POST-DISTILLATION HYBRID DATASET CURATION VIA DIFFERENTIATED VALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation produces compact synthetic training sets, while coreset selection preserves real examples with natural instance-level variation. This complementarity motivates post-distillation Hybrid Dataset Curation (HDC), where a compact training subset is constructed from a fixed distilled set and real candidates. However, curating such hybrid subsets is challenging because both sources play different roles in training. Existing strategies either score both sources with the same agnostic criterion or determine the real-distilled composition before sample valuation, which limits adaptive hybrid subset construction. To address these limitations, we propose Differentiated Hybrid Curation (DHC), a framework for post-distillation HDC based on differentiated valuation. DHC first constructs a balanced and semantically aligned real-distilled candidate pool to make cross-source comparison more tractable. It then introduces the Differentiated Identity Score (DIS), which compares candidate responses under distilled and real experts to provide a sample-wise source-aware valuation signal. Finally, DHC combines DIS with gradient-based representativeness and diversity-aware updates to construct class-balanced hybrid subsets without prescribing a fixed real-distilled ratio. Experiments on ImageWoof, ImageNette, and ImageNet-1K show that DHC improves over strong coreset selection and dataset distillation baselines in most settings and remains competitive with hybrid curation baselines.

1 INTRODUCTION

The rapid progress of deep learning has been driven largely by increasingly large training datasets (Tala Talaei Khoei, 2025; Wan et al., 2024), yet such scale makes dataset storage, transfer, and repeated model training costly under resource constraints. Dataset reduction addresses these costs by constructing compact training sets that retain as much utility as possible under limited data budgets. Existing methods mainly follow two paradigms, Coreset Selection (CS) and Dataset Distillation (DD).

CS and DD are individually incomplete yet complementary. CS selects existing samples from a given real training set and can prioritize informative or difficult examples (Xia et al., 2024; Guo et al., 2022), although its representation remains confined to samples already present in the given dataset. DD instead optimizes synthetic samples to concentrate information relevant to training under compact data budgets (Lee et al., 2022; Zhang et al., 2025), but its tendency to capture class prototypes may underrepresent variation among individual instances (Guo et al., 2024). These complementary properties motivate **Hybrid Dataset Curation (HDC)**, which constructs a compact dataset comprising real and distilled samples.

Existing HDC methods follow two workflows distinguished by the order of selection and distillation. SelMatch (Lee & Chung, 2024) selects real samples for subsequent distillation, whereas CCFS (Chen et al., 2025b) selects real samples after obtaining a distilled set. The stronger performance reported by CCFS over SelMatch in their overlapping evaluations indicates that effective hybrid construction need not be coupled with distillation. This decoupling also enables one-time distillation with repeated downstream reuse. We therefore focus on **post-distillation HDC**, where a compact hybrid subset is selected from a fixed distilled set and accessible labeled real candidates without modifying or rerunning distillation.

In post-distillation HDC, a natural approach is to apply conventional coreset selection (Sener & Savarese, 2017; Toneva et al., 2018; Xia et al., 2023; Zhang et al., 2024) to rank real and distilled

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

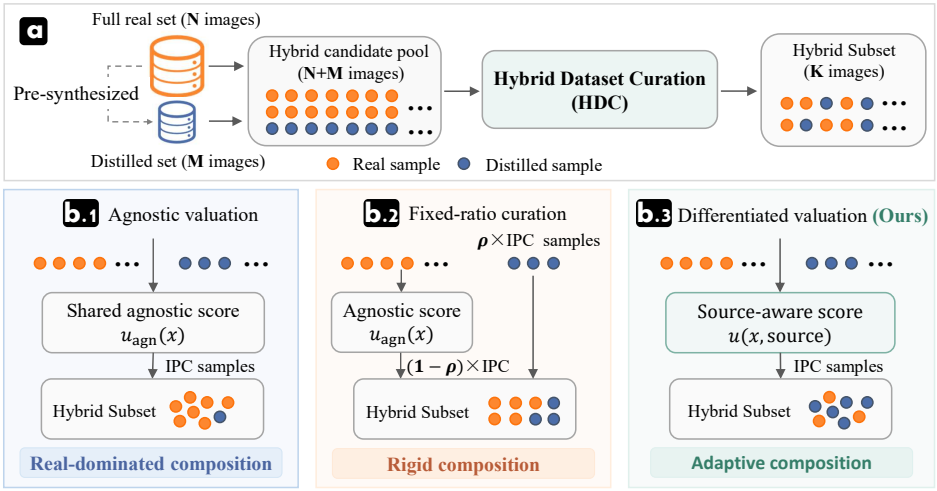


Figure 1: **Post-distillation HDC and its limitations**. **Top:** Overview of HDC, which constructs a compact hybrid subset from a full real set and a pre-synthesized distilled set. **Bottom:** Comparison of three representative strategies for post-distillation HDC. **(b.1)** Agnostic valuation ranks all candidates with a shared score, which can yield a real-dominated composition. **(b.2)** Fixed-ratio curation prescribes the real-distilled subset composition and selects the real samples with an agnostic score, yielding a rigid composition. **(b.3)** Differentiated valuation incorporates source information into sample-wise scoring, enabling an adaptively composed hybrid subset.

candidates using the same source-agnostic criterion. The resulting selection is sensitive to candidate-pool composition and can underuse distilled samples when real candidates dominate, causing *source-composition bias*, as illustrated in Figure 1(b.1). Even after source cardinality is controlled, a shared score may overlook the distinct roles of real and distilled samples, resulting in *valuation mismatch*. CCFS circumvents both limitations by allocating the budget between real and distilled sources before selecting the real component. However, this predefined allocation prevents candidates from competing across sources according to their individual utility, leading to a rigid composition as illustrated in Figure 1(b.2). These observations motivate sample-wise source-aware valuation that jointly evaluates real and distilled candidates and allows the source composition to emerge from selection.

To address these limitations, we propose **Differentiated Hybrid Curation (DHC)**, a framework for source-aware, sample-wise differentiated valuation in post-distillation HDC. DHC first builds a structured hybrid candidate pool by pairing a pre-synthesized distilled set with its real nearest neighbors, making cross-source comparison more tractable. It then introduces the Differentiated Identity Score (DIS), a sample-wise source-aware signal that estimates each candidate's position on a prototype-to-instance spectrum by contrasting gradient responses from a prototype expert trained on the distilled set and an instance expert trained on the corresponding real set. DHC combines DIS with gradient-based representativeness and diversity-aware updates to jointly select real and distilled candidates. The final source composition thereby emerges from selection rather than a predefined allocation, as illustrated in Figure 1(b.3).

In summary, our main contributions are threefold:

- We formalize post-distillation Hybrid Dataset Curation (HDC) as a selection problem over a fixed heterogeneous pool of real and distilled candidates. We identify source-composition bias and valuation mismatch under source-agnostic selection, together with the rigidity of predefined source allocation.
- We propose DHC, a post-distillation curation framework that constructs a balanced and semantically aligned candidate pool and introduces DIS for sample-wise source-aware valuation. Without prescribing a source ratio, DHC jointly selects real and distilled candidates so that their final composition emerges from selection.
- Extensive experiments on ImageWoof, ImageNette, and ImageNet-1K show that DHC improves over strong coreset selection and dataset distillation baselines in most settings and remains competitive with CCFS.

2 PROBLEM FORMULATION

2.1 POST-DISTILLATION HYBRID DATASET CURATION

Let $\mathcal{D}_{\text{full}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a labeled real training set with C classes, and let $\mathcal{D}_{\text{distill}}$ denote a pre-synthesized distilled set derived from $\mathcal{D}_{\text{full}}$. We consider a post-distillation setting in which synthesis is completed before curation and the distilled samples are neither modified nor regenerated. The raw hybrid candidate pool is

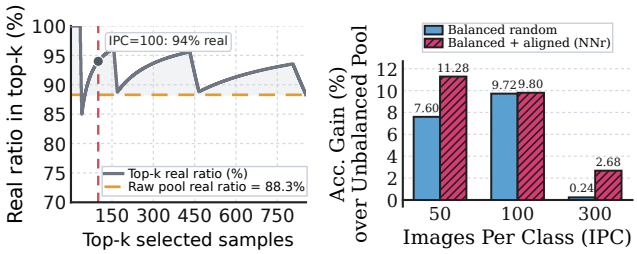
$$\mathcal{C}_{\text{raw}} = \mathcal{D}_{\text{full}} \cup \mathcal{D}_{\text{distill}}. \tag{1}$$

Post-distillation HDC aims to construct a compact training subset $\mathcal{S}_{\text{hybrid}} \subseteq \mathcal{C}_{\text{raw}}$ under a class-balanced budget, with $|\mathcal{S}_{\text{hybrid}}^{(c)}| = \text{IPC}$ for each class c, where IPC denotes images per class.

2.2 CHALLENGES IN POST-DISTILLATION HDC

Post-distillation HDC must jointly determine candidate utility and source composition. We consider the challenges arising from two direct strategies, namely applying a shared coreset criterion across sources and allocating the subset budget between sources before selection.

Source-composition Bias. A natural baseline is to score all candidates in $\mathcal{C}_{\text{raw}}$ using a conventional coreset criterion and select those with the highest scores. We term this strategy *agnostic valuation*, as the same scoring function $u_{\text{agn}}(x)$ is applied to real and distilled samples without modeling their distinct information roles. This formulation includes difficulty-based (Toneva et al., 2018), representativeness-based (Xia et al., 2023), and gradient-based (Killamsetty et al., 2021) criteria applied uniformly across the hybrid pool.



(a) Top-k source composition. (b) Gains from balanced pools.

Figure 2: **Agnostic valuation is sensitive to source composition.** (a) In the raw pool, top-k Forgetting selection tracks the real-dominated source composition for one ImageWoof class. (b) Source-balanced candidate pools yield accuracy gains over raw pool under the same agnostic valuation.

