
Computational Budget Should Be Considered in Data Selection

Weilin Wan¹, Weizhong Zhang^{2*}, Cheng Jin¹

¹College of Computer Science and Artificial Intelligence, Fudan University

²School of Data Science, Fudan University

wlw23@m.fudan.edu.cn, {weizhongzhang, jc}@fudan.edu.cn

Abstract

Data selection improves computational efficiency by choosing informative subsets of training samples. However, existing methods ignore the compute budget, treating data selection and importance evaluation independently of compute budget constraints. Yet empirical studies show no algorithm can consistently outperform others (or even random selection) across varying budgets. We therefore argue that compute budget must be integral to data-selection strategies, since different budgets impose distinct requirements on data quantity, quality, and distribution for effective training. To this end, we propose a novel Computational budget-Aware Data Selection (CADS) method and naturally formulate it into a bilevel optimization framework, where the inner loop trains the model within the constraints of the computational budget on some selected subset of training data, while the outer loop optimizes data selection based on model evaluation. Our technical contributions lie in addressing two main challenges in solving this bilevel optimization problem: the expensive Hessian matrix estimation for outer-loop gradients and the computational burden of achieving inner-loop optimality during iterations. To solve the first issue, we propose a probabilistic reparameterization strategy and compute the gradient using a Hessian-free policy gradient estimator. To address the second challenge, we transform the inner optimization problem into a penalty term in the outer objective, further discovering that we only need to estimate the minimum of a one-dimensional loss to calculate the gradient, significantly improving efficiency. To accommodate different data selection granularities, we present two complementary CADS variants: an example-level version (CADS-E) offering fine-grained control and a source-level version (CADS-S) aggregating samples into source groups for scalable, efficient selection without sacrificing effectiveness. Extensive experiments show that our method achieves performance gains of up to 14.42% over baselines in vision and language benchmarks. Additionally, CADS achieves a 3-20x speedup compared to conventional bilevel implementations, with acceleration correlating positively with compute budget size.

1 Introduction

Model training costs have been increasing rapidly, and as training data accumulates over time, much of it becomes redundant. This makes data selection a crucial approach to reduce computational burden and enhance model performance [52, 59]. Numerous studies have tackled the challenge of selecting informative subsets of training samples, employing diverse criteria to identify the most impactful data for training [2]. Most existing methods fall into two categories. The first category proposes a measurement metric and selects the top-K samples based on the scores derived from this

*Corresponding Author.

metric [25, 63, 70]. Some methods further incorporate similarity measures, such as cosine similarity, to ensure diversity among the selected samples [63]. The second category aims to match certain distributions, such as gradients, embeddings, or other indicative features, to better represent the overall dataset [26, 27, 41, 69].

However, despite numerous advances in data selection methods, a critical gap remains: most approaches overlook the computational budget as a key factor shaping the selection process. Empirical studies [18, 22] reveal that optimal model performance depends on balancing training data volume, model scale, and available compute budget, highlighting that data selection is inherently linked to computational constraints. Moreover, recent studies [11, 54, 62] reveal that sophisticated selection strategies often fail to consistently outperform random selection across varied experimental settings. [67] further demonstrate that when computational budgets are explicitly factored in, previously effective data selection approaches rarely remain optimal. *We thus argue that computational budget must be integral to data selection strategies*, as it determines the appropriate quantity, quality, and distribution of training data, making it a first-order design decision rather than a fixed hyperparameter.

Our argument is further supported by neural network learning dynamics. Models rely on different granularities of knowledge at various training stages, reflected in distinct data subsets. Rahaman et al. [50] showed that networks exhibit a spectral bias, learning low-frequency features before higher-frequency ones. Under limited computational budgets, focusing on data rich in low-frequency features can optimize learning, whereas larger budgets allow leveraging diverse, higher-frequency information. We verify these effects empirically in Section 3.

In this paper, we propose a novel Computational Budget-Aware Data Selection (CADS) method, which we naturally formulate within a bilevel optimization framework. In this framework, the inner loop trains the model under the constraints of the computational budget on a selected subset of training data, while the outer loop focuses on optimizing data selection based on the evaluation of the model trained in the inner loop. Our technical contributions target two main challenges: the expensive estimation of the Hessian matrix for gradients in the outer loop and the computational burden associated with achieving optimality in the inner loop during iterations. To solve the first challenge, we introduce a probabilistic reparameterization strategy and utilize a Hessian-free policy gradient estimator to compute the gradient efficiently. For the second challenge, we reformulate the inner optimization problem as a penalty term in the outer objective function. Notably, we find that estimating the minimum of a one-dimensional loss is sufficient for calculating the gradient, significantly enhancing computational efficiency.

To meet varying data selection demands in practice, we implement CADS in two variants addressing different operational requirements: CADS-E operates at example-level granularity for precise control, while CADS-S assigns weights to source groups for improved scalability. Experiments reveal performance improvements of up to 14.42% over baselines across vision and language tasks. CADS delivers 3-20× speedup versus conventional bilevel methods, with acceleration scaling proportionally to compute budget size.

Summary of contributions:

- We highlight the crucial role of computational budget in data selection, advocating it as a first-order design factor rather than a fixed hyperparameter, thus addressing a key gap by linking data selection to computational constraints.
- We propose a novel bilevel optimization framework for Computational Budget-Aware Data Selection (CADS), where the inner loop trains the model under budget constraints on a selected data subset, and the outer loop optimizes data selection based on the trained model's evaluation.
- To efficiently solve this bilevel problem, we develop techniques including a probabilistic reparameterization for gradient estimation avoiding expensive Hessian computation, and a reformulation of the inner problem as a penalty term in the outer objective. We further show that estimating a one-dimensional loss minimum suffices for gradient calculation, greatly improving computational efficiency.
- Extensive empirical evaluations demonstrate that our CADS framework achieves superior performance across multiple datasets and training settings, validating the effectiveness of computational budget-aware data selection.

2 Related work

2.1 Data selection

Traditional data selection methods and related concepts. Data selection methods aim to identify subsets of training samples that maximize informativeness or diversity and minimize redundancy [2]. Approaches leveraging submodular optimization and clustering have shown success in various tasks [40, 51, 61]. Coreset selection techniques specifically construct small representative subsets that retain essential properties of the full dataset, often using importance weights or gradient similarity [26, 51]. Learning-based methods integrate selection into training through training dynamics analysis [44, 59], bilevel optimization [27, 74], gradient properties [21, 23], or uncertainty measures [3]. Dataset distillation methods construct synthetic examples to compress large datasets, enabling faster or more efficient training by matching gradient statistics [6, 10, 60, 68]. While highly compact, distilled sets often struggle to match real-data diversity or scale with heterogeneous corpora characteristic of large-scale model training. Curriculum learning [4] dynamically orders or selects data according to difficulty, often progressing from easy to challenging samples [17, 55]. A related body of work uses importance sampling and adaptive reweighting to prioritize data points that contribute most to model improvement or convergence [1, 4, 36], demonstrating efficiency gains especially in large-scale or imbalanced data regimes.

Data selection for LLM fine-tuning. Unlike traditional methods, data selection for LLMs often employs simpler heuristic rules due to the unprecedented scale of both models and datasets. Efficient subset selection has become crucial in LLM fine-tuning [2, 39, 42, 49]. LIMA demonstrated strong performance with just 1,000 human-curated examples [72], while automated approaches have emerged to replace manual curation. These include natural language indicators with BlendSearch [5], semantic intention tagging [37], model-based quality filtering [7, 31], instruction difficulty metrics [30], and uncertainty-driven active learning [28]. Recent advances explore gradient-based techniques, including clustered coreset selection [69] and graph methods [71], alongside simple yet effective heuristics like length-based prioritization [70]. Novel approaches leverage small model training trajectories to guide selection for larger models [64] and sparse autoencoders for diversity-driven selection [63]. DEITA provides a unified framework balancing quality, complexity, and diversity [34]. These methods achieve comparable or superior performance to full-dataset training while reducing computational requirements, demonstrating that strategically selected subsets can match or exceed naive scaling.

The compute-aware gap. Existing methods typically operate with predefined data budgets without considering computational constraints. Recent studies [11, 54, 62] reveal that sophisticated selection strategies often fail to consistently outperform random selection across varied experimental settings. [67] further demonstrate that when computational budgets are explicitly factored in, previously effective data selection approaches rarely remain optimal. This computational perspective motivates our approach to data selection that explicitly accounts for training budget constraints.

2.2 Bilevel optimization

Algorithmic advances Early work on bilevel optimization in deep learning focused on implicit differentiation and Hessian–vector products [12, 16, 45] as well as iterative differentiation with truncated back-propagation [38, 53, 73]. These methods scale poorly because they require nested loops and second-order information. Stochastic variants [9, 15, 19, 20, 24] cut per-iteration cost but still depend on Jacobian/Hessian evaluations. Finite-difference estimators [56, 65] remove explicit Hessians yet introduce instability. Most recently, fully first-order, penalty-based formulations [8, 29, 66] recast the inner optimality conditions as constraints, eliminating higher-order derivatives and enabling training with only standard gradients. This advancement enables bilevel algorithms to be applied to increasingly larger-scale problems.

Practical applications Bilevel optimization has become a fundamental tool in machine learning, initially enabling hyperparameter optimization by jointly optimizing model parameters and hyperparameters [14]. Key applications include learning-rate and weight-decay tuning [13, 35]. In neural architecture search, DARTS frames architecture parameters as outer variables and weights as inner variables, achieving efficient architecture discovery [33]. Resource-aware tasks benefit from probabilistic bilevel formulations for memory-constrained coreset selection [74], and gradient-matching

methods for learning compact synthetic datasets [60]. Extending to large-scale settings, ScaleBiO applies first-order bilevel data reweighting at the scale of billions of tokens in language modeling, demonstrating improved performance with manageable computational overhead [43]. These developments highlight bilevel optimization’s versatility in structuring complex learning problems, enabling efficient parameter tuning, architecture search, and data management, and suggest continued expansion into broader applications.

