POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Green AI Training via Dataset Subset Selection: The GAIT Framework

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Davide Esposito

Advisor: Prof. Marcello Restelli

Co-advisors: Alessandro Lavelli, Alessio Russo

Academic year: 2025-2026

---

1. Introduction

Recurring training has become a practical necessity in modern machine learning systems: models are rebuilt or updated as data evolve, pipelines change, and monitoring reveals degradation. In this lifecycle view, training time and energy consumption are not merely engineering metrics - they become operational costs that accumulate across cycles and translate into environmental impact through associated emissions [7, 14, 15].

This thesis addresses the problem of reducing the computational footprint of supervised deep learning training, with a focus on *training time*, *energy consumption*, and the resulting emissions. The optimization lever is *data-driven*: if the effective training set can be reduced while preserving most of the task-relevant information, then the number of optimization updates decreases and end-to-end training cost drops without modifying the downstream training pipeline.

We work in a *stationary* regime: subset selection is performed once, before training, and the task-relevant distribution is assumed stable over the horizon of a training cycle. The selection stage must be easy to integrate, must not require invasive access to training internals, and must remain strictly cheaper than full training in time, energy, and memory.

Contributions. The thesis makes two distinct and complementary contributions.

The first is **GAIT** (Green AI Training), a modular decision-support *framework* for cost-aware subset selection. GAIT's role is not to define how subsets are selected, but to make the *consequences* of selection decisions comparable and auditable: it standardizes how training time and energy are projected across candidate subsets via hardware-calibrated profiling, and how savings are reported in a baseline-referenced, reproducible form. Critically, GAIT is *selector-agnostic* - it operates as a pipeline layer that wraps any compliant selection method and produces versionable cost artifacts regardless of the underlying selector. This decoupling is the key architectural contribution: it separates the *what to select* problem from the *how much does it cost* problem, and does so without requiring access to training internals or modifying the downstream pipeline.

The second contribution is a **new training-free, one-shot, supervised subset selection method** designed to operate effectively within

the GAIT regime. The method builds a target-aware hardness prior from local neighborhoods in a fixed representation space and draws diverse subsets via weighted sampling without replacement, with optional geometry-based regularizers for robustness at aggressive pruning rates. Together, the two contributions form an integrated system: GAIT provides the measurement and accountability layer, while the proposed selection method provides a competitive, low-overhead selector that operates within GAIT's constraints.

2. Related Work: Dataset Subset Selection

Dataset subset selection (often called *coreset selection*) aims to retain a small subset of training examples that preserves most of the predictive utility of the full dataset. It is worth distinguishing it from adjacent regimes: *active learning* constructs a labeled dataset via sequential oracle queries rather than compressing an existing one; *dataset distillation* synthesizes new training examples rather than selecting originals; and *dataset pruning* typically emphasizes removing noisy or harmful samples rather than approximating full-data utility under an explicit budget. This thesis operates strictly in the supervised coreset selection regime: a fully labeled pool is given and the objective is to select a size-k subset that preserves downstream predictive performance.

A useful way to read the coreset literature is through three axes: the signal source (training-free vs. training-dependent), the schedule (one-shot vs. iterative), and the mechanism used to convert a signal into a size-constrained subset. Recent surveys emphasize that, in end-to-end pipelines, selection overhead - representation extraction, neighborhood/similarity construction, and optimization - can dominate the savings from training on fewer samples if not controlled [11].

Under the constraints of this thesis, three points are central. First, *training-dependent* methods (EL2N [13], forgetting-based scoring, GradMatch [5], bilevel objectives) can be effective at small budgets because the signal is close to the learning objective, but they couple selection to the training stack and increase overhead via warm-up training or gradient computation; in reported benchmarks, the most expensive methods require runtimes on the order of hours or days [11], making them incompatible with the production constraints targeted here. Second, *training-free* selectors avoid this coupling but typically rely on geometry-only signals, coverage, density and pairwise similarity that are agnostic to the task target; at moderate budgets, this agnosticism limits their utility compared to methods that exploit label information. Third, many training-free selectors rely on similarity structures that become infeasible if implemented densely ($O(n^2)$ memory/time), motivating sparse neighborhood designs and lightweight sampling mechanisms.

