Look but Don't Touch: Gradient Informed Selection Training

David Mayo*
MIT
dmayo2@mit.edu

Chris Zhang*
MIT
chriszh@mit.edu

Boris Katz
MIT
boris@mit.edu

Andrei Barbu
MIT
abarbu@mit.edu

Brian Cheung
MIT
cheungb@mit.edu

Abstract

The amount of data available for training foundation models is far greater than our amount of compute. In many domains, this will likely always be the case. Further, not all data is equally valuable for learning, and the learning value of data changes over the course of training. To optimize learning in this setting, several active data selection methods have been proposed; however, they either incur significant additional computational costs or offer limited performance benefits. We propose Gradient Informed Selection Training (GIST), an active data selection method that selects a core subset of examples from mini-batches based on their gradient alignment with a small, fixed holdout set taken from the training set. At each training step, GIST computes per-example gradients and selects only those that are most aligned with the holdout gradient, thereby guiding model updates toward better generalization. On the large, noisy web-scraped image dataset Clothing-1M, GIST trains in 3x faster wall clock time, using 6x fewer steps, and achieves 4% higher final accuracy than RHO-LOSS and uniform data selection.

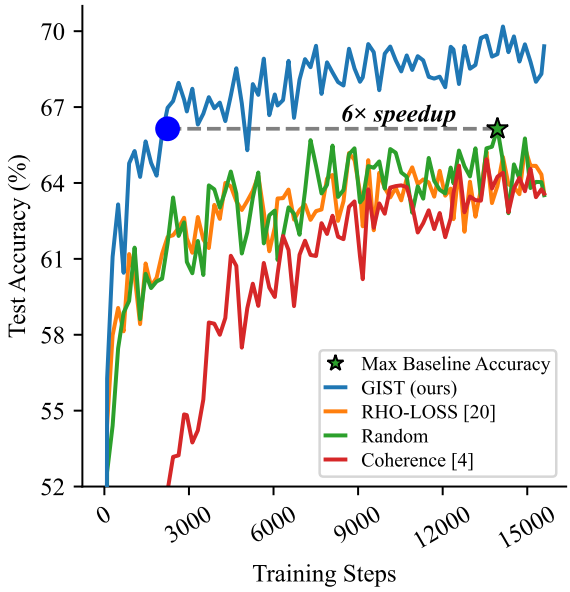

Figure 1. **Acceleration in training for large-scale noisy datasets (Clothing-1M)**. GIST trains in 6x fewer steps and achieves at least 4% accuracy gain across all other baselines (RHO-LOSS, random selection, gradient coherence).

1. Introduction

Modern machine learning models are growing at an unprecedented rate in both size and capability. Foundation models such as CLIP [25] and GPT [3] demonstrate that scaling model size and dataset size can significantly improve generalization across diverse tasks. In particular, we focus on vision model training, where the signal-to-noise ratio in data is often lower than in text or structured domains. Real-world vision datasets, such as those collected for applications like autonomous driving, frequently suffer from label noise, occlusions, near-duplicate frames, and other imperfections that are harder to clean at scale [11, 19, 34]. These challenges make vision tasks especially sensitive to the quality and selection of training data.

Large models can sometimes compensate by learning from vast amounts of data, but this strategy comes at substantial computational cost and can slow convergence or degrade final performance. Selectively focusing on higher-quality or more informative examples offers a promising way to address this inefficiency.

It is well understood that not all training examples contribute equally to learning. Techniques such as curriculum learning, where examples are presented in a progression from easy to difficult [2], emphasize the role of the order and quality of training data. However, in practice, the improvements achievable through curriculum learning are limited, both in terms of final accuracy and the practical difficulty of properly ranking samples and pacing their introduction during training [27].

This motivates a shift toward active data selection, which offers a more adaptive approach to prioritizing examples.

*Equal contribution

Recent methods such as RHO-LOSS [20] and InfoBatch [24] have shown that focusing on valuable examples during training can accelerate convergence and improve generalization, particularly in noisy settings. However, we hypothesize that current methods still fall short of realizing the full potential of active selection. Our goal is to push closer to this upper bound and quantitatively demonstrate that a better "curriculum" exists—specifically, that there is a more effective way to prioritize data to enhance model learning.

In this work, we provide the following contributions:

- We propose an active data selection approach, **Gradient Informed Selection Training (GIST)**, designed to identify and prioritize the most valuable training examples based on gradient information. Through experiments on noisy datasets such as Clothing1M [32], we show that even in the presence of significant label noise, a better training curriculum exists. Our method outperforms prior state-of-the-art methods (i.e. RHO-LOSS), demonstrating that a substantial performance gap still remains between current selection strategies and the optimal subset selection.
- We analyze the performance of data selection methods across a range of **selection fractions**, defined as the proportion of examples chosen from each large batch to form the smaller training batch. We find that GIST outperforms RHO-LOSS at all selection fractions greater than 0.3. At 0.6 GIST trains in 3x faster wall clock time, using 6x fewer steps, and reaches 4% higher final accuracy.

2. Related Work

2.1. Data Selection Approaches

Curriculum learning is one of the earliest structured data selection strategies. First introduced by Bengio et al. [2], it involves presenting training examples in a progression from easy to difficult, mimicking the way humans learn. This staged exposure has been shown to accelerate convergence, improve generalization, and enhance model robustness across various domains including vision and language [14, 22, 31, 35, 36].

Data subset selection refers to methods aimed at identifying smaller, representative subsets of a larger dataset to reduce computational cost while retaining performance. It differs from data cleaning, which primarily removes or corrects erroneous and mislabeled samples [23, 26] and active data selection, which dynamically adjusts the training data based on feedback during training. Subset selection emphasizes the representativeness of the chosen samples, often trading off exploration (diversity of samples) against exploitation (selecting samples closely aligned with the current model's needs), a balance extensively studied within reinforcement learning paradigms [28].

