semantic consistency using the Image-Text Alignment Score (S_{ITA}), Subset-Invariance Score (S_{SI}), and Classification-Structure Consistency Score (S_{CSC}), respectively. The selection stage integrates the three complementary sample-level scores and performs coverage-aware selection across classes and subsets while reducing redundancy to construct the final coreset.

The three sample-level scores instantiate the first three principles, while the Coverage-aware Coreset Selection stage addresses distributional coverage. The resulting coreset is then used for downstream training. The following subsections detail the three scoring criteria and the selection procedure. A summary of the main notation, together with pseudocode for GenDS, is provided in Appendix F.

4.2 Subset-Invariance Score

Guided by Principle (2), which is motivated by the cross-subset divergence term, we prioritize samples with strong cross-subset invariance. Such samples are less dependent on subset-specific information and are more likely to remain informative across training subsets (Arjovsky et al., 2019).

Subset Partition. Measuring subset invariance requires a subset assignment d_i for each training sample. When explicit subset annotations are available, as in Office-Home (Venkateswara et al., 2017), we use them directly. Otherwise, inspired by prior work on latent domain discovery (Matsuura & Harada, 2020; Bose et al., 2024), we infer latent subsets by clustering style representations extracted from a frozen CLIP vision encoder (Radford et al., 2021). We then apply Variational Fair Clustering (Ziko et al., 2021) to prevent clusters from being dominated by particular classes, and use the resulting clusters as subset assignments. Implementation details are provided in Appendix B.



(a) Subset partition (b) Sample-level multi-dimensional scoring (c) Coverage-aware coreset selection (d) Training and testing

Figure 1: Overview of the proposed Generalizable Data Selection (GenDS) framework.

Subset-Invariance Scoring. We characterize sample-level subset invariance by measuring how much subset-specific information remains recoverable after suppressing subset-dependent information in the learned representation. To this end, we first learn adapted representations that preserve category-relevant information while reducing their dependence on subset identity. Following CLIP-Adapter (Gao et al., 2024), we obtain adapted visual features $\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x})$ and textual class prototypes $\tilde{\boldsymbol{f}}_{\mathrm{T}}(\mathbf{t}_c)$. Let G_y denote the category predictor formed by the adapted textual prototypes, and let G_{d_1} denote the subset discriminator operating on the adapted visual features. Following domain-adversarial learning (Ganin et al., 2016), the adapted representations are learned through

$$\min_{\theta_{\text{adapter}}} \max_{G_{d_1}} \mathbb{E}_{(\mathbf{x},y,d)\sim P_S} \left[\mathcal{L}_{\text{InfoNCE}}\big(G_y(\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x})), y\big) - \lambda_d \mathcal{L}_{\text{CE}}\big(G_{d_1}(\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x})), d\big) \right], \tag{2}$$

where $P_S = \frac{1}{M}\sum_{m=1}^{M} P_m$ denotes the uniform subset mixture, and θ_{adapter} the adapter parameters.

We then assess the residual recoverability of subset identity from the adapted representation. Since G_{d_1} participates in the adversarial objective, its chance-level performance does not necessarily imply that subset-specific information has been removed from the representation. Prior work has shown that such information may remain recoverable by separately trained classifiers even when the adversarial discriminator fails to distinguish the target attribute (Elazar & Goldberg, 2018; Ravfogel et al., 2020). We therefore freeze the learned representation and train an independent subset probe G_{d_2} to evaluate the remaining recoverability of subset identity. Let $p_{i,j}$ denote the probability that G_{d_2} assigns sample $\mathbf{x}_i$ to subset j. We define the **Subset-Invariance Score** (S_{SI}) as

$$S_{\text{SI}}(\mathbf{x}_i) = -\sum_{j=1}^{M} p_{i,j} \log p_{i,j}. \tag{3}$$

A higher S_{SI} indicates greater subset-prediction uncertainty, suggesting stronger subset invariance.

4.3 Image-Text Alignment Score

Guided by Principle (1), we prioritize samples whose visual content is semantically consistent with their class labels, as they provide more reliable supervision. We define the **Image-Text Alignment Score** (S_{ITA}) as the semantic alignment between each image and its textual prototype in the adapted CLIP representation space (Hessel et al., 2021). For a sample $\mathbf{x}_i$ with label y_i, we compute

$$S_{\text{ITA}}(\mathbf{x}_i) = \cos\left(\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x}_i), \tilde{\boldsymbol{f}}_{\mathrm{T}}(\mathbf{t}_{y_i})\right), \tag{4}$$

where $\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x}_i)$ and $\tilde{\boldsymbol{f}}_{\mathrm{T}}(\mathbf{t}_{y_i})$ denote the adapted visual feature and textual prototype, respectively. A higher S_{ITA} indicates stronger semantic alignment between the image and its class.

4.4 CLASSIFICATION-STRUCTURE CONSISTENCY SCORE

Guided by Principle (3), we consider the ideal joint error term $\lambda_{\mathcal{H},(P_\pi,Q)}$, which reflects whether a common hypothesis can perform well on both the training mixture P_π and the test distribution Q. Since Q is unavailable during data selection, this term cannot be directly evaluated. We therefore seek a training-side characterization based on the observed subsets. Motivated by multi-source adaptation analyses that consider target mixtures of source distributions (Mansour et al., 2008; Hoffman et al., 2018), we use agreement among subset-specific labeling functions as a training-side signal of classification-structure consistency across the observed subsets.

For each subset $\mathcal{D}_i$, we train a subset-specific classifier $h_{\mathcal{D}_i}$ to approximate its labeling function g_{tr_i}. Given a sample $\mathbf{x}_k$, these classifiers produce M predictive distributions, and we measure their pairwise consistency. We define the **Classification-Structure Consistency Score (S_{CSC})** as

$$S_{\mathrm{CSC}}(\mathbf{x}_k) = 1 - \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathrm{JS}\left(h_{\mathcal{D}_i}(\mathbf{x}_k), h_{\mathcal{D}_j}(\mathbf{x}_k)\right), \tag{5}$$

where $\mathrm{JS}(\cdot,\cdot)$ denotes the Jensen–Shannon (JS) divergence (Lin, 1991). A higher S_{CSC} indicates stronger agreement among subset-specific classifiers, suggesting a more consistent classification structure across training subsets. The theoretical motivation for using pairwise classification consistency as a training-side characterization of the ideal joint error term is provided in Appendix C.

4.5 COVERAGE-AWARE CORESET SELECTION

We first integrate the three scores into a unified utility score. For each sample $\mathbf{x}_i$, we define

$$U(\mathbf{x}_i) = \mathrm{Norm}\big(S_{\mathrm{ITA}}(\mathbf{x}_i)\big) \cdot \mathrm{Norm}\big(S_{\mathrm{SI}}(\mathbf{x}_i)\big) \cdot \mathrm{Norm}\big(S_{\mathrm{CSC}}(\mathbf{x}_i)\big), \tag{6}$$

where $\mathrm{Norm}(\cdot)$ denotes the normalization applied to each score, and the multiplicative aggregation favors samples that perform consistently well across the three complementary criteria.

Guided by Principle (4), we promote distributional coverage across both classes and subsets through coverage-aware budget allocation and redundancy control. Let N_c denote the number of training samples in class c. We first assign a class-wise budget $k_c = \lfloor rN_c \rfloor$, which preserves the original class proportions. Since the test distribution Q is unavailable, the mixture weights π^* corresponding to the closest training mixture cannot be identified. We therefore avoid severe underrepresentation of any observed subset by assigning each subset within class c a balanced base quota $k_{c,d} = \lfloor k_c/M \rfloor$. Under these quotas, candidates within each class-subset partition (c,d) are ranked in descending order by $U(\mathbf{x}_i)$ and selected greedily with redundancy control. A candidate $\mathbf{x}_i$ is selected only if

$$\max_{\mathbf{x}_k \in \mathcal{A}_{c,d}} \cos\left(\tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x}_i), \tilde{\boldsymbol{f}}_{\mathrm{I}}(\mathbf{x}_k)\right) < \tau, \tag{7}$$

where $\mathcal{A}_{c,d}$ denotes the samples already selected from the corresponding class-subset partition, and τ is a redundancy threshold. Selection continues until $|\mathcal{A}_{c,d}| = k_{c,d}$ or no eligible candidates remain.