3 An exploratory experiment to illustrate the impact of compute constraints on data choice

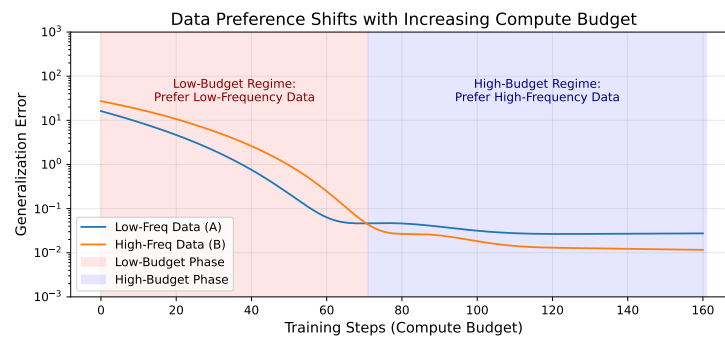


Figure 1: Validation loss of models trained on low- and high-frequency data under varying compute budgets. Two regimes emerge: low-budget favors low-frequency data, while high-budget favors high-frequency data, showing compute’s effect on data choice.

The spectral bias of neural networks—favoring learning low-frequency features before higher-frequency ones—has been well documented in prior work [50]. Their findings highlight that models naturally prioritize simpler, low-frequency information early in training. In this section, we build upon this insight to explore the relationship between computational budget and data selection more explicitly. To visualize this connection, we adopt a simple synthetic data generation scheme designed for interpretability.

Specifically, inspired by [50], data are sampled from the function $y = x + \frac{\sin(\pi x)}{\pi x}$ (where the fraction equals 1 when $x = 0$) with additive Gaussian noise. The low-frequency dataset (Model A) samples exclusively at $\sin(\pi x) = 0$ points, capturing linear patterns. The high-frequency dataset (Model B) samples uniformly across the domain, preserving spectral complexity. Both datasets contain 50,000 samples each to control for size effects, isolating differences to spectral properties alone. We employ a three-layer fully connected network with ReLU activations; architectural details appear in the appendix. Figure 1 presents validation loss trajectories across varying computational budgets. With limited compute, the low-frequency model demonstrates superior generalization. As compute increases, the high-frequency model achieves better performance, confirming the advantage of richer data when resources permit. These findings establish two distinct regimes in data preference determined by computational constraints, necessitating budget-aware data selection methods that dynamically adapt to available compute—a central focus of this work.

4 Methods

This section introduces **Compute-Aware Data Selection (CADS)**, a single objective that jointly optimizes (i) model parameters and (ii) both *how much* and *which* data to process under a fixed compute budget. We first formalise compute-constrained selection as a bilevel optimization problem and review a direct policy-gradient baseline. The remainder of the section presents two CADS variants: CADS-E, operating at the example level, and CADS-S, operating at the source level.

4.1 Compute-constrained data selection

To address the limitations of traditional data selection methods, we redefine the problem with the computational budget as an explicit constraint. Given a fixed compute limit C , we aim to identify a subset of data—along with its size—that maximizes validation accuracy while ensuring that the total number of sample-processing steps does not exceed C . This can be formalized as the following bilevel optimization problem:

$$\min_{\mathbf{m}} \mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m})) \quad \text{s.t. } \boldsymbol{\theta}_C(\mathbf{m}) = \text{Train}(\mathbf{m}, C), \quad (1)$$

where $\mathbf{m}$ is a binary selection vector indicating which samples to include in training, and $\boldsymbol{\theta}_C$ represents the model parameters derived from training with budget C using the selected subset $\mathbf{m}$. A natural approach to this problem is bilevel optimization [12, 13, 16, 35, 38, 45, 53]. To optimize $\mathbf{m}$, we compute the gradient of the validation loss with respect to $\mathbf{m}$, given by $\nabla_{\mathbf{m}} \mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m})) = \frac{\partial \mathcal{L}_{\text{val}}}{\partial \boldsymbol{\theta}_C} \cdot \frac{\partial \boldsymbol{\theta}_C}{\partial \mathbf{m}}$. Calculating $\frac{\partial \boldsymbol{\theta}_C}{\partial \mathbf{m}}$ requires assessing how model parameters vary with $\mathbf{m}$, which can be approached in two ways: implicit differentiation at the optimality condition, requiring costly second-order derivatives, or unrolling the training trajectory and backpropagating, which incurs even higher computational costs. Given our emphasis on efficiency within compute constraints, both methods are impractical, especially because the assumption $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{inner}}(\boldsymbol{\theta}, \mathbf{m}) = 0$ may not hold. The limited budget C prevents reaching a true local minimum, leaving the inner gradient non-zero and making implicit differentiation unreliable. Inspired by [74], we propose a learnable sampling distribution parameterized by $\mathbf{s}$ and optimize $\mathbf{s}$ using policy gradients instead of directly optimizing $\mathbf{m}$. This approach allows us to navigate the bilevel problem without relying on gradients from the inner optimization under compute constraints. In this formulation, we define $\mathbf{m} \in \{0, 1\}^N$ as a binary mask selecting a subset from the training corpus $\mathcal{D}$, with $p(\mathbf{m} | \mathbf{s})$ representing the corresponding sampling distribution. Given a strict computational budget of C forward-backward passes, our objective is to identify the sampler that minimizes the expected validation loss after training for exactly C steps on the masked data:

$$\min_{\mathbf{s}} \mathbb{E}_{\mathbf{m} \sim p(\mathbf{m} | \mathbf{s})} [\mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m}))] \quad \text{s.t. } \boldsymbol{\theta}_C(\mathbf{m}) = \text{Train}(\mathbf{m}, C), \quad (2)$$

where $\boldsymbol{\theta}_C(\mathbf{m})$ represents the model parameters obtained by training on the selected subset $\mathbf{m}$ with the fixed compute budget C . Unlike classical bilevel formulations, where the inner problem is solved to convergence, this formulation explicitly constrains training to C steps, closely matching practical real-world training scenarios.

4.2 Bilevel-CADS: policy-gradient baseline

A direct yet costly way to optimize Equation (2) is to treat the selector as a stochastic policy and apply REINFORCE. The sampler is factorised over examples:

$$p(\mathbf{m} | \mathbf{s}) = \prod_{i=1}^N s_i^{m_i} (1 - s_i)^{1-m_i} \quad (3)$$

where $s_i \in [0, 1]$ is the inclusion probability for example i . The policy gradient of the outer objective is then

$$\nabla_{\mathbf{s}} \mathbb{E}_{p(\mathbf{m} | \mathbf{s})} [\mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m}))] = \mathbb{E}_{p(\mathbf{m} | \mathbf{s})} [\mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m})) \nabla_{\mathbf{s}} \log p(\mathbf{m} | \mathbf{s})] \quad (4)$$

In practice we approximate the expectation with K Monte-Carlo samples:

$$\nabla_{\mathbf{s}} \mathbb{E}_{p(\mathbf{m} | \mathbf{s})} [\mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m}))] \approx \frac{1}{K} \sum_{k=1}^K \mathcal{L}_{\text{val}}(\boldsymbol{\theta}_C(\mathbf{m}_k)) \nabla_{\mathbf{s}} \log p(\mathbf{m}_k | \mathbf{s}), \quad \mathbf{m}_k \sim p(\mathbf{m} | \mathbf{s}) \quad (5)$$

Each gradient update therefore entails training K independent models for C compute units each, yielding a total cost of $K \times C$ per outer step. With even modest values (e.g. $K=5$), this baseline quickly becomes prohibitive and motivates the more efficient relaxations introduced in Sections 4.3.

4.3 Penalty-based single-level relaxation

To address the prohibitive computational cost of the policy gradient approach, inspired by [29, 43], we reformulate the bilevel problem as a single-level optimization through a penalty-based relaxation.

Instead of repeatedly training models to exhaustion for each sampled mask, we introduce an auxiliary variable $\mathbf{u}$ that acts as a proxy for the fully-trained parameters θ_C and penalise the discrepancy between the training losses of θ and $\mathbf{u}$:

$$\min_{\mathbf{s}, \theta} \mathcal{L}_{\text{penalty}}^{\alpha}(\theta, \mathbf{s}) \triangleq \mathbb{E}_{p(\mathbf{m}|\mathbf{s})} \left[\mathcal{L}_{\text{val}}(\theta) + \alpha (\mathcal{L}_{\text{trn}}(\theta, \mathbf{m}) - \mathcal{L}_{\text{trn}}(\mathbf{u}, \mathbf{m}))^2 \right] \quad (6)$$

with

$$\mathbf{u} = \text{Train}(\theta_0, \mathbf{m}; K), \quad K = \frac{C}{|\mathbf{m}|}.$$

In this formulation, $\mathcal{L}_{\text{trn}}(\mathbf{u}, \mathbf{m})$ serves as a target training loss that reflects the performance of a model trained for a limited number of steps K under the given compute budget C .

Technical difficulty in solving problem (6). Unlike classical bilevel settings where the inner problem reaches (or approximates) optimality, here the target loss corresponds to an intermediate, compute-constrained solution that generally does not satisfy stationary conditions (i.e., non-zero gradients). This makes direct optimization challenging: naively updating the auxiliary variable $\mathbf{u}$ jointly with θ may violate the compute budget constraint or require costly iterative inner optimization. To avoid the need for explicitly unrolling K training steps for each candidate $\mathbf{m}$, we approximate $\mathcal{L}_{\text{trn}}(\mathbf{u}, \mathbf{m})$ with a scale-dependent surrogate $l(|\mathbf{m}|)$, which can be efficiently estimated. In practice, $l(|\mathbf{m}|)$ is found to be well-approximated by a log-linear function of the subset size $|\mathbf{m}|$, capturing how the achievable training loss scales with computational budget. We fit this log-space interpolation using training loss measurements collected at multiple subset sizes, balancing data efficiency and approximation quality. A detailed discussion of the surrogate $l(\cdot)$ is provided in the appendix. Subsequently, substituting the approximation yields the single-level objective:

$$\mathcal{L}_{\text{CADS}}^{\alpha}(\theta, \mathbf{s}) = \mathbb{E}_{p(\mathbf{m}|\mathbf{s})} \left[\mathcal{L}_{\text{val}}(\theta) + \alpha (\mathcal{L}_{\text{trn}}(\theta, \mathbf{m}) - l(|\mathbf{m}|))^2 \right]. \quad (7)$$

Finally, by approximating the penalty objective $\mathcal{L}_{\text{penalty}}^{\alpha}$ (6) with the CADS objective $\mathcal{L}_{\text{CADS}}^{\alpha}$ (7), we can efficiently solve the budget-constrained minimization in Problem (6).