The gap this thesis addresses is therefore not a choice between the two families, but the absence of a *training-free selector that incorporates a supervised, target-aware hardness signal* while remaining compatible with one-shot, overhead-bounded operation. This thesis positions itself explicitly in that regime, and GAIT provides the framework layer that makes cost accountability portable across any selector operating within it.

3. GAIT Framework

GAIT (Green AI Training) is a modular, decision-support framework that standardizes how subset candidates are generated, profiled, and compared *under a fixed training context*. GAIT is designed as a deterministic pipeline stage producing *versionable artifacts* to ensure traceability, reproducibility, and auditability of cost-aware decisions.

Operating regime (framework boundaries). GAIT is context-conditioned: its outputs are meaningful only with respect to (i) the dataset (and any fixed representation used for selection), (ii) the training configuration (model, batch size, optimizer/recipe, epoch budget, precision, evaluation cadence), and (iii) the target execution environment (hardware/software stack and available telemetry). Within this regime, GAIT targets selection protocols that are one-shot, training-free by default, and low overhead, so that selection remains subordinate to the savings enabled at training time.

Modules and artifacts. GAIT is organized into three modules, each emitting a versioned

artifact that acts as the interface to downstream stages:

- **Selection** generates candidate subsets across budgets and exports a *subset specification* (indices/IDs, subset size, method and seed, representation/preprocessing identifiers).
- **Profiler** calibrates environment-specific predictors of training time and energy on the *target machine* and exports a *curve bundle* (fitted curves, probe protocol metadata, environment fingerprint, and fit-quality indicators).
- **Estimation** combines subset specifications and the curve bundle to export an *estimation report* containing projected time/energy (optional emissions), baseline-referenced Δ-estimates, and conservative risk signals.

Profiler: calibration by probe runs. Floating-point operations (FLOPs) alone are often insufficient to predict realized time/energy because throughput depends on the concrete hardware/software stack and on system effects (data pipeline, scheduling, warm-up). GAIT therefore adopts calibration: the Profiler runs a small number of short *probe trainings* at selected dataset fractions, keeping the training configuration fixed and changing only the number of samples per epoch. Since early epochs can be affected by transient overhead (initialization, compilation, caching), GAIT separates a short *burn-in* phase from steady training and aggregates totals as burn-in contribution plus steady contribution. The resulting curve bundle is reusable across budgets within the same context, and its fit-quality indicators (coverage/error summaries and extrapolation flags) provide a lightweight confidence layer for downstream estimates.

Large-scale validation: accurate projections with minimal calibration. To test calibration in a realistic long regime and to highlight amortization, we evaluate GAIT on a compute-intensive setting: **ResNet50** fine-tuning (*all layers unfrozen*) on **Food101** (∼ 100,000 images) for **50 epochs**. Budgets are generated via `RandomSelector` to isolate *cost estimation* from selection quality. The Profiler uses only **two probe points** ($M = 2$), each a **4-epoch** micro-training ($E_{\text{burn}} = 1$ burn-in epoch plus 3 steady epochs), to calibrate a curve bundle for the full configuration. Table 1 reports predicted vs. measured totals. Despite the multi-hour horizon, errors are already below ∼6% at 10% budget and rapidly decrease as the budget increases; from 20% onward, time error stays below 1% and energy error becomes very small. This matches a *scale effect*: as utilization increases, fixed overheads are amortized and time/energy scale more smoothly with subset size, making low-variance curves highly accurate.