If the class-wise budget k_c is not fully utilized, we supplement the selected set using a redundancy-aware criterion inspired by Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998). Let $\mathcal{A}_c = \bigcup_{d=1}^{M} \mathcal{A}_{c,d}$. Remaining candidates are iteratively selected according to $U_{\mathrm{RA}}(\mathbf{x}_i) = U(\mathbf{x}_i)\big(1 - S_{\cos}(\mathbf{x}_i)\big)$, where $S_{\cos}(\mathbf{x}_i)$ denotes the maximum similarity between $\mathbf{x}_i$ and the samples already selected in $\mathcal{A}_c$, until $|\mathcal{A}_c| = k_c$. Finally, the resulting coreset is $\mathcal{D}'_{\mathrm{tr}} = \bigcup_{c=1}^{C} \mathcal{A}_c$.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Following the task settings adopted in prior data selection studies (Xia et al., 2023; Tan et al., 2023; Yang et al., 2025; Griffin et al., 2026), we focus on image classification tasks in this work.

Datasets. We conduct experiments on three standard image classification benchmarks (CIFAR-100 (Krizhevsky et al., 2009), ImageNet-1K (Deng et al., 2009), Tiny-ImageNet (Le & Yang, 2015)) and three multi-domain benchmarks (VLCS (Fang et al., 2013), Office-Home (Venkateswara et al., 2017), miniDomainNet (Zhou et al., 2021)). See Appendix D for detailed dataset statistics.

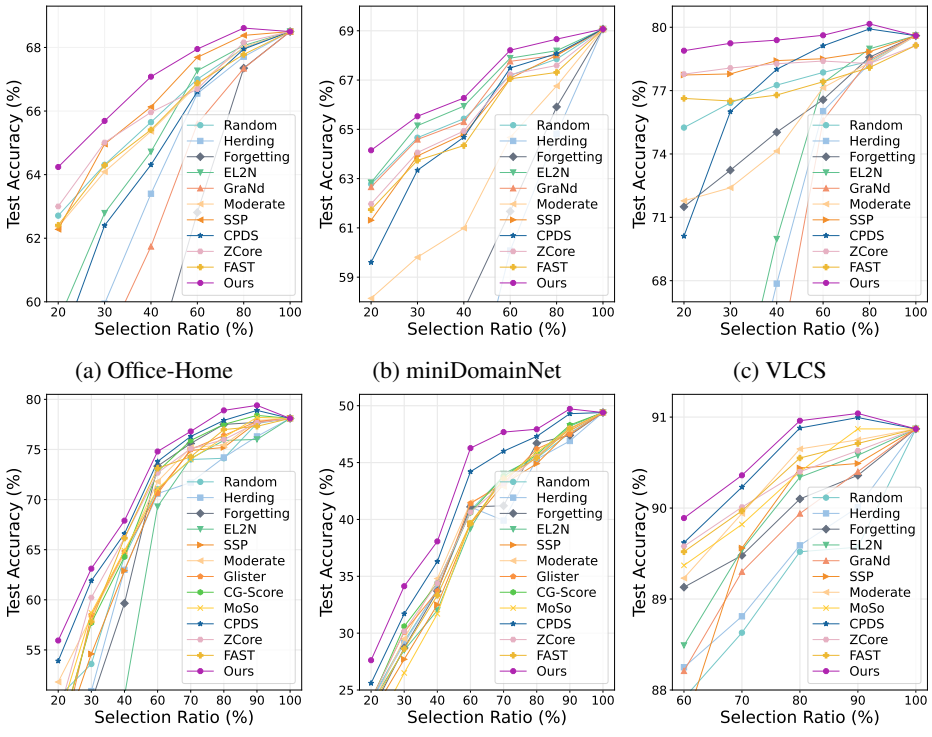
Baselines. We compare GenDS against multiple representative data selection methods: (1) Random; (2) Herding (Welling, 2009); (3) Forgetting (Toneva et al., 2019); (4) GraNd (Paul et al., 2021); (5) EL2N (Paul et al., 2021); (6) Glister (Killamsetty et al., 2021b); (7) SSP (Sorscher et al., 2022); (8) Moderate-DS (Xia et al., 2023); (9) CG-Score (Ki et al., 2023); (10) MoSo (Tan et al., 2023); (11) CPDS (Yang et al., 2025); (12) ZCore (Griffin et al., 2026); (13) FAST (Cui et al., 2026).

Implementation. We use CLIP (Radford et al., 2021) as the backbone for the scoring phase and employ CLIP-Adapter (Gao et al., 2024) for fine-tuning. More details are provided in Appendix E.

5.2 Main Results

Following the evaluation protocols of previous data selection works on standard image classification benchmarks (Sorscher et al., 2022; Xia et al., 2023; Yang et al., 2025), we report Top-1 accuracy on CIFAR-100 and Tiny-ImageNet, and Top-5 accuracy on ImageNet-1K. For multi-domain benchmarks, we adopt the leave-one-domain-out protocol (Muandet et al., 2013; Gulrajani & Lopez-Paz, 2021; Zhou et al., 2023b), where one domain is held out for testing and the remaining domains are used for training. We report Top-1 accuracy averaged across all held-out domains.

As shown in Figure 2, our method consistently achieves the best performance across all six datasets under varying selection ratios, especially at low ratios such as 20% and 30%. It performs well on multi-domain benchmarks with significant distribution shifts while maintaining strong results on standard image classification benchmarks, demonstrating broad applicability. Moreover, at high ratios such as 80%, the coresets even outperform the full dataset, suggesting that our method also acts as an effective data cleaning mechanism that reduces data scale without sacrificing performance.



(a) Office-Home (b) miniDomainNet (c) VLCS

(d) CIFAR-100 (e) Tiny-ImageNet (f) ImageNet-1K

Figure 2: Comparison of our method with various data selection baselines on (a) Office-Home, (b) miniDomainNet, (c) VLCS, (d) CIFAR-100, (e) Tiny-ImageNet, and (f) ImageNet-1K.

5.3 Robustness Evaluation

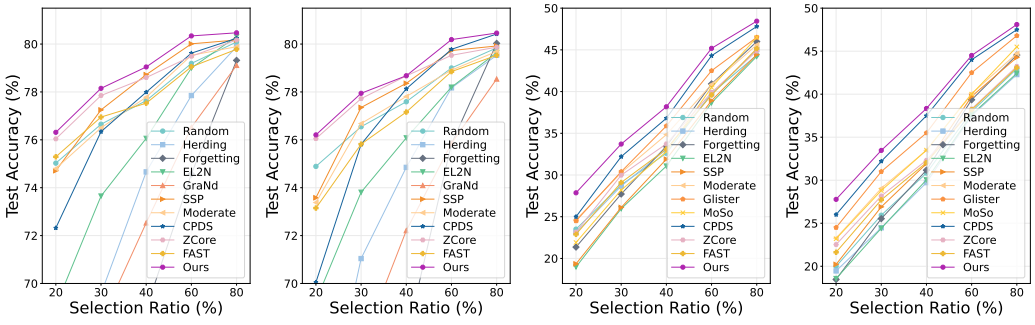
We evaluate the robustness of our method under noisy-label and corrupted-image settings. For noisy labels, following prior works (Tanaka et al., 2018; Wei et al., 2024), we introduce symmetric noise

by randomly flipping 20% of the training labels on Office-Home, CIFAR-100, and Tiny-ImageNet. For corrupted images, we construct realistic image corruption scenarios (Hendrycks & Dietterich, 2019; Xia et al., 2023), including Gaussian noise, random occlusion, resolution reduction, fog, and motion blur, at corruption rates of 5% and 20%. Illustrative examples are provided in Appendix H.

