# A CLIP-POWERED FRAMEWORK FOR ROBUST AND GENERALIZABLE DATA SELECTION

**Suorong Yang[1,2*], Peng Ye[2,3], Wanli Ouyang[2], Dongzhan Zhou[2 †], Furao Shen[1 †]**
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2] Shanghai Artificial Intelligence Laboratory
[3] The Chinese University of Hong Kong

## ABSTRACT

Large-scale datasets have been pivotal to the advancements of deep learning models in recent years, but training on such large datasets inevitably incurs substantial storage and computational overhead. Meanwhile, real-world datasets often contain redundant and noisy data, imposing a negative impact on training efficiency and model performance. Data selection has shown promise in identifying the most representative samples from the entire dataset, which aims to minimize the performance gap with reduced training costs. Existing works typically rely on single-modality information to assign importance scores for individual samples, which may lead to inaccurate assessments, especially when dealing with noisy or corrupted samples. To address this limitation, we propose a novel CLIP-powered data selection framework that leverages multimodal information for more robust and generalizable sample selection. Specifically, our framework consists of three key modules—dataset adaptation, sample scoring, and selection optimization—that together harness extensive pre-trained multimodal knowledge to comprehensively assess sample influence and optimize the selection results through multi-objective optimization. Extensive experiments demonstrate that our approach consistently outperforms existing state-of-the-art baselines on various benchmark datasets. Notably, our method effectively removes noisy or damaged samples from the dataset, enabling it to achieve even higher performance with less data. This indicates that it is not only a way to accelerate training but can also improve overall data quality. The implementation is available at `https://github.com/Jackbrocp/clip-powered-data-selection`.

## 1 INTRODUCTION

Recent advancements in deep learning have been propelled by increasingly large and complex models that utilize vast datasets to achieve start-of-the-art performance Liu et al. (2024b); Touvron et al. (2023). However, this success normally comes with considerable costs for data storage and computational resources, which may even limit the deployment of models to specialized infrastructure and hinder their scalability across different applications. Moreover, real-world datasets often contain redundancy and noise, which can degrade the training efficiency and performance.

To alleviate the data redundancy issue and improve the training efficiency, there are typically two kinds of methods, i.e., dynamic pruning and data selection. Dynamic pruning methods Raju et al. (2021); Qin et al. (2024) aim to reduce training costs by dynamically selecting only the most influential samples from the dataset during training. Despite their effectiveness in accelerating training, they still face the cost of large-scale data storage, and their selected samples often lack the ability to generalize well across different training processes and architectures. In contrast, data selection methods Paul et al. (2021); Yang et al. (2023b); Tan et al. (2024); Xia et al. (2023) pre-select a fixed subset of the essential data points before the training begins, allowing the model to achieve performance comparable to that obtained with the full dataset. By focusing on the most critical data points, these methods ensure better generalization ability across various training scenarios.

[*] This work was done during his internship at Shanghai Artificial Intelligence Laboratory.
[†] Corresponding authors.

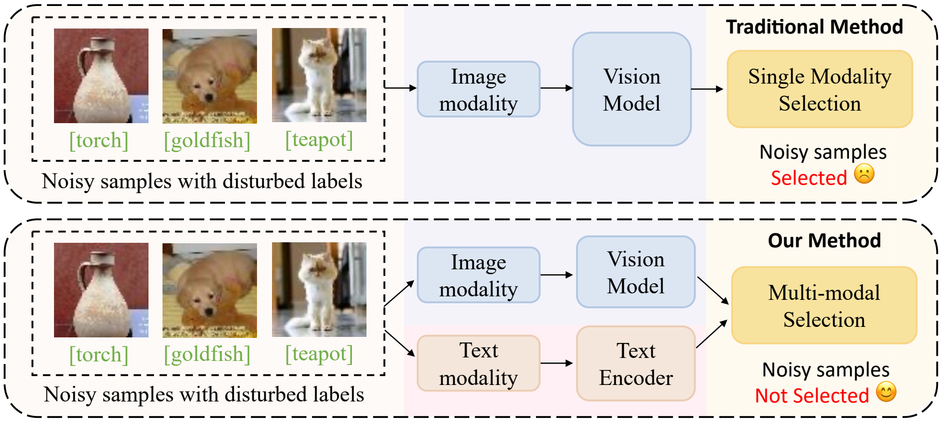


Figure 1: Benefit of multimodal selection. Traditional single-modality approaches (upper part) may struggle with noisy and corrupted data, while our multimodal method (lower part) identifies class-representative samples with diversity while effectively filtering out noise and corrupted data.

Existing data selection methods typically employ carefully designed criteria via three perspectives: importance scores Paul et al. (2021); Tan et al. (2024), image data distribution Zheng et al. (2023); Xia et al. (2023), and optimization-based functions Killamsetty et al. (2021b); Yang et al. (2023b). Although achieving promising results, these methods exhibit certain limitations. On one hand, relying solely on single-modality image information can lead to ambiguities, especially when noisy samples are present, which may result in inaccurate assessments of a sample's effect. For instance, some methods employ difficulty-based criteria to select data; however, distinguishing between truly difficult samples and noisy ones based solely on image modalities presents a significant challenge. On the other hand, existing methods typically select samples with the highest or lowest scores, while the interaction between high-score and low-score samples within a group can significantly influence the overall performance, which is known as the ”group effect” Koh et al. (2019); Yang et al. (2023b). Thus, a more beneficial approach is to leverage the power of multimodal information and evaluate the collective effect of the sample group.

In this paper, we propose a CLIP-powered data selection framework that employs multimodal information for more robust and generalizable data selection, where the category text serves as a strong complement to the image modality and promotes overall performance. The framework consists of three modules, namely dataset adaptation, sampling scoring, and selection optimization module. First, the dataset adaptation module integrates image and text adapters to facilitate the transfer of pretraining knowledge to the target data. Subsequently, the sample scoring module calculates the Semantic Alignment Score (SAS) and Sample Diversity Score (SDS) based on the adapted multimodal features, which measure the image-text alignment and the variability of local patterns. Using these two scores together can select semantically representative samples while maintaining their inherent diversity. Further, in order to address the group effect, we introduce a selection optimization module to identify the ideal subsets w.r.t. the expected selection ratio through a multi-objective optimization strategy. By leveraging the multi-modal information and carefully designed supervision signals, our framework enables the selection of high-quality samples in a flexible and efficient manner.

Comprehensive evaluation across various benchmark datasets demonstrates that our approach effectively improves the performance of selected datasets, especially on large-scale datasets such as ImageNet-1k Deng et al. (2009). Moreover, the selected datasets exhibit superior cross-architecture generalization across ResNet-18/50 He et al. (2016), ViT Dosovitskiy et al. (2020), VGG-16 Simonyan & Zisserman (2014), etc. Notably, since most existing methods are not robust to more complex and realistic scenarios, we further validate the strong robustness of our method in more challenging scenes. For instance, our proposed method can achieve an 8.13% improvement in accuracy on CIFAR-100 and a 4.41% improvement on Tiny-ImageNet compared to the leading baselines. Meanwhile, superior performance is achieved with very high efficiency compared to other baselines.

The contributions of this work can be summarized as follows: 1) We analyze the drawbacks of previous works that rely solely on image modalities in depth, and propose a new CLIP-powered data selection framework that leverages multimodal features for robust and generalizable data selection for the first time. 2) Our framework comprises dataset adaptation and sample scoring modules to foster multi-modality knowledge transfer and comprehensive sample importance evaluation. This dual-modality design effectively removes noisy and corrupted samples from the dataset. 3) A selection optimization module is designed to identify the optimal subsets w.r.t. the expected selection ratios through multi-objective optimization, which effectively addresses the group effect while maintaining high efficiency. 4) Experimental results show that our method outperforms previous SOTA approaches in terms of performance, cross-architecture generalization, and robustness to noisy and corrupted images. Meanwhile, our approach achieves the best trade-off in performance and selection efficiency, establishing a strong baseline of data selection for future research.