In our setting, $\mathcal{C}_{\text{raw}}$ contains substantially more real candidates than distilled candidates, which can bias a shared top-k ranking toward real samples. We examine this effect on ImageWoof using Forgetting (Toneva et al., 2018) as u_{agn} and selecting the top-k candidates within each class, where k equals the IPC budget. Keeping the score and selector fixed, we compare $\mathcal{C}_{\text{raw}}$ with two source-balanced pools. The balanced-random pool $\mathcal{C}_{\text{bal-rand}} = \mathcal{D}_{\text{rand}} \cup \mathcal{D}_{\text{distill}}$ uses randomly sampled real candidates, whereas the balanced-aligned pool $\mathcal{C}_{\text{bal-NNr}} = \mathcal{D}_{\text{NNr}} \cup \mathcal{D}_{\text{distill}}$ uses the real nearest neighbors of the distilled samples in embedding space. Both contain equal numbers of real and distilled candidates per class. Figure 2(a) illustrates that selection from the raw pool is strongly skewed toward real samples. Figure 2(b) further shows that source-balanced pools improve downstream accuracy over the raw unbalanced pool under the same scoring and selection procedure, particularly at low IPC. We term this undesirable dependence of selection on candidate-pool source proportions *source-composition bias*. Additional supporting results using Graph Cut (Iyer et al., 2021) as u_{agn} are reported in Appendix E.1.

Valuation Mismatch. Controlling source composition does not remove another assumption of agnostic valuation, namely that real and distilled candidates can be compared by a shared scoring criterion. However, because the two sources play different roles in training, scores assigned by this criterion may not be directly comparable across sources. We refer to this limited cross-source score comparability as *valuation mismatch*. To examine whether source identity affects valuation, we introduce a simple source-aware valuation:

$$u_\gamma(x) = z(u_{\text{agn}}(x)) + \gamma \cdot \mathbf{1}[x \in \mathcal{D}_{\text{distill}}],$$

where $z(\cdot)$ denotes within-class normalization and $\mathbf{1}[\cdot]$ is the indicator function. The parameter γ introduces a global preference for distilled samples when positive and real samples when negative, while $\gamma = 0$ recovers agnostic valuation. We evaluate this diagnostic on $\mathcal{C}_{\text{bal-NNr}}$ using Forgetting as u_{agn} and varying only γ. Figure 3 shows that an appropriate nonzero γ improves over agnostic valuation at the evaluated IPC budgets, while the preferred value of γ varies across budgets. The complete source-preference sweep across IPC budgets is provided in Appendix E.2. These results indicate that source identity remains relevant to candidate utility after controlling source composition. However, the source-aware component of u_γ is constant within each source and cannot capture sample-specific differences in how source identity affects utility.

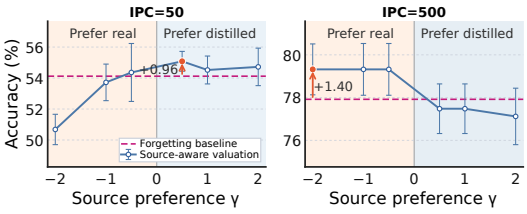


Figure 3: **Source-aware valuation outperforms agnostic valuation.** We vary the source-preference parameter γ in u_γ on the balanced-aligned pool with Forgetting as u_{agn}, and the dashed line denotes pure Forgetting. Source-aware valuation improves over the agnostic baseline, while the best value of γ differs between low and high IPC.

Rigidity of predefined source allocation. An alternative to agnostic valuation is to divide the subset budget between sources before selection. CCFS (Chen et al., 2025b) follows this strategy by prescribing an IPC-dependent real-distilled ratio and then selecting the real component. This allocation prevents the larger real pool from determining source composition through a shared ranking and avoids relying on a source-agnostic score for cross-source allocation. However, it fixes the source budgets before candidates are evaluated across sources. Consequently, real and distilled candidates cannot compete for the total budget according to their sample-wise utility, and the final composition remains constrained by the predefined allocation.

2.3 Principles for Hybrid Subset Construction

The preceding analysis yields three requirements for post-distillation HDC. First, candidate valuation should operate on a controlled hybrid pool so that source cardinality does not dominate selection. Second, valuation should account for source identity at the sample level rather than relying on either a shared source-agnostic score or a constant source-level preference. Third, real and distilled candidates should be evaluated jointly under the subset budget, allowing their final composition to emerge from sample-wise selection rather than a predefined source allocation.

These requirements do not replace established objectives in coreset selection. Representativeness remains necessary for retaining candidates that reflect the overall training objective, while diversity limits redundant selections (Killamsetty et al., 2021; Sener & Savarese, 2017). Neither objective, however, specifies how utility should be compared across real and distilled candidates. We therefore introduce sample-wise source-aware valuation to account for cross-source differences and combine it with established representativeness and diversity criteria during subset construction.

3 Differentiated Hybrid Curation

3.1 Overview

Building on the design principles in Section 2.3, we propose **Differentiated Hybrid Curation (DHC)** for post-distillation HDC. As shown in Figure 4, DHC constructs a balanced and semantically aligned candidate pool, computes sample-wise source-aware valuation with the Differentiated Identity Score (DIS), and combines it with representativeness and diversity-aware updates for class-balanced subset construction. The following subsections describe these steps in order.

3.2 Balanced and Semantically Aligned Candidate Pool

Section 2.2 shows that valuation in a raw hybrid pool can be affected by the real-distilled source composition. DHC therefore constructs a controlled candidate pool before sample valuation. Balancing

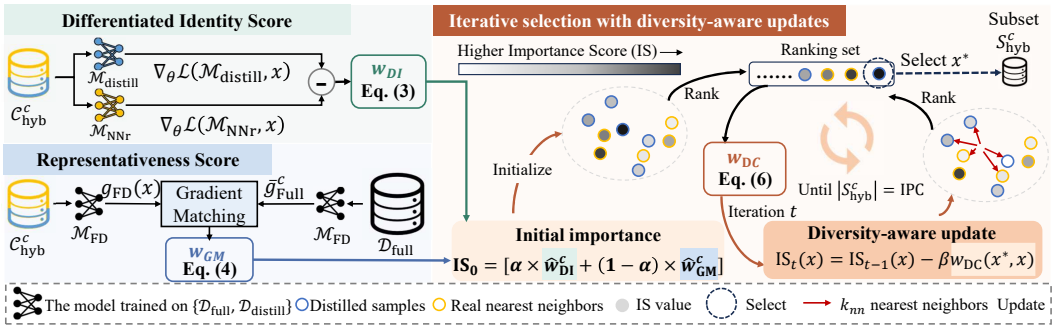


Figure 4: **Overview of Differentiated Hybrid Curation.** DHC operates on the balanced-aligned pool $\mathcal{C}_{\text{hyb}} = \mathcal{D}_{\text{distill}} \cup \mathcal{D}_{\text{NNr}}$, where $\mathcal{D}_{\text{NNr}}$ contains real nearest neighbors of the pre-synthesized distilled set. For each candidate, DHC computes the Differentiated Identity Score w_{DI} and gradient-based representativeness w_{GM}, then combines them to initialize the importance score IS_0. It iteratively selects the highest-score unselected sample within each class and updates the scores of its k_{nn} nearest neighbors using the diversity-aware penalty w_{DC} until the IPC budget is reached.

source cardinality removes the quantity confounder, but random real samples may still introduce large content variation relative to the distilled samples. To reduce this additional variation, DHC uses real nearest neighbors when forming the balanced pool.

Let $\mathcal{D}_{\text{distill}}$ denote the pre-synthesized distilled set with IPC samples per class. For each class c, we construct $\mathcal{D}^{(c)}_{\text{NNr}}$ by retrieving real samples from $\mathcal{D}^{(c)}_{\text{full}}$ that are nearest to the distilled samples in the CLIP feature space (Radford et al., 2021). Similarity is measured by cosine similarity between CLIP embeddings. Each distilled sample is matched to the nearest unassigned real sample, skipping real samples that have already been selected. The resulting real nearest-neighbor set $\mathcal{D}_{\text{NNr}}$ satisfies $|\mathcal{D}^{(c)}_{\text{NNr}}| = |\mathcal{D}^{(c)}_{\text{distill}}| = \text{IPC}$ for each class c. The DHC candidate pool is defined as

$$\mathcal{C}_{\text{hyb}} = \mathcal{D}_{\text{distill}} \cup \mathcal{D}_{\text{NNr}}. \tag{2}$$

This pool controls source cardinality within each class and reduces unrelated content differences between real and distilled candidates, providing the basis for the sample-wise source-aware valuation.

3.3 Differentiated Identity Score for Source-Aware Valuation

Within the controlled candidate pool, DHC must account for source-dependent differences at the sample level. We therefore introduce the Differentiated Identity Score (DIS), denoted by w_{DI}.

DIS estimates where each candidate lies on a prototype-to-instance spectrum through two source-specific experts. We train a prototype expert $\mathcal{M}_{\text{distill}}$ on $\mathcal{D}_{\text{distill}}$ and an instance expert $\mathcal{M}_{\text{NNr}}$ on $\mathcal{D}_{\text{NNr}}$. For each candidate $x_i \in \mathcal{C}_{\text{hyb}}$, DIS is defined as

$$w_{\text{DI}}(x_i) = \|\nabla_\theta \mathcal{L}(\mathcal{M}_{\text{NNr}}, x_i)\|_2 - \|\nabla_\theta \mathcal{L}(\mathcal{M}_{\text{distill}}, x_i)\|_2. \tag{3}$$

A larger gradient norm reflects a stronger update induced by a sample under the corresponding expert (Ash et al., 2019; Coleman et al., 2019). A positive $w_{\text{DI}}(x_i)$ indicates a stronger response from the instance expert and is interpreted as relatively more prototype-like, while a negative value suggests a more instance-like response pattern. As a sample-wise source-aware signal, DIS distinguishes candidates within the same source but does not independently measure subset quality. Additional DIS statistics, distribution plots, and a directionality ablation are provided in Appendix C.2.