Optimization. Stochastic estimates of the following gradients are obtained with a single forward/backward pass per sampled subset:

$$\nabla_{\theta} \mathcal{L}_{\text{CADS}}^{\alpha}(\theta, \mathbf{s}) = \mathbb{E}_{p(\mathbf{m}|\mathbf{s})} \left[\nabla_{\theta} \mathcal{L}_{\text{val}} + 2\alpha (\mathcal{L}_{\text{trn}} - l(|\mathbf{m}|)) \nabla_{\theta} \mathcal{L}_{\text{trn}} \right], \quad (8)$$

$$\nabla_{\mathbf{s}} \mathcal{L}_{\text{CADS}}^{\alpha}(\theta, \mathbf{s}) = \mathbb{E}_{p(\mathbf{m}|\mathbf{s})} \left[(\mathcal{L}_{\text{val}} + \alpha (\mathcal{L}_{\text{trn}} - l(|\mathbf{m}|))^2) \nabla_{\mathbf{s}} \log p(\mathbf{m} | \mathbf{s}) \right]. \quad (9)$$

Any first-order optimizer (e.g. Adam) can then update $(\theta, \mathbf{s})$ jointly.

4.4 CADS-E: example-level selection

In the example-level approach, CADS learns to select individual examples by assigning a parameter $s_i \in [0, 1]$ to each example i in the training corpus. We model the sampling distribution $p(\mathbf{m}|\mathbf{s})$ as independent Bernoulli variables for each example as the same distribution defined in (3). The gradient of $\log p(\mathbf{m}|\mathbf{s})$ with respect to s_i is straightforward:

$$\frac{\partial}{\partial s_i} \log p(\mathbf{m}|\mathbf{s}) = \frac{m_i}{s_i} - \frac{1 - m_i}{1 - s_i}. \quad (10)$$

During optimization, each subset m_i represents a binary mask over the training examples. Algorithm 1 provides the pseudocode implementation.

4.5 CADS-S: source-level selection

In the source-level variant, instead of parameterizing selection at the level of individual samples, we aggregates samples into broader source groups and assigns weights at this higher level. Specially, the dataset is partitioned into n sources $\{\mathcal{D}_i\}_{i=1}^n$, and CADS learns a sampling ratio vector $\mathbf{r} \in [0, 1]^n$. We let $p(\mathbf{r} | \mathbf{s})$ be an independent truncated Gaussian for each dimension:

$$p(\mathbf{r} | \mathbf{s}) = \prod_{j=1}^n \frac{\phi((r^j - s^j)/\sigma)}{\sigma [\Phi((1 - s^j)/\sigma) - \Phi((-s^j)/\sigma)]}, \quad (11)$$

Algorithm 1 CADS-E (example-level)

Require: budget C , initial θ_0, s_0 , subset count K

- 1: **while** not converged **do**
- 2: Sample K subsets $\{m_k\} \sim p(m | s)$
- 3: **for** each m_k **do**
- 4: Compute $\mathcal{L}_{\text{trn}}(\theta, m_k)$
- 5: Compute $\mathcal{L}_{\text{val}}(\theta)$
- 6: Compute approximate loss $l(|m_k|)$
- 7: **end for**
- 8: Estimate $\mathcal{L}_{\text{CADS}}$ using (7)
- 9: Compute gradients $\nabla_{\theta} \mathcal{L}$ and $\nabla_s \mathcal{L}$
- 10: Update $\theta \leftarrow \theta - \eta_{\theta} \nabla_{\theta} \mathcal{L}$
- 11: Update $s \leftarrow s - \eta_s \nabla_s \mathcal{L}$
- 12: Update $s \leftarrow s - \eta_s \nabla_s \mathcal{L}$
- 13: Update $s \leftarrow s - \eta_s \nabla_s \mathcal{L}$
- 14: **end while**
- 15: **return** s^*, θ^*

Algorithm 2 CADS-S (source-level)

Require: budget C , sources $\{\mathcal{D}_i\}$, initial θ_0, s_0 , subset count K , variance σ

- 1: **while** not converged **do**
- 2: Sample K ratio vectors $\{r_k\} \sim p(r | s)$
- 3: **for** each r_k **do**
- 4: Form $\mathcal{D}_{r_k}$ by subsampling sources
- 5: Compute $\mathcal{L}_{\text{trn}}(\theta, \mathcal{D}_{r_k})$
- 6: Compute $\mathcal{L}_{\text{val}}(\theta)$
- 7: Compute approximate loss $l(|\mathcal{D}_{r_k}|)$
- 8: **end for**
- 9: Estimate $\mathcal{L}_{\text{CADS}}$ using (7)
- 10: Compute gradients $\nabla_{\theta} \mathcal{L}$ and $\nabla_s \mathcal{L}$
- 11: Update $\theta \leftarrow \theta - \eta_{\theta} \nabla_{\theta} \mathcal{L}$
- 12: Update $s \leftarrow s - \eta_s \nabla_s \mathcal{L}$
- 13: Update variance $\sigma \leftarrow 0.99 \sigma$
- 14: **end while**
- 15: **return** s^*, θ^*

where ϕ and Φ denote the standard normal PDF and CDF. This is a truncated Gaussian centered at s^j with scale σ . We truncate to $[0, 1]$ to obtain valid sampling values. The denominator normalizes the distribution, ensuring the integral equals 1 after truncation. To stabilize optimization, we anneal σ using $\sigma_t = 0.99^t \sigma_0$. This gradually concentrates probability mass around s^j , leading to consistent sampling results. We visualize the distribution under different s and σ values in the appendix. Algorithm 2 provides pseudo-code of CADS-S.

5 Experiments

In this section, we present a series of experiments aimed at rigorously evaluating the performance and robustness of our proposed methods across diverse settings and conditions. Detailed experimental settings and implementation details are provided in the appendix.

5.1 Small-scale validation

Our goal is to verify that, under a fixed compute budget of 20,000 sample usages (20 full-epoch trainings on a 1,000-sample MNIST subset), Bilevel-CADS and CADS-E can identify near-optimal training subsets. Specifically, we evaluate four approaches across four initial subset sizes spanning from small to large fractions of the total 1,000 samples (200 to 800 samples); experiments under other compute budgets are presented in the appendix. We compare:

- **Random selection:** directly train and test using randomly chosen subsets at these sizes.
- **PBCS [74]:** selects optimal subsets matching the target sizes following original paper settings, then trains and evaluates models.
- **Bilevel-CADS and CADS-E:** start with initial subsets of similar sizes, then execute their respective algorithms to refine and obtain final subsets for model training.

As shown in Table 1, Bilevel-CADS consistently achieves the highest accuracy across all initial subset sizes, significantly outperforming random selection and other methods. Meanwhile, Figure 2 visualizes how the Bilevel-CADS optimization process converges to a stable subset size near 500 samples regardless of the initial subset size, indicating robustness and effective subset refinement driven by the compute budget constraint.

5.2 Performance with heterogeneous data sources

Small-scale experiments on CIFAR-10. To validate the effectiveness of CADS-S, we conducted experiments on the CIFAR-10 dataset. Specifically, we reserved 10% of the training set (5,000

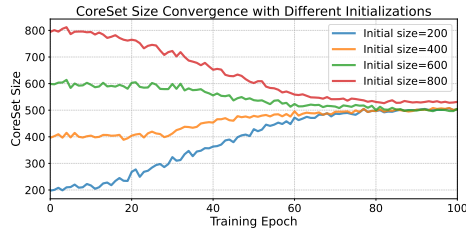


Figure 2: Subset size evolution during optimization with Bilevel-CADS starting from different initial subset sizes.

Table 1: Accuracy (%) comparison of different data selection methods with various initial subset sizes on MNIST.

Method	Initial subset size				Average
	200	400	600	800	
Random	88.05	90.62	90.75	89.91	89.83
PBCS	89.81	92.14	92.64	92.98	91.89
CADS-E	91.48	92.28	92.34	92.90	92.25
Bilevel-CADS	92.44	92.90	92.85	93.09	92.80

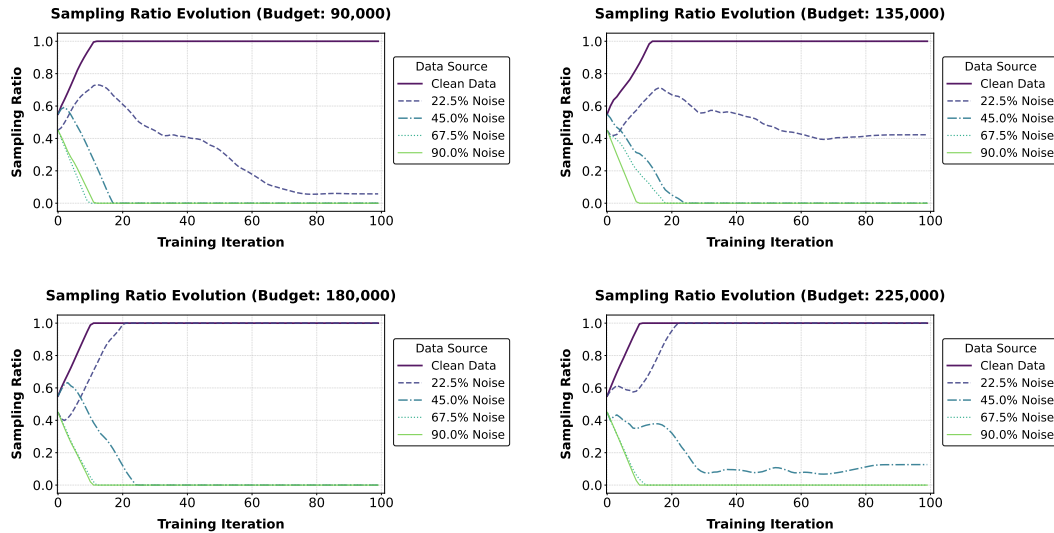


Figure 3: Evolution of the sampling ratio over training iterations under different compute budgets: 90,000, 135,000, 180,000, and 225,000 samples respectively.

samples) as a validation set and partitioned the remaining 45,000 samples into five equal groups of 9,000 each, simulating heterogeneous data sources with noise levels of 0%, 22.5%, 45.0%, 67.5%, and 90.0%, respectively. Our objective was to assess how CADS-S performs compared to training on the single best data source, and the full dataset. The compute budget was varied between 90,000 and 225,000 sample usages.