## 2 Related Work

Date-efficient deep learning generally incorporates dynamic data pruning Qin et al. (2024); Raju et al. (2021), static data selection Xia et al. (2023); Tan et al. (2024); Yang et al. (2023b), data distillation Lei & Tao (2023); Du et al. (2023); Zhang et al. (2023); Cazenavette et al. (2022), and data condensation Liu et al. (2023); Yang et al. (2023a). Following the static data selection, we propose a method capable of identifying representative and diverse samples across various selection ratios. Data selection, or static dataset pruning, aims to identify and select the most representative samples from training datasets before training begins. Training on the selected subsets can achieve comparable performance to that obtained with the full dataset while reducing training and storage costs. Current data selection methods can be broadly divided into three categories: importance criteria Paul et al. (2021); Tan et al. (2024), dataset distribution-based methods Xia et al. (2023); Zheng et al. (2023), and optimization-based methods Nohyun et al. (2023); Yang et al. (2023c).

**Selection with importance criteria** is the most popular type. These methods typically involve calculating importance scores for each sample and selecting samples based on these scores. For instance, EL2N and GraNd score Paul et al. (2021) measure the importance by calculating the expectation of the $\ell_2$-norm error vector and the expectation of the gradient norm, respectively. MoSo Tan et al. (2024) calculates the change of the optimal empirical risk when removing a specific sample from the training set. Forgetting Toneva et al. (2018) tracks the frequency with which a sample is misclassified after being correctly classified during training. Similarly, Memorization Feldman & Zhang (2020) assesses the impact of a sample's presence or absence on the model's ability to predict it correctly. While importance criteria-based methods are often computationally efficient, their performance may be affected by the group effect Yang et al. (2023b;c) and may not generalize well to complex, real-world scenarios Xia et al. (2023).

**Dataset distribution-based methods** select samples by analyzing the geometric distribution of datasets. For instance, Herding Welling (2009) determines the sample importance according to samples' distance to their corresponding class centers. The work Ramalingam et al. (2023) applies greedy k-center to select the coreset with good data coverage. D2 Maharana et al. (2023) calculates and updates the difficulty scores of each sample by incorporating the difficulty of its neighboring examples. Moderate-DS Xia et al. (2023) choose samples with closer distances to the median score, while CCS Zheng et al. (2023) balances the data distribution and the sample importance in selection.

**Optimization-based methods** select samples through various optimization techniques, such as temporal dual-depth scoring Zhang et al. (2024), gradient matching Mirzasoleiman et al. (2020b); Killamsetty et al. (2021a), scalable self-supervised pruning metric Sorscher et al. (2022), influence function Yang et al. (2023b); Pooladzandi et al. (2022), bi-level optimization Killamsetty et al. (2021b), facility location function Mirzasoleiman et al. (2020a); Yang et al. (2023d), and submodularity Iyer et al. (2021); Nohyun et al. (2023); Kothawade et al. (2022); Wei et al. (2015). In contrast to these methods that only rely on image information, we leverage multimodal messages for data selection, which incorporates the semantic alignment between image data and corresponding category information. This contributes to comprehensive assessments of sample effectiveness, particularly in complex scenarios where samples may be corrupted or wrongly labeled.

## 3 The Proposed Method

Our proposed method is summarized in Figure 2. The approach involves the use of the pretrained vision-language foundation model CLIP to construct multimodal feature spaces. Nevertheless,

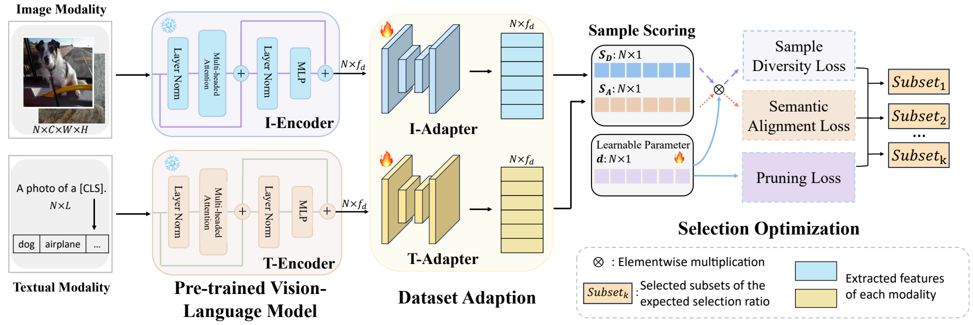


Figure 2: Our proposed method consists of dataset adaptation, sampling scoring, and selection optimization modules. The dataset adaptation module is used to learn dataset-specific knowledge. The sample scoring module computes two scores, $\boldsymbol{S}_A$ and $\boldsymbol{S}_D$, to assess sample importance, based on which the selection optimization identifies the optimal subsets w.r.t. the expected selection ratios.

there may exist domain shifts or discrepancies between the pretrained training dataset and the target dataset Liu et al. (2024a); Alijani et al. (2024). To facilitate dataset adaptation and enhance the learning of dataset-specific knowledge, we incorporate dimension-preserving adapters for both modalities. Following this, two scores are derived to comprehensively assess the sample importance, i.e., the Semantic Alignment Score (SAS), denoted as $\boldsymbol{S}_A$, and the Sample Diversity Score (SDS), denoted as $\boldsymbol{S}_D$. Furthermore, rather than solely based on sample scores for selection, we design a multi-objective optimization to identify the optimal subsets w.r.t. the expected selection ratios, which effectively mitigates the group effect. We provide the detailed methodology in the subsequent sections.

### 3.1 Dataset Adaptation

To alleviate the domain shifts and discrepancies between the pretrained and target datasets, we incorporate dimension-preserving adapters to perform adaptation on the target dataset. The image and text adapters are denoted as $A_I$ and $A_T$, respectively. Both adapters are fine-tuned for knowledge transfer while the pretrained CLIP weights are frozen. To maintain high efficiency, both adapters utilize simple MLP.

Specifically, the fine-tuning process employs the InfoNCE loss Parulekar et al. (2023); Oord et al. (2018), which maximizes the mutual information between the image and text representations. The text representation describes the category information using the prompt: “A photo of [CLS]”, where the token [CLS] represents the corresponding category. The loss ensures that the adapters effectively align and capture the relevant features from both modalities. Furthermore, it enhances the model's ability to distinguish between positive and negative pairs, thereby improving the robustness and accuracy of the deep representations for the specific dataset.

### 3.2 Sample Scoring

For classification datasets, the learning process for training samples is intrinsically linked to acquiring knowledge of the corresponding categories. A training sample that more accurately represents its category is typically more effective in training deep networks. In this way, the **Semantic Alignment Score (SAS)** is designed to assess the semantic similarity between training samples and their corresponding categories. Specifically, since image and text features reside within the same embedding space Radford et al. (2021), the SAS is derived by calculating the cosine similarity between the embedded image and corresponding textual deep descriptions. Accordingly, the SAS for the $i$-th sample is defined as:

$$\boldsymbol{S}_{Ai} = \cos(A_I(E_I(\boldsymbol{I}_i)), A_T(E_T(\boldsymbol{T}_i))), \tag{1}$$

where $\boldsymbol{I}_i$ is the $i$-th sample, $\boldsymbol{T}_i$ is the textual description of the corresponding category for $\boldsymbol{I}_i$, $E_I$ and $E_T$ are pretrained image and text encoders, respectively.