3.4 Hybrid Subset Construction

DHC jointly selects from all candidates in $\mathcal{C}_{\text{hyb}}$ without source-specific quotas, allowing the final real-distilled composition to emerge from sample-wise selection. Following Section 2.3, DIS provides source-aware differentiation, representativeness measures alignment with the training objective, and diversity discourages redundant selections. DIS and representativeness are sample-wise quantities used to initialize candidate importance, whereas diversity depends on previously selected samples and is imposed through iterative score updates.

Gradient-based Representativeness. To complement DIS with an established subset-selection criterion, we use a gradient-matching score w_{GM} for representativeness, inspired by gradient matching principles from both dataset distillation and coreset selection Zhao et al. (2020); Killamsetty et al. (2021). We train a model $\mathcal{M}_{\mathrm{FD}}$ on $\mathcal{D}_{\mathrm{full}} \cup \mathcal{D}_{\mathrm{distill}}$ to obtain gradients under a shared model. For each candidate x_i of class c, we compute $g_{\mathrm{FD}}(x_i) = \nabla_\theta \mathcal{L}(\mathcal{M}_{\mathrm{FD}}, x_i)$. We also compute the class-mean gradient over the full real dataset $\bar{g}^{(c)}_{\mathrm{full}} = \frac{1}{|\mathcal{D}^{(c)}_{\mathrm{full}}|} \sum_{x_j \in \mathcal{D}^{(c)}_{\mathrm{full}}} g_{\mathrm{FD}}(x_j)$. The representativeness score is defined as

$$w_{\mathrm{GM}}(x_i) = \frac{g_{\mathrm{FD}}(x_i) \cdot \bar{g}^{(c)}_{\mathrm{full}}}{\|g_{\mathrm{FD}}(x_i)\|_2 \cdot \|\bar{g}^{(c)}_{\mathrm{full}}\|_2}. \tag{4}$$

A larger $w_{\mathrm{GM}}(x_i)$ indicates stronger alignment with the class-level update direction.

Initial Sample Importance. We combine DIS and representativeness to initialize sample importance. For each class c, we apply min-max normalization to obtain $\hat{w}^{(c)}_{\mathrm{DI}}$ and $\hat{w}^{(c)}_{\mathrm{GM}}$. The initial importance score is

$$\mathrm{IS}_0(x_i) = \alpha \hat{w}^{(c)}_{\mathrm{DI}}(x_i) + (1-\alpha)\hat{w}^{(c)}_{\mathrm{GM}}(x_i), \tag{5}$$

where $\alpha \in [0,1]$ controls the relative contribution of DIS and representativeness. DIS provides sample-wise source-aware differentiation, while w_{GM} accounts for alignment with the class-level training objective.

Iterative Selection with Diversity-aware Updates. Diversity is a subset-level factor because the redundancy of a candidate depends on what has already been selected Maharana et al. (2024). We therefore instantiate diversity through a pairwise similarity penalty rather than adding it to IS_0. For candidates x_i and x_j, we compute CLIP embeddings $\phi(x_i)$ and $\phi(x_j)$, and define

$$w_{\mathrm{DC}}(x_i, x_j) = \frac{\phi(x_i) \cdot \phi(x_j)}{\|\phi(x_i)\|_2 \cdot \|\phi(x_j)\|_2}. \tag{6}$$

A larger $w_{\mathrm{DC}}(x_i, x_j)$ indicates stronger semantic similarity and therefore higher redundancy if both samples are selected.

Within each class, DHC iteratively selects the candidate with the highest current importance score until IPC samples are selected. At iteration t, after selecting x^*, DHC removes it from the remaining candidate set and penalizes its unselected class-wise nearest neighbors according to their redundancy with x^*. For each unselected neighbor $x_j \in \mathcal{N}^{(c)}_{k_{\mathrm{nn}}}(x^*)$, the score update is

$$\mathrm{IS}_t(x_j) = \mathrm{IS}_{t-1}(x_j) - \beta w_{\mathrm{DC}}(x^*, x_j), \tag{7}$$

where $\mathcal{N}^{(c)}_{k_{\mathrm{nn}}}(x^*)$ denotes the k_{nn} nearest neighbors of x^* within class c, and β controls the diversity penalty strength. Scores of other unselected candidates remain unchanged. We set $k_{\mathrm{nn}} = 10$ in all experiments. The selected samples across all classes form the final hybrid subset $\mathcal{S}_{\mathrm{hybrid}}$, whose real-distilled composition is determined by the joint selection process.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct comprehensive experiments on the full ImageNet-1K dataset (224×224) (Russakovsky et al., 2015) and two of its subsets representing distinct difficulty levels, namely ImageNette and ImageWoof (Howard, 2019). ImageNette comprises 10 visually dissimilar classes that are relatively easy to distinguish. In contrast, ImageWoof consists of 10 visually similar dog breeds, presenting a more challenging classification task. We evaluate ImageWoof at $\mathrm{IPC} \in \{50, 100, 300, 500\}$, ImageNette at $\mathrm{IPC} \in \{100, 300\}$, and ImageNet-1K at $\mathrm{IPC} = 10$.

Baselines. We compare DHC with three groups of dataset reduction methods. Results without explicit citations are obtained from our implementation under the same experimental protocol.

- **Coreset Selection (CS):** We re-implement representative methods covering difficulty-based selection (Forgetting (Toneva et al., 2018)), representativeness-based selection (Moderate_DS (MDS) (Xia et al., 2023)), diversity-based selection (K-center Greedy (KCG) (Sener & Savarese, 2017)), and submodular selection balancing representativeness and diversity (Graph Cut (Iyer et al., 2021)). We also include Random selection and the Nearest-Neighbor Set ($\mathcal{D}_{\mathrm{NNr}}$) alone as baselines.
- **Dataset Distillation (DD):** We include DM (Zhao & Bilen, 2023), IDC-1 (Kim et al., 2022), Minimax (Gu et al., 2024), and DiT+MGD3 (Chan-Santiago et al., 2025) from their original publications. We also re-implement DiT-IGD (Chen et al., 2025a) as a strong generative distillation baseline.
- **Hybrid Dataset Curation (HDC):** We re-implement CCFS (Chen et al., 2025b), the state-of-the-art post-distillation HDC method, while SelMatch (Lee & Chung, 2024) is not directly comparable under this protocol because it couples selection with distillation.

Distilled Set Generation. We use DiT-IGD (Chen et al., 2025a) to generate the pre-synthesized distilled sets for all post-distillation hybrid curation methods. For ImageNet-1K, we follow the original DiT-IGD protocol and use a ConvNet-6 surrogate for ResNet-18 evaluation. For ImageWoof and ImageNette, we use the same architecture for the surrogate and final evaluation model, such as using a ResNet-18 surrogate for ResNet-18 evaluation.

DHC Configuration. The balancing hyperparameter α (Eq. 5) and diversity penalty strength β (Eq. 7) are tuned for each IPC setting. Detailed settings, sensitivity analysis, and evaluation using a common hyperparameter rule across architectures are provided in Appendices B.3 and D.1.

Evaluation Protocol. For all methods, we curate class-balanced subsets for the given images per class (IPC), train the specified model architecture from scratch on the curated subsets, and report top-1 test accuracy (%) on the test set. Test architectures include ConvNet-6 and ResNet-18 (He et al., 2016). We also report full-training-set performance as a reference. Unless otherwise specified, we report mean accuracy and standard deviation over five different seeds.

Implementation Details. We implement DHC in PyTorch and conduct all experiments on NVIDIA RTX 4090 GPUs. For final evaluation, training on all curated subsets, including those from our re-implemented baselines, follows the official hyperparameters and schedule from the DiT-IGD repository for ImageWoof and ImageNette. For ImageNet-1K, we follow CDA (Yin & Shen, 2024a) and use pre-generated soft labels for training on the final hybrid subsets. For re-implemented baselines, we strive to follow their original hyperparameter settings to the best of our ability.

4.2 Comparison with State-of-the-Art Methods

We present the main results in Table 1, Table 2, and Figure 5. CS and DD methods construct purely real and purely synthetic subsets using the corresponding full real dataset, respectively. CCFS and DHC are evaluated in the post-distillation setting, where hybrid subsets are constructed from a pre-synthesized distilled set and the corresponding full real dataset.

DHC achieves strong overall performance across datasets, architectures, and IPC budgets. On ImageWoof, DHC obtains the best performance in most settings, including all ResNet-18 settings. For example, at IPC 50 under ConvNet-6, DHC outperforms the strongest coreset selection and dataset distillation baselines by 5.44% and 1.82%, respectively. Under ResNet-18, DHC consistently outperforms CCFS across all IPC budgets. On ImageNet-1K with IPC 10, DHC also achieves the highest accuracy among all compared methods, reaching 46.71%.

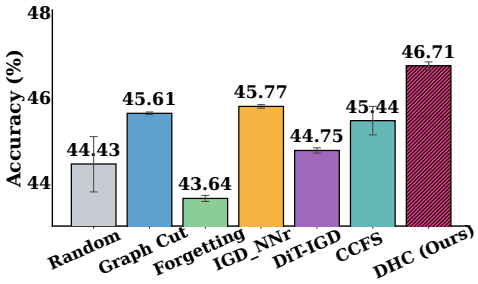


Figure 5: **Performance on ImageNet-1K.** Standard deviation is reported over three runs.

On ImageNette, DHC achieves the highest accuracy under ConvNet-6 at IPC 300 and remains competitive with CCFS across the other settings. Together, these results demonstrate strong performance across the evaluated datasets and settings.

Additional analyses of candidate real set and long-tailed settings are provided in the appendix.

Table 1: **Performance comparison with representative dataset reduction methods on ImageWoof.** To ensure fair comparisons, the distilled sets required by both our re-implementation of CCFS and our method are synthesized using the same distillation approach, DiT-IGD. A dash (-) indicates the result was not reported. The best results are marked in **bold**, and the second-best are underlined. "IGD_NNr" indicates the real nearest neighbors of the DiT-IGD (Chen et al., 2025a) distilled data, evaluated as a standalone coreset.