There have been significant advances in data subset selection methods [6, 13, 15, 21, 30]. More recently, research has shifted toward active data selection, where the training set is dynamically filtered or prioritized based on model feedback [7, 9, 17, 20, 24, 33]. The Reducible Holdout Loss (RHO-LOSS), proposed by Mindermann et al. [20], prioritizes examples with high reducible loss: those whose errors can still be improved with further training. This is done through a teacher-student framework: the teacher estimates the irreducible loss of each sample, while the student updates on the most promising ones. RHO-LOSS has been shown to increase training efficiency and improve generalization, and has been extended to large-scale multimodal training in recent work [29]. In this work, we propose a simpler, one-network method for active data selection as opposed to Rho-LOSS's two model teacher-student framework.

2.2. Gradient-Based Methods

Gradient-based data selection approaches aim to retain only those examples whose gradient contributions are most representative or beneficial [16, 18]. Grad-Match, introduced by Killamsetty et al., selects subsets of training data by minimizing the difference between the full-batch gradient and that computed on a subset [16]. The method identifies a small group of samples whose average gradient closely matches that of the complete dataset, enabling reduced computational overhead while maintaining strong performance.

Gradient coherence, proposed by Chatterjee and Zielinski [4, 5], measures the alignment between gradients of individual samples during training. High coherence means that examples push the model in similar directions, which should more strongly support generalization; low coherence, by contrast, signals conflicting dataset features leading to overfitting. In this work, we also experiment with using coherence to actively select examples that collectively steer the model towards having robust learning dynamics during training.

3. Gradient Informed Selection Training (GIST)

The success of methods such as RHO-LOSS [20] demonstrates prioritizing training examples expected to yield the highest learning gain can significantly improve model performance, especially in noisy or complex datasets. Yet, current strategies fall short of the performance upper bound achievable if we could always train on the most beneficial examples. This work asks: can we design a method that more effectively identifies high-value training samples to move closer to this ideal?

To approach the theoretical ideal of training on the most informative examples, we propose using a small, fixed holdout set drawn from the training distribution as a stable proxy

for the test set. Our method selects training examples whose gradients are best aligned with the average gradient from this holdout set. This alignment encourages model updates that generalize better, without relying on any information from the test set.

Figure 2 provides visual intuition: rather than relying on test set gradients (which are unavailable during training), GIST selects training examples based on how closely their gradients align with the direction of a holdout set gradient. Because the holdout set is never directly used for updates, it encourages generalization without overfitting. By curating this holdout set (e.g., selecting clean or non-ambiguous examples) we can further steer learning toward robustness. While this may risk excluding some long-tail but correct datapoints, we find in practice that the trade-off improves both training efficiency and resilience to label noise. Additionally, we observe that using only last-layer gradients for similarity comparison retains model performance while significantly reducing computational overhead. See Algorithm 1 for the formal procedure and Figure 3 for a visual overview.

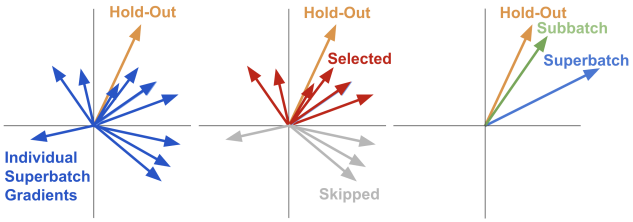

Figure 2. Visual intuition for GIST. Each blue arrow represents the gradient direction, projected into 2 dimensions, for individual training examples within a large batch (**superbatch**). We expect the resultant selected **subbatch** average gradient (represented by the green arrow) of training examples to better approximate the ideal gradient, compared to the gradient computed across the entire superbatch.

4. Evaluation on Clothing-1M

We evaluated GIST and existing baselines on the Clothing-1 Million dataset, chosen for its noisy real-world challenges. It's the most widely accepted benchmark for image recognition with noisy labels [1].

We choose the following baselines to benchmark against:

- **Random:** A simple baseline where subbatches are selected uniformly at random from each superbatch. This serves as a lower-bound reference for selection-based methods.
- **RHO-LOSS:** A teacher-student framework that prioritizes examples with high reducible loss—those for which the model is still expected to improve [20]. It has shown state-of-the-art results in noisy settings at low selection fractions, making it a strong baseline on Clothing1M.

Algorithm 1 Gradient Informed Selection Training (GIST)

Require: Small holdout set $\mathcal{D}_{\text{ho}}$ (class-balanced, $\ll$ total training size), batch size n_b, superbatch size $n_B > n_b$, learning rate η

Initialize model parameters θ^0 and time step $t = 0$
for $t = 0, 1, 2, \ldots$ **do**
  Sample a minibatch $\mathcal{D}_{\text{ho}}^{(t)} \subset \mathcal{D}_{\text{ho}}$
  HOLDOUTGRAD $\leftarrow$ last-layer gradient of $L(\mathcal{D}_{\text{ho}}^{(t)}; \theta^t)$
  Randomly sample a superbatch $\mathcal{B}_t$ of size n_B
  for each (x_i, y_i) in $\mathcal{B}_t$ **do**
    TRAINGRAD$[i] \leftarrow$ last-layer gradient of $L(y_i \mid x_i; \theta^t)$
    SIMILARITY$[i] \leftarrow \cos\big(\text{TRAINGRAD}[i], \text{HOLDOUTGRAD}\big)$
  end for
  $b_t \leftarrow$ top-n_b examples in $\mathcal{B}_t$ with highest SIMILARITY
  $g_t \leftarrow$