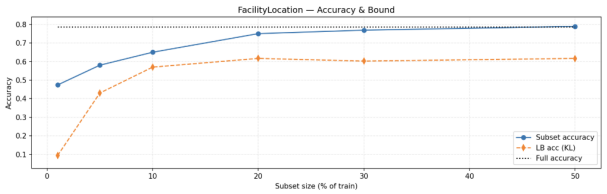

Figure 1: Illustration of GAIT's anchored risk envelope: subset accuracy (solid), conservative lower envelope from the KL-based proxy via Pinsker (dashed), and full-pool reference (dotted) for Facility Location across budgets.

Estimation: projections, deltas, and conservative risk. Given a candidate subset S (budget $m = |S|$) and a target training plan, Estimation queries the curve bundle to project $\widehat{T}(S)$ and $\widehat{\mathcal{E}}(S)$ (and optional emissions via a deployment-provided carbon intensity), then reports baseline-referenced Δ-estimates against the full pool V. Beyond cost, GAIT reports a conservative *distributional risk proxy* to flag candidates that may be non-representative. The proxy is distributional (not algorithmic): it controls expectation shifts due to distribution mismatch and can be instantiated through total variation bounds, using a KL→TV conversion via Pinsker's inequality when convenient [1]. When a trusted full-pool reference performance is available, GAIT can convert the proxy into an *anchored* conservative envelope; Fig. 1 shows an example.

4. Problem Formulation and Operating Assumptions

Let $D = \{(x_i, y_i)\}_{i=1}^{n}$ be a labeled training pool of i.i.d. samples from an unknown distribution P over $\mathcal{X} \times \mathcal{Y}$, with index set $V = \{1, \ldots, n\}$. For a budget $k \ll n$, the subset selection problem aims to select $S \subseteq V$, $|S| = k$, such that

Budget (%)	t_{real} (s)	t_{pred} (s)	e_{real} (kWh)	e_{pred} (kWh)	APE_t (%)	APE_e (%)
10	6391	6055	0.2744	0.2582	5.26	5.89
20	9116	9063	0.4097	0.3975	0.58	2.98
30	12045	12069	0.5431	0.5367	0.20	1.18
50	18022	18091	0.8175	0.8154	0.39	0.26
75	25553	25617	1.1643	1.1636	0.25	0.06

Table 1: Large-scale simulation (Food101, ResNet50 fine-tuning, 50 epochs): Profiler accuracy after calibrating the curve bundle with only $M = 2$ probe runs (4 epochs each). Errors are absolute percentage errors (APE).

a model trained on the induced subset $D_S = \{(x_i, y_i) : i \in S\}$ achieves performance comparable to training on the full pool. Operationally, the thesis considers a *budget grid* and produces multiple candidates that must be comparable under a fixed training context.
A standard formal reading is the (idealized) bilevel objective:

$$S_k^\star \in \arg\min_{S\subseteq V:\ |S|=k} R_P(\theta_S), \qquad \theta_S \in \arg\min_{\theta} \widehat{R}_S(\theta), \tag{1}$$

where $R_P(\theta) = E_{(X,Y)\sim P}[\ell(\theta; X, Y)]$ is the population risk and $\widehat{R}_S$ is the empirical risk on D_S. Since P is unknown and the bilevel objective is intractable, practical methods rely on surrogates computed from representations, similarities, or lightweight supervised signals [9, 11].

Operating assumptions (GAIT-aligned). The operative assumptions of this thesis reflect production boundaries and motivate both GAIT and the proposed method: (i) **model-agnostic integration**: selection does not require architecture-specific hooks or access to model internals; (ii) **training-free selection**: selection avoids gradients, training losses, or intermediate optimization states; (iii) **one-shot execution**: the subset is computed once prior to training (not iteratively refined); (iv) **low overhead and memory-safety**: selection must remain negligible relative to the intended training savings and avoid dense $O(n^2)$ similarity kernels. These constraints intentionally restrict the method space, but improve deployability in standardized pipelines where modifying training code or running expensive selection routines is often infeasible.