Key Insights from Fig. 3: It clearly illustrates that, under lower compute budgets, our algorithm predominantly favors clean data, assigning near-zero weights to the noisy data sources. However, with an increase in the compute budget, the algorithm gradually elevates the weights of data sources with lower noise levels.

At a compute budget of 225,000 samples, the algorithm incorporates both the clean data and data from the source with 22.5% noise, while also selecting a negligible amount (<0.1) from the source with 45% noise. This phenomenon highlights that as the compute budget expands, even data containing some noise can retain significant value in the training process.

In stark contrast, data sources with noise levels greater than 60% are consistently excluded from selection, regardless of the compute budget in play. These findings suggest that as the compute budget grows, the algorithm becomes increasingly adept at leveraging noisy data, emphasizing the importance of diverse data sources in enhancing model performance.

Scaling to larger datasets To further evaluate the scalability and effectiveness of our approach, we conduct experiments on the DomainNet dataset [47], which contains over 0.5 million images across six diverse visual domains: Real, Sketch, Clipart, Painting, Infograph, and Quickdraw. Our setup

Table 2: Accuracy (%) comparison of different data selection methods with various compute budget.

Method	Compute budget				Average
	90,000	135,000	180,000	225,000	
Best-source	56.31	61.59	63.72	65.08	61.65
Full-dataset	44.34	49.80	53.34	57.03	51.13
CADS-S	57.05	62.07	65.96	67.22	63.08

Table 3: Accuracy (%) comparison of different data selection methods with various initial subset sizes on DomainNet.

Method	Initial sampling ratio (%)				Average
	20	40	60	80	
Uniform-sampling	29.73	37.48	39.70	37.96	36.22
CADS-S	44.15	44.22	44.23	44.49	42.27

uses all domains as data sources; however, we treat the Real domain as a high-quality data source and only include a subset of 10,000 samples from it to simulate limited access to the most reliable data. The other five domains are used in full. The total compute budget allocated for training corresponds approximately to training on the full dataset for 10 epochs. Table 3 reports the test accuracy of different data selection methods under varying initial sampling ratios. Our proposed CADS-S method consistently improves over uniform sampling across all ratios. Figure 4 illustrates the evolution of sampling ratios during the optimization process, demonstrating how CADS-S dynamically adjusts data source importance over epochs to better allocate the computational budget.

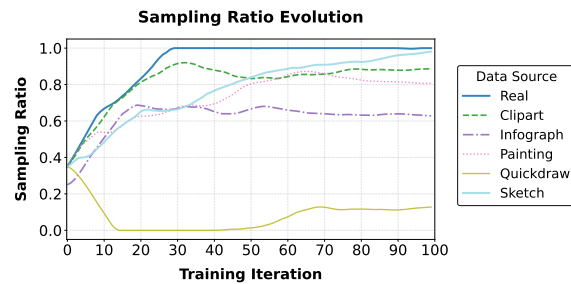


Figure 4: Sampling ratio evolution on DomainNet

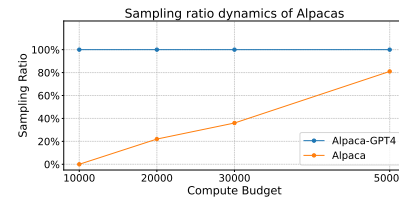


Figure 5: Sampling ratio dynamics of Alpaca-GPT4 and Alpaca datasets under different compute budgets optimized by CADS-S.

5.3 Performance on instruction tuning tasks

Small-scale experiments. We evaluated our CADS-S approach on instruction fine-tuning tasks using the GPT-2 model [48]. The dataset consists of two heterogeneous data sources: (i) **Alpaca-GPT4** [46], a high-quality dataset from GPT-4-generated instructions, from which we sample 1,000 examples to simulate the rarity of premium data sources in practical scenarios; (ii) **Alpaca** [57], a larger dataset containing 9,000 examples, representing standard quality data. The total compute budget varies within the range $[1 \times 10^4, 5 \times 10^4]$ sample usages. Other settings remain consistent with those described in Section 5.2. As shown in Tab 4, CADS-S consistently outperforms baseline methods, across all compute budgets studied. This validates the proposed method’s ability to allocate computational resources between heterogeneous data sources to optimize model performance.

Table 4: Perplexity comparison of different data selection methods with various compute budget.

Method	Compute budget			
	1e+4	2e+4	3e+4	5e+4
Best-source	8.01	8.02	8.21	8.87
Full-dataset	8.26	8.01	7.90	7.77
CADS-S	8.01	7.92	7.85	7.77

Table 5: Perplexity comparison of different data selection methods with various initial subset sizes on additional datasets.

Method	Initial sampling ratio (%)				Average
	20	40	60	80	
Uniform-sampling	7.90	7.52	7.43	7.39	7.56
CADS-S	6.97	6.95	6.97	6.90	6.95

Scaling to additional datasets. We further evaluated CADS across 13 instruction-following fine-tuning datasets, sampling 10,000 examples from each, including AlpacaGPT4 [46], SlimOrca [32], Alpaca [57], GPTeacher [58], and multilingual Alpaca variants, totaling around 130,000 samples.

Detailed dataset setups and preprocessing are provided in the appendix. As shown in Tab 5, our method consistently outperforms baseline approaches in perplexity across different initial sampling ratios, demonstrating its robustness and wide applicability.

5.4 Computational Cost Analysis of Established Baselines

To provide a clear comparison of computational efficiency, we establish a unified framework parameterized by:

- **C**: The total compute budget required for a single, standard training run on the full dataset. This serves as our baseline unit of cost.
- **N**: The total number of samples in the full dataset.
- **T**: The number of training epochs, defined as $T = C/N$.

All computational overheads are measured relative to the baseline cost **C** of training on the full dataset.

A Note on Pre-training Cost: For methods that require a pre-trained model to compute sample-wise statistics (e.g., gradients, forgetting events), we assume the cost of obtaining this model is $0.2C$. This represents a standard approximation for the early-stage training necessary to yield meaningful statistics, without requiring full convergence.

Table 6: Computational Cost Analysis of Established Baselines.

Method	Total Computational Overhead
Random Selection	C (no additional cost)
Forgetting Scores	1.2C (0.2C for pre-training and forgetting computation + C for final training)
GraNd	1.2C (0.2C for pre-training and gradient computation + C for final training)
EL2N	1.2C (0.2C for pre-training and error computation + C for final training)
CADS (Ours)	$(5/A + 2\gamma)C$ where A is amortization factor
Influence Functions	$2C + O(N^2)$ (full training + expensive Hessian computations + final training)
Data Shapley	11C–101C (C for final training + 10C–100C for Monte Carlo approximation)
Probabilistic Bilevel Optimization	50C–100C (requires 50–100 outer iterations)
Greedy Coreset	$\sim (N^2/T + 1)C$ where K is coreset size

Summary: CADS achieves superior efficiency among bilevel methods with overhead $(5/A + 2\gamma)C$. With $A=5$ amortization, total cost reduces to $2\gamma C$, $\gamma < 1$. Traditional methods like Forgetting Scores offer moderate 1.2C cost but lack budget-aware optimization and deliver inferior effectiveness. Advanced bilevel methods become prohibitively expensive (50C+), while CADS provides an optimal balance between computational efficiency and budget-aware capabilities.

Note: The core contribution of our work is introducing computational budget constraints into coreset selection, rather than optimizing selection time alone. Once learned, a coreset can be reused multiple times across different training scenarios, model architectures, and experiments. Therefore, the computational investment in coreset selection can be amortized over multiple uses, making the selection efficiency a secondary consideration compared to the quality of the resulting coreset under budget constraints.

6 Conclusion

We introduced CADS, a compute-aware data-selection framework that casts subset choice as bilevel optimization under explicit budget constraints. By combining a probabilistic reparameterization with a Hessian-free policy-gradient estimator and a penalty-based surrogate that reduces inner-loop optimization to a one-dimensional loss, CADS efficiently balances data quantity, quality, and distribution. Our example-level (CADS-E) and source-level (CADS-S) variants deliver up to 14.42% accuracy gains and 3–20× speedups over fixed-budget baselines on vision and language benchmarks. Future work will extend CADS to jointly adapt model size, dataset scale, and compute budget for fully resource-adaptive training.

Acknowledgments and Disclosure of Funding

This work was supported by High-Quality Development Project of Shanghai Municipal Commission of Economy and Informatization (Grant No. 2024-GZL-RGZN-02010), AI for Science Foundation of Fudan University (FudanX24AI028), and the National Nature Science Foundation of China (62472097).