As shown in Table 1 and Figure 3, our method consistently outperforms all baselines under both settings. For noisy samples, S_{ITA} identifies image-label semantic inconsistencies and assigns lower scores to unreliable samples, while S_{CSC} and S_{SI} further favor samples with stable prediction structures and stronger cross-subset invariance. Notably, corrupted samples are not necessarily harmful: mildly corrupted samples may still retain useful semantic information for representation learning, and our method can effectively preserve such samples. For example, on Tiny-ImageNet at an 80% selection ratio, the accuracy reaches 48.44% and 48.10% under 5% and 20% corruption, respectively, compared with 47.72% on clean samples.

Table 1: Experimental results (accuracy (%), mean ± std) on Office-Home (R), CIFAR-100, and Tiny-ImageNet with noisy labels. Office-Home (R) denotes the Sim-to-Real setting, where the coreset is selected from the Art, Clipart, and Product domains and tested on the Real-World domain. 20% of the training labels are randomly flipped to other classes to simulate symmetric label noise.

Method / Selection Ratio	Office-Home (R)		CIFAR-100		Tiny-ImageNet	
	20%	30%	20%	30%	20%	30%
Random	69.50 ± 0.72	72.06 ± 0.76	34.47 ± 0.64	43.26 ± 1.21	17.78 ± 0.44	23.88 ± 0.42
Herding	64.02 ± 0.62	71.35 ± 0.16	42.29 ± 1.75	50.52 ± 3.38	18.98 ± 0.44	24.23 ± 0.29
Forgetting	56.71 ± 0.78	62.83 ± 0.67	36.53 ± 1.11	45.78 ± 1.04	13.20 ± 0.38	21.79 ± 0.43
GraNd	52.18 ± 1.33	62.00 ± 1.08	31.72 ± 0.67	42.80 ± 0.30	18.28 ± 0.32	23.72 ± 0.18
EL2N	42.53 ± 0.86	55.44 ± 0.80	29.82 ± 1.19	33.62 ± 2.35	13.93 ± 0.69	18.57 ± 0.31
SSP	70.31 ± 0.53	73.93 ± 0.91	31.08 ± 0.78	41.87 ± 0.63	15.10 ± 0.73	21.01 ± 0.36
Moderate-DS	68.03 ± 0.60	70.88 ± 0.67	40.25 ± 0.12	48.53 ± 1.60	19.64 ± 0.40	24.96 ± 0.30
CPDS	71.38 ± 1.07	73.83 ± 0.27	46.05 ± 0.21	58.34 ± 0.36	24.50 ± 0.17	30.69 ± 0.35
ZCore	71.61 ± 0.64	73.67 ± 0.71	35.68 ± 2.16	45.32 ± 2.05	18.58 ± 0.67	24.50 ± 0.51
FAST	71.10 ± 0.15	73.26 ± 0.31	33.72 ± 0.98	43.26 ± 1.88	18.61 ± 0.10	23.49 ± 0.53
Ours	**72.58** ± 0.29	**75.17** ± 0.34	**50.54** ± 1.41	**58.97** ± 0.46	**25.35** ± 0.21	**31.18** ± 0.55



(a) Office-Home (R), 5% (b) Office-Home (R), 20% (c) Tiny-ImageNet, 5% (d) Tiny-ImageNet, 20%

Figure 3: Comparison of robustness to corrupted images under different corruption rates.

5.4 Ablation Study

Ablation on Components. Table 2 evaluates the individual components of GenDS. The full combination consistently performs best, confirming the contribution of each component. S_{SI} shows a larger effect on Office-Home, where subset distributions differ substantially, while S_{ITA} contributes more than S_{SI} and S_{CSC} on Tiny-ImageNet, where cross-subset distribution shift is smaller.

Additional Ablations. See Appendix I for additional ablation studies on the role of the text modality and the effectiveness of the multiplicative score fusion strategy in Eq. 6.

Table 2: Component ablation of GenDS on Office-Home (R) and Tiny-ImageNet.

Method	Office-Home (R)			Tiny-ImageNet		
	20%	30%	40%	20%	30%	40%
S_{ITA}	75.90 $\pm$ 0.21	76.89 $\pm$ 0.25	78.52 $\pm$ 0.09	26.76 $\pm$ 0.42	33.43 $\pm$ 0.33	37.60 $\pm$ 0.21
S_{SI}	76.31 $\pm$ 0.12	78.10 $\pm$ 0.07	79.31 $\pm$ 0.45	23.59 $\pm$ 0.32	29.11 $\pm$ 0.37	34.98 $\pm$ 0.23
S_{CSC}	75.12 $\pm$ 0.53	76.93 $\pm$ 0.35	78.26 $\pm$ 0.29	24.79 $\pm$ 0.17	29.47 $\pm$ 0.37	35.36 $\pm$ 0.16
Redundancy reduction	76.19 $\pm$ 0.64	77.79 $\pm$ 0.51	79.07 $\pm$ 0.41	24.35 $\pm$ 0.11	30.03 $\pm$ 0.08	34.59 $\pm$ 0.14
w/o S_{ITA}	77.39 $\pm$ 0.22	78.70 $\pm$ 0.03	79.80 $\pm$ 0.14	26.86 $\pm$ 0.26	32.60 $\pm$ 0.14	36.77 $\pm$ 0.83
w/o S_{SI}	77.23 $\pm$ 0.26	78.49 $\pm$ 0.13	79.50 $\pm$ 0.29	27.41 $\pm$ 0.19	33.62 $\pm$ 0.66	37.94 $\pm$ 0.47
w/o S_{CSC}	77.26 $\pm$ 0.30	78.78 $\pm$ 0.05	79.89 $\pm$ 0.05	27.54 $\pm$ 0.19	33.85 $\pm$ 0.21	37.63 $\pm$ 0.28
w/o Redundancy reduction	76.43 $\pm$ 0.35	77.75 $\pm$ 0.32	78.85 $\pm$ 0.19	27.32 $\pm$ 0.36	33.13 $\pm$ 0.52	37.55 $\pm$ 0.30
Ours	**78.52** $\pm$ 0.29	**79.14** $\pm$ 0.21	**80.22** $\pm$ 0.33	**27.62** $\pm$ 0.14	**34.13** $\pm$ 0.11	**38.07** $\pm$ 0.56

5.5 Additional Analyses

Sensitivity Analyses. The redundancy threshold τ controls the degree of redundancy removal, while K determines the number of clusters used for latent subset discovery on standard image classification benchmarks. As shown in Table 3, performance remains relatively stable across the evaluated values of τ and K, indicating that our method is robust to these hyperparameters.

Table 3: Sensitivity to the redundancy threshold τ on Office-Home (R) and CIFAR-100, and to the number of latent clusters K on CIFAR-100, at a 40% selection ratio.

τ / Dataset	Office-Home (R)	CIFAR-100		
		$K = 2$	$K = 3$	$K = 5$
0.70	79.81	67.58	67.09	67.06
0.75	80.12	67.86	67.42	67.18
0.80	80.22	67.90	67.61	67.44
0.85	79.80	67.82	67.22	66.89
0.90	79.77	67.19	67.59	66.95

Runtime Analysis. We evaluate all data selection methods on Tiny-ImageNet using the same NVIDIA RTX 4090 GPU and report the runtime for generating subsets at selection ratios ranging from 20% to 90%. GenDS takes 4.97 minutes, about 26.3% less time than CPDS (6.74 minutes), demonstrating favorable efficiency among training-based and CLIP-based methods. Although GenDS is slower than training-free approaches, its one-time selection cost can be amortized since the resulting coreset can be reused across different training configurations.

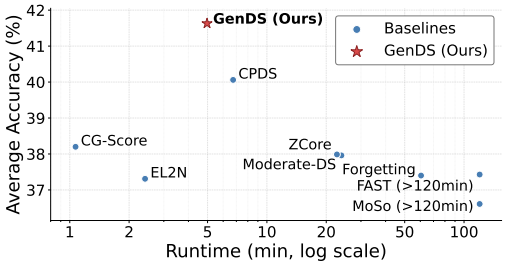


Figure 4: Runtime and accuracy comparison of data selection methods under the same setting.