5. Methodology

This thesis proposes a supervised subset-selection family designed to satisfy the operating constraints introduced in Section 4: *training-free*, *one-shot*, and *low-overhead* (time/memory) while remaining applicable to both regression and classification. The method is defined as a modular pipeline: (i) build a target-aware *hardness prior* from local neighborhoods in a fixed representation space, (ii) optionally regularize it with geometry-only components for robustness at aggressive pruning, and (iii) draw a size-k subset via *weighted sampling without replacement* (WRSWoR).

Problem setup. Let $D = \{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from an unknown stationary distribution P over $\mathcal{X} \times \mathcal{Y}$, with index set $V = \{1, \dots, n\}$. Given a budget $k \ll n$, the ideal objective is to find $S \subseteq V$, $|S| = k$, such that training on D_S yields performance close to training on D. Since training-dependent bilevel formulations are incompatible with the constraints above, we instead construct a *training-free scoring prior* $\pi \in \Delta^{n-1}$ and sample S without replacement according to π.

Representation space and neighbor queries. Let $\Phi : \mathcal{X} \to R^d$ be a fixed representation mapping (standardized features or frozen embeddings) and define $z_i = \Phi(x_i)$. We compute neighbors in a standardized selection space

$$\tilde{z}_i := \text{Std}(z_i), \qquad d_{ij} := \|\tilde{z}_i - \tilde{z}_j\|_2. \tag{2}$$

Let $\mathcal{N}_i$ be the set of the K nearest neighbors of i (excluding i), with distances $\{d_{ij}\}_{j\in\mathcal{N}_i}$. All signals below reuse this sparse KNN structure (no dense $n \times n$ kernels).

V1: target-aware local hardness prior. We define a local, target-conditioned hardness score using kernelized KNN neighborhoods. Let h_i be a data-dependent bandwidth given by the distance to the K_σ-th neighbor:

$$h_i := d_{i,(K_\sigma)}, \qquad 1 \le K_\sigma \le K. \tag{3}$$

Define normalized Gaussian kernel weights

$$\begin{aligned} \widetilde{\omega}_{ij} &:= \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h_i+\varepsilon}\right)^2\right), \\ \omega_{ij} &:= \frac{\widetilde{\omega}_{ij}}{\sum_{r\in\mathcal{N}_i}\widetilde{\omega}_{ir}+\varepsilon}. \end{aligned} \tag{4}$$

This corresponds to local kernel smoothing of conditional moments (Nadaraya–Watson-style) [12, 16].

Regression. For $y \in R$, estimate a local conditional mean and variance:

$$\widehat{\mu}_i := \sum_{j\in\mathcal{N}_i}\omega_{ij}y_j, \qquad \widehat{v}_i := \sum_{j\in\mathcal{N}_i}\omega_{ij}(y_j-\widehat{\mu}_i)^2. \tag{5}$$

Large $\widehat{v}_i$ indicates locally ambiguous neighborhoods, hence higher *target-driven hardness*.

Classification. For $y \in \{1,\dots,C\}$, estimate a local class distribution $\widehat{p}_{ic} := \sum_{j\in\mathcal{N}_i}\omega_{ij}I[y_j = c]$ and define hardness via an uncertainty functional (e.g., entropy, Gini, or margin).

To stabilize scaling across datasets, we apply a rank normalization $s_i := \mathrm{Rank01}(\widehat{v}_i) + \varepsilon$ (or the analogous uncertainty score), and define the V1 weights and prior

$$\rho_i^{\mathrm{V1}} := s_i, \qquad \pi_i^{\mathrm{V1}} := \frac{\rho_i^{\mathrm{V1}}}{\sum_{r=1}^n \rho_r^{\mathrm{V1}}}. \tag{6}$$

One-shot subset construction via WRSWoR. Given nonnegative weights ρ (or prior π), we draw a size-k subset *without replacement*, so that higher weights imply higher inclusion likelihood while avoiding duplicates [2]. We implement WRSWoR using the Gumbel–Top-k construction [6]:

$$\begin{aligned} g_i &\sim \mathrm{Gumbel}(0,1), \\ \mathrm{key}_i &:= \log(\pi_i+\varepsilon) + g_i, \\ S &:= \text{Top-}k(\{\mathrm{key}_i\}_{i=1}^n)\,. \end{aligned} \tag{7}$$

This yields a one-shot selector: no iterative greedy optimization, no gradients, and no partial/full training to compute scores.