References

- [1] G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.
- [2] A. Albalak, Y. Elazar, S. M. Xie, S. Longpre, N. Lambert, X. Wang, N. Muennighoff, B. Hou, L. Pan, H. Jeong, et al. A survey on data selection for language models, 2024. URL <https://arxiv.org/abs/2402.16827>, 2022.
- [3] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *ArXiv*, abs/1906.03671, 2019.
- [4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- [5] Y. Cao, Y. Kang, and L. Sun. Instruction mining: High-quality instruction data selection for large language models. *ArXiv*, abs/2307.06290, 2023.
- [6] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10708–10717, 2022.
- [7] L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, et al. Alpargasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.
- [8] L. Chen, Y. Ma, and J. Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. 2023.
- [9] T. Chen, Y. Sun, Q.-W. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- [10] Z. Deng and O. Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. *ArXiv*, abs/2206.02916, 2022.
- [11] H. Diddee and D. Ippolito. Chasing random: Instruction selection strategies fail to generalize. *ArXiv*, abs/2410.15225, 2024.
- [12] J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.
- [13] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International conference on machine learning*, pages 1165–1173. PMLR, 2017.
- [14] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, 2018.
- [15] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv: Optimization and Control*, 2018.
- [16] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.
- [17] G. Hachohen and D. Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019.
- [18] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [19] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *ArXiv*, abs/2007.05170, 2020.

- [20] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, 2020.
- [21] T. B. Johnson and C. Guestrin. Training deep models faster with robust, approximate importance sampling. In *Neural Information Processing Systems*, 2018.
- [22] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [23] A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, 2018.
- [24] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Neural Information Processing Systems*, 2021.
- [25] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *ArXiv*, abs/2005.00700, 2020.
- [26] K. Killamsetty, G. Ramakrishnan, A. De, and R. Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021.
- [27] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, R. I. U. of Texas at Dallas, I. I. of Technology Bombay Institution One, and I. Two. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI Conference on Artificial Intelligence*, 2020.
- [28] P.-N. Kung, F. Yin, D. Wu, K.-W. Chang, and N. Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2023.
- [29] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.
- [30] M. Li, Y. Zhang, Z. Li, J. Chen, L. Chen, N. Cheng, J. Wang, T. Zhou, and J. Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *North American Chapter of the Association for Computational Linguistics*, 2023.
- [31] W. Lian et al. Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023.
- [32] W. Lian, G. Wang, B. Goodson, E. Pentland, A. Cook, C. Vong, and "Teknum". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023.
- [33] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. *ArXiv*, abs/1806.09055, 2018.
- [34] W. Liu, W. Zeng, K. He, Y. Jiang, and J. He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *ArXiv*, abs/2312.15685, 2023.
- [35] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020.
- [36] I. Loshchilov and F. Hutter. Online batch selection for faster training of neural networks. *ArXiv*, abs/1511.06343, 2015.
- [37] K. Lu, H. Yuan, Z. Yuan, R. Lin, J. Lin, C. Tan, and C. Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. *ArXiv*, abs/2308.07074, 2023.
- [38] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.
- [39] M. Marone and B. V. Durme. Data portraits: Recording foundation model training data. *ArXiv*, abs/2303.03919, 2023.
- [40] B. Mirzasoileiman, A. Badanidiyuru, and A. Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *International Conference on Machine Learning*, pages 1358–1367. PMLR, 2016.
- [41] B. Mirzasoileiman, J. A. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 2019.

- [42] Y. Oren, S. Sagawa, T. B. Hashimoto, and P. Liang. Distributionally robust language modeling. *ArXiv*, abs/1909.02060, 2019.
- [43] R. Pan, J. Zhang, X. Pan, R. Pi, X. Wang, and T. Zhang. Scalebio: Scalable bilevel optimization for llm data reweighting. *arXiv preprint arXiv:2406.19976*, 2024.
- [44] M. Paul, S. Ganguli, and G. K. Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.
- [45] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.
- [46] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *ArXiv*, abs/2304.03277, 2023.
- [47] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.
- [48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.
- [49] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. F. J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. M. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. K. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d’Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. G. Johnson, B. A. Hechtman, L. Weidinger, I. Gabriel, W. S. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. W. Ayoub, J. Stanway, L. L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021.
- [50] N. Rahaman, A. Baratin, D. Arpit, F. Dräxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. C. Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2018.
- [51] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.
- [52] B. Settles. Active learning literature survey. 2009.
- [53] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.
- [54] M. Shen. Rethinking data selection for supervised fine-tuning. *ArXiv*, abs/2402.06094, 2024.
- [55] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526 – 1565, 2021.
- [56] D. Sow, K. Ji, and Y. Liang. On the convergence theory for hessian-free bilevel algorithms. In *Neural Information Processing Systems*, 2021.
- [57] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [58] Teknium. Gpteacher: General instruction dataset. <https://huggingface.co/datasets/teknium/GPTeacher-General-Instruct>, 2023. Accessed: 2025-05-07.
- [59] M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon. An empirical study of example forgetting during deep neural network learning. *ArXiv*, abs/1812.05159, 2018.
- [60] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.
- [61] K. Wei, R. Iyer, and J. Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR, 2015.

- [62] T. Xia, B. Yu, K. Dang, A. Yang, Y. Wu, Y. Tian, Y. Chang, and J. Lin. Rethinking data selection at scale: Random selection is almost all you need. *ArXiv*, abs/2410.09335, 2024.
- [63] X. Yang, S. Nie, L. Liu, S. Gururangan, U. Karn, R. Hou, M. Khabsa, and Y. Mao. Diversity-driven data selection for language model tuning through sparse autoencoder. *ArXiv*, abs/2502.14050, 2025.
- [64] Y. Yang, S. Mishra, J. N. Chiang, and B. Mirzasoleiman. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *ArXiv*, abs/2403.07384, 2024.
- [65] Y. Yang, P. Xiao, and K. Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *ArXiv*, abs/2312.03807, 2023.
- [66] M. Ye, B. Liu, S. J. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *ArXiv*, abs/2209.08709, 2022.
- [67] J. O. Yin and A. M. Rush. Compute-constrained data selection. *ArXiv*, abs/2410.16208, 2024.
- [68] R. Yu, S. Liu, and X. Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:150–170, 2023.
- [69] J. Zhang, Y. Qin, R. Pi, W. Zhang, R. Pan, and T. Zhang. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data. *ArXiv*, abs/2407.15235, 2024.
- [70] H. Zhao, M. Andriushchenko, F. Croce, and N. Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *ArXiv*, abs/2402.04833, 2024.
- [71] Y. Zhao, L. Du, X. Ding, Y. Ouyang, H. Wang, K. Xiong, J.-F. Gao, Z. Sun, D. Xu, Y. Qing, D. Li, B. Qin, and T. Liu. Beyond similarity: A gradient-based graph method for instruction tuning data selection. *ArXiv*, abs/2502.11062, 2025.
- [72] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206, 2023.
- [73] X. Zhou, Y. Lin, R. Pi, W. Zhang, R. Xu, P. Cui, and T. Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*, pages 27203–27221. PMLR, 2022.
- [74] X. Zhou, R. Pi, W. Zhang, Y. Lin, Z. Chen, and T. Zhang. Probabilistic bilevel coreset selection. In *International conference on machine learning*, pages 27287–27302. PMLR, 2022.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the main claims of the paper, including the key contributions and scope of the work.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have included a separate "Limitations" section in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The paper clearly states all assumptions for each theoretical result and provides complete and correct proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper provides detailed descriptions of the experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: we provide detailed instructions for accessing the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides detailed training and test configurations in the Appendix to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All small-scale experiments have considered statistical significance. Due to computational resource limitations, large-scale experiments could not provide full statistical significance analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed information on the computing resources used for each experiment in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research fully adheres to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discuss both potential positive societal impacts and negative societal impacts of the work in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve the release of models or datasets that pose a high risk of misuse or dual-use. Therefore, no specific safeguards were necessary for responsible release.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We provide detailed information in the appendix about all existing assets used in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are introduced in the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects; therefore, no participant instructions, screenshots, or compensation details are applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects or crowdsourcing, so there were no risks to participants and no Institutional Review Board (IRB) approvals were required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methods developed in this research do not involve the use of large language models (LLMs) as important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Supplemental Material: Computational Budget Should Be Considered in Data Selection

In this appendix, we provide comprehensive additional materials to supplement the main text. The contents include:

- **Broader impacts (Section A):** A discussion on the broader implications of our research.
- **Experimental setting and implementation details (Section B):** Detailed information on the experimental settings and implementation details.
- **Log-linear surrogate of compute-constrained training loss (Section C):** Explanation of approximating $\mathcal{L}_{\text{trn}}(\mathbf{u}, \mathbf{m})$ by the scale-dependent surrogate $l(|\mathbf{m}|)$, its log-space interpolation, and fitting procedure.
- **Visualizations on truncated Gaussian distribution (Section D):** Detailed visualization on the truncated Gaussian distribution.
- **Additional experimental results (Section E):** MNIST experiments varying the total compute budget to compare random, PBCS, CADS-E, and Bilevel-CADS. The results show that our method consistently achieves the highest accuracy (Table 9).
- **Time efficiency analysis (Section F):** Benchmarking of average training and sampling runtimes on representative hardware.
- **Scalability to other bilevel optimization problems (Section G):** Discussion on the applicability conditions, loss estimation accuracy, and a validation framework for applying CADS to new domains.
- **Limitations and future work (Section H):** Analysis of the limitations in our current framework, and plans for future improvements.
- **Licenses for existing assets (Section I):** Acknowledgment and respect for the licenses and terms of use of datasets and code libraries utilized in our research.