For selected datasets, the reduced data volume may limit the diversity of the selected data, which is important for training datasets Yang et al. (2024a). To address this, we introduce another diversity

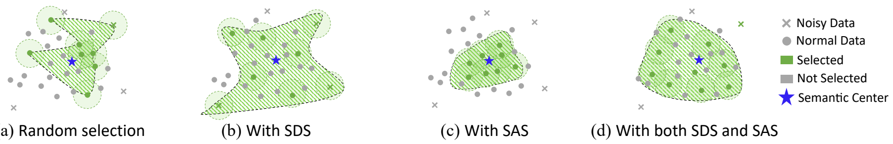


Figure 3: Illustration of the effectiveness of SAS and SDS. The circle and cross represent the normal and noisy samples, respectively. Different colors correspond to the selection results. SDS (b) selects diverse samples but may include noises. SAS (c) could avoid the noisy samples but potentially miss broader category information. Using both scores (d) can select category-representative while maintaining high diversity.

perspective to comprehensively assess the effect of training data. The **Sample Diversity Score (SDS)** is defined as the average distance between each sample and its $k$ neighbor samples of the same class:

$$\boldsymbol{S}_{Di} = \frac{1}{k} \sum_{\boldsymbol{I}_j \in \mathrm{KNN}(\boldsymbol{I}_i)} \|A_I(E_I(\boldsymbol{I}_i)) - A_I(E_I(\boldsymbol{I}_j))\|, \tag{2}$$

where we use the KNN algorithm to obtain the neighbor samples for each sample, the distance metric is based on the $\ell_2$ norm, and $k$ is usually set to 10% of the number of samples per class. In this way, the SDS can be understood as the local density of training samples in the feature space. If a sample has a larger number of neighbors with closer distances (i.e., lower SDS), its training efficacy may be more easily substituted by other neighbor samples. Therefore, selecting samples with higher SDS is generally more advantageous.

The effects of SAS and SDS are depicted in Figure 3. SDS contributes to the diversity of samples, but the selected data points may contain noise (Figure 3(b)). On the other hand, SAS can select samples that are semantically appropriate, as these points are basically around the sample center (Figure 3(c)). However, these samples may be too concentrated and thus lack diversity. When using SDS and SAS together (Figure 3(d)), we can cover the entire category space with fewer data and select samples that are both semantically representative and diverse, thereby boosting the effectiveness of data selection.

### 3.3 Selection Optimization

Instead of relying on combinatorial optimization functions for sample selection Killamsetty et al. (2021b), our method determines the selection through SGD multi-objective optimization, which improves computational efficiency and accelerates convergence. Specifically, we introduce a sample-wise parameter $\boldsymbol{d}$ to denote the selection decision, where elements of 1 indicate the selection while 0 indicates otherwise. Although binary parameters are difficult to optimize in neural networks due to the absence of gradient, we employ the $\mathrm{sigmoid}(\cdot)$ function to push the continuous values of $\boldsymbol{d}$ towards approximate binarization. After optimization, $\boldsymbol{d}$ is strictly binarized to explicitly indicate the final sample selection. Initially, $\boldsymbol{d}$ is initialized with all 1s.

To guide the optimization process, we introduce three loss items. The first item, $\mathcal{L}_{sa}$, is designed to prioritize samples with high SAS since these samples are more representative of their corresponding categories, which is defined as follows:

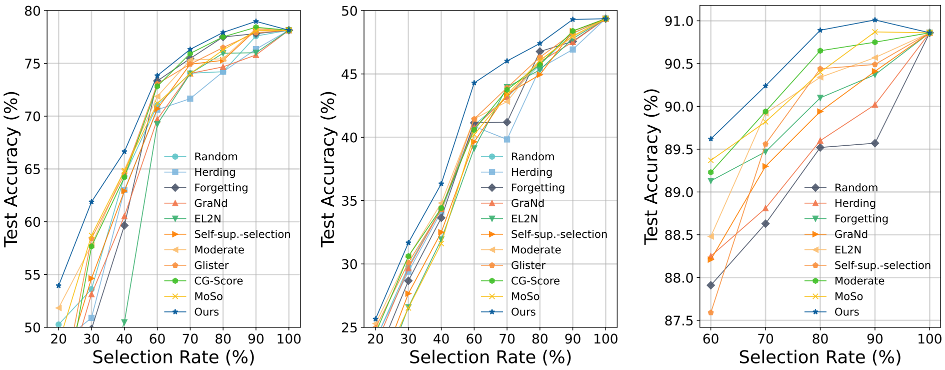
$$\mathcal{L}_{sa} = -\frac{1}{N} \sum_{i}^{N} \mathrm{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Ai}, \tag{3}$$

where $N$ is the total number of samples. $\mathcal{L}_{sa}$ punishes samples with low SAS and encourages the selection of samples with better semantic alignment.

In addition, we introduce another loss term, $\mathcal{L}_{sd}$, to encourage the selection of more diverse samples characterized by higher SDS, which is defined as:

$$\mathcal{L}_{sd} = -\frac{1}{N} \sum \mathrm{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Di}. \tag{4}$$

To mitigate the group effect, we optimize the selected datasets w.r.t. specific selection ratios, aiming to identify the optimal subsets. We introduce a selection loss term, $\mathcal{L}_s$, to ensure the selection



(a) CIFAR-100

(b) Tiny-ImageNet

(c) ImageNet-1k

Figure 4: Illustrations of comparing our method with various data selection baselines on CIFAR-100 (a), Tiny-ImageNet (b), and ImageNet-1k (c).

process adheres to the target ratio. However, deriving exact selection rates from continuous real-valued parameter optimization is difficult. While strictly binarized values facilitate explicit sample selection, they are challenging to optimize through gradient backpropagation. To address this, we utlize the straight-through estimator (STE) Bengio et al. (2013) to estimate the actual selection rate and derive gradients. STE allows gradients to pass through the discrete decisions during backpropagation, effectively combining the benefits of both continuous and binary parameters for efficient optimization and accurate sample selection. In this way, $\mathcal{L}_s$ is defined as:

$$\mathcal{L}_s = \sqrt{\left[\frac{1}{N}\sum_i \text{STE}\left[\mathbb{I}\left(\text{sigmoid}(\boldsymbol{d}_i)_{\text{i}} > 0.5\right)\right] - s_r\right]^2}, \tag{5}$$

where $\mathbb{I}$ is an indicator function, and $s_r$ denotes the expected selection ratio. The loss term guides the parameter $\boldsymbol{d}$ toward near-binary values, ensuring the count of ones aligns with the expected sample size. Since the selection is guided by adaptive optimization, the final selection ratio may deviate slightly from the target. To minimize this deviation, we constrain $\mathcal{L}_s$ with a threshold $\theta$, which in our work is set to $5 \times 10^{-4}$, ensuring the actual selection ratio differs by less than $\pm 0.05\%$ from the expected value. We also provide a theoretical analysis of $\theta$ on the actual selection ratio gap in Appendix A. Finally, the overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{sa} + \alpha\mathcal{L}_{sd} + \beta\mathcal{L}_s, \tag{6}$$

where $\alpha$ and $\beta$ are coefficients that adjust for numerical differences among the loss terms and can be set conveniently. The complete workflow is outlined in Algorithm 1 in Appendix B.

**Complexity Analysis** The proposed method comprises three main components. 1) The dataset adaptation involves fine-tuning the image and text adapters. Since the adapters consist of simple linear layers, the number of parameters is small, and both the forward passes and backward passes are computationally efficient. 2) In the sample scoring process, the complexity of calculating the SAS and SDS is $O(N)$ and $O(K * f_d)$, respectively, where $K$ is the number of categories and $f_d$ is the feature dimension, typically 512. The complexity of the KNN algorithm is $O(N_k * f_d)$, where $N_k$ is the number of samples per class. Given that $K$ and $f_d$ are constants and $N_k$ is usually much smaller than $N$, the overall complexity of this process is approximately $O(N)$. 3) The selection optimization of $\boldsymbol{d}$ does not involve deep models and is a numerical optimization process. The complexity is proportional to the number of parameters, i.e., $O(|w|) = O(N)$.