Cross-architecture Adaptability Analysis. We evaluate the cross-architecture adaptability of the selected coresets across different network backbones. See Appendix G for experimental details.

6 Conclusion

In this paper, we revisit data selection beyond the conventional ERM-oriented perspective and study it from the perspective of generalization. Guided by the four individual terms in a domain generalization error bound, we formulate corresponding principles for data selection and design GenDS to operationalize them through complementary scoring and coverage-aware selection. Extensive experiments demonstrate its effectiveness across standard image classification and multi-domain benchmarks and across varying selection ratios. We hope this perspective can inspire future data selection methods that more explicitly account for generalization beyond the observed training data.

AI USE STATEMENT

We used generative AI tools for language polishing and improving manuscript readability, as well as to support literature retrieval and discovery, such as identifying potentially relevant prior work and search directions. AI-generated suggestions served only as an aid to literature search; all cited works were independently verified by the authors against their original sources for bibliographic accuracy and relevance to the corresponding claims. We did not rely on generative AI to establish the existence, content, or validity of any cited work. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

We provide descriptions of the proposed method, datasets, and experimental settings in the main paper and appendix. The appendix further provides theoretical motivation and detailed analysis related to the design of the scoring criteria. Ablation studies and sensitivity analyses are provided in the main paper and appendix to document our empirical evaluation.

REFERENCES

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804v4*, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. In *NeurIPS*, 2020.

Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *WACV*, 2024.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, 2022.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *ICLR*, 2020.

Jin Cui, Boran Zhao, Jiajun Xu, Jiaqi Guo, Shuo Guan, and Pengju Ren. Fast: Topology-aware frequency-domain distribution matching for coreset selection. In *CVPR*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *EMNLP*, 2018.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *NeurIPS*, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 2016.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 2024.

Priya Goyal, Benjamin Lefaudeux, Mannat Singh, Jeremy Reizenstein, Vinicius Reis, Min Xu, Matthew Leavitt, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. `https://github.com/facebookresearch/vissl`, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of machine learning research*, 2012.

Brent A Griffin, Jacob Marks, and Jason J Corso. Zero-shot coreset selection via iterative subspace sampling. In *WACV*, 2026.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. In *CVPRW*, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *NeurIPS*, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nohyun Ki, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *ICLR*, 2023.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *ICML*, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI*, 2021b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Muquan Li, Dongyang Zhang, Qiang Dong, Xiurui Xie, and Ke Qin. Adaptive dataset quantization. In *AAAI*, 2025.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 1991.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *NeurIPS*, 2008.

Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *ICML*, 2022.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *NeurIPS*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *ICLR*, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *NeurIPS*, 2022.

Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. In *NeurIPS*, 2023.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *ICLR*, 2019.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *NeurIPS*, 1991.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. In *NeurIPS*, 2024.

Max Welling. Herding dynamical weights to learn. In *ICML*, 2009.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2023.

Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. In *NeurIPS*, 2019.

Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, and Furao Shen. A clip-powered framework for robust and generalizable data selection. In *ICLR*, 2025.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2021.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *ICLR*, 2023.

Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization. In *ICCV*, 2023a.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 2021.

Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, 2023b.

Xiao Zhou, Renjie Pi, Weizhong Zhang, Yong Lin, Zonghao Chen, and Tong Zhang. Probabilistic bilevel coreset selection. In *ICML*, 2022.

Imtiaz Masud Ziko, Jing Yuan, Eric Granger, and Ismail Ben Ayed. Variational fair clustering. In *AAAI*, 2021.

Appendix

Table of Contents

A The Distribution Shift Between the ImageNet Training Set and Test Set **14**

B Latent subset partition **15**

C Theoretical Motivation for the Classification-Structure Consistency Score (S_{CSC}) **15**

C.1 Bounding the Ideal Joint Error 15

C.2 Extension to Mixture Distributions and Training-Side Structures 16

C.3 From Classification-Structure Consistency to a Sample-Level Score 17

D Datasets **18**

E Implementation Details **19**

F Detailed Algorithmic Framework of GenDS **19**

G Cross-architecture Adaptability Evaluation **20**

H Examples of Image Corruption Types **21**

I Additional Ablation Studies **22**

J Limitations **22**

A The Distribution Shift Between the ImageNet Training Set and Test Set

To further examine this phenomenon at scale, we quantify the distribution shift in the standard ImageNet (Deng et al., 2009) train-test split. Specifically, we randomly draw 10,000 training and 10,000 test samples (denoted as train-test), together with two disjoint sets of 10,000 training samples each (train-train) as a reference. Features are extracted using a ResNet-50 (He et al., 2016) and a CLIP (Radford et al., 2021) vision encoder. We then compute the mean pairwise L2 distance, Wasserstein-2 distance, maximum mean discrepancy (MMD) (Gretton et al., 2012), and linear separability (i.e., an empirical proxy for $\mathcal{H}$-divergence obtained by training a linear classifier to distinguish the two feature sets). As shown in Figure 5, all metrics are consistently larger for train-test than for train-train, revealing a mild but clear distribution shift even within this canonical i.i.d. benchmark.

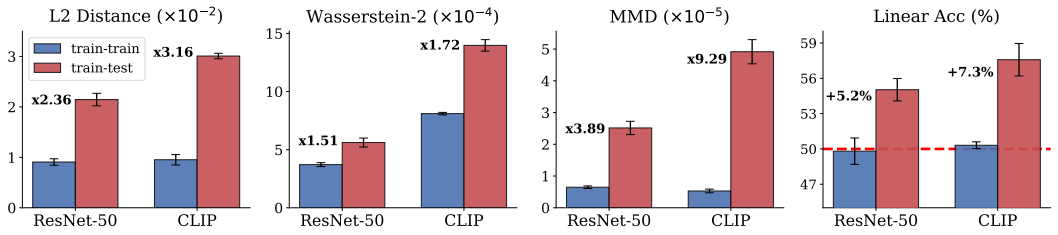


Figure 5: Train-Train vs. Train-Test comparison.

B LATENT SUBSET PARTITION

Given an image $\mathbf{x}_i$, we feed it into the CLIP vision encoder (Radford et al., 2021) and collect patch-token features from all Transformer blocks. To focus on local style statistics, we discard the CLS token and retain only patch tokens for style estimation. We divide all Transformer blocks into J consecutive groups. For the j-th group, the patch-token features from all blocks in this group are concatenated along the token dimension to form $\boldsymbol{\phi}_j(\mathbf{x}_i)$. We then compute the dimension-wise mean and standard deviation over the patch-token dimension and define the final style feature as:

$$\boldsymbol{f}_{\text{style}}(\mathbf{x}_i) = \left[\boldsymbol{\mu}(\boldsymbol{\phi}_1(\mathbf{x}_i)) \,; \boldsymbol{\sigma}(\boldsymbol{\phi}_1(\mathbf{x}_i)) \,; \cdots \,; \boldsymbol{\mu}(\boldsymbol{\phi}_J(\mathbf{x}_i)) \,; \boldsymbol{\sigma}(\boldsymbol{\phi}_J(\mathbf{x}_i))\right], \tag{8}$$

where $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ denote the dimension-wise mean and standard deviation computed over the token dimension, respectively, and $[\cdot;\cdot]$ denotes vector concatenation.

We compress the style features with PCA and perform fair K-means following Variational Fair Clustering (Ziko et al., 2021), which augments K-means with a KL-based regularizer. We optimize:

$$\min_{\{s_i\},\{\boldsymbol{c}_k\}_{k=1}^K} \sum_{i=1}^{N} \|\boldsymbol{z}_i - \boldsymbol{c}_{s_i}\|_2^2 + \lambda_{\text{fair}} \sum_{k=1}^{K} D_{\text{KL}}\big(U \,\|\, P_k\big), \tag{9}$$

where $\{s_i\}$ and $\{\boldsymbol{c}_k\}$ denote the cluster assignments and centroids, respectively, $\boldsymbol{z}_i$ denotes the PCA-compressed style feature, P_k is the empirical class distribution within cluster k, and U is the global class distribution of the dataset. This KL regularization encourages each cluster to align with the global class prior, making the discovered clusters more likely to reflect style variations rather than class-discriminative information. We treat the resulting fair clusters as a latent subset partition.