V1+SoftCCS: stratified smoothing for high pruning. At aggressive pruning rates, score-only priors can become overly concentrated and reduce coverage. Inspired by coverage-centric selection [17], SoftCCS performs a lightweight smoothing directly in probability space. We partition samples into B strata via quantiles of $\{\pi_i^{\mathrm{V1}}\}$, let $\mathcal{I}_b$ be the indices in stratum b, and define a stratum-uniform distribution $u_i := 1/|\mathcal{I}_{b(i)}|$. The mixed prior is

$$\pi_i^{\mathrm{V1+SoftCCS}} := (1-\lambda)\,\pi_i^{\mathrm{V1}} + \lambda\,u_i. \tag{8}$$

Here $\lambda \in [0,1]$. The resulting prior is then used for WRSWoR sampling via (7).

V1+DensityMix: radius-based density/coverage regularization. To mitigate redundancy in dense regions, we introduce a geometry-only regularizer derived from KNN radii. Let $r_i^{(K)} := d_{i,(K)}$ denote the distance from $\tilde{z}_i$ to its K-th neighbor; radii act as an inverse proxy for local density [8]. Optionally, we calibrate the neighborhood order to the budget k by selecting $K^\star(k)$ such that a uniformly sampled subset of size k has a target expected coverage. If $\mathcal{N}_i^{(K)}$ is the KNN set of i, then for uniform S with $|S| = k$,

$$E\Big[\mathrm{Cov}^{(K)}(S)\Big] = 1 - \frac{\binom{n-K}{k}}{\binom{n}{k}}, \tag{9}$$

and $K^\star(k)$ is the smallest K (up to a cap) achieving a desired coverage level. We then set $r_i(k) := d_{i,(K^\star(k))}$ (or use a fixed $K^\star$ if calibration is disabled).

We map radii to a smooth density score (Gaussian centered at the mean radius), normalize it to a prior $\pi^{\mathrm{dens}}(k)$, and mix with V1:

$$\pi_i^{\mathrm{V1+DMix}}(k) := (1-\eta)\,\pi_i^{\mathrm{V1}} + \eta\,\pi_i^{\mathrm{dens}}(k). \tag{10}$$

Here $\eta \in [0,1]$. Sampling again uses (7). This variant improves robustness under high pruning by discouraging overconcentration in high-density regions while preserving the target-driven hardness signal.

Computational footprint. All variants reuse a sparse KNN graph and avoid dense $O(n^2)$ similarities. Conditioned on a precomputed K_{need}-NN structure, V1 and

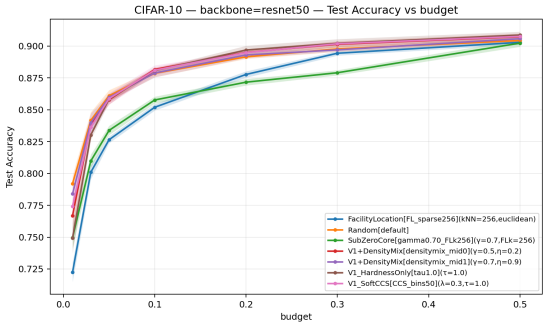

Figure 2: CIFAR-10: test accuracy vs. budget (ResNet50 features), mean over $R = 10$ runs with 95% CI.

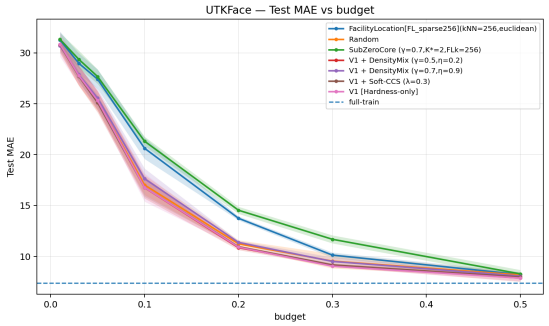

Figure 3: UTKFace (ResNet50): test MAE vs. budget, mean over $R = 10$ runs with 95% CI.