## 4 Experiment

### 4.1 Experimental Setup

**Baselines.** We compare our proposed method with ten most representative SOTA baselines, i.e., (1) Random, (2) MoSo Tan et al. (2024), (3) Glister Killamsetty et al. (2021b), (4) Herding Welling

Table 1: Test accuracy (%) on Tiny-ImageNet. VGG-16 and DenseNet-121 are exploited.

| Method / Selection Ratio | VGG-16 | | | | Densenet-121 | | | |
|---|---|---|---|---|---|---|---|---|
| | 70% | 80% | 90% | 100% | 70% | 80% | 90% | 100% |
| Random | 47.39±2.72 | 49.38±0.23 | 51.15±0.64 | 57.23±1.08 | 59.55±0.20 | 60.78±0.18 | 61.03±0.22 | 62.22±0.23 |
| EL2N | 48.30±2.95 | 48.75±1.65 | 49.01±1.31 | 57.23±1.08 | 59.61±0.00 | 60.38±0.04 | 61.16±0.47 | 62.22±0.23 |
| GraNd | 50.79±1.26 | 46.84±1.38 | 54.73±0.49 | 57.23±1.08 | 59.62±0.02 | 60.84±0.09 | 61.10±0.05 | 62.22±0.23 |
| MoSo | 50.47±1.01 | 50.12±0.83 | 50.07±0.43 | 57.23±1.08 | 59.27±0.33 | 59.86±0.07 | 60.00±0.37 | 62.22±0.23 |
| Herding | 48.59±0.07 | 45.77±0.12 | 50.77±1.24 | 57.23±1.08 | 59.00±0.28 | 60.03±0.35 | 61.15±0.12 | 62.22±0.23 |
| Glister | 48.74±2.29 | 50.05±0.02 | 49.42±1.81 | 57.23±1.08 | 59.98±0.01 | 60.62±0.34 | 61.28±0.18 | 62.22±0.23 |
| CG-Score | 48.73±2.70 | 48.49±1.88 | 49.62±1.08 | 57.23±1.08 | 59.74±0.15 | 60.55±0.20 | 61.14±0.11 | 62.22±0.23 |
| Self-sup. prototypes | 48.38±1.38 | 49.98±1.49 | 54.71±0.84 | 57.23±1.08 | 59.56±0.03 | 60.22±0.12 | 60.91±0.29 | 62.22±0.23 |
| Forgetting | 47.50±2.43 | 48.59±1.77 | 49.82±0.62 | 57.23±1.08 | 58.54±0.15 | 60.39±0.46 | 61.12±0.10 | 62.22±0.23 |
| Moderate-DS | 50.78±0.93 | 49.31±0.41 | 49.25±0.77 | 57.23±1.08 | 59.41±0.18 | 60.42±0.14 | 61.44±0.11 | 62.22±0.23 |
| Ours | **53.40±3.20** | **52.25±0.58** | **56.34±2.93** | 57.23±1.08 | **60.12±0.06** | **60.93±0.03** | **61.59±0.03** | 62.22±0.23 |

Table 2: Experimental results (accuracy, %, mean ± std) on CIFAR-100 and Tiny-ImageNet with noisy labels. 20% of labels are disturbed. We also report the numerical analysis of the proportion (%) of noisy data in the selected CIFAR-100 datasets.

| Method / Selection Ratio | CIFAR-100 (label noise) | | Tiny-ImageNet (label noise) | | Noise ratios | |
|---|---|---|---|---|---|---|
| | 20% | 30% | 20% | 30% | 20% | 30% |
| Random | 34.47±0.64 | 43.26±1.21 | 17.78±0.44 | 23.88±0.42 | 20.80 | 19.83 |
| MoSo | 31.01±0.67 | 43.73±0.14 | 21.55±0.37 | 27.80±0.16 | 7.78 | 8.82 |
| Moderate-DS | 40.25±0.12 | 48.53±1.60 | 19.64±0.40 | 24.96±0.30 | 0.30 | 0.31 |
| Glister | 28.51±1.46 | 43.16±1.31 | 21.61±0.19 | 25.45±0.23 | 21.21 | 21.95 |
| Herding | 42.29±1.75 | 50.52±3.38 | 18.98±0.44 | 24.23±0.29 | 35.00 | 30.56 |
| Forgetting | 36.53±1.11 | 45.78±1.04 | 13.20±0.38 | 21.79±0.43 | 23.00 | 21.76 |
| GraNd | 31.72±0.67 | 42.80±0.30 | 18.28±0.32 | 23.72±0.18 | 5.00 | 5.14 |
| EL2N | 29.82±1.19 | 33.62±2.35 | 13.93±0.69 | 18.57±0.31 | 22.00 | 21.80 |
| Self-sup. prototypes | 31.08±0.78 | 41.87±0.63 | 15.10±0.73 | 21.01±0.36 | 21.70 | 20.21 |
| CG-Score | 6.82±1.60 | 20.07±0.45 | 8.35±0.65 | 15.31±0.90 | 45.09 | 39.69 |
| Ours | **46.05±0.21** | **58.34±0.36** | **26.09±0.12** | **33.13±0.25** | **0.25** | **0.32** |

(2009), (5) Forgetting Toneva et al. (2018), (6) GraNd and (7) EL2N Paul et al. (2021), (8) Self-sup.-selection (SSP) Sorscher et al. (2022), (9) CG-Score Nohyun et al. (2023), and (10) Moderate-DS Xia et al. (2023).

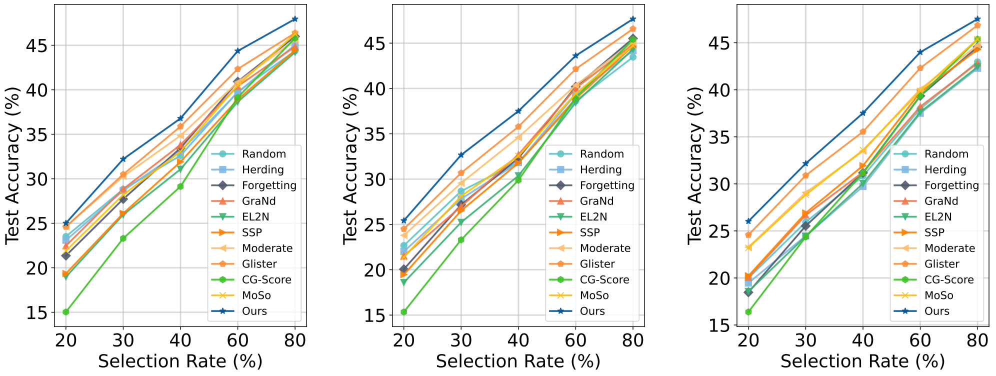
**Parameter settings.** The parameters in our proposed method can be easily set. The coefficient $\alpha$ is proportional to the expected selection rate $s_r$, balancing the importance of dataset diversity, i.e., $\alpha$ can be set equivalent to $s_r$. The coefficient $\beta$ is set to 2 across datasets to adjust the numerical differences among loss items. For more details, please refer to Appendix C.

### 4.2 Comparison with the State-of-the-arts

**Performance Comparisons** Consistent with prior works Xia et al. (2023); Sorscher et al. (2022), we report top-1 accuracy on CIFAR-100 and Tiny-ImageNet, and top-5 accuracy on ImageNet-1k. Note that the methods Glister and CG-Score are not compared on ImageNet-1k due to the heavy computation costs. Specifically, Glister obtains iterative solving of the bi-level optimization problem Killamsetty et al. (2021b), while CG-Score involves the calculation of large Gram matrix inversions, making them impractical for large-scale datasets.

As illustrated in Figure 4, our method consistently achieves the best accuracy across all datasets. Particularly, on the more challenging Tiny-ImageNet and ImageNet-1k datasets, our approach outperforms other methods by a notable margin. While existing approaches yield relatively marginal accuracy improvements on the small-scale CIFAR-100 dataset, the gains brought by our method are more substantial. Additionally, with relatively high selection ratios on these datasets, such as 90%, our selected datasets exhibit nearly lossless or even higher performance compared with the original full datasets and other baselines. The results indicate that our method can not only reduce training costs but may also serve as a way for data cleaning.

**Generalization Comparisons on Different Architectures** In this section, we evaluate the generalization effectiveness of our selected datasets on deep architectures different from those used in the selection process. Specifically, we employ the VGG-16 and DenseNet-121 models to train on the selected datasets from Tiny-ImageNet. As shown in Table 1, the results indicate that our method surpasses all baseline methods on both architectures, demonstrating its desirable architectural gen-



(a) 5% Corruption Rate (b) 10% Corruption Rate (c) 20% Corruption Rate

Figure 5: Comparison of robustness to corrupted images.