In our implementation, we set the number of clusters to $K = 2$ and the fairness coefficient to $\lambda_{\text{fair}} = 0.05$. To prevent overly strong regularization from dominating the optimization in early iterations, we adopt a warm-up and gradual ramp-up strategy. Specifically, the fairness weight is set to zero during the first T_{warm} iterations, allowing the clustering objective to first learn a meaningful structure based on feature similarity. After the warm-up phase, the fairness constraint is gradually strengthened by increasing its weight from 0 to λ_{fair}, leading to a more stable optimization process. Let t denote the current iteration. The time-dependent fairness weight is defined as

$$\lambda_t = \begin{cases} 0, & t < T_{\text{warm}}, \\ \lambda_{\text{fair}} \cdot \dfrac{t - T_{\text{warm}} + 1}{T_{\max} - T_{\text{warm}}}, & t \geq T_{\text{warm}}, \end{cases} \tag{10}$$

where $T_{\max}$ is the total number of iterations.

C THEORETICAL MOTIVATION FOR THE CLASSIFICATION-STRUCTURE CONSISTENCY SCORE (S_{CSC})

In this section, we provide the theoretical motivation connecting the ideal joint error term $\lambda_{\mathcal{H},(P_\pi,Q)}$ with our proposed sample-level Classification-Structure Consistency Score (S_{CSC}). The key idea is to characterize classification-structure compatibility using the observed training subsets and derive a tractable training-side proxy without requiring access to the unseen test distribution Q.

C.1 BOUNDING THE IDEAL JOINT ERROR

For an arbitrary training mixture $P_\pi = \sum_{i=1}^{M} \pi_i P_i$, where $\pi \in \Delta_M$, we consider the corresponding ideal joint error between the training mixture P_π and the test distribution Q, defined as:

$$\lambda_{\mathcal{H},(P_\pi,Q)} = \min_{h\in\mathcal{H}} \left[\epsilon_{P_\pi}(h) + \epsilon_Q(h)\right]. \tag{11}$$

This term characterizes whether there exists a common hypothesis in $\mathcal{H}$ that can simultaneously achieve low risk on the training mixture P_π and the unseen test distribution Q. A small $\lambda_{\mathcal{H},(P_\pi,Q)}$ therefore requires the classification structures of the two distributions to be sufficiently compatible with respect to the hypothesis space under consideration.

To make this structural interpretation explicit, let g_π and g_{te} denote the labeling functions associated with P_π and Q, respectively. For this analysis, we consider deterministic labeling functions and hypotheses represented by one-hot label vectors, and set $d_g(u,v) = \frac{1}{2}\|u - v\|_1$. For one-hot vectors,

this distance equals the zero–one disagreement, while its extension to probability vectors is a norm-induced metric. We further consider the realizable setting in which $g_\pi, g_{te} \in \mathcal{H}$ and each labeling function incurs zero risk on its associated distribution, i.e., $\epsilon_{P_\pi}(g_\pi) = 0$ and $\epsilon_Q(g_{te}) = 0$. Under these assumptions, the expected risks can be written as:

$$\epsilon_{P_\pi}(h) = \mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\big(h(\mathbf{x}), g_\pi(\mathbf{x})\big)\right], \quad \epsilon_Q(h) = \mathbb{E}_{\mathbf{x}\sim Q}\left[d_g\big(h(\mathbf{x}), g_{te}(\mathbf{x})\big)\right]. \tag{12}$$

Since the ideal joint error is minimized over $\mathcal{H}$, substituting either labeling function into the objective provides a valid upper bound. Taking $h = g_\pi$ gives:

$$\lambda_{\mathcal{H},(P_\pi,Q)} \le \epsilon_{P_\pi}(g_\pi) + \epsilon_Q(g_\pi) = \mathbb{E}_{\mathbf{x}\sim Q}\left[d_g\big(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\big)\right]. \tag{13}$$

Similarly, taking $h = g_{te}$ gives:

$$\lambda_{\mathcal{H},(P_\pi,Q)} \le \epsilon_{P_\pi}(g_{te}) + \epsilon_Q(g_{te}) = \mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\big(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\big)\right]. \tag{14}$$

Combining the two cases yields:

$$\lambda_{\mathcal{H},(P_\pi,Q)} \le \min\left\{\mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\big(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\big)\right], \mathbb{E}_{\mathbf{x}\sim Q}\left[d_g\big(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\big)\right]\right\}. \tag{15}$$

This relation shows that the discrepancy between the training-side and test-side labeling structures is associated with the compatibility underlying the ideal joint error. However, neither g_{te} nor samples from Q are available during data selection. We therefore focus on the P_π-side upper bound and seek a characterization of its structural discrepancy using only the observed training subsets.

C.2 Extension to Mixture Distributions and Training-Side Structures

Under the subset partition setting in Section 3.3, let $\{g_{tr_i}\}_{i=1}^M$ denote the labeling functions associated with the M training subsets, with $M \ge 2$, where $g_{tr_i}(\mathbf{x})$ denotes the corresponding label distribution for $\mathbf{x}$. Rather than assuming that the labeling function of a mixture distribution is directly given by a weighted sum of the subset-specific labeling functions, we use their convex combinations only as auxiliary representations of the classification structures spanned by the observed subsets.

For any coefficient vector $\beta \in \Delta_M$, define the corresponding aggregate classification structure as:

$$\bar{g}_\beta(\mathbf{x}) = \sum_{i=1}^M \beta_i g_{tr_i}(\mathbf{x}). \tag{16}$$

Consider two arbitrary coefficient vectors $\pi, \alpha \in \Delta_M$. Using the simplex identities $\sum_{i=1}^M \pi_i = 1$ and $\sum_{j=1}^M \alpha_j = 1$, their difference can be expressed as:

$$\begin{aligned}\bar{g}_\pi(\mathbf{x}) - \bar{g}_\alpha(\mathbf{x}) &= \left(\sum_{i=1}^M \pi_i g_{tr_i}(\mathbf{x})\right)\left(\sum_{j=1}^M \alpha_j\right) - \left(\sum_{j=1}^M \alpha_j g_{tr_j}(\mathbf{x})\right)\left(\sum_{i=1}^M \pi_i\right) \\ &= \sum_{i=1}^M \sum_{j=1}^M \pi_i \alpha_j \left(g_{tr_i}(\mathbf{x}) - g_{tr_j}(\mathbf{x})\right).\end{aligned} \tag{17}$$

For a norm-induced metric $d_g(\cdot,\cdot)$, applying the triangle inequality gives:

$$\begin{aligned}d_g\left(\bar{g}_\pi(\mathbf{x}), \bar{g}_\alpha(\mathbf{x})\right) &\le \sum_{i=1}^M \sum_{j=1}^M \pi_i \alpha_j d_g\left(g_{tr_i}(\mathbf{x}), g_{tr_j}(\mathbf{x})\right) \\ &\le \max_{i\ne j} d_g\left(g_{tr_i}(\mathbf{x}), g_{tr_j}(\mathbf{x})\right).\end{aligned} \tag{18}$$

Thus, pairwise disagreement among the subset-specific labeling functions controls the variation between arbitrary convex combinations of the observed classification structures.

We define the average pairwise disagreement among the subset-specific labeling functions as:

$$\overline{\mathcal{D}}_g(\mathbf{x}) = \frac{2}{M(M-1)} \sum_{i=1}^M \sum_{j=i+1}^M d_g\left(g_{tr_i}(\mathbf{x}), g_{tr_j}(\mathbf{x})\right). \tag{19}$$

It follows that:

$$\max_{i\neq j} d_g\left(g_{tr_i}(\mathbf{x}), g_{tr_j}(\mathbf{x})\right) \leq \sum_{i=1}^{M}\sum_{j=i+1}^{M} d_g\left(g_{tr_i}(\mathbf{x}), g_{tr_j}(\mathbf{x})\right) = \frac{M(M-1)}{2}\overline{\mathcal{D}}_g(\mathbf{x}). \tag{20}$$

Substituting this relation into equation 18 yields:

$$d_g\left(\bar{g}_\pi(\mathbf{x}), \bar{g}_\alpha(\mathbf{x})\right) \leq \frac{M(M-1)}{2}\overline{\mathcal{D}}_g(\mathbf{x}), \qquad \forall \pi, \alpha \in \Delta_M. \tag{21}$$

Therefore, small average pairwise disagreement among the subset-specific labeling functions implies that different convex combinations of the observed classification structures remain close. This result holds for arbitrary mixture weights π and α and does not require assuming that the underlying mixture coefficients are uniform.