Model	ConvNet-6				ResNet-18			
IPC (images per class)	50	100	300	500	50	100	300	500
Full	85.72±0.78				89.52±0.41			
Random	42.60±0.96	53.92±1.06	70.00±0.36	78.13±1.60	53.00±2.22	66.40±1.34	79.60±0.96	83.96±0.32
Graph Cut (Iyer et al., 2021)	48.04±0.79	51.84±1.18	68.96±1.68	77.80±1.31	56.64±2.39	67.32±1.29	79.80±1.43	84.36±0.85
KCG (Sener & Savarese, 2017)	32.20±1.50	39.00±1.56	64.60±2.27	77.68±1.57	35.24±1.32	46.28±2.38	73.04±1.46	82.40±1.22
Forgetting (Toneva et al., 2018)	42.44±0.79	54.88±0.64	70.80±1.19	77.28±0.79	48.71±0.82	60.64±1.18	78.28±0.48	83.00±0.46
MDS (Xia et al., 2023)	40.20±1.07	51.24±1.71	70.04±0.61	77.88±0.83	59.00±1.93	68.20±1.26	80.32±0.99	84.20±0.57
IGD_NNr	50.88±1.21	58.76±0.56	71.20±0.99	78.00±0.80	62.20±0.44	70.20±0.88	79.40±0.75	83.60±0.40
DM (Zhao & Bilen, 2023)	43.80±1.10	50.10±0.90	-	-	53.90±0.70	64.90±0.70	-	-
IDC-1 (Kim et al., 2022)	42.60±0.90	51.00±1.10	-	-	54.50±1.00	57.70±0.80	-	-
Minimax (Gu et al., 2024)	50.70±1.80	57.10±1.90	-	-	60.50±0.50	67.40±0.70	-	-
DiT+MGD3 (Chan-Santiago et al., 2025)	54.50±1.60	60.10±1.20	-	-	58.30±1.40	68.80±0.70	-	-
DiT-IGD (Chen et al., 2025a)	54.48±1.30	61.24±0.74	70.04±0.86	73.80±0.68	66.04±0.89	71.32±1.56	78.60±0.63	80.76±1.33
CCFS (Chen et al., 2025b)	54.84±0.66	62.20±0.31	73.56±0.84	**79.48±0.53**	63.04±1.05	68.68±0.85	78.80±0.57	81.92±0.39
DHC (Ours)	**56.32±0.53**	**62.60±0.85**	**74.24±0.74**	78.52±0.96	**66.64±0.85**	**71.92±0.77**	**80.80±1.05**	**84.48±0.93**

Table 2: **Performance comparison with representative dataset reduction methods on ImageNette.** We follow the same settings and notations as stated in Table 1.

Model	ConvNet-6		ResNet-18	
IPC	100	300	100	300
Full	94.84±0.45		95.56±0.27	
Random	81.36±1.39	89.48±0.99	83.80±0.49	90.16±0.34
Graph Cut	84.36±1.34	89.52±0.61	83.84±0.73	89.96±0.51
KCG	67.80±1.77	83.88±1.01	70.84±1.25	87.48±0.81
Forgetting	76.96±1.20	88.36±0.59	78.76±1.34	89.56±0.97
MDS	80.68±2.05	88.72±0.55	84.40±0.97	90.32±0.39
IGD_NNr	85.36±0.15	89.88±0.32	85.60±0.49	89.76±0.65
DM	78.50±0.80	-	81.50±0.40	-
IDC-1	80.60±1.10	-	81.70±0.80	-
Minimax	81.10±0.30	-	81.30±0.70	-
DiT-IGD	84.72±0.53	88.80±0.40	84.84±0.51	88.36±0.64
CCFS	**88.72±0.97**	90.20±0.42	**87.24±0.81**	**91.20±0.57**
DHC (Ours)	86.28±0.92	**90.52±0.48**	85.40±0.33	90.68±0.57

4.3 Ablation Study and Analysis

Component Ablation. We evaluate the contribution of the DHC components on ImageWoof, including the Differentiated Identity Score (w_{DI}), gradient-based representativeness (w_{GM}), and diversity-aware update (w_{DC}).

As shown in Table 3, combining w_{DI} with w_{GM} outperforms either score alone, showing that source-aware valuation and representativeness provide complementary sample-wise information. The variant without DIS, $w_{\mathrm{GM}} + w_{\mathrm{DC}}$, is a strong baseline, especially at larger IPC budgets. The full model achieves the best performance across all IPC settings, indicating that DIS remains beneficial when combined with representativeness and diversity-aware selection. The gain from w_{DC} is larger at IPC 300 and 500, consistent with the need to reduce redundancy when more samples are selected.

Analysis on Selection within the Balanced Hybrid Pool. To isolate the effect of the subset construction algorithm from candidate pool construction, we compare DHC with representative coreset selection methods on the same balanced hybrid pool $\mathcal{D}_{\mathrm{distill}} \cup \mathcal{D}_{\mathrm{NNr}}$. All methods select class-balanced subsets from this fixed pool and are evaluated with ConvNet-6 on ImageWoof.

As shown in Figure 6, DHC outperforms all compared coreset selection methods across all IPC budgets. The advantage is most visible at low budgets, where DHC reaches 56.32% at IPC 50,

Table 3: **Ablation study of DHC components.** We evaluate the contribution of source-aware valuation (w_{DI}), gradient-based representativeness (w_{GM}), and diversity-aware updates (w_{DC}) on ImageWoof using ConvNet-6.

w_{DI}	w_{GM}	w_{DC}	IPC		
			100	300	500
●	○	○	62.00±0.55	70.32±0.50	73.76±1.20
○	●	○	61.28±1.59	71.96±0.48	76.00±1.10
●	●	○	62.60±0.85	73.24±0.48	76.24±0.60
○	●	●	62.40±1.25	73.32±0.52	77.80±0.52
●	●	●	**62.60±0.85**	**74.24±0.74**	**78.52±0.96**

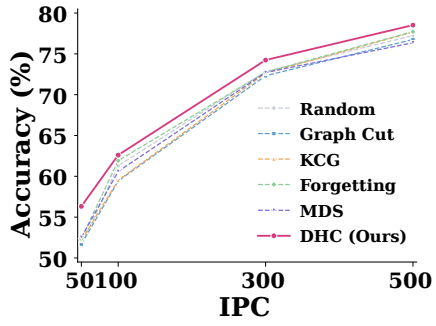


Figure 6: **Comparison with coreset selection on the balanced hybrid pool.** All methods select class-balanced subsets from $\mathcal{D}_{\mathrm{distill}} \cup \mathcal{D}_{\mathrm{NNr}}$, followed by ConvNet-6 training.

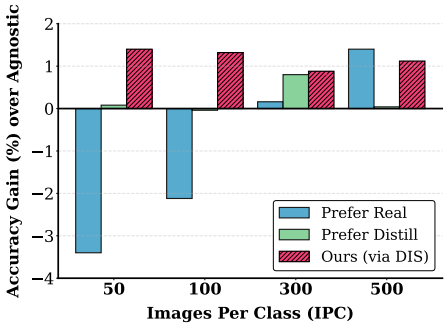


Figure 7: **Effect of sample-wise source-aware valuation**. We compare DIS with fixed source-preference heuristics and report accuracy gains over the agnostic baseline.

exceeding the second-best method by 3.80%. At larger budgets, all methods improve with increasing IPC, and DHC consistently achieves the highest accuracy. These results validate DHC's subset construction beyond the effect of the balanced hybrid pool.

Analysis on Differentiated Identity Score To evaluate the role of DIS in DHC, we replace DIS with three fixed source-preference strategies while keeping the remaining subset construction procedure unchanged. Let R and D denote the source weights for real and distilled candidates. Agnostic sets $(R, D) = (1, 1)$, Prefer Real sets $(R, D) = (1, 0)$, and Prefer Distill sets $(R, D) = (0, 1)$.

Figure 7 reports accuracy gains relative to the agnostic baseline. DIS yields positive gains across all IPC budgets and performs better than fixed heuristics at low and medium budgets. The fixed heuristics show budget-dependent behavior, where preferring distilled candidates is more effective at smaller budgets, while preferring real candidates becomes stronger at IPC 500. These results support the effectiveness of DIS as the source-aware valuation component in DHC.

5 Conclusion

In this work, we studied post-distillation Hybrid Dataset Curation, where compact training subsets are constructed from real and distilled candidates. We identified source-composition bias and valuation mismatch under agnostic valuation, together with the rigidity of predefined source allocation. To address these limitations, we proposed Differentiated Hybrid Curation, which first constructs a balanced and semantically aligned candidate pool and then instantiates source-aware valuation with the Differentiated Identity Score. Together with gradient-based representativeness and diversity-aware updates, DHC constructs class-balanced hybrid subsets without prescribing a fixed real-distilled subset ratio. Experiments on ImageWoof, ImageNette, and ImageNet-1K show that DHC performs strongly against coreset selection, dataset distillation, and hybrid curation baselines. Future work may extend DHC to broader data modalities and develop more efficient source-aware valuation for large-scale curation.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing and proofreading after the manuscript had been completed. The tools were used to suggest improvements in grammar, clarity, concision, and readability. All AI-assisted suggestions were reviewed and revised by the authors. The authors verified that the final text accurately reflects the research, results, and claims. We take full responsibility for the final content of this work, including all text and claims.

We have not used generative AI tools for research ideation, problem formulation, method design, experimental design, data collection, data analysis, code development, figure generation, or citation selection. These tasks were performed by the authors.

ETHICS STATEMENT

This work aims to improve data-efficient training by constructing compact hybrid subsets from real and distilled candidates. Reducing the amount of data used for training can lower storage and computation costs, which may benefit research and deployment under limited resources. The potential risks mainly come from the underlying datasets. If the real or distilled candidates contain biases, spurious correlations, or harmful content, the curated subset may preserve these issues. Therefore, applications in sensitive domains should include dataset auditing, fairness evaluation, and task-specific safety checks.

REPRODUCIBILITY STATEMENT

We include implementation details in Section 4.1 in the main body and Section B in the Appendix. The code of our method will be released.