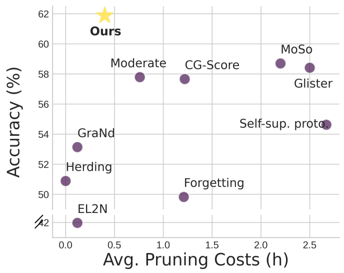
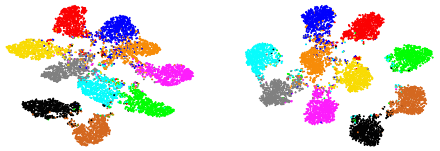


Figure 6: Effectiveness *vs.* efficiency on C-100. Results are reported with R-50 under 30% selection ratio on a 4-2080TI GPU.



(a) baseline
DI=$3.0 \times 10^{-5}$

(b) 90% dataset
DI=$4.3 \times 10^{-5}$

Figure 7: Visualization of the test set distribution. DI: Dunn Index.

eralization ability. This suggests that the selected datasets are broadly applicable, irrespective of the specific network architecture.

**Training Efficiency Comparisons** To evaluate the selection efficiency of various methods, we present an analysis of the balance between effectiveness and efficiency. As shown in Figure 6, our method presents the best performance with desirable efficiency. Herding, EL2N, and GraNd obtain the lowest selection costs because they rely on predefined metrics or select samples in very early training. Our method is slightly slower than theirs but exhibits higher accuracy. Compared with the optimization-based approaches, like MoSo and Glister, our method enjoys both lower costs and better performance. The results verify the effectiveness of our method in balancing selection efficiency and accuracy.

### 4.3 Comparison of Robustness

**Robustness on Noisy Labels** Real-world datasets often involve label noise, where some sample labels are incorrectly flipped, resulting in mislabeled data. Unfortunately, creating clean and diverse datasets is time-consuming and expensive. Therefore, it is necessary to evaluate the performance of selection methods under such complex scenarios. In this study, we introduce symmetric noises Li et al. (2022) to generate mislabeled data on both C-100 and T-ImageNet, with a 20% noise rate.

As can be seen in Table 2, our approach exhibits superior robustness to noisy labels, significantly outperforming other baselines by a large margin. Specifically, our approach yields improvements of over 10.12% on CIFAR-100 and 4.41% on Tiny-ImageNet compared to previous leading methods. Additionally, Table 2 shows the distribution of noisy data within the selected datasets, where our method selects only 0.24% noisy samples, considerably fewer than other baselines. This reduction in noise underscores the potential of our method to improve overall data quality.

Table 3: Performance and saved costs (%) on ImageNet-1k across Swin-T, ViT-B, and ViT-L on a 4-A100-GPU serve.

| Model | 80% | 90% | Full Data |
|---|---|---|---|
| ViT-B | 81.13 | **81.46** | **81.46** |
| ViT-L | 84.37 | **84.74** | 84.59 |
| Swin-T | 78.05 | **78.63** | 78.31 |
| Saved (%) | 20.62% | 10.31% | - |

Table 4: Generalization evaluation on more challenging ImageNet-1k benchmark datasets.

| Dataset | Model | 80% | 90% | Full Data |
|---|---|---|---|---|
| ImageNet-Hard | R-18 | 10.89 | **11.33** | 10.85 |
| | R-50 | 14.75 | **14.98** | 14.75 |
| ImageNet-A | R-18 | 1.65 | **2.04** | 1.12 |
| | R-50 | 3.17 | **3.31** | 3.09 |
| ImageNet-R | R-18 | 32.99 | **33.70** | 33.03 |
| | R-50 | 36.60 | **37.11** | 36.16 |

We argue that the robustness of our approach can be attributed to $\boldsymbol{S}_A$ in Eq. 1, which assesses the semantic alignment between image content and its labels. In the presence of label noise, this alignment is disrupted, resulting in a lower SAS, which in turn reduces the likelihood of such samples being selected during optimization. In contrast, most baseline methods rely solely on image features for selection, which may result in performance degradation when faced with noisy labels. In some cases, the performance of these methods is even worse than that of random selection. While Moderate also selects a relatively low proportion of noisy data, its performance is worse than ours. This discrepancy highlights the effectiveness of our method in making more strategic selections in noisy environments, thereby not only minimizing noises but also optimizing the selected datasets.

**Robustness on Corrupted Images** We further evaluate the performance of our proposed method on real-world noise corruptions that are frequently encountered Singh et al. (2024); Wei et al. (2024). To simulate such corruptions, we employ the following five types of realistic noises Xia et al. (2023), including Gaussian noise, random occlusion, resolution, fog, and motion blur. The corruption rate is set to 5%, 10%, and 20%, respectively.

As shown in Figure 5, compared with prior baselines, our approach consistently presents greater robustness to corrupted images across varying corruption rates, demonstrating strong generalization in these challenging scenarios. Notably, even at a high corruption rate of 20%, our method maintains desirable generalization performance. This robustness is primarily attributed to the integration of text modality into the selection process, alongside the image modality. The SAS defined in Eq. 1 measures the alignment between the image features and their corresponding category features. When images are corrupted, this alignment is disrupted, thereby reducing the SAS and correspondingly decreasing the likelihood of selecting those images. In contrast, methods such as Forgetting tend to prioritize difficult training samples, potentially making corrupted images more likely to be selected, as these images are typically more difficult to correctly classify. As a result, these methods are less robust to corrupted images, leading to a deterioration in generalization performance.

### 4.4 Dataset Selection Improves Generalization

**Visualization Analysis** To demonstrate the generalization of the selected datasets, we train two models using the original dataset and the selected dataset (90% selection ratio), respectively, and obtain their embedding results on the CIFAR-10 test set. To visualize the dataset distribution, we apply t-SNE to the embeddings generated by both models. The visualization in Figure 7 shows that the model trained on the selected dataset produces better embedding results: a better inter-cluster separation and intra-cluster compactness. For a quantitative analysis, we use the Dunn Index (DI) Ncir et al. (2021) to evaluate the clustering results (the higher, the better). After removing 10% of the data, the DI increases by 43%, presenting better clustering results.

**Generalization to More Advanced Architectures** We further employ the selected datasets to train more advanced ViT-based architectures, including Swin Transformer, ViT-Base, and ViT-Large. From Table 3, and corroborated by the results from previous sections, our selected datasets consistently achieve lossless performance across both CNN-based and Transformer-based architectures with reduced training costs. This demonstrates that our approach obtains highly generalizable datasets applicable to a wide range of network architectures.

**Generalization to More Challenging Benchmark Datasets** To further evaluate the generalization and robustness of models trained on our selected datasets, we conduct experiments using ResNet-18 and ResNet-50 models, training on both the full datasets and our selected datasets. These models are then tested on more challenging benchmarks, including ImageNet-Hard Taesiri

Table 5: Evaluation of our components on CIFAR-100 (C-100) and Tiny-ImageNet (T-IN).

| | w/o adapter | w/o $\mathcal{L}_{sa}$ | w/o $\mathcal{L}_{sd}$ | w/o $\mathcal{L}_{s}$ | w/o adp&$\mathcal{L}_{sa}$ | w/o adp&$\mathcal{L}_{sd}$ | w/o adp&$\mathcal{L}_{s}$ | w/o adp&$\mathcal{L}_{sd}$&$\mathcal{L}_{sa}$ | Ours |
|---|---|---|---|---|---|---|---|---|---|
| C-100 | 78.20±0.18 | 78.42±0.46 | 78.85±0.05 | 78.48±0.32 | 78.11±0.16 | 78.21±0.07 | 77.10±0.29 | 77.47±0.31 | **78.98±0.09** |
| T-IN | 46.68±0.12 | 46.79±0.39 | 49.14±0.09 | 46.01±0.38 | 47.23±0.06 | 46.70±0.33 | 45.79±0.11 | 45.69±0.10 | **49.30±0.12** |

et al. (2024), ImageNet-R Hendrycks et al. (2021a), and ImageNet-A Hendrycks et al. (2021b). The results, shown in Table 4, demonstrate that models trained on our selected data consistently exhibit superior generalization and robustness on these harder ImageNet benchmarks compared to those trained on the original datasets. Notably, this improved performance is achieved with reduced training costs, further highlighting the efficacy of our approach.