V1+DensityMix run in $O(nK + n \log k)$ time and $O(n)$ extra memory, while SoftCCS adds an $O(n \log n)$ stratification step. Overall memory is dominated by representation storage $O(nd)$ and sparse neighborhoods $O(nK_{\text{need}})$, matching GAIT's low-overhead contract.

6. Experiments

The experimental study evaluates the proposed training-free, one-shot selection family under two complementary objectives: (i) quantify the *utility–budget* trade-off on representative supervised benchmarks under GAIT-compliant constraints, and (ii) measure the *energy savings* achieved by data-centric training reduction on an end-to-end pipeline. We report some of the main results obtained during the experimental campaign.

Experimental setting. Selection is performed in a shared representation space extracted from a frozen ImageNet backbone (ResNet50 and VGG16), standardized and reduced via PCA (90% explained variance). Downstream utility is measured by training a fixed-capacity MLP head with early stopping on a held-out validation split; all non-selection hyperparameters are held constant across methods. All compared selectors satisfy the same comparability contract: *training-free*, *one-shot*, and *memory-safe* (sparse kNN graph, no dense $O(n^2)$ kernels). Baselines are **Random** sampling, **Facility Location**, and **SubZeroCore** [10]. Results are reported as mean $\pm$ 95% CI over $R = 10$ repetitions.

Feature-Space Utility Results

Regime-level interpretation. Results across both tasks reveal three consistent regimes. At *extreme scarcity* ($b \leq 0.05$), behavior is dataset-dependent: on CIFAR-10 random sampling is hardest to beat [4], while on CIFAR-100 coverage-oriented methods (SubZeroCore) lead because preserving geometric support across 100 classes is critical at very small subset sizes [17]; UTKFace already shows advantages for calibrated variants even at low budgets, consistent with a continuous regression target providing a more stable hardness signal. In the *transition regime* ($b \approx 0.10$), lightweight calibration (SoftCCS or mild DensityMix) provides its largest relative benefit by mitigating over-concentration of the hardness prior. From *moderate budgets* ($b \geq 0.20$) onward, the V1 family is consistently at or near the top across all benchmarks and backbones: the supervised hardness signal becomes the dominant driver once the subset retains broad representational support. The advantage is most pronounced on UTKFace, where geometry-only coverage objectives are poorly aligned with the continuous regression target (V1 HardnessOnly: MAE 9.04 at b=0.30 vs. 10.13 for Facility Location and 11.69 for SubZeroCore).

Energy Trade-off: End-to-End Results

Feature-space results measure predictive utility, but GAIT's primary motivation is reducing training cost. We complement the utility analysis with a dedicated CIFAR-10 end-to-end study: each method selects a subset at budget b, a ResNet50 is trained from scratch for a fixed horizon, and training-stage energy is measured

Table 2: CIFAR-10 (ResNet50) test accuracy (%) as mean ± 95% CI over $R = 10$ repetitions. Best per budget in bold. DensityMix uses γ=0.7, η=0.9. SoftCCS uses λ=0.3.

Budget	Random	FacilityLocation	SubZeroCore	V1 [hardness-only]	V1+SoftCCS	V1+DensityMix
1%	**79.21±0.62**	72.25±0.75	74.95±0.83	74.94±1.17	77.43±0.64	76.69±1.25
3%	**84.17±0.65**	80.09±0.55	80.98±0.51	83.02±0.37	83.63±0.46	83.84±0.31
5%	**86.11±0.44**	82.65±0.27	83.38±0.36	85.84±0.27	85.88±0.43	85.78±0.38
10%	87.86±0.30	85.20±0.25	85.77±0.27	88.07±0.15	88.11±0.38	**88.17±0.16**
20%	89.16±0.16	87.77±0.25	87.17±0.25	**89.69±0.33**	89.49±0.09	89.51±0.23
30%	89.78±0.22	89.44±0.18	87.91±0.20	**90.22±0.33**	90.18±0.15	90.14±0.24
50%	90.44±0.21	90.29±0.18	90.23±0.28	**90.87±0.26**	90.76±0.07	90.73±0.26