We next relate this training-side quantity to the inaccessible test labeling structure. Let g_π denote the actual labeling function associated with P_π, which is not assumed to equal $\bar{g}_\pi$, and let g_{te} denote the labeling function associated with Q. For any $\alpha \in \Delta_M$, the triangle inequality gives:

$$d_g\left(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\right) \leq d_g\left(g_\pi(\mathbf{x}), \bar{g}_\pi(\mathbf{x})\right) + d_g\left(\bar{g}_\pi(\mathbf{x}), \bar{g}_\alpha(\mathbf{x})\right) + d_g\left(\bar{g}_\alpha(\mathbf{x}), g_{te}(\mathbf{x})\right). \tag{22}$$

Combining equation 21 with equation 22 yields:

$$d_g\left(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\right) \leq d_g\left(g_\pi(\mathbf{x}), \bar{g}_\pi(\mathbf{x})\right) + \frac{M(M-1)}{2}\overline{\mathcal{D}}_g(\mathbf{x}) + d_g\left(\bar{g}_\alpha(\mathbf{x}), g_{te}(\mathbf{x})\right). \tag{23}$$

Taking expectation with respect to the training mixture P_π gives:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\left(g_\pi(\mathbf{x}), g_{te}(\mathbf{x})\right)\right] \leq\ & \mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\left(g_\pi(\mathbf{x}), \bar{g}_\pi(\mathbf{x})\right)\right] \\ & + \frac{M(M-1)}{2}\mathbb{E}_{\mathbf{x}\sim P_\pi}\left[\overline{\mathcal{D}}_g(\mathbf{x})\right] \\ & + \mathbb{E}_{\mathbf{x}\sim P_\pi}\left[d_g\left(\bar{g}_\alpha(\mathbf{x}), g_{te}(\mathbf{x})\right)\right].\end{aligned} \tag{24}$$

Together with the P_π-side inequality in equation 15, this decomposition identifies the expected pairwise disagreement among the subset-specific labeling functions as a training-side component associated with the structural compatibility underlying $\lambda_{\mathcal{H},(P_\pi,Q)}$.

The first term reflects the discrepancy between the actual labeling structure of the training mixture and the corresponding convex aggregation of the observed subset-specific structures. The last term captures the residual mismatch between the unseen test labeling structure and a convex combination of the observed training structures. These residual terms are generally not directly observable during data selection. In contrast, $\overline{\mathcal{D}}_g(\mathbf{x})$ involves only training-subset labeling functions and can be approximated using subset-specific classifiers. Thus, pairwise consistency characterizes a training-side structural component of the bound, while the residual terms account for the remaining mismatch.

This decomposition motivates cross-subset pairwise consistency as a training-side proxy for classification-structure compatibility. Specifically, small pairwise disagreement keeps convex combinations of the observed subset-specific structures close despite unknown mixture weights, providing a test-agnostic characterization of structural stability across training subsets.

C.3 From Classification-Structure Consistency to a Sample-Level Score

Motivated by the preceding analysis, we measure the classification-structure consistency of a sample through pairwise agreement among subset-specific labeling functions. For a sample $\mathbf{x}_k$, we define:

$$S_{\mathrm{CSC}}(\mathbf{x}_k) = 1 - \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\delta\left(g_{tr_i}(\mathbf{x}_k), g_{tr_j}(\mathbf{x}_k)\right), \tag{25}$$

where $\delta(\cdot,\cdot)$ denotes a discrepancy measure between two predictive distributions. The equal weighting over distinct subset pairs should not be interpreted as assuming uniform training or test mixture weights. Since the test-dependent weights are unavailable during data selection, equal pairwise ag-

gregation instead provides a permutation-invariant and test-agnostic summary of the classification consistency across all observed subsets.

The theoretical analysis above uses the norm-induced metric $d_g(\cdot,\cdot)$ to establish the structural relations in equation 18–equation 23. The practical score retains the resulting pairwise-consistency structure while allowing $\delta(\cdot,\cdot)$ to be instantiated by a suitable discrepancy measure between predictive distributions. Therefore, the preceding inequalities provide theoretical motivation for the pairwise aggregation itself, rather than a direct bound for an arbitrary practical choice of δ.

In practice, the underlying labeling functions $\{g_{tr_i}\}_{i=1}^{M}$ are inaccessible. We therefore approximate them using subset-specific classifiers $\{h_{\mathcal{D}_i}\}_{i=1}^{M}$ trained on the corresponding subsets and instantiate δ using the Jensen–Shannon (JS) divergence (Lin, 1991). Operationally, the score is evaluated as:

$$S_{\mathrm{CSC}}(\mathbf{x}_k) = 1 - \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathrm{JS}\left(h_{\mathcal{D}_i}(\mathbf{x}_k), h_{\mathcal{D}_j}(\mathbf{x}_k)\right). \tag{26}$$

A higher S_{CSC} indicates stronger agreement among the subset-specific classifiers on the corresponding sample, suggesting that its predictive structure is more stable across the observed training subsets. Accordingly, S_{CSC} provides a tractable training-side characterization of the classification-structure compatibility motivated by $\lambda_{\mathcal{H},(P_\pi,Q)}$.

D DATASETS

The detailed statistics for standard image classification and multi-domain benchmarks are summarized in Table 4 and Table 5, respectively.

Table 4: Detailed statistics of standard image classification benchmarks.

Dataset	Classes	Train	Test	Resolution
CIFAR-100	100	50,000	10,000	32×32
Tiny-ImageNet	200	100,000	10,000	64×64
ImageNet-1K	1,000	1,281,167	50,000	$224 \times 224^*$

Table 5: Dataset statistics and domain composition for multi-domain benchmarks.

Dataset	Domain	Images	Classes	Total
VLCS	Caltech101 (C)	1,415	5	10,729
	LabelMe (L)	2,656		
	VOC2007 (V)	3,376		
	SUN09 (S)	3,282		
Office-Home	Art (A)	2,427	65	15,588
	Clipart (C)	4,365		
	Product (P)	4,439		
	Real-World (R)	4,357		
miniDomainNet†	Clipart (C)	18,703	126	140,006
	Painting (P)	31,202		
	Real-World (R)	65,609		
	Sketch (S)	24,492		

Note: (1)* ImageNet-1K images are of various sizes but are typically preprocessed to 224×224 inputs for training and evaluation. For Tiny-ImageNet and ImageNet-1K, we use the official validation set as the test set for evaluation. (2)† miniDomainNet is a subset of DomainNet (Peng et al., 2019), constructed by sampling images from the original dataset according to the official split files (Zhou et al., 2021), while preserving its core complexity.