REFERENCES

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep Learning Through the Lens of Example Difficulty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889. Curran Associates, Inc., 2021.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4750–4759, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3739–3748, 2023.

Jeffrey A Chan-Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. Mgd3: Mode-guided dataset distillation using diffusion models. *International Conference on Machine Learning*, 2025.

Mingyang Chen, Jiawei Du, Bo Huang, Yi Wang, Xiaobo Zhang, and Wei Wang. Influence-guided diffusion for dataset distillation. In *International Conference on Learning Representation*, 2025a.

Yanda Chen, Gongwei Chen, Miao Zhang, Weili Guan, and Liqiang Nie. Curriculum coarse-to-fine selection for high-ipc dataset distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20437–20446, 2025b.

Yilan Chen, Wei Huang, and Lily Weng. Provable and efficient dataset distillation for kernel ridge regression. *Advances in Neural Information Processing Systems*, 37:88739–88771, 2024.

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. *arXiv preprint arXiv:1906.11829*, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15793–15803, 2024.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications*, pp. 181–195, Cham, 2022. Springer International Publishing.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *International Conference on Learning Representation*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7713–7722, 2024.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. `https://github.com/fastai/imagenette`, March 2019. URL `https://github.com/fastai/imagenette`.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pp. 722–754. PMLR, 2021.

Krishnateja Killamsetty, S Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pp. 11102–11118. PMLR, 2022.

Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, pp. 12352–12364. PMLR, 2022.

Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *International Conference on Machine Learning*, 2024.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. In *International Conference on Learning Representation*, 2024.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *International Conference on Learning Representation*, 2021.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

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 *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16709–16718, 2024.

Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4797–4806, 2025.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5809–5818, 2024.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9390–9399, 2024.

Aditi Singh Tala Talaei Khoei. Data reduction in big data: a survey of methods, challenges and future directions. *International Journal of Data Science and Analytics*, 20(3):1643–1682, 2025.

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 *International Conference on Learning Representation*, 2018.

Zhijing Wan, Zhixiang Wang, Cheukting Chung, and Zheng Wang. A survey of dataset refinement for problems in computer vision datasets. *ACM computing surveys*, 56(7):1–34, 2024.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12196–12205, 2022.

Max Welling. Herding dynamical weights to learn. In *International Conference on Machine Learning*, pp. 1121–1128, 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 *The Eleventh International Conference on Learning Representations*, pp. 1–20, 2023.

Xiaobo Xia, Jiale Liu, Shaokun Zhang, Qingyun Wu, Hongxin Wei, and Tongliang Liu. Refined coreset selection: Towards minimal coreset size under model performance constraints. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=yb5xV8LFDq.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024a.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=PlaZD2nGCl.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36: 73582–73603, 2023.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, and Joey Tianyi Zhou. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26223–26232, June 2024.

Xin Zhang, Jiawei Du, Ping Liu, and Joey Tianyi Zhou. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. *International Conference on Learning Representation*, 2025.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representation*, 2023.

Yawen Zou, Guang Li, Duo Su, Zi Wang, Jun Yu, and Chao Zhang. Dataset distillation via vision-language category prototype. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

A NOTATION SUMMARY

For clarity, Table 4 summarizes the main notation used in the paper. We use superscript (c) to denote the subset or quantity associated with class c.

Table 4: Summary of main notation.

Notation	Description
Datasets, candidate pools, and selected subsets	
$\mathcal{D}_{\text{full}}$	Full real dataset with C classes
$\mathcal{D}_{\text{distill}}$	Pre-synthesized distilled set derived from $\mathcal{D}_{\text{full}}$
$\mathcal{D}_{\text{NNr}}$	Real nearest-neighbor set paired with $\mathcal{D}_{\text{distill}}$
$\mathcal{D}_{\text{rand}}$	Class-balanced random real subset sampled from $\mathcal{D}_{\text{full}}$
$\mathcal{C}_{\text{raw}}$	Raw post-distillation pool $\mathcal{D}_{\text{full}} \cup \mathcal{D}_{\text{distill}}$
$\mathcal{C}_{\text{bal-rand}}$	Balanced-random pool $\mathcal{D}_{\text{rand}} \cup \mathcal{D}_{\text{distill}}$
$\mathcal{C}_{\text{bal-NNr}}$	Balanced-aligned diagnostic pool $\mathcal{D}_{\text{NNr}} \cup \mathcal{D}_{\text{distill}}$
$\mathcal{C}_{\text{hyb}}$	DHC candidate pool $\mathcal{D}_{\text{distill}} \cup \mathcal{D}_{\text{NNr}}$
$\mathcal{S}_{\text{hybrid}}$	Final class-balanced hybrid subset constructed by DHC
IPC	Images per class in the selected subset
Samples, models, and features	
x_i, y_i	Input sample and its label
x^*	Candidate selected at the current iteration
C	Number of classes
N	Number of samples in $\mathcal{D}_{\text{full}}$
$\phi(\cdot)$	CLIP feature embedding function used for nearest-neighbor retrieval and diversity
$\mathcal{M}_{\text{distill}}$	Prototype expert trained on $\mathcal{D}_{\text{distill}}$
$\mathcal{M}_{\text{NNr}}$	Instance expert trained on $\mathcal{D}_{\text{NNr}}$
$\mathcal{M}_{\text{FD}}$	Shared model trained on $\mathcal{D}_{\text{full}} \cup \mathcal{D}_{\text{distill}}$
$g_{\text{FD}}(x_i)$	Gradient of candidate x_i under $\mathcal{M}_{\text{FD}}$
$\bar{g}^{(c)}_{\text{full}}$	Class-mean gradient over $\mathcal{D}^{(c)}_{\text{full}}$
Valuation scores and subset construction	
$u_{\text{agn}}(x)$	Agnostic coreset score applied uniformly to real and distilled candidates
$u_\gamma(x)$	Source-aware diagnostic valuation with source-preference parameter γ
γ	Source-preference parameter in the diagnostic valuation u_γ
$z(\cdot)$	Within-class normalization used in diagnostic valuation
$\mathbf{1}[\cdot]$	Indicator function
$w_{\text{DI}}(x_i)$	Differentiated Identity Score for sample-wise source-aware valuation
$w_{\text{GM}}(x_i)$	Gradient-based representativeness score
$w_{\text{DC}}(x_i, x_j)$	Pairwise semantic similarity used in the diversity-aware update
$\hat{w}^{(c)}_{\text{DI}}, \hat{w}^{(c)}_{\text{GM}}$	Class-wise normalized DIS and representativeness scores
$\text{IS}_0(x_i)$	Initial importance score from normalized w_{DI} and w_{GM}
$\text{IS}_t(x_i)$	Importance score after the t-th diversity-aware update
$\mathcal{N}^{(c)}_{k_{\text{nn}}}(x^*)$	k_{nn} nearest neighbors of selected sample x^* within class c
Hyperparameters	
α	Weight balancing DIS and representativeness in IS_0
β	Diversity penalty strength in the score update
k_{nn}	Number of nearest neighbors penalized after each selection step

B Implementation and Reproduction Details

B.1 Balanced and Semantically Aligned Candidate Pool Construction

Algorithm 1 summarizes the construction of the balanced and aligned candidate pool used by DHC. The procedure retrieves one real nearest neighbor for each distilled sample within the same class and prevents duplicated real candidates in $\mathcal{D}_{\mathrm{NNr}}$.

Algorithm 1 Balanced and Aligned Candidate Pool Construction

Input: Full real dataset $\mathcal{D}_{\mathrm{full}}$, pre-synthesized distilled set $\mathcal{D}_{\mathrm{distill}}$, CLIP image encoder $\phi(\cdot)$
Output: Real nearest-neighbor set $\mathcal{D}_{\mathrm{NNr}}$ and hybrid candidate pool $\mathcal{C}_{\mathrm{hyb}}$

Initialize $\mathcal{D}_{\mathrm{NNr}} \leftarrow \emptyset$
for each class $c = 1, \ldots, C$ **do**
  Initialize $\mathcal{D}_{\mathrm{NNr}}^{(c)} \leftarrow \emptyset$
  Compute CLIP embeddings $\phi(x)$ for $x \in \mathcal{D}_{\mathrm{full}}^{(c)} \cup \mathcal{D}_{\mathrm{distill}}^{(c)}$
  for each distilled sample $\tilde{x} \in \mathcal{D}_{\mathrm{distill}}^{(c)}$ **do**
    Rank real samples $x \in \mathcal{D}_{\mathrm{full}}^{(c)}$ by cosine similarity between $\phi(\tilde{x})$ and $\phi(x)$
    Select the nearest real sample $x^\star$ that is not already in $\mathcal{D}_{\mathrm{NNr}}^{(c)}$
    $\mathcal{D}_{\mathrm{NNr}}^{(c)} \leftarrow \mathcal{D}_{\mathrm{NNr}}^{(c)} \cup \{x^\star\}$
  end for
  $\mathcal{D}_{\mathrm{NNr}} \leftarrow \mathcal{D}_{\mathrm{NNr}} \cup \mathcal{D}_{\mathrm{NNr}}^{(c)}$
end for
$\mathcal{C}_{\mathrm{hyb}} \leftarrow \mathcal{D}_{\mathrm{distill}} \cup \mathcal{D}_{\mathrm{NNr}}$

The resulting pool satisfies $|\mathcal{D}_{\mathrm{NNr}}^{(c)}| = |\mathcal{D}_{\mathrm{distill}}^{(c)}|$ for each class c.

B.2 Baseline Reproduction Details

Coreset selection baselines. For coreset selection baselines, we follow the experimental setup of Guo *et al.* (Guo et al., 2022). We pre-train the backbone model for 10 epochs before selection, and the pre-trained architecture is matched to the final target model. We evaluate Random, Forgetting (Toneva et al., 2018), K-center Greedy (Sener & Savarese, 2017), Moderate_DS (Xia et al., 2023), and Graph Cut (Iyer et al., 2021).