### 4.5 Ablation Study

**Effect of Dataset Adaptation** To assess the effect of dataset adaptation, instead of using fine-tuned image and text adapters, we directly utilize the per-trained CLIP model to derive the scores $\boldsymbol{S}_A$ and $\boldsymbol{S}_D$. The experimental results, presented in Table 5 with a 90% selection ratio, show a significant decline in accuracy, with drops exceeding 2% on Tiny-ImageNet. Thus, dataset adaptation is essential for effectively transferring the model's generalization ability to target datasets. This is particularly crucial for datasets that differ substantially from the pre-training datasets, such as CIFAR, where variations in image sizes and domain are distinct.

**Effect of Text Modality** Previous works Xia et al. (2023) use the average image features as the prototype and calculate the Euclidean distance between the embedded image and the corresponding prototype, which is used to filter noisy labels Wu et al. (2021). To evaluate the effect of our text modality, we also leverage this distance to assess the noisy samples and replace the text feature in Eq. 1. With a 20% noisy ratio, the accuracy drops from 46.05% to 16.39% with a 20% selection ratio and from 58.34% to 38.35% with a 30% selection ratio, validating the effectiveness of introducing textual modality. Due to the limited space, full results can be seen in Appendix L.

**Effect of Loss Terms** In Table 5, we evaluate the effect of each loss term and their combination in Eq. 6. The overall loss function achieves the highest accuracy. When $\mathcal{L}_{sa}$ is omitted, the selection process tends to prefer more diverse samples, but some class-representative samples may not be selected, which deteriorates the model performance severely. Without $\mathcal{L}_{sd}$, the selection emphasizes the most category-representative samples. Although the resulting performance drop is slightly smaller, the diversity of the selected datasets is compromised. Therefore, the incorporation of $\mathcal{L}_{sd}$ ensures a balanced representation of the selected dataset. Without $\mathcal{L}_{s}$, since we can not obtain the binarized selection decisions w.r.t. the expected selection ratios, we directly sort the scores in $\boldsymbol{d}$ and select the samples with higher scores. However, this degrades our method into a totally score-based selection and fails to address the group effect, leading to a noticeable drop in performance.

## 5 Discussion and Conclusion

**Limitation and future work.** In this section, we discuss some potential limitations and future work for our method. 1). While the pretrained CLIP model contributes to selecting the most representative samples, the potential biases in the CLIP model may propagate to the selected dataset Alabdulmohsin et al. (2024). Future work may explore bias mitigation strategies by fine-tuning CLIP with bias-aware loss functions. 2). Our work primarily focuses on vision-based datasets for data selection, using textual modalities to guide sample selection, where both modalities are balanced. However, datasets with high modality imbalance may bring challenges since it is difficult to assess the modality alignment. Thus, future work could consider leveraging dynamic modality weight allocation mechanisms and augmentation strategies for modality-imbalanced dataset selection.

This paper proposes a novel CLIP-powered data selection framework that leverages multimodal information for more robust and generalizable sample selection. To achieve this, our proposed framework incorporates three modules: dataset adaptation, sample scoring, and selection optimization modules. These modules assess the data effectiveness in model training and optimize the selection results w.r.t. the expected results. As a result, our framework is capable of selecting the most representative samples with high diversity. Extensive experiments demonstrate the effectiveness and efficiency of our approach, especially in terms of generalization performance on large-scale datasets and robustness in more challenging scenarios, such as noisy and corrupted images.

## 6 Acknowledgement

This work is supported by the STI 2030-Major Projects of China under Grant 2021ZD0201300, the Fundamental Research Funds for the Central Universities under Grant 2024300394, the National Natural Science Foundation of China under Grant 62276127. This work is supported by the Shanghai Municipal Science and Technology Major Project. This work is supported by Shanghai Artificial Intelligence Laboratory.

## References

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*, 2024.

Shadi Alijani, Jamil Fayyad, and Homayoun Najjaran. Vision transformers in domain adaptation and domain generalization: a study of robustness. *Neural Computing and Applications*, pp. 1–29, 2024.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

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

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, Aug 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3758, 2023.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

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

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

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/a78482ce76496fcf49085f2190e675b4-Paper.pdf`.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A unified framework of parameterized submodular information measures for targeted data subset selection and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, 2022.

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

Shiye Lei and Dacheng Tao. A comprehensive survey to dataset distillation. *arXiv preprint arXiv:2301.05603*, 2023.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 316–325, 2022.

Fan Liu, Tianshu Zhang, Wenwen Dai, Wenwen Cai, Xiaocong Zhou, and Delong Chen. Few-shot adaptation of multi-modal foundation models: A survey. *arXiv preprint arXiv:2401.01736*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

Songhua Liu, Jingwen Ye, Runpeng Yu, and Xinchao Wang. Slimmable dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3759–3768, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*, 2023.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020a.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020b.

Chiheb-Eddine Ben Ncir, Abdallah Hamza, and Waad Bouaguel. Parallel and scalable dunn index for the validation of big data clusters. *Parallel Computing*, 102:102751, 2021.

Ki Nohyun, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. Infonce loss provably learns cluster-preserving representations. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1914–1961. PMLR, 2023.

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

Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, pp. 17848–17869. PMLR, 2022.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, xu Zhao Pan, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=C61sk5LsK6`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ravi S Raju, Kyle Daruwalla, and Mikko Lipasti. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Srikumar Ramalingam, Pranjal Awasthi, and Sanjiv Kumar. A weighted k-center algorithm for data subset selection. *arXiv preprint arXiv:2312.10602*, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2505–2515, 2024.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. *Advances in Neural Information Processing Systems*, 36, 2024.

Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36, 2024.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pp. 1954–1963. PMLR, 2015.

Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. *arXiv preprint arXiv:2409.19696*, 2024.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128, 2009.

Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 62–71, 2021.

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

Enneng Yang, Li Shen, Zhenyi Wang, Tongliang Liu, and Guibing Guo. An efficient dataset condensation plugin and its application to continual learning. *Advances in Neural Information Processing Systems*, 36, 2023a.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. In *International Conference on Learning Representations*, 2023b.

Suorong Yang, Hongchao Yang, Suhan Guo, Furao Shen, and Jian Zhao. Not all data matters: An end-to-end adaptive dataset pruning framework for enhancing model performance and efficiency. *arXiv preprint arXiv:2312.05599*, 2023c.

Suorong Yang, Suhan Guo, Jian Zhao, and Furao Shen. Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study. *Pattern Recognition*, 148:110204, 2024a.

Suorong Yang, Furao Shen, and Jian Zhao. Entaugment: Entropy-driven adaptive data augmentation framework for image classification. In *European Conference on Computer Vision*, pp. 197–214. Springer, 2024b.

Yu Yang, Hao Kang, and Baharan Mirzasoleiman. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*, pp. 39314–39330. PMLR, 2023d.

Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11950–11959, 2023.

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

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

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

## A Analysis on the Threshold $\theta$

We present the relationship between the value of $\theta$ and the actual selection ratio gap. Suppose that we have $\mathcal{L}_s = \sqrt{\left[\frac{1}{N}\sum_i \mathrm{STE}\left[\mathbb{I}\left(\mathrm{sigmoid}(\boldsymbol{d}_i)_{\mathrm{i}} > 0.5\right)\right] - s_r\right]^2} \leq \theta$. Then, we have,

$$\frac{\|\boldsymbol{d}\|_1}{N} \leq s_r \pm \theta. \tag{7}$$

In this way, the actual selection ratio can be constrained into the $\theta$ gap. We set $\theta$ as $5 \times 10^{-4}$. For instance, on CIFAR datasets with 50000 samples, if the expected selected number is 40000, the actual selected number can be between 39975 and 40025.

## B The Specific Algorithm Workflow

To better understand the workflow of our proposed method, we present the detailed algorithm in Algorithm 1 below.