A Broader impacts

This work introduces CADS, a compute-aware data selection algorithm that dynamically adapts the training data budget according to available computational resources. By optimizing data efficiency relative to compute constraints, CADS significantly reduces the computational overhead of training deep learning models without sacrificing performance. This advancement enables more accessible and environmentally sustainable machine learning research and deployment, particularly in scenarios constrained by limited hardware resources. The method's potential to lower energy consumption during model training aligns with broader community efforts towards greener AI, contributing to reductions in carbon footprint and operational costs. Additionally, CADS facilitates democratization of deep learning by empowering researchers and practitioners with modest resources to train competitive models. However, as with any automation that accelerates model training, there exist potential risks, including the possibility of speeding up the development of models with harmful biases, privacy vulnerabilities, or malicious intent if not applied responsibly. We stress the importance of ethical use and compliance with community norms and regulatory standards when deploying such techniques. Overall, CADS represents a step forward in efficient and responsible machine learning practices, promoting sustainability and equitable access within the deep learning community.

B Experimental setting and implementation details

In this section, we detail the protocols and implementation specifics underlying all our evaluations. All experiments were run on a single machine equipped with an NVIDIA A100 80 GB GPU under CUDA 12.6 and NVIDIA driver 470.199.02, using PyTorch 2.5.1.

B.1 Exploratory experiment on spectrum bias

To isolate the effect of spectral content on generalization, we generate two synthetic datasets with identical noise level $\sigma = 0.1$ but different frequency composition.

Data generation. We sample $N = 50,000$ inputs for each group; for the low-frequency dataset (Group A) we set $y = x + \mathcal{N}(0, \sigma^2)$, and for the high-frequency dataset (Group B) we set $y = x + \frac{\sin(\pi x)}{\pi x} + \mathcal{N}(0, \sigma^2)$. A fixed validation set of 10,000 points is used to track generalization error.

Model and training. Each model is a three-layer MLP (100–100 hidden units, ReLU). We train for 160 steps using Adam (learning rate 3×10^{-4}) and batch size 1000. After every parameter update, we compute the MSE on the validation set and log the result. The different positions on the x-axis in Figure 1 correspond to these intermediate results recorded during the training process.

B.2 Small-scale validation

We validate CADS on a reduced-scale MNIST classification task using a simple convolutional network. All experiments use a fixed training set of 1,000 examples and batch size 1,000. We optimize under an epoch-based budget of 20 epochs. The selection parameters $\mathbf{s} \in \mathbb{R}^{|D|}$ are initialized as $\mathbf{s} = v \cdot \mathbf{1}$, $v \in \{0.2, 0.4, 0.6, 0.8\}$, $\mathbf{1} \in \mathbb{R}^{|D|}$ is the all-ones vector. We solve the bilevel problem with Adam for both the network parameters θ (learning rate 5×10^{-3}) and the selection weights $\mathbf{s}$ (learning rate 5×10^{-2}). The outer loop runs for 300 iterations with variance reduction and gradient clipping enabled.

Network architecture. We employ a simple convolutional model: two convolutional layers with 5×5 kernels and channel counts $\{32, 64\}$, each followed by ReLU and optional 2×2 max-pooling; the feature map is flattened ($4 \times 4 \times 64 = 1024$ units) and passed through a fully connected layer of 128 units with ReLU; a final linear layer outputs class logits. Input normalization by fixed mean and standard deviation is applied when enabled. No dropout is used.

B.3 Heterogeneous data sources: small-scale experiments on CIFAR-10

We evaluate CADS-S on a grouped CIFAR-10 variant with five demographic groups and up to 90% label noise. All experiments use the full training set (no limit) with batch size 256. We allocate an epoch-based compute budget of 2 to 5 epochs. The selection parameter $\mathbf{s}$ is initialized to $0.5 \cdot \mathbf{1}^{|D|}$. We solve the bilevel problem with Adam for both network parameters θ (learning rate 5×10^{-3}) and selection weights $\mathbf{s}$ (learning rate 5×10^{-2}), over 100 outer iterations with variance reduction and gradient clipping. We adopt the standard ResNet-18 backbone for all runs.

B.4 Heterogeneous data sources: scaling to larger datasets

We evaluate CADS-S on DomainNet with 6 groups using the full training set and batch size 1024. We allocate an epoch-based compute budget of 10 epochs. The selection parameters $\mathbf{s} \in \mathbb{R}^{|D|}$ are initialized as $\mathbf{s} = v \cdot \mathbf{1}$, $v \in \{0.2, 0.4, 0.6, 0.8\}$, $\mathbf{1} \in \mathbb{R}^{|D|}$ is the all-ones vector. We solve the bilevel problem with Adam for both network parameters θ (learning rate 5×10^{-3}) and selection weights $\mathbf{s}$ (learning rate 5×10^{-2}), over 100 outer iterations with variance reduction and gradient clipping. We adopt the standard ResNet-18 backbone for all runs.

DomainNet dataset. DomainNet comprises over 0.5 million images across six visual domains: Real, Sketch, Clipart, Painting, Infograph, and Quickdraw (As shown in Fig. 6). In our setup, we treat the Real domain as a high-quality source but include only 10,000 samples from it to simulate limited access; the other five domains are used in full.

B.5 Instruction tuning tasks: small-scale experiments

We evaluated our CADS-S approach on instruction fine-tuning tasks using the GPT-2 model [48]. The dataset consists of two heterogeneous data sources: (i) **Alpaca-GPT4** [46], a high-quality dataset from GPT-4-generated instructions, from which we sample 1,000 examples to simulate the rarity

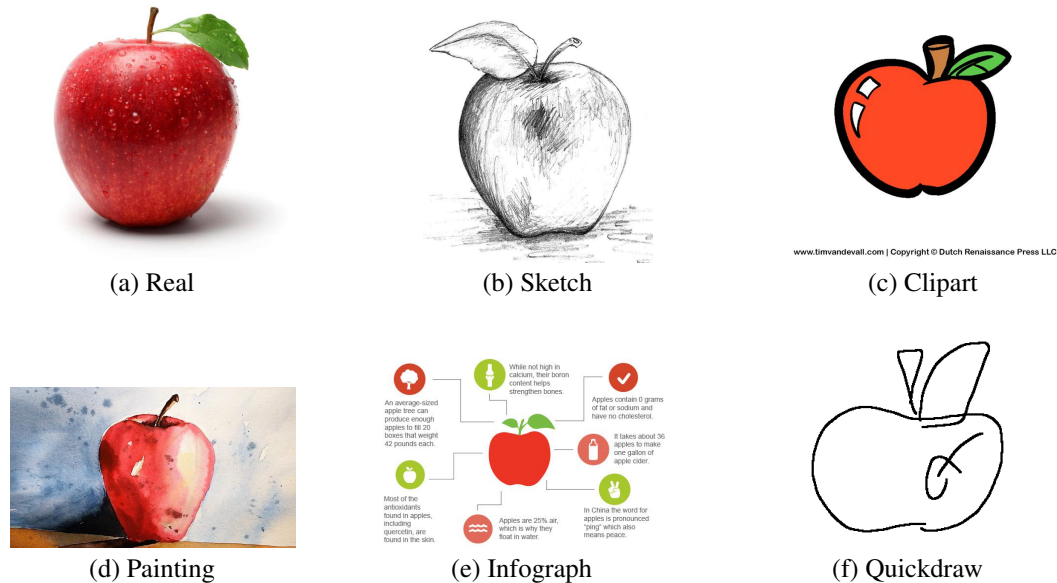


Figure 6: Representative examples from the DomainNet domains: Real, Sketch, Clipart (top row); Painting, Infograph, Quickdraw (bottom row).

of premium data sources in practical scenarios; (ii) **Alpaca** [57], a larger dataset containing 9,000 examples, representing standard-quality data. The total compute budget varies within the range $[1 \times 10^4, 5 \times 10^4]$ sample usages. The selection weight s is initialized to $0.5 \cdot \mathbf{1}^{|D|}$. We solve the bilevel problem with AdamW for both model parameters θ (learning rate 1×10^{-5}) and selection weights s (learning rate 5×10^{-2}). We adopt the GPT-2 backbone with maximum sequence length 1024 for all runs.

Data preprocessing. Raw instruction–response samples are loaded from JSON or JSONL files and consolidated into a flat record list. A pretrained tokenizer is initialized and extended to include five special markers: a padding token, an end-of-sequence token, and three role-demarcation tokens (`<system>`, `<user>`, `<assistant>`). Each record is then serialized into a single text sequence by:

1. Emitting the system token, followed by the system-level directive, and two line breaks.
2. Emitting the user token, followed by the instruction text (and any optional input), and two line breaks.
3. Emitting the assistant token, followed by the target response, and terminating with the end-of-sequence token.

The concatenated text is first tokenized without truncation to measure its length; any example that exceeds the maximum allowed token count is discarded. The remaining examples are split into training, validation, and test subsets according to either explicitly provided sizes or a default 90/5/5 ratio. At training time, each sequence is tokenized with truncation to the maximum length, producing token identifiers and attention masks. All token positions corresponding to the system and user segments are assigned an ignore index in the label sequence so that only assistant-response tokens contribute to the loss. A bespoke collation routine then pads every batch to the length of its longest sequence, using the padding token for inputs, zeros for masks, and the ignore index for both prompt and padding positions in the labels.

B.6 Instruction tuning tasks: scaling to additional datasets

We extended our evaluation to 13 distinct instruction-following corpora, drawing ten thousand examples from each. These included the original AlpacaGPT4 benchmark[46], the SlimOrca set[32], the Alpaca collection[57], the GPTeacher suite[58], and nine multilingual variants of the Alpaca data.

In total, approximately 130 000 samples were processed with the identical tokenization, formatting, filtering, and batch-collation pipeline described above. The complete list of datasets, along with their sizes and licenses, is summarized in Table 7.

Dataset	Size	License
AlpacaGPT4 [46]	52K	Apache-2.0
SlimOrca [32]	518K	MIT
GPTTeacher [58]	89K	MIT
Alpaca [57]	52K	Apache-2.0
Alpaca-es ²	52K	CC-BY-4.0
Alpaca-de ³	50K	Apache-2.0
Alpaca-ja ⁴	52K	CC-BY-NC-SA-4.0
Alpaca-ko ⁵	50K	CC-BY-NC-4.0
Alpaca-ru ⁶	30K	CC-BY-4.0
Alpaca-it ⁷	52K	CC-BY-NC-SA-4.0
Alpaca-fr ⁸	55K	Apache-2.0
Alpaca-zh ⁹	49K	CC-BY-4.0
Alpaca-pt ¹⁰	52K	CC-BY-NC-4.0

Table 7: Summary of high-quality instruction-tuning datasets (top) and multilingual Alpaca variants (bottom).