Dataset distillation baselines. For dataset distillation baselines, results explicitly cited in the main tables are quoted from their original publications. We re-implement DiT-IGD (Chen et al., 2025a) as a strong generative distillation baseline and also use it to generate the pre-synthesized distilled sets required by post-distillation hybrid curation methods. For ImageWoof and ImageNette, final evaluation follows the official hyperparameters and schedule from the DiT-IGD repository. For ImageNet-1K, we follow CDA-style training and use pre-generated soft labels for training on the final hybrid subsets.

Hybrid curation baseline. We reproduce CCFS (Chen et al., 2025b), the strongest post-distillation HDC baseline considered in our experiments, as the strongest post-distillation hybrid curation baseline. For a consistent comparison under the post-distillation setting, the distilled sets used by CCFS and DHC are generated with DiT-IGD under the same experimental protocol. Following the original CCFS design, we use Forgetting scores (Toneva et al., 2018) for real-sample selection and search the distillation portion over $\{0.1, 0.2, \ldots, 0.9\}$ for each IPC setting. We use the default setting of $J = 3$ curriculum phases, and reduce it to $J = 2$ when the candidate pool is insufficient to select the target number of real samples. Our implementation is based on the official CCFS codebase (Chen et al., 2025b), with adaptations for ImageWoof and ImageNette.

Final evaluation protocol. For all curated subsets from DHC and re-implemented baselines, we train the final target model from scratch and report top-1 test accuracy. Unless otherwise specified, we report mean accuracy and standard deviation over five training seeds. All experiments are conducted on NVIDIA RTX 4090 GPUs.

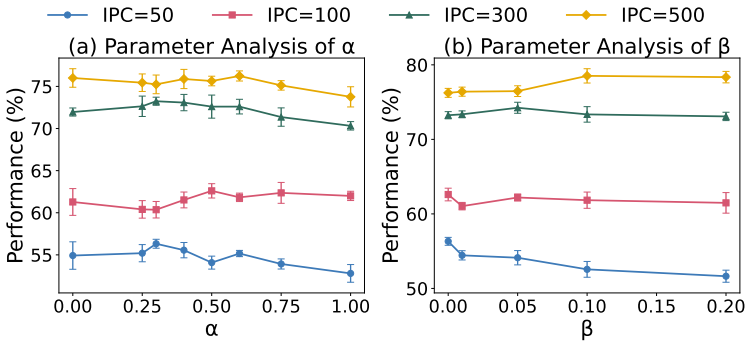


Figure 8: **Hyperparameter analysis on ImageWoof.** We analyze the sensitivity of DHC to **(a)** the weight α for balancing DIS and representativeness and **(b)** the diversity penalty strength β.

Table 5: **Hyperparameter settings for α and β.** We report the values used for each dataset, architecture, and IPC setting. The symbol "-" denotes settings that were not evaluated in our study.

Dataset	Parameter	ConvNet-6				ResNet-18			
		IPC50	IPC100	IPC300	IPC500	IPC50	IPC100	IPC300	IPC500
ImageWoof	α	0.30	0.50	0.30	0.60	0.90	1.00	0.50	0.30
	β	0.00	0.00	0.05	0.10	0.00	0.00	0.05	0.10
ImageNette	α	-	0.25	0.95	-	-	1.00	0.25	-
	β	-	0.00	0.20	-	-	0.00	0.20	-

B.3 Hyperparameter Settings

DHC has two main hyperparameters. The weight α controls the relative contribution of the Differentiated Identity Score and gradient-based representativeness in IS_0, while β controls the diversity penalty strength in the iterative score update. We set $k_{\mathrm{nn}} = 10$ in all experiments.

Figure 8 shows the sensitivity of DHC to α and β on ImageWoof with ConvNet-6. Table 5 reports the hyperparameter values used for ImageWoof and ImageNette across architectures and IPC settings.

C Additional Validation of Method Design

C.1 Candidate-Pool Construction and Diagnostics

Candidate Real Set Ablation. We conduct a controlled ablation on the real component of the hybrid candidate pool. The distilled set is fixed to the DiT-IGD distilled set, and we vary only the real candidate set paired with it. The compared real sets include the full real training set, a class-balanced random subset, a class-balanced Forgetting subset, the farthest real samples from the distilled samples in the CLIP feature space, and the nearest real samples used by DHC. Except for the full real set, all real candidate sets have the same class-wise cardinality as $\mathcal{D}_{\mathrm{distill}}$. We then apply the same DHC subset construction procedure to each resulting hybrid candidate pool.

Table 6 shows that using $\mathcal{D}_{\mathrm{NNr}}$ yields the highest accuracy at both IPC budgets. The random and Forgetting subsets outperform the full real set in this controlled setting, which supports controlling source cardinality before sample-wise valuation. Compared with random, Forgetting, and farthest-neighbor real sets, $\mathcal{D}_{\mathrm{NNr}}$ further improves accuracy, supporting the use of semantically aligned real counterparts when constructing the DHC candidate pool.

Representativeness and diversity diagnostics. To further characterize the real candidate sets, we examine their geometry in the CLIP feature space. We use the average distance to the corresponding full-data class centroid as a proxy for representativeness, where a smaller value indicates greater class centrality. We measure candidate dispersion using the average within-class pairwise distance, where

Table 6: **Ablation study on the candidate real set composition.** We fix the DiT-IGD distilled set and vary the real candidate set used to form the hybrid pool. All settings use DHC for final subset construction on ImageWoof with ConvNet-6. Except for Full, the real candidate set has the same class-wise size as $\mathcal{D}_{\text{distill}}$.

Candidate Pool		IPC	
Distilled Set	Real Set	100	300
DiT-IGD	Full	58.96±1.55	71.16±1.01
	Random	61.40±1.22	72.40±1.41
	Forgetting	61.76±0.65	73.40±0.88
	Farthest neighbors	61.20±1.30	70.87±0.47
	$\mathcal{D}_{\text{NNr}}$	**62.60±0.85**	**74.24±0.74**

Table 7: **Geometric diagnostics for real candidate sets.** We report the average distance to the corresponding full-data class centroid and the average within-class pairwise distance for each real candidate set used in the ImageWoof ConvNet-6 experiments. Lower centroid distance indicates greater class centrality, while higher pairwise distance indicates greater candidate dispersion.

IPC	Real candidate set	Centroid dist. ↓	Pairwise dist. ↑
100	Random	0.5169	0.7347
	Forgetting	0.5101	0.7233
	Farthest	0.7067	**0.9477**
	$\mathcal{D}_{\text{NNr}}$	**0.4150**	0.5531
300	Random	0.5169	0.7347
	Forgetting	0.5111	0.7253
	Farthest	0.6423	**0.8815**
	$\mathcal{D}_{\text{NNr}}$	**0.4295**	0.5833

a larger value indicates a more dispersed candidate set. These aggregate statistics are used only for diagnosis and are not directly optimized by DHC.

As shown in Tables 6 and 7, among the source-balanced candidate sets, the farthest-neighbor set has the largest pairwise distance but the lowest downstream accuracy at both IPC budgets. Candidate dispersion alone therefore does not ensure an effective hybrid subset. In contrast, $\mathcal{D}_{\text{NNr}}$ has the smallest centroid distance and yields the highest accuracy. These results support using nearest real neighbors to construct a class-centered and semantically aligned candidate pool before sample-wise valuation.

Diversity after subset construction. Because nearest-neighbor construction reduces unrelated content variation, it may also produce locally similar candidates around the distilled samples. We therefore examine whether the final DHC subset remains concentrated in these regions. Table 8 compares the average within-class pairwise CLIP distance of the distilled set, the uncurated balanced-aligned pool $\mathcal{C}_{\text{hyb}}$, and the final DHC subset.

At IPC 100, the three distances are similar, with the final subset exhibiting slightly lower dispersion. At IPC 300, however, the final DHC subset has a larger pairwise distance than both the distilled set and the uncurated hybrid pool. This budget-dependent behavior is consistent with the diversity-aware update becoming more consequential as more samples are selected. Overall, the results indicate that nearest-neighbor pool construction does not force the final subset to maximize similarity to the distilled samples, while DHC can preserve or increase semantic dispersion when redundancy becomes more pronounced at larger budgets.

Table 8: **Diversity before and after hybrid subset construction.** We report the average within-class pairwise CLIP distance on ImageWoof with ConvNet-6. The hybrid pool is $\mathcal{C}_{\mathrm{hyb}} = \mathcal{D}_{\mathrm{distill}} \cup \mathcal{D}_{\mathrm{NNr}}$.

Subset	IPC 100	IPC 300
$\mathcal{D}_{\mathrm{distill}}$	**0.5689**	0.5646
$\mathcal{C}_{\mathrm{hyb}}$	0.5684	0.5824
DHC subset	0.5583	**0.6028**

Table 9: **DIS statistics across datasets, budgets, and architectures.** Each entry reports median DIS / positive-DIS ratio. "Cand." and "Sel." denote candidates and selected samples, respectively. Positive-DIS ratio is the fraction of samples with $w_{\mathrm{DI}}(x) > 0$.

Dataset	Architecture	IPC	Real cand.	Distilled cand.	Sel. real	Sel. distilled
ImageWoof	ConvNet-6	50	$-11.6684/0.0140$	$9.3916/0.9660$	$-7.5866/0.0228$	$8.5373/0.9599$
ImageWoof	ConvNet-6	300	$-5.2503/0.0883$	$0.8788/0.7007$	$-4.2624/0.1016$	$1.1815/0.7281$
ImageNette	ConvNet-6	100	$-3.3631/0.1120$	$1.8339/0.9080$	$-1.9623/0.1629$	$1.5420/0.8973$
ImageNette	ConvNet-6	300	$-2.0255/0.2167$	$0.7801/0.7967$	$-1.7515/0.2436$	$1.2152/0.8286$
ImageWoof	ResNet-18	50	$-6.5774/0.0220$	$5.9575/0.9940$	$-0.5904/0.3583$	$6.5188/0.9979$

C.2 Direct Validation of the Differentiated Identity Score

DIS statistics across settings. We directly examine whether DIS captures the expected source-dependent response patterns beyond its contribution to downstream accuracy. For each setting, Table 9 reports the median DIS and the fraction of samples with positive DIS separately for real and distilled candidates and for their selected subsets. Across datasets, IPC budgets, and architectures, real candidates have negative median DIS and low positive-DIS ratios, whereas distilled candidates have positive median DIS and substantially higher positive-DIS ratios. This consistent shift supports the interpretation of DIS as a sample-wise source-aware signal.