**Algorithm 1** The general workflow.

```
Input: dataset $\mathcal{D}$, total number of training samples $N$, total number of epochs $T$, selection ratio
    $s_r$, a threshold $\theta$, the pretrained image and text encoders are $E_I$ and $E_T$, the fine-tuned image
    and text adapters are $A_I$ and $A_T$, respectively.
$\boldsymbol{d} \leftarrow \mathbf{1}$
$\boldsymbol{s} \leftarrow \mathbf{0}$
for i=0:N-1 do
    Calculate the SAS $\boldsymbol{S}_A$ according to Eq. 1
    Calculate the $K$ neighbors $\boldsymbol{x'}$ of $\boldsymbol{x}_i$ using the KNN algorithm
    Calculate the SDS $\boldsymbol{S}_D$ according to Eq. 2
end for
for t=0:T-1 do
    Calculate the loss $\mathcal{L}_{sa}$ according to Eq. 3
    Calculate the loss $\mathcal{L}_{sd}$ according to Eq. 4
    Calculate the pruning loss $\mathcal{L}_s$ according to Eq. 5
    Calculate the overall loss $\mathcal{L}$ according to Eq. 6
    Update $\boldsymbol{d}$ based on the loss $\mathcal{L}$ and SGD with momentum
    if $\mathcal{L}_s \leq \theta$ then
        break
    end if
end for
Return $\boldsymbol{d}$
```

## C Implementation Details

**Datasets and Network Architectures.** Consistent with previous works Tan et al. (2024); Xia et al. (2023); Zheng et al. (2023), we evaluate the effectiveness of our proposed method on various popularly used benchmark datasets, including CIFAR-10/100 Krizhevsky et al. (2009), Tiny-ImageNet Chrabaszcz et al. (2017), and ImageNet-1k Deng et al. (2009). To evaluate the generalization performance of our selected datasets, we study the effectiveness of our proposed method on a wide range of network architectures, including ResNet-18/50 He et al. (2016), Vision Transformer (ViT) Dosovitskiy et al. (2020), Swin-Transformer Liu et al. (2021), VGG-16 Simonyan & Zisserman (2014), and DenseNet-121 Huang et al. (2017).

**Training on the Selected Datasets** Closely following previous works Xia et al. (2023); Yang et al. (2023c); Sorscher et al. (2022), for experiments on CIFAR-10/100, we adopt a batch size of 128, an SGD optimizer with a momentum of 0.9, weight decay of $5e-4$, an initial learning rate of 0.1, and a total training epoch of 200. The learning rate is divided by 5 after the 60th, the 120th, and the 160th epoch. For experiments on Tiny-ImageNet, we adopt a batch size of 256, an SGD optimizer with a momentum of 0.9, a weight decay of 1r-4, an initial learning rate of 0.1, and a total epoch of 90. The learning rate is divided by 10 after the 30th and the 60th epoch. For experiments on ImageNet-1k, following Xia et al. (2023); Sorscher et al. (2022); Yang et al. (2024b), the VISSL library Goyal et al. (2021) is exploited. We adopt a base learning rate of 0.01, a batch size of 256, an SGD optimizer with a momentum of 0.9, a weight decay of 1e-3, and a total epoch of 105. All experiments are conducted by three individual runs with different random seeds, while on ImageNet-1k, due to the huge training costs, the experiment in each case is performed once. Unless specified, the network architecture used is the ResNet-50 model. All hyperparameters and experimental settings for training on different selected datasets are kept the same.

**Fine-tuning the Adapters** We adopt the initial learning rate of $1 \times 10^{-4}$, the Adam optimizer with a step size of 30 epochs, a decay factor of 0.1, and a total epoch of 30. The batch size is set to 256 on CIFAR-10/100, 64 on Tiny-ImageNet, and 512 on ImageNet-1k.

**Selection Optimization** We adopt an SGD optimizer with a momentum of 0.9, an initial learning rate of $1 \times 10^{-3}$, and a total training iteration of $1 \times 10^{5}$. Since this optimization merely involves numerical optimization on parameter $\boldsymbol{d}$ using vanilla SGD without introducing any deep networks, the optimization can be very efficiently completed.

## D EXPERIMENTAL RESULTS ON CIFAR-10

Table 6: Test accuracy (%) on CIFAR-10 with ResNet-50.

| Method / Selection ratio | 20% | 30% | 40% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|
| Random | 84.12±1.53 | 90.34±0.39 | 92.71±0.38 | 94.43±0.37 | 95.02±0.29 | 95.55±0.14 | 95.89±0.11 | 96.12±0.12 |
| EL2N | 70.32±0.74 | 87.48±0.80 | 89.23±0.61 | 94.43±0.27 | 95.17±0.27 | 95.55±0.18 | 96.01±0.20 | 96.12±0.12 |
| MoSo | 83.33±0.47 | 89.17±0.14 | 92.47±0.14 | 94.69±0.20 | 95.50±0.00 | 95.93±0.01 | 96.26±0.02 | 96.12±0.12 |
| GraNd | 79.23±0.84 | 87.88±0.90 | 92.17±0.73 | 94.14±0.47 | 95.19±0.12 | 95.35±0.38 | 95.96±0.05 | 96.12±0.12 |
| Glister | 79.23±0.55 | 87.88±0.49 | 92.17±0.34 | 95.03±0.13 | 95.61±0.05 | 95.98±0.17 | 96.34±0.02 | 96.12±0.12 |
| Herding | 78.42±0.78 | 87.77±0.66 | 89.40±0.54 | 89.12±0.35 | 92.11±0.13 | 93.92±0.36 | 95.50±0.13 | 96.12±0.12 |
| CG-Score | 80.50±1.23 | 89.35±0.87 | 92.73±0.37 | 95.19±0.23 | 95.87±0.17 | 95.99±0.16 | 96.16±0.15 | 96.12±0.12 |
| Forgetting | 67.58±1.05 | 88.12±1.40 | 93.61±0.87 | 95.17±0.25 | 95.85±0.20 | 95.46±0.27 | 95.85±0.37 | 96.12±0.12 |
| Moderate-DS | 81.75±0.38 | 90.94±0.27 | 92.79±0.31 | 94.69±0.24 | 95.26±0.30 | 95.73±0.19 | 96.17±0.15 | 96.12±0.12 |
| Self-sup. prototypes | 84.60±1.01 | 90.07±1.14 | 92.64±0.93 | 94.42±0.72 | 94.98±0.61 | 95.87±0.53 | 95.95±0.44 | 96.12±0.12 |
| Ours | **85.70**±0.06 | **91.10**±0.26 | **93.89**±0.13 | **94.85**±0.06 | **95.43**±0.21 | **96.11**±0.18 | **96.53**±0.12 | 96.12±0.12 |

Due to the space constraint on the main paper, we provide the experimental results on CIFAR-10 in Table 6. It can be seen that our proposed method achieves superior performance across various selection ratios on CIFAR-10. Since CIFAR-10 is relatively simple to classify for ResNet-50 models, various methods show only marginal accuracy differences, particularly at higher selection ratios. However, our method consistently outperforms existing baselines by a relatively larger margin. This highlights the effectiveness of our multimodal data selection strategy in improving model performance.

## E GENERALIZATION ON VISION TRANSFORMER

To further demonstrate the generalization performance of our method on various architectures, we employ the Vision Transformer. The implementation is based on the public Github repository . Specifically, we utilize the ViT-small to train on the selected datasets. The experimental results are presented in Table 7. It can be observed that our method achieves the best performance compared to other baselines using ViT, validating the superior generalization performance on ViT architectures. Combined with the results on VGG-16 and DenseNet-121 in Section 4.2, we demonstrate that our selected datasets obtain a wide range of applications regardless of specific architectures.

https://github.com/kentaroy47/vision-transformers-cifar10

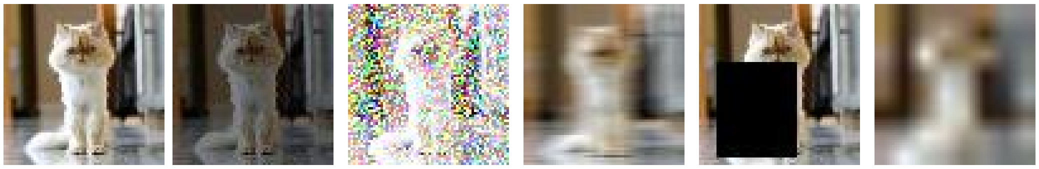
Table 7: The test accuracy (%) on CIFAR-10 with ViT-small.