C Log-linear surrogate of compute-constrained training loss

C.1 Data Collection

We collect estimates of the K -step (“compute-constrained”) training loss $\mathcal{L}_{\text{trn}}(\theta; |m|)$ at a range of subset sizes $|m|$. By default we choose nine relative fractions $\{0.01, 0.02, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9\}$ of the full training set (clamped to at least 50 examples on MNIST or the configured batch-size otherwise). For each size we train for a total compute budget N (so that each inner training run sees roughly the same total number of gradient steps), and record the final training loss. We then summarize each size by its empirical ℓ_j .

C.2 Surrogate model choice

We set $\varepsilon = 10^{-10}$ and fit the transformed log-loss $\log(l(|m|) + \varepsilon)$ using two options:

- **Linear:** $\log(l(|m|) + \varepsilon) = k|m| + b$.
- **Cubic spline interpolation:** fit a cubic spline through the points $(|m_j|, \log(l(|m_j|) + \varepsilon))$, yielding a piecewise-cubic function $\hat{f}(|m|)$.

²<https://huggingface.co/datasets/bertin-project/alpaca-spanish>

³https://huggingface.co/datasets/mayflowergmbh/alpaca-gpt4_de

⁴https://huggingface.co/datasets/fujiki/japanese_alpaca_data

⁵https://huggingface.co/datasets/Bingsu/ko_alpaca_data

⁶https://huggingface.co/datasets/IlyaGusev/ru_turbo_alpaca

⁷https://huggingface.co/datasets/mchl-labs/stambecco_data_it

⁸<https://huggingface.co/datasets/jpacifico/French-Alpaca-dataset-Instruct-55K>

⁹<https://huggingface.co/datasets/llm-wizard/alpaca-gpt4-data-zh>

¹⁰<https://huggingface.co/datasets/dominguesm/alpaca-data-pt-br>

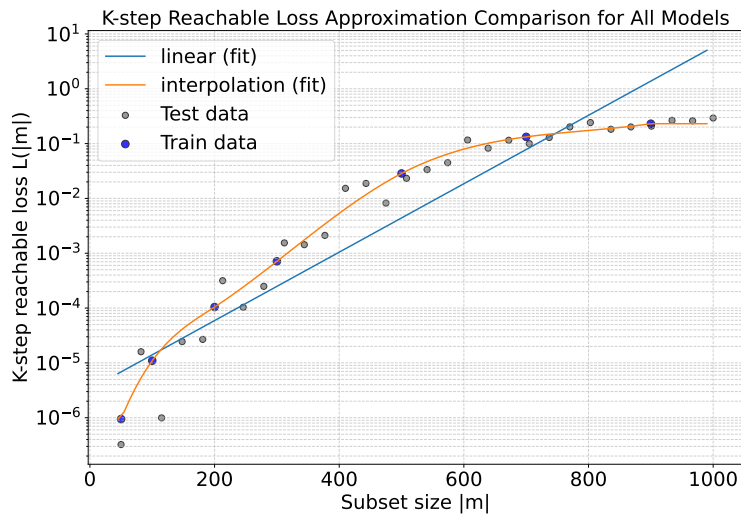


Figure 7: All-models comparison of the K-step reachable loss as a function of subset size $|m|$. Blue circles denote observed mean loss on training subsets; gray squares denote observations on held-out test subsets. Solid curves show the fitted approximation functions for each candidate model. The vertical axis is plotted on a logarithmic scale.

After fitting in log-space, we recover the original-scale surrogate via $l(|m|) = \exp(\hat{f}(|m|))$. Figure 7 overlays these fitted curves on the empirical training-loss measurements. The cubic spline interpolation achieves a noticeably lower MSE, motivating its use when maximum fidelity is desired, while the linear surrogate offers simplicity and a closed-form gradient. **Since CADS never computes $\partial l(|m|)/\partial |m|$, we exclusively employ the cubic spline interpolation due to its superior fit.**

C.3 Analysis of Sampling Efficiency for Loss Estimation

To evaluate the sample efficiency of our loss estimation method, we conducted an analysis to determine the minimum number of subset samples (K) required for reliable interpolation.

Experimental Design On the CIFAR-10 dataset with a ResNet-18 model, we evaluated the loss estimator’s performance under varying sampling densities, where $K \in \{4, 5, 6, 7, 8\}$. For each value of K , we sampled K distinct subset sizes to train corresponding models, yielding K pairs of (size, loss) data points. These points were used to fit the loss estimator. Subsequently, we tested each fitted estimator on a held-out set of 100 separately sampled (size, loss) pairs to measure its interpolation accuracy.

Results The results, summarized in Table 8, demonstrate that the estimator’s performance improves significantly up to $K = 5$ sampling points, after which the returns diminish. This indicates that our method is highly sample-efficient. Notably, increasing K further (e.g., $K = 7$) can slightly degrade performance, likely due to overfitting the estimator to a specific set of sample points, which hinders its ability to generalize to unseen subset sizes.

D Visualizations on truncated Gaussian distribution

In this section we provide visual intuition for the source-level sampling prior by plotting the truncated Gaussian density $p(r | s)$ on the interval $[0, 1]$ for a variety of center values s and scales σ . Figures 8, 9, and 10 illustrate how annealing σ concentrates the distribution around its mean.

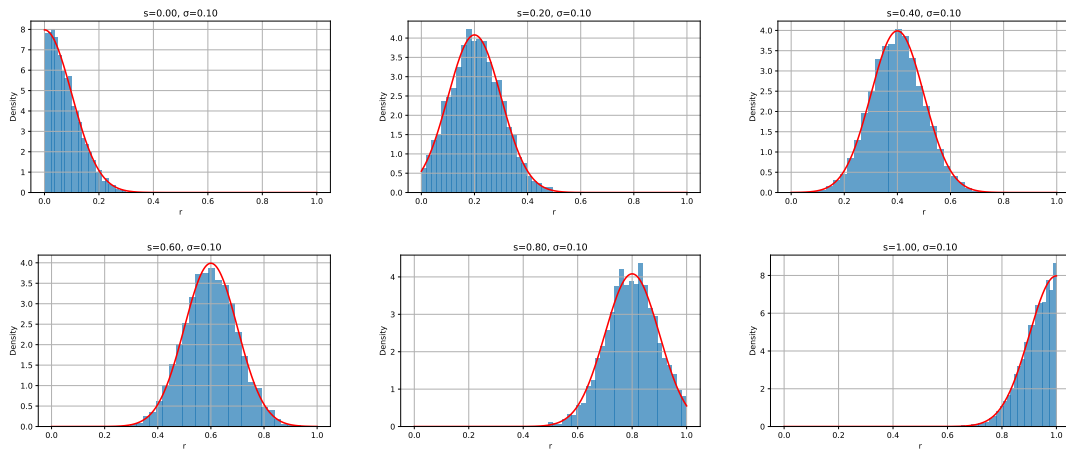


Figure 8: Sampling prior $p(r | s)$ on $[0, 1]$ for various center values s with scale $\sigma = 0.1$.

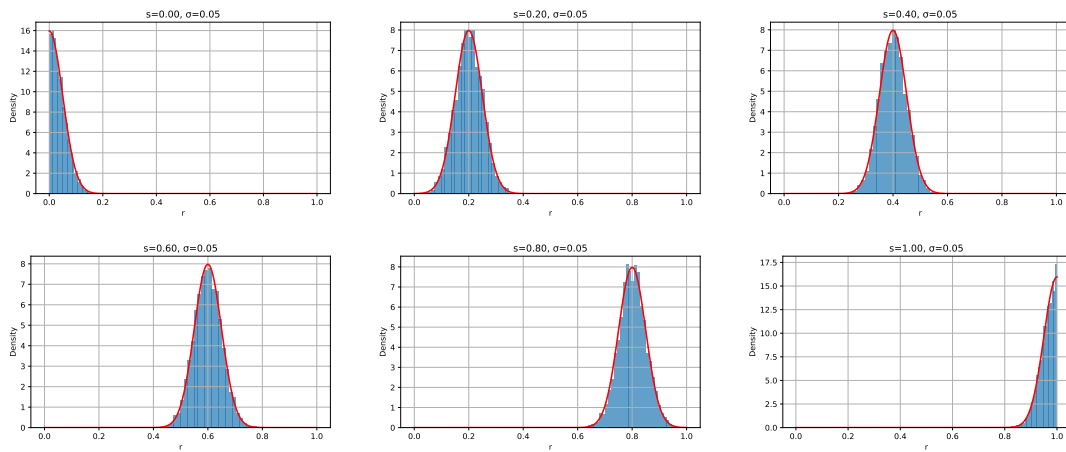


Figure 9: Sampling prior $p(r | s)$ on $[0, 1]$ for various center values s with scale $\sigma = 0.05$.