Selection does not merely preserve the aggregate candidate distributions. Selected real samples shift toward higher DIS in every evaluated setting, while selected distilled samples remain predominantly positive but may shift in either direction. This behavior is consistent with DIS being combined with representativeness and diversity-aware updates rather than used as a standalone ranking score.

Figure 9 visualizes the corresponding candidate and selected distributions for two representative ImageWoof classes under ConvNet-6 at IPC 50. In both classes, real and distilled candidates exhibit systematically shifted DIS distributions, although their supports may overlap. After selection, samples from both sources remain represented and the distributions become more concentrated.

Directionality of DIS. We further test whether the direction encoded by DIS is consequential for selection. Within each class of the same balanced-aligned pool, we reverse the normalized DIS term while keeping the candidate pool, representativeness score, diversity update, α, β, and downstream training protocol unchanged. Table 10 shows that the original DIS direction outperforms its reversal at every evaluated IPC budget. The differences are largest at IPC 50 and IPC 100 and become small at larger budgets, indicating that the DIS direction is most consequential under compact data budgets. These results support the use of positive DIS as the prototype-oriented direction within the combined selection score, without implying that DIS alone determines sample utility.

C.3 Selected Source Composition

We analyze the source composition of selected hybrid subsets on ImageWoof with ConvNet-6. For each method and IPC budget, we report the numbers of selected real and distilled samples aggregated over all classes under its best-performing configuration. This comparison focuses on how the final real-distilled composition differs between CCFS, which controls source composition through its curation protocol, and DHC, whose composition results from sample-wise valuation and subset construction.

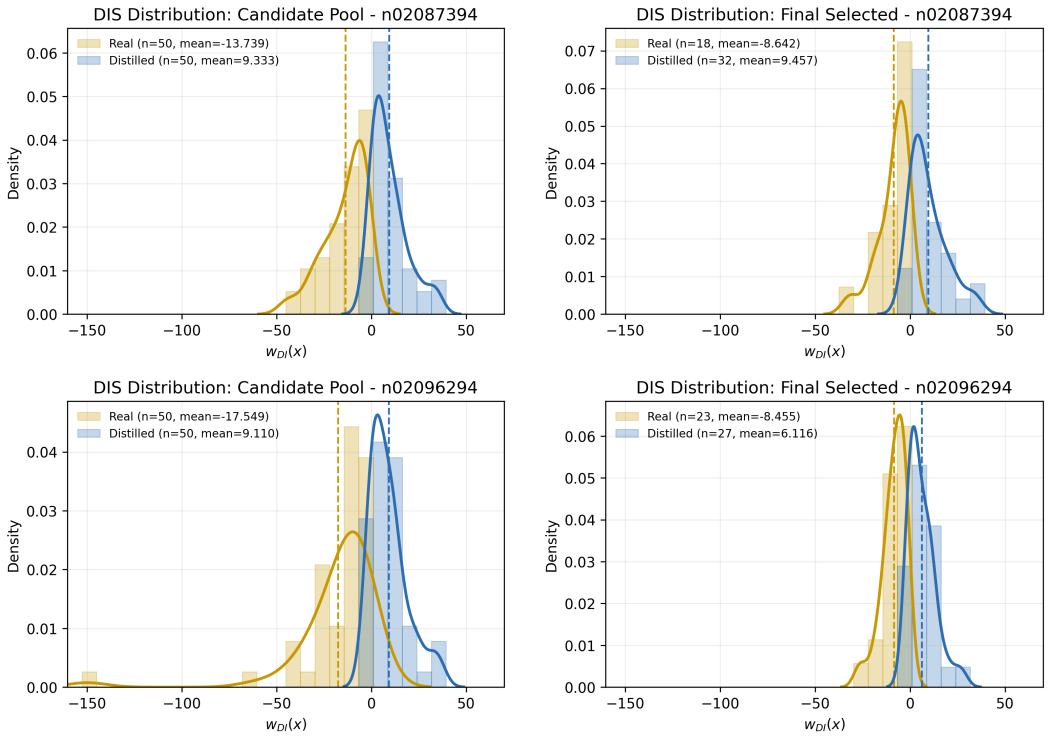


Figure 9: **DIS distributions before and after subset construction.** The left and right columns show the balanced-aligned candidate pool and final DHC subset, respectively, for two representative ImageWoof classes under ConvNet-6 at IPC 50. Shaded histograms and solid curves show the source-specific distributions, while dashed vertical lines denote their means.

Table 10: **Directionality of DIS on ImageWoof with ConvNet-6.** Reverse DIS negates the class-wise normalized DIS term while leaving all other components and settings unchanged.

Strategy	IPC 50	IPC 100	IPC 300	IPC 500
Original DIS	**56.32±0.53**	**62.60±0.85**	**74.24±0.74**	**78.52±0.96**
Reverse DIS	52.93±0.41	59.91±1.18	74.20±1.47	78.13±0.77

Table 11 shows that DHC produces IPC-dependent source compositions without prescribing a fixed real-distilled ratio. Compared with CCFS, DHC selects a larger real portion at IPC 50, IPC 300, and IPC 500, while producing a similar distilled-dominant composition at IPC 100. These results indicate that DHC does not simply follow the source composition used by CCFS. Instead, its final source mixture changes with the budget through sample-wise source-aware valuation and subset construction.

D Practicality and Robustness

D.1 Default Hyperparameter Generalization

We evaluate whether DHC remains effective without tuning α and β separately for each architecture and budget. We use a simple budget-aware default with $\alpha = 0.6$ throughout, $\beta = 0$ at low budgets, and $\beta = 0.1$ at larger budgets. Specifically, we set $\beta = 0$ for IPC 50 and 100 and $\beta = 0.1$ for IPC 300 and 500 on ImageWoof, using the same rule for ConvNet-6 and ResNet-18.

As shown in Table 12, the default rule retains competitive performance relative to tuned DHC across ImageWoof budgets and is competitive with CCFS under both architectures. These results show that

Table 11: **Selected source composition on ImageWoof with ConvNet-6.** Each entry reports the number of selected real and distilled samples, aggregated over all classes. The value in parentheses denotes the real-sample ratio.

Method	IPC			
	50	100	300	500
CCFS	150/350 (30.0%)	300/700 (30.0%)	1200/1800 (40.0%)	2500/2500 (50.0%)
DHC (Ours)	183/317 (36.6%)	297/703 (29.7%)	1480/1520 (49.3%)	2629/2371 (52.6%)

Table 12: **Performance with shared default hyperparameters.** Results are top-1 test accuracy (%). DHC (Default) uses the same budget-aware rule across both ImageWoof architectures and all IPC settings.

Dataset	Architecture	Method	IPC 50	IPC 100	IPC 300	IPC 500
ImageWoof	ConvNet-6	CCFS	54.84±0.66	62.20±0.31	73.56±0.84	79.48±0.53
		DHC (Tuned)	56.32±0.53	62.60±0.85	74.24±0.74	78.52±0.96
		DHC (Default)	55.16±0.37	61.83±0.50	73.85±0.30	78.52±0.96
	ResNet-18	CCFS	63.04±1.05	68.68±0.85	78.80±0.57	81.92±0.39
		DHC (Tuned)	66.64±0.85	71.92±0.77	80.80±1.05	84.48±0.93
		DHC (Default)	65.04±0.80	70.11±1.02	80.27±0.57	83.27±0.50

DHC can use a shared budget-aware rule across architectures and IPC settings, reducing the need for architecture-specific and per-budget tuning, while tuning can still improve performance.

D.2 Computational Cost and Memory

Table 13: **Post-distillation curation cost on ImageWoof.** We report wall-clock time at IPC 50 and 500 and peak GPU memory.

Method	Time (s)		Peak memory (MB)
	IPC 50	IPC 500	
Random	1.3	1.3	0.0
Forgetting	369.5	356.7	8268.0
Graph Cut	404.9	451.5	8268.0
CCFS	7929.2	7630.8	13168.8
DHC (Ours)	3969.8	12238.2	8251.7

We report post-distillation curation cost on ImageWoof with ConvNet-6, measured on a single NVIDIA RTX 4090 GPU. The measurement includes candidate-pool construction, score computation, and subset selection after the distilled artifacts are available. We exclude distilled-set generation and final model training to isolate the curation cost evaluated under our post-distillation protocol.

Table 13 shows that DHC incurs greater curation time than conventional coreset methods due to source-specific expert training and gradient-based valuation. Relative to CCFS, DHC is faster at IPC 50 but slower at IPC 500, where training the two source-specific experts dominates its cost. DHC uses less peak GPU memory than CCFS in this setting.

D.3 Cross-Architecture Usability

We further examine whether a curated subset remains useful when the target training architecture changes. In this experiment, the distilled set generation and hybrid curation procedure use ConvNet-6, while the test model is ResNet-18. This setting tests whether the curated subset can be reused across architectures without rerunning the full curation pipeline. For reference, we also report DHC with architecture-matched curation and evaluation under ResNet-18.

Table 14: **Cross-architecture generalization on ImageWoof.** The distilled set generation and hybrid curation pipeline use ConvNet-6, while the final model is ResNet-18. The last row reports DHC with architecture-matched curation and evaluation under ResNet-18.