E Implementation Details

Sample-level Multi-dimensional Scoring We employ CLIP (Radford et al., 2021) as the backbone for multi-dimensional scoring, using a ViT-B/32 visual encoder and a Transformer text encoder. Both encoders are kept frozen, while lightweight adapters are optimized during training. We employ a shared image adapter, a general text adapter, and subset-specific text adapters, all implemented as lightweight two-layer MLPs with a hidden dimension of 256 and an output dimension of 512. Residual adaptation is applied to both modalities, with both residual weights set to 0.7. We use the prompt template "A photo of a [class]" to construct textual class prototypes. For standard image classification benchmarks, we set the latent clustering number to $K = 2$, yielding $M = K = 2$ training subsets. The adapters are trained for 30 epochs with Adam (Kingma & Ba, 2015) using a base learning rate of 1×10^{-4}. The subset discriminator G_{d_1} uses a learning rate of 0.5×10^{-4}, and the adversarial loss weight is set to 1.0. Following the DANN-style gradient reversal strategy (Ganin et al., 2016), we apply a three-epoch warm-up and schedule the gradient reversal coefficient as $\eta(p) = \frac{2}{1+\exp(-6p)} - 1$, where p denotes the normalized training progress. The independent probe G_{d_2} is trained on the frozen adapted representations for one third of the adapter training epochs. We use a linear classifier for multi-domain benchmarks and a one-hidden-layer MLP (512–256–M) for standard benchmarks. For classification-structure consistency scoring, each subset-specific classifier $h_{\mathcal{D}_i}$ uses the shared adapted visual representation together with textual prototypes generated by its dedicated subset-specific text adapter. The general and subset-specific text adapters are jointly optimized with the shared image adapter. Predictions are computed from cosine-similarity logits with a softmax temperature of $\tau' = 0.07$.

Coverage-aware Coreset Selection Considering the varying levels of redundancy across datasets, we adopt dataset-specific similarity thresholds τ for redundancy reduction, set to 0.8 for miniDomainNet, Office-Home, VLCS, CIFAR-100 and ImageNet-1K, and 0.9 for Tiny-ImageNet. Across all datasets, we independently apply rank normalization to each of the three sample-level scores over all samples for consistent score scaling, mapping them into $[0, 1]$.

Training on the Selected Coresets Following previous work (Sorscher et al., 2022; Xia et al., 2023; Yang et al., 2025), we adopt conventional training protocols for standard image classification benchmarks. For CIFAR-100, we use a batch size of 128 and train for 200 epochs with SGD (momentum 0.9, weight decay 5×10^{-4}, initial learning rate 0.1). The learning rate is decayed by a factor of 5 at epochs 60, 120, and 160. For Tiny-ImageNet, we use a batch size of 256 and train for 90 epochs with SGD (momentum 0.9, weight decay 1×10^{-4}, initial learning rate 0.1). The learning rate is decayed by a factor of 10 at epochs 30 and 60. For ImageNet-1K, following prior work and the VISSL library (Goyal et al., 2021; Sorscher et al., 2022; Xia et al., 2023; Yang et al., 2025), we use a batch size of 256 for 105 epochs with SGD (momentum 0.9, weight decay 1×10^{-3}, base learning rate 0.01). For multi-domain benchmarks (VLCS, Office-Home, and miniDomainNet), we perform training for 40 epochs using SGD with a momentum of 0.9 and a weight decay of 5×10^{-4}. Specifically, we set the batch size to 128 and the initial learning rate to 0.001 for both VLCS and Office-Home, while for miniDomainNet, we employ a batch size of 512 with an initial learning rate of 0.01. All experiments are repeated three times with different random seeds, except for ImageNet-1K, which is evaluated once due to computational constraints, following prior work (Xia et al., 2023; Yang et al., 2025). Unless otherwise specified, ResNet-50 (He et al., 2016) is adopted as the default backbone. Specifically, for standard image classification benchmarks, the backbone is randomly initialized and trained from scratch, whereas for multi-domain benchmarks, we instead initialize it with pre-trained weights and perform fine-tuning.

F Detailed Algorithmic Framework of GenDS

We summarize the main notation used in the theoretical analysis and the GenDS framework, and provide pseudocode for the overall procedure. Table 6 lists the key symbols used throughout the paper. As summarized in Algorithm 1, following subset partitioning, GenDS consists of Sample-level Multi-dimensional Scoring and Coverage-aware Coreset Selection. The scoring stage evaluates predictive utility, cross-subset invariance, and semantic consistency using S_{ITA}, S_{SI}, and S_{CSC}, respectively, while the selection stage integrates these scores and promotes distributional coverage through balanced budget allocation across classes and subsets, together with redundancy reduction.

Algorithm 1: GenDS: Generalizable Data Selection

Input : Training set $\mathcal{D}_{tr}$, selection ratio r, similarity threshold τ, pre-trained CLIP encoders, subset annotations if available, and latent cluster count K otherwise.
Output: Selected coreset $\mathcal{D}'_{tr}$.

```
// Subset Partition
```
Freeze the pre-trained CLIP encoders;
if *subset annotations are unavailable* **then**
  Extract style representations via Eq. 8;
  Infer latent subset assignments by Variational Fair Clustering with K clusters;
end
```
// Phase 1: Sample-level Multi-dimensional Scoring
```
Learn adapted representations with adversarial subset discriminator G_{d_1} via Eq. 2;
Compute $S_{\mathrm{ITA}}(\mathbf{x}_i)$ via Eq. 4;
Train independent subset probe G_{d_2} and compute $S_{\mathrm{SI}}(\mathbf{x}_i)$ via Eq. 3;
Train subset-specific classifiers $\{h_{\mathcal{D}_d}\}_{d=1}^{M}$ and compute $S_{\mathrm{CSC}}(\mathbf{x}_i)$ via Eq. 5;
```
// Phase 2: Coverage-aware Coreset Selection
```
Compute the composite utility $U(\mathbf{x}_i)$ via Eq. 6;
for *each class* c **do**
  Compute the class budget $k_c = \lfloor rN_c \rfloor$;
  Allocate the balanced base quota $k_{c,d} = \lfloor k_c/M \rfloor$ to each subset d;
  for *each subset* d **do**
    Rank samples by $U(\mathbf{x}_i)$;
    Select up to $k_{c,d}$ samples subject to the redundancy criterion in Eq. 7;
  end
  Fill the remaining class budget using redundancy-aware selection;
end
Construct $\mathcal{D}'_{tr}$ from the selected samples;
return $\mathcal{D}'_{tr}$;

Table 6: Summary of the main notation used in the theoretical analysis and the GenDS framework.

Notation	Description
$\mathcal{D}_{tr}, \mathcal{D}'_{tr}$	Full training set and selected coreset.
$\mathcal{D}_i, P_i$	The i-th training subset and its corresponding distribution.
P, Q	Training distribution and test distribution.
M, K	Number of training subsets and number of clusters for latent subset discovery.
π, P_π	Mixture weights and corresponding training mixture.
π^*, P^*	Mixture weights and training mixture closest to Q.
γ, ρ	Discrepancy between P^* and Q, and maximum pairwise discrepancy among training subsets.
$\lambda_{\mathcal{H},(P_\pi,Q)}$	Ideal joint error between P_π and Q.
G_{d_1}, G_{d_2}	Adversarial subset discriminator and independent subset probe.
$h_{\mathcal{D}_i}$	Classifier trained on subset $\mathcal{D}_i$.
$S_{\mathrm{ITA}}, S_{\mathrm{SI}}, S_{\mathrm{CSC}}$	Image-Text Alignment Score, Subset-Invariance Score, and Classification-Structure Consistency Score.
$U(\mathbf{x}_i)$	Composite sample utility.
$k_c, k_{c,d}$	Class-level selection budget and class-subset base quota.
τ	Similarity threshold for redundancy reduction.

G Cross-architecture Adaptability Evaluation

We evaluate the cross-architecture adaptability of the selected coresets by training different downstream architectures on them. As shown in Table 7, our method surpasses all baselines across all evaluated selection ratios on both VGG-16 (Simonyan & Zisserman, 2015) and DenseNet-121 (Huang et al., 2017). To further assess this adaptability across different architecture families,

we also evaluate a Vision Transformer (ViT) (Dosovitskiy et al., 2021). Specifically, we train a ViT-Small model on the selected CIFAR-10 (Krizhevsky et al., 2009) coreset. The experimental results, presented in Table 8, show that our method achieves the best performance among all baselines across the evaluated selection ratios from 60% to 90%. Combined with the results on VGG-16 and DenseNet-121, these findings demonstrate that the selected coresets transfer well across different downstream architectures, including both CNNs and Vision Transformers.

Table 7: Test accuracy (%) on Tiny-ImageNet using VGG-16 and DenseNet-121.