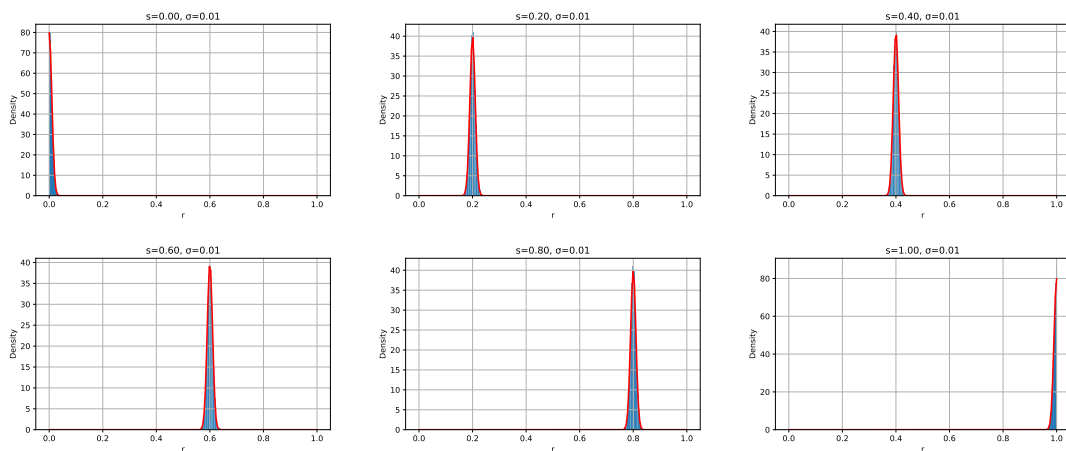


Figure 10: Sampling prior $p(r | s)$ on $[0, 1]$ for various center values s with scale $\sigma = 0.01$.

Table 8: Mean Square Error (MSE) of the loss estimator with respect to the number of sampling points (K). Performance saturates at $K = 5$, demonstrating high sample efficiency.

Sampling Points (K)	Mean Square Error
4	0.057668
5	0.016326
6	0.018594
7	0.020031
8	0.017905

E Additional experimental results

Additional result on MNIST The purpose of this experiment is to investigate the effect of compute budget on data selection within the MNIST dataset. We vary the compute budget in terms of total forward counts (10 000, 20 000, 50 000 and 100 000). At each budget we compare four sampling strategies—random, PBCS, CADS-E and our Bilevel-CADS. As shown in Table 9, our Bilevel-CADS consistently outperforms all baselines in all compute budgets.

Table 9: Results on MNIST with various compute budgets.

Method	Compute budget				Average
	10,000	20,000	50,000	100,000	
Random	87.36	88.05	87.71	87.24	87.59
PBCS	89.61	90.48	90.96	90.41	90.37
CADS-E	90.62	91.48	92.45	93.17	91.93
Bilevel-CADS	90.70	92.44	93.09	93.61	92.46

F Time efficiency analysis

F.1 Computational complexity compare to standard bilevel optimization

We compare the per-iteration cost of our single-epoch inner loop against the full-training inner loop used in bilevel-CADS, and then account for the one-time cost of fitting the loss estimator $l(|\mathbf{m}|)$.

Let

- $|\mathcal{D}|$ be the size of the training set,
- $T_{\text{ep}} = O(|\mathcal{D}|)$ the cost of one full epoch (one forward+backward pass),
- N the number of epochs used by bilevel-CADS in its inner optimizer,
- K the number of subsets sampled per outer iteration,
- $|\mathbf{m}|$ the average subset size,
- M the total number of outer iterations in a run.

Bilevel-CADS inner solve (per outer step)

$$\text{Cost}_{\text{bilevel}} = K \times N \times T_{\text{ep}} = O(KN|\mathcal{D}|).$$

Our CADS-E/S inner loop (per outer step, ignoring estimator)

$$\text{Cost}_{\text{CADS}}^{\text{base}} = K \times T_{\text{ep}} = O(K|\mathcal{D}|).$$

Estimator fitting overhead To fit the mapping $l(|\mathbf{m}|)$ we train 8 models for N epochs each (As detailed in D), incurring a one-time cost

$$\text{Cost}_{\text{est}} = 8 \times N \times T_{\text{ep}}.$$

Amortized over M outer steps, this adds

$$\frac{\text{Cost}_{\text{est}}}{M} = \frac{8 N T_{\text{ep}}}{M}$$

to each iteration.

Total CADS cost (amortized)

$$\text{Cost}_{\text{CADS}} = K \times T_{\text{ep}} + \frac{8 N T_{\text{ep}}}{M} = T_{\text{ep}} \left(K + \frac{8 N}{M} \right).$$

Speed-up factor

$$\frac{\text{Cost}_{\text{bilevel}}}{\text{Cost}_{\text{CADS}}} = \frac{K N T_{\text{ep}}}{T_{\text{ep}} \left(K + \frac{8 N}{M} \right)} = \frac{K N}{K + 8 N/M}.$$

In our experiments $K = 5$ and $M = 100$, giving

$$\frac{\text{Cost}_{\text{bilevel}}}{\text{Cost}_{\text{CADS}}} = \frac{1}{1/N + 0.016}.$$

In practice, larger N yields an even greater speed-up.

F.2 Empirical Validation

We measure end-to-end wall-clock time of bilevel-CADS and CADS-E on MNIST as a function of compute budget C (total epochs). We sweep C over $\{5, 50, 100, 200, 500, 1000, 2000, 5000\}$ and record the runtime of each method under identical hardware. Figure 11 confirms a roughly linear scaling for both methods, with CADS-E exhibiting a much smaller slope.

F.3 Analysis of Selection Algorithm Runtime

A potential concern regarding our method is that the runtime of the selection algorithm may scale with the number of sampled subsets (K), which could offset the computational efficiency gains. We address this concern by highlighting several key aspects of our approach.

Minimal Sampling for Practical Implementation Our policy gradient approach for coreset selection requires only a small number of samples (K) to ensure training stability. For instance, the comparable PBCS method [74] utilizes $K = 1$ for all its experiments. Our results show that using $K = 2$ is sufficient to achieve strong performance while maintaining high computational efficiency. In cases where smaller K values might lead to higher gradient variance (e.g., in more complex problems), this can be mitigated using established variance reduction techniques. A common approach is to use a self-critical baseline, which adjusts the reward signal as follows:

$$\tilde{\mathcal{L}}(S_i) = \mathcal{L}(S_i) - \frac{1}{K} \sum_{j=1}^K \mathcal{L}(S_j). \quad (12)$$

Amortized Cost of Coreset Selection It is important to emphasize that the core contribution of this work is the introduction of computational budget constraints into the coreset selection process, rather than optimizing the selection runtime itself. Once generated, a coreset is a reusable asset that can be applied across various training scenarios, different model architectures, and subsequent experiments. Consequently, the initial computational investment in the selection process is amortized over these multiple uses. This makes the selection efficiency a secondary consideration compared to the quality and utility of the resulting budget-constrained coreset.

G Scalability to Other Bilevel Optimization Problems

CADS is designed to address a fundamental challenge in bilevel optimization: scenarios where the inner-level problem's convergence is hindered by computational budget constraints. Many practical bilevel problems suffer from this issue, where traditional methods that assume unlimited computational resources for inner-level convergence often fall short. We have also successfully applied a CADS-like method to resource-intensive diffusion models, demonstrating its potential beyond the scope of this work. The scalability of CADS to other bilevel problems primarily depends on the following factors:

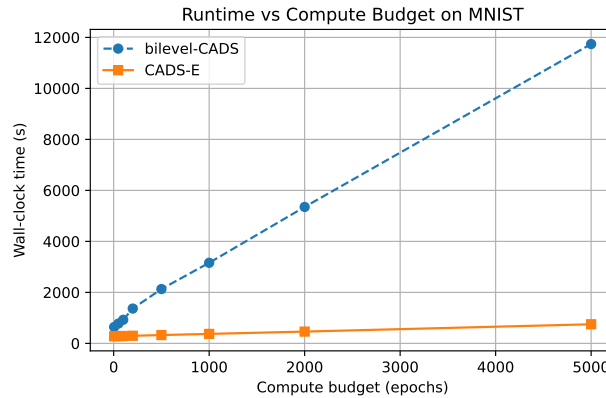


Figure 11: Wall-clock runtime vs. compute budget C for bilevel-CADS (dashed) and CADS-E (solid), MNIST.

Applicability Conditions The core premise of CADS is most relevant when the inner-level optimization is computationally expensive and cannot be fully converged. Our approach provides a framework to handle such resource-constrained scenarios, which are common in real-world applications but often overlooked by conventional methods.

Loss Estimation Accuracy The effectiveness of CADS in a new domain hinges on the accuracy of the loss estimator. As demonstrated in our analysis of loss estimation generalization, the proposed estimator maintains relatively stable predictive performance across the diverse experimental domains explored in this paper. This stability suggests that similar performance may be achievable in other domains with comparable data characteristics.

Validation Framework for New Domains To adapt CADS to a new bilevel optimization problem, we propose the following validation framework:

- **Loss Estimator Generalization Assessment.** Before full-scale implementation, it is crucial to assess whether the data distribution characteristics of the new domain can be reliably predicted. This can be achieved by conducting experiments similar to our generalization evaluation to validate the feasibility of loss estimation.
- **Domain-Adaptive Estimator Design.** While our current estimator is effective for data selection problems, different bilevel scenarios might benefit from specialized estimators. The architecture of the loss estimator should be tailored to the specific requirements and data characteristics of the target domain to ensure optimal performance.

H Limitations and Future Work

Similar to existing methods, our approach is based on standard training methodologies, but in the industry, large language models incorporate numerous engineering optimizations. Consequently, our budget is likely proportional to theirs, rather than perfectly aligned. For instance, in sparse mixture of experts (MoE), the budget is determined by model capacity and routing complexity, whereas we measure performance by the number of forward passes. This measurement may not fully correspond with the budget metrics used in industry. However, we firmly believe that our methods are applicable in industrial contexts; we simply need to adjust the budget measurements accordingly.

I Licenses for existing assets

We rely on several public datasets and open-source libraries, and all original authors are properly credited and their licenses fully respected. The vision benchmarks we employ are MNIST (public domain), CIFAR-10 (MIT License) and DomainNet (CC BY-NC-SA 4.0). Our instruction-tuning

corpora are summarized in Table 7, with licenses ranging from Apache 2.0 to various Creative Commons and MIT terms. On the software side, we build atop PyTorch and torchvision (BSD 3-Clause), SciPy (BSD), and HuggingFace Transformers (GPT-2, Apache 2.0); ResNet models are taken from torchvision (BSD 3-Clause). We confirm that every dataset and code dependency is used in accordance with its license and citation requirements.