Table 3: UTKFace (ResNet50) test MAE (years; lower is better) as mean ± 95% CI over $R = 10$ repetitions. Best (lowest) per budget in bold. DensityMix uses γ=0.7, η=0.9. SoftCCS uses λ=0.3.

Budget	Random	FacilityLocation	SubZeroCore	V1 [hardness-only]	V1+SoftCCS	V1+DensityMix
1%	30.78±0.70	31.27±0.81	31.33±0.63	30.84±1.22	30.80±0.49	**30.74±0.77**
3%	27.79±0.99	28.98±0.99	29.36±0.85	27.76±0.93	**27.70±0.48**	27.75±0.97
5%	25.28±0.96	27.41±0.96	27.67±0.70	25.31±0.91	**25.11±0.72**	25.19±1.05
10%	17.02±0.97	20.63±1.04	21.32±0.45	16.77±1.38	16.77±0.90	**16.76±1.06**
20%	11.24±0.36	13.75±0.19	14.54±0.34	10.91±0.07	10.96±0.03	**10.86±0.23**
30%	9.54±0.71	10.13±0.26	11.69±0.39	**9.04±0.09**	9.16±0.17	9.14±0.13
50%	8.19±0.24	8.22±0.24	8.29±0.39	**7.88±0.30**	7.98±0.17	7.88±0.37

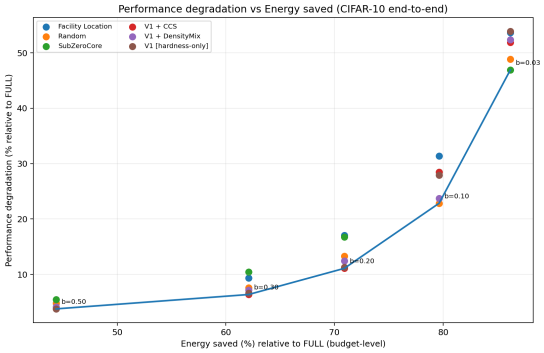

Figure 4: Performance degradation (% relative to FULL) vs. energy saved (% relative to FULL) for all methods and budgets, CIFAR-10 end-to-end (ResNet50 from scratch). Points closer to the bottom-left are more favorable. Budget levels are annotated.

via CodeCarbon. At fixed budget, training energy is approximately method-invariant (determined by subset size and training recipe), so method comparisons focus on the performance column while energy is interpreted as a function of budget.

Pareto trade-off. Figure 4 shows the performance–energy Pareto plot. The trade-off curve is strongly convex: moving from b=0.50 to b=0.20 yields large additional energy savings (≈16% → ≈52%) at a moderate performance cost (≈4% → ≈11%), while pushing to b=0.10 and below increases degradation steeply for diminishing energy returns. At the most actionable operating point b=0.20, V1 HardnessOnly incurs −10.35% degradation at 52% energy savings. Since training energy is approximately method-invariant at fixed budget, the ratio is most informative as a cross-budget summary. At b=0.20, V1 HardnessOnly reaches 0.200 pp/% energy saved (vs. 0.274 for SubZeroCore and 0.353 for Facility Location); at b=0.30, V1+SoftCCS achieves the best ratio overall (0.178). Together, these results identify $b \in [0.20, 0.30]$ as the practically relevant operating range: substantial energy savings with contained performance cost. The evaluation grid also covers extreme pruning rates ($b \leq 0.05$) to characterize selection failure modes, but the practical relevance of those regimes depends on application-specific accuracy requirements that are outside the scope of this work.