Curate with	Method	Target model: ResNet-18		
		IPC 50	IPC 100	IPC 300
ConvNet-6	Random	53.00±2.22	66.40±1.34	79.60±0.96
	Forgetting	56.72±1.16	66.48±1.18	78.72±1.00
	$\mathcal{D}_{\mathrm{NNr}}$	59.80±1.11	69.20±1.12	79.88±1.20
	DiT-IGD	63.24±1.32	70.16±1.14	77.64±0.53
	CCFS (Chen et al., 2025b)	63.38±1.25	70.88±0.59	80.48±0.27
	DHC (Ours)	63.58±1.00	71.04±0.74	80.60±0.42
ResNet-18	DHC (Ours)	66.64±0.85	71.92±0.77	80.80±1.05

Table 15: **Performance comparison on long-tailed ImageWoof.** We construct an imbalanced training set with an imbalance ratio of 100 and curate 10% samples per class.

Ratio Per Class	Random	Graph Cut	Forgetting	MDS	$\mathcal{D}_{\mathrm{NNr}}$	DiT-IGD	DHC (Ours)
10%	20.87±0.90	24.53±0.34	19.67±0.25	21.60±0.16	23.00±0.49	24.67±0.47	**26.07±0.19**

Table 14 shows that the ConvNet-6 curated DHC subset remains competitive when transferred to ResNet-18. Among all methods curated with ConvNet-6, DHC achieves the highest accuracy at IPC 50, 100, and 300. This indicates that the curated subset can retain utility when transferred between the evaluated convolutional architectures.

The architecture-matched ResNet-18 DHC results provide an upper reference for this transfer setting. Compared with this reference, the transferred ConvNet-6 curated DHC subset incurs a gap of 3.06% at IPC 50, 0.88% at IPC 100, and 0.20% at IPC 300. The small gap at medium and larger budgets suggests that DHC has practical usability potential, since a subset curated once with a lightweight model can remain effective for a stronger target architecture.

D.4 Evaluation under Long-Tailed Data

We further evaluate DHC under class imbalance. Following the exponential imbalance protocol in (Cui et al., 2019), we construct a long-tailed version of ImageWoof with an imbalance ratio of 100, where the number of training samples ranges from 1300 in the head class to 13 in the tail class. For each class, we curate 10% of its available training samples and evaluate the trained model on the original ImageWoof test set.

As shown in Table 15, DHC achieves the highest accuracy among the compared methods in this setting. The result provides an additional stress test showing that DHC remains effective in this long-tailed setting.

E Additional Diagnostics for Post-Distillation HDC

We provide additional diagnostics for the two empirical questions discussed in Section 2.2. These results are intended to test whether the observations in Section 2.2 are specific to the displayed settings.

E.1 Candidate-Pool Source Composition

Section 2.2 uses Forgetting as the agnostic valuation method. We further evaluate Graph Cut as another source-agnostic selection method and compare the same three candidate-pool constructions. Figure 10 reports the accuracy gain over the raw unbalanced pool. The balanced-random pool is not uniformly beneficial and can underperform the raw pool at IPC=50. In contrast, the balanced-aligned pool based on $\mathcal{D}_{\mathrm{NNr}}$ consistently improves over the raw pool across the displayed IPC budgets. This

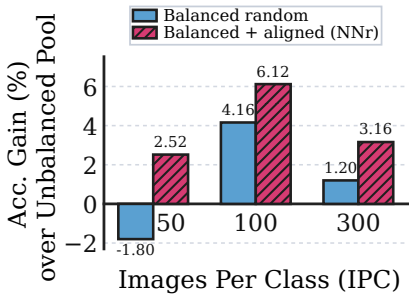


Figure 10: **Candidate-pool diagnostic with Graph Cut as agnostic selection.** We report the accuracy gain over the raw unbalanced pool on ImageWoof with ConvNet-6.

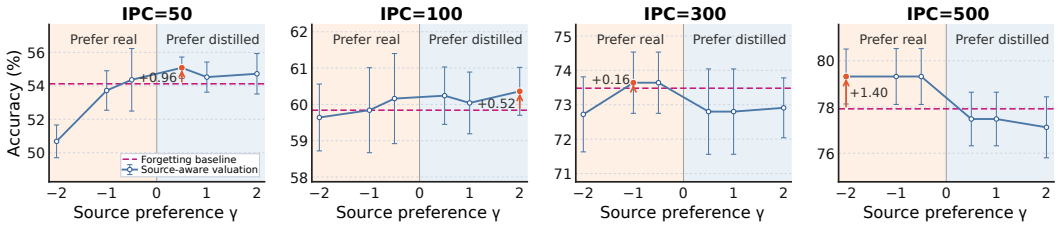


Figure 11: **Source-preference sweep across IPC budgets.** We sweep the source-preference parameter γ on $\mathcal{C}_{\text{bal-NNr}}$. The preferred source-level correction varies with the IPC budget, indicating that a fixed global source preference is insufficient for post-distillation HDC.

supports the Section 2.2 observation that controlling source cardinality alone is insufficient, and that semantic alignment provides more consistent gains in hybrid candidate-pool construction.

E.2 Source-Level Valuation Mismatch

Section 2.2 reports the source-preference sweep at two IPC budgets. Figure 11 provides the full sweep over IPC $\in \{50, 100, 300, 500\}$. The best source-preference value changes across budgets. Positive γ is preferred at IPC 50 and IPC 100, while negative γ becomes competitive or preferred at IPC 300 and IPC 500. This confirms that source information is useful for hybrid valuation, but a single fixed source preference is not stable across budgets. It further motivates the sample-wise source-aware valuation used by DHC.

F Related Work

F.1 Coreset Selection

Coreset Selection (CS) aims to select a compact subset of real samples that preserves the utility of the full training set. Existing CS methods differ mainly in how they estimate the utility of candidate samples. Representativeness-based methods select samples that approximate the data distribution or class-level structure, using feature similarity, prototype matching, or gradient alignment (Welling, 2009; Xia et al., 2023; Killamsetty et al., 2021). Difficulty-based methods assign larger values to samples that are hard to learn, uncertain, or frequently forgotten during training (Toneva et al., 2018; Paul et al., 2021; Baldock et al., 2021; He et al., 2024). Diversity-based methods encourage coverage of the data distribution by avoiding redundant selections, often through distance-based or coverage-based criteria (Sener & Savarese, 2017; Zheng et al., 2023). Submodular-based methods further combine representativeness and diversity in a unified subset objective, such as graph-cut or facility-location formulations (Iyer et al., 2021; Maharana et al., 2024; Yu et al., 2025).

These methods provide important tools for data-efficient learning and motivate the representativeness and diversity criteria used in DHC. However, standard CS is designed for homogeneous real datasets, where all candidates share the same source and can be ranked by a common scoring criteria. In

post-distillation HDC, the candidate pool contains both distilled and real samples, which can play different roles in training. Directly applying a CS score to all candidates therefore yields an agnostic valuation strategy, as discussed in Section 2.2. DHC builds on the representativeness and diversity principles of CS, but augments them with sample-wise source-aware valuation for heterogeneous real-distilled candidate pools.

F.2 Dataset Distillation

Dataset Distillation (DD) compresses a full training set into a small set of synthetic samples that can train a model efficiently. Early DD methods optimize synthetic data by matching gradients, features, training trajectories, or kernel-induced statistics between synthetic and real data (Zhao et al., 2020; Zhao & Bilen, 2023; Cazenavette et al., 2022; Wang et al., 2022; Nguyen et al., 2021; Chen et al., 2024). Recent studies improve scalability and visual fidelity through decoupled optimization and generative distillation frameworks (Yin et al., 2023; Shao et al., 2024; Yin & Shen, 2024b; Shen et al., 2025; Cazenavette et al., 2023; Gu et al., 2024; Su et al., 2024; Zou et al., 2025). Diffusion-based distillation further improves the realism and semantic coverage of synthetic samples, and DiT-IGD (Chen et al., 2025a) is used in our experiments to generate the pre-synthesized distilled sets.

DD is complementary to CS. While CS preserves natural real instances, DD provides compact distilled samples that often encode class-level or prototype-like information. However, distilled samples may underrepresent fine-grained instance variation or natural data complexity under strict storage budgets (Guo et al., 2024; Sun et al., 2024). This complementarity motivates post-distillation HDC, where a pre-synthesized distilled set is curated together with real samples. DHC does not modify the distillation algorithm itself. Instead, it treats the distilled set as given and studies how to select a compact hybrid subset from both distilled and real candidates.

F.3 Hybrid Dataset Curation

Hybrid Dataset Curation (HDC) seeks to construct compact training subsets by combining distilled and real samples. Recent methods differ mainly in how real samples are selected relative to the distilled set. SelMatch (Lee & Chung, 2024) follows a select-then-distill strategy, where real samples are selected before distillation and the distilled data are optimized to complement them. This makes SelMatch related to hybrid curation, but its selection stage is performed before the final distilled set is available. In contrast, our work focuses on a post-distillation setting, where the distilled set is already synthesized and the remaining question is how to curate a hybrid subset from distilled and real candidates.

CCFS (Chen et al., 2025b) follows a distill-then-select strategy, where a distilled set is first obtained and real samples are then selected through a curriculum coarse-to-fine procedure. Since CCFS selects real samples after distillation, it is closer to our post-distillation setting and serves as the most directly related HDC baseline in our experiments. CCFS also reports stronger results than SelMatch on overlapping settings, which further motivates using CCFS as the main hybrid curation baseline. However, CCFS controls the real-distilled composition through its curation pipeline and IPC-dependent ratios. Such source-level control can mitigate source imbalance, but it does not assign differentiated values to individual candidates within the same source.

DHC instead performs sample-wise source-aware valuation over a heterogeneous real-distilled candidate pool after distillation. It combines DIS with gradient-based representativeness and semantic diversity, avoiding a fixed real-distilled subset composition before valuation. DHC differs from SelMatch in when curation occurs and from CCFS in how source composition is determined.

G Limitations

DHC introduces additional computation for source-aware valuation. DIS requires training source-specific experts, and representativeness estimation requires gradient computation under an auxiliary model. Although these costs are manageable in our experiments, more efficient source-aware valuation is important for larger-scale curation. Our experiments focus on image classification. Extending DHC to broader data modalities and structured prediction tasks may require new definitions of source-aware valuation and diversity-aware subset construction.