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

## F Implementation Details of Training ViT-based Models on ImageNet-1k

We train ViT-Base and ViT-Large on the selected ImageNet-1k datasets based on the implementation of He et al. (2021), and we train Swin Transformer based on the implementation of Liu et al. (2021). Specifically, we fine-tune the pre-trained ViT-B and ViT-L with a batch size of 32, a base learning rate of $5e-4$, a layer decay of 0.65, and a weight decay of 0.05. We fine-tune the Swin Transformer with a weight decay of $1e-8$, a base learning rate of $2e-5$, a warmup learning rate of $2e-8$, and a batch size of 64.

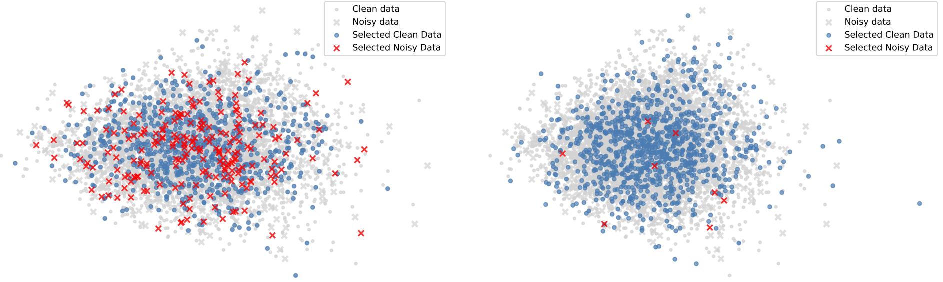
## G Illustration of Image Corruption Types



(a) Origin (b) fog (c) Gaussian (d) motion blur (e) occlution (f) resolution

Figure 8: Illustration of the image corruption types, including fog, Gaussian noise, motion blur, random occlution, and resolution.

## H Visualization of Our Method in Noisy Conditions



(a) Random selection (b) Ours

Figure 9: Illustration of the selected data in noisy conditions. The noisy ratio and selection ratio are 20%.

Table 8: Comparison of actual selection costs (h) of various methods.

| Method | Random | Herding | MoSo | Moderate | Glister | CG-Score | Forgetting | GraNd | EL2N | SSP | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CIFAR-10 | 0 | 0 | 2.21 | 0.75 | 2.55 | 1.24 | 1.23 | 0.12 | 0.12 | 2.69 | 0.40 |
| Tiny-ImageNet | 0 | 0 | 3.17 | 1.50 | 5.75 | 0.15 | 2.16 | 0.21 | 0.21 | 5.14 | 0.76 |

## I Comparison of Actual Selection Costs

In Section 4.2, we present the comparison of tradeoffs in effectiveness and efficiency. In this section, to better understand the actual high efficiency of our method, we further present a direct comparison of the actual selection costs of different selection methods. The devices used are 4 NVIDIA RTX2080TI GPUs and an Inter(R) CPU E5-2678 @ 2.50GHz. We report the average costs across five independent runs and various selection ratios in Table 8, consisting of the costs of both fine-tuning the adapter and selection optimization. Consistent with the complexity analysis in Section 3, our approach achieves a superior efficiency compared to other baselines.

## J Tradeoff between the selection costs and accuracy

In this section, we present the tradeoff between accuracy and training costs across various selection ratios on CIFAR-100. The devices used are 2 NVIDIA RTX2080TI GPUs and an Intel(R) CPU E5-2678 @ 2.50GHz. The results are presented in Figure 10, where the baseline model uses the original dataset for training. It can be seen that with the increase in the selection ratios, the training costs approximate that of the baseline model. Notably, when the selection ratios are above 80%, the selected datasets can achieve lossless accuracy with lower training costs.

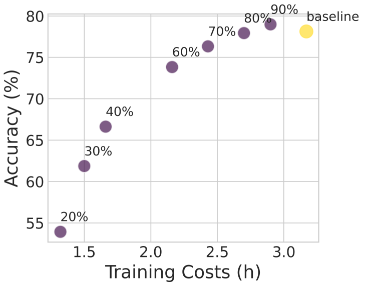


Figure 10: Relationship between accuracy and actual training costs.

It is important to emphasize that the practical gains from this approach are particularly significant in scenarios requiring the training of numerous models, such as neural architecture search (NAS) Ren et al. (2021); Zoph et al. (2018). In such applications, the reduction in training time and computational resources can result in substantial efficiency improvements, further amplifying the impact of our data selection framework.

## K More Analysis on the Threshold $\theta$

Table 9: Effect of the threshold $\theta$ on Tiny-ImageNet.

| $\theta$ | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|
| $5 \times 10^{-4}$ | 44.29% | 46.02% | 47.41% | 49.30% | 49.36% |
| $5 \times 10^{-5}$ | 44.25% | 45.98% | 47.86% | 49.17% | 49.36% |

Threshold $\theta$ adjusts the gap between the actual and expected selection ratios. Employing a smaller $\theta$ makes weights $\boldsymbol{d}$ closer to the expected selection ratios while may increase the training costs for convergence. In our main results, we typically set $\theta$ as $5 \times 10^{-4}$. In this section, we employ a smaller threshold on Tiny-ImageNet, i.e., $5 \times 10^{-5}$, which indicates that the difference in the final sample numbers is less than 5. In Table 9, we select 60%-90% of the data from Tiny-ImageNet. It can be seen that the performance is robust to the change in $\theta$. With a tighter bound, the performance exhibits minimal differences.

Table 10: Comparison of noise reduction performance using text features (Ours) vs. average image features under varying noise and selection ratios. Noise proportion means the introduced noise ratio in the selected datasets.

| | Noise Ratio (%) | 20 | | 50 | | 70 | |
|---|---|---|---|---|---|---|---|
| | Selection Ratio (%) | 20 | 30 | 20 | 30 | 20 | 30 |
| Avg. Image Feat. | Noise Proportion (%) | 16.39 | 25.35 | 20.00 | 29.95 | 20.22 | 30.16 |
| | Accuracy (%) | 28.42 | 38.35 | 16.56 | 23.19 | 11.18 | 14.61 |
| Ours | Noise Proportion (%) | **0.24** | **0.32** | **0.43** | **0.68** | **0.80** | **4.30** |
| | Accuracy (%) | **46.05** | **58.34** | **52.56** | **60.72** | **51.50** | **56.80** |

## L Effectiveness of Text Modality

Based on single-modal features, existing methods Xia et al. (2023) use the average image features as the prototype and calculate the Euclidean distance between the embedded image and the corresponding prototype. This distance is used to filter noisy labels. To further evaluate the effectiveness of our multi-modal framework, we use the average image feature to replace the text feature in Eq. 1, which is then used for selection optimization in Eq. 6.

In Table 10, we evaluate the noise robustness across various noise and selection ratios. It can be seen that using average image features leads to higher noise ratios compared to our method. This validates the effectiveness of complementary textual information provided by text features, leading to superior denoising performance.

## M Numerical Analysis on the Selected Datasets

Table 11: Actual selection ratios on both Tiny-ImageNet and CIFAR-100.

| Expected Ratios | 20% | 30% | 40% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|
| Tiny-ImageNet | 20.03%↑ | 30.03%↑ | 39.98%↓ | 59.97%↓ | 70.01%↑ | 80.02%↑ | 89.99%↓ |
| CIFAR-100 | 19.95%↓ | 30.02%↑ | 40.00%− | 60.04%↑ | 70.03%↑ | 80.04%↑ | 90.04%↑ |

We present the numerical analysis of the actual selection ratios on Tiny-ImageNet in Table 11. It can be seen that the actual selection ratios may be slightly higher or lower than the expected values, with minimal deviations. These deviations are within the theoretical gap bounds. Notably, when the actual selection ratios are below the expected ones, our method **selects fewer samples compared to other baselines**. Despite this reduction in training data volume, our method still achieves the best performance, highlighting the effectiveness of the selection strategy. Consequently, the performance improvements achieved are not due to the slight increases (e.g., less than 0.4% on average) in sample numbers but rather due to the strategic selection of informative samples, demonstrating the efficacy of our selection algorithm in optimizing the selected samples.