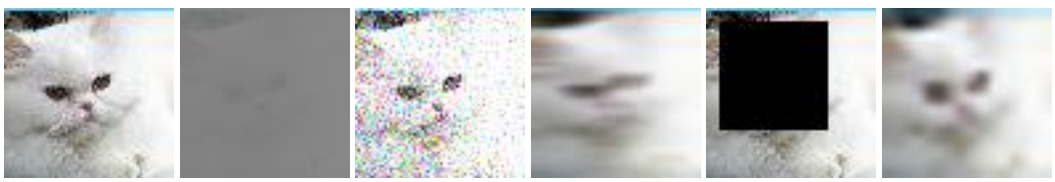
Method / Sel. Ratio	VGG-16			DenseNet-121		
	70%	80%	90%	70%	80%	90%
Random	47.39 ± 2.72	49.38 ± 0.23	51.15 ± 0.64	59.55 ± 0.20	60.78 ± 0.18	61.03 ± 0.22
Herding	48.59 ± 0.07	45.77 ± 0.12	50.77 ± 1.24	59.00 ± 0.28	60.03 ± 0.35	61.15 ± 0.12
Forgetting	47.50 ± 2.43	48.59 ± 1.77	49.82 ± 0.62	58.54 ± 0.15	60.39 ± 0.46	61.12 ± 0.10
GraNd	50.79 ± 1.26	46.84 ± 1.38	54.73 ± 0.49	59.62 ± 0.02	60.84 ± 0.09	61.10 ± 0.05
EL2N	48.30 ± 2.95	48.75 ± 1.65	49.01 ± 1.31	59.61 ± 0.00	60.38 ± 0.04	61.16 ± 0.47
Glister	48.74 ± 2.29	50.05 ± 0.02	49.42 ± 1.81	59.98 ± 0.01	60.62 ± 0.34	61.28 ± 0.18
SSP	48.38 ± 1.38	49.98 ± 1.49	54.71 ± 0.84	59.56 ± 0.03	60.22 ± 0.12	60.91 ± 0.29
Moderate-DS	50.78 ± 0.93	49.31 ± 0.41	49.25 ± 0.77	59.41 ± 0.18	60.42 ± 0.14	61.44 ± 0.11
CG-Score	48.73 ± 2.70	48.49 ± 1.88	49.62 ± 1.08	59.74 ± 0.15	60.55 ± 0.20	61.14 ± 0.11
MoSo	50.47 ± 1.01	50.12 ± 0.83	50.07 ± 0.43	59.27 ± 0.33	59.86 ± 0.07	60.00 ± 0.37
CPDS	53.40 ± 3.20	52.25 ± 0.58	56.34 ± 2.93	60.12 ± 0.06	60.93 ± 0.03	61.59 ± 0.03
Ours	**55.84** ± 3.33	**54.03** ± 2.43	**57.35** ± 3.33	**61.05** ± 0.56	**61.69** ± 0.15	**61.72** ± 0.08

Table 8: Test accuracy (%) on CIFAR-10 using ViT-Small.

Method / Sel. Ratio	60%	70%	80%	90%	100%
Random	78.98 ± 0.28	80.30 ± 0.36	81.33 ± 0.10	82.63 ± 0.18	84.00 ± 0.32
Herding	76.08 ± 0.19	78.53 ± 0.45	80.31 ± 0.01	82.08 ± 0.02	84.00 ± 0.32
EL2N	79.35 ± 0.09	80.73 ± 0.08	81.62 ± 0.08	82.90 ± 0.09	84.00 ± 0.32
GraNd	79.22 ± 0.06	80.59 ± 0.19	81.53 ± 0.18	82.72 ± 0.08	84.00 ± 0.32
Glister	78.33 ± 0.01	79.84 ± 0.27	81.33 ± 0.05	82.65 ± 0.09	84.00 ± 0.32
CG-Score	79.09 ± 0.29	80.96 ± 0.05	82.02 ± 0.23	82.91 ± 0.05	84.00 ± 0.32
Forgetting	77.87 ± 0.32	80.86 ± 0.08	81.90 ± 0.31	82.69 ± 0.20	84.00 ± 0.32
SSP	79.24 ± 0.16	80.34 ± 0.21	81.66 ± 0.25	82.86 ± 0.19	84.00 ± 0.32
Moderate-DS	79.54 ± 0.19	81.28 ± 0.13	81.98 ± 0.16	82.61 ± 0.27	84.00 ± 0.32
MoSo	79.45 ± 0.11	80.27 ± 0.23	81.82 ± 0.15	82.92 ± 0.34	84.00 ± 0.32
CPDS	80.21 ± 0.10	82.13 ± 0.23	82.32 ± 0.29	84.26 ± 0.05	84.00 ± 0.32
Ours	**81.20** ± 0.10	**82.48** ± 0.20	**84.02** ± 0.05	**84.53** ± 0.06	84.00 ± 0.32

H Examples of Image Corruption Types

Figure 6 illustrates representative examples of five image corruption types considered in our robustness evaluation: fog, Gaussian noise, motion blur, random occlusion, and resolution degradation.



(a) Original (b) Fog (c) Gaussian (d) Motion blur (e) Occlusion (f) Resolution

Figure 6: Illustration of the image corruption types, including fog, Gaussian noise, motion blur, random occlusion, and resolution degradation.

Table 9: Ablation on text modality across selection budgets on Office-Home (R) and Tiny-ImageNet.

Method	Office-Home (R)				Tiny-ImageNet			
	20%	30%	40%	60%	20%	30%	40%	60%
Moderate-DS (Visual Prototype, ResNet)	73.8	76.2	77.5	79.6	23.5	29.5	34.8	41.5
CPDS (Image–Text, CLIP)	65.5	72.0	75.7	78.4	25.6	31.7	36.3	44.2
GenDS (Visual Prototype, ResNet)	75.7	77.6	78.3	79.4	26.5	32.5	36.5	44.8
GenDS (Visual Prototype, ViT)	74.7	76.9	78.4	79.4	26.9	32.5	37.1	44.8
GenDS (Image–Text, CLIP)	**78.5**	**79.1**	**80.2**	**80.6**	**27.6**	**34.1**	**38.1**	**46.3**

Table 10: Comparison of additive and multiplicative fusion on Office-Home (R) and Tiny-ImageNet.

Fusion Strategy	Office-Home (R)			Tiny-ImageNet		
	20%	30%	40%	20%	30%	40%
Additive	77.3	78.6	79.7	26.4	32.2	36.6
Multiplicative	**78.5**	**79.1**	**80.2**	**27.6**	**34.1**	**38.1**

I Additional Ablation Studies

Ablation on Text Modality. Following prior work (Xia et al., 2023), we construct an image-only GenDS variant using visual prototypes to study the role of text. Given an image encoder $\boldsymbol{f}_{\mathrm{I}}$, we define the class-wise visual prototype as $\mu_c = \mathrm{Norm}\left(\frac{1}{|\mathcal{D}_c|}\sum_{\mathbf{x}_i \in \mathcal{D}_c} \boldsymbol{f}_{\mathrm{I}}(\mathbf{x}_i)\right)$. We similarly define $\mu_{c,d}$ over samples of class c in subset d. These prototypes replace image–text alignment and prompt-based classifiers, with the remaining pipeline unchanged. As shown in Table 9, GenDS remains competitive even without CLIP, demonstrating its applicability beyond CLIP-based representations. Incorporating CLIP-based image–text representations further improves performance.

Ablation on Score Fusion Strategy. We further compare additive and multiplicative fusion for integrating the three scoring criteria. For additive fusion, the normalized scores are summed with equal weights, while multiplicative fusion follows Eq. 6. As shown in Table 10, multiplicative fusion consistently achieves better performance across different selection budgets on both Office-Home (R) and Tiny-ImageNet, supporting our choice of multiplicative fusion.

J Limitations

Our current framework primarily targets offline data selection with a fixed candidate pool. In streaming or continuously evolving settings, the subset structure and sample utility may change over time, requiring dynamic or incremental updates to both the scoring process and coreset construction. Extending our data selection framework and its generalization-oriented analysis to such online settings is an important direction for future work.