7. Conclusions and Future Work

This thesis addresses the problem of reducing the computational footprint of supervised training without modifying the downstream architecture or requiring invasive access to training dynamics. The answer is organized around two complementary contributions that together form an integrated system.

GAIT provides the framework layer: a selector-

agnostic, modular pipeline that makes training time and energy comparable across candidate subsets through hardware-calibrated profiling, baseline-referenced Δ-estimates, and conservative distributional risk signals. Its key architectural property - decoupling cost accountability from the choice of selector - makes it applicable to any training-free, one-shot method that respects the operating constraints, and ensures that cost-aware decisions are reproducible and auditable across training cycles.

The proposed selection method provides a competitive selector within the GAIT regime: a target-driven hardness prior built from local neighborhoods in a fixed representation space, with optional geometry-only regularizers for robustness at aggressive pruning, and one-shot subset construction via weighted sampling without replacement. Experimental results across CIFAR-10, CIFAR-100, and UTKFace show that the V1 family is consistently at or near the top from $b \geq 0.20$ onward, where the supervised hardness signal becomes the dominant driver of utility. End-to-end energy measurements on CIFAR-10 confirm the practical value of the approach: at b=0.20, V1 HardnessOnly achieves ≈52% energy savings with only ≈10.35% performance degradation. Interpreting the ratio $(-\Delta_{\text{perf}})/\Delta_E$ as a cross-budget efficiency summary, V1 HardnessOnly reaches 0.200 pp/% at b=0.20 and V1+SoftCCS achieves the best ratio overall (0.178) at b=0.30, confirming that the supervised hardness signal translates into a more favourable efficiency profile across the production-relevant range. The practically relevant operating range, where energy savings are substantial and performance cost remains contained, is $b \in [0.20, 0.30]$.

Open questions and future directions. Three directions follow naturally from the current work. The most immediate extension is moving beyond stationarity: in non-stationary retraining settings where covariate shift or concept drift occur, subset selection must be coupled with explicit monitoring and trigger policies, and cost reporting must extend over a horizon of repeated retrains to account for cumulative time/energy/emissions [3]. A second direction concerns the selection policy itself: the current method is a single selector, but production systems often require a *policy family* adapted per context - drift type, labeling availability, reliability constraints - including continual or online settings where subsets must be maintained incrementally over streaming data. Finally, extending GAIT to multi-modal tasks (text, time series, audio, fused inputs) requires modality-compatible representations and selection interfaces so the same workflow - subset specification, cost deltas, conservative risk - remains applicable beyond image and tabular data.

8. Bibliography

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[2] Pavlos S. Efraimidis and Paul G. Spirakis. Weighted random sampling with a reservoir. *Inf. Process. Lett.*, 97(5):181–185, 2006.

[3] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[4] Xinyu Guo, Weijie Zhao, and Yuchen Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications (DEXA)*, Lecture Notes in Computer Science, pages 181–195. Springer, 2022.

[5] Krishna Sai Abhishek Killamsetty, S. Durga, Devansh Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Gradmatch: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5464–5474. PMLR, 2021.

[6] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pages 3499–3508. PMLR, 2019.

[7] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, 11:31866–31879, 2023.

[8] Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[9] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 6950–6960. PMLR, 2020.

[10] Brian B. Moser, Tobias Christian Nauen, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Joachim Folz, and Andreas Dengel. Subzerocore: A submodular approach with zero training for coreset selection. *CoRR*, abs/2509.21748, 2025.

[11] Brian B. Moser, Arundhati S. Shanbhag, Stanislav Frolov, Federico Raue, Joachim Folz, and Andreas Dengel. A coreset selection of coreset selection literature: Introduction and recent advances. *CoRR*, abs/2505.17799, 2025.

[12] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[13] Mansheej Paul, Surya Ganguli, Gintare Karolina Dziugaite, Karthik Kashinath, and Deepak Agarwal. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 20596–20607, 2021.

[14] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Commun. ACM*, 63(12):54–63, 2020.

[15] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics, 2019.

[16] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[17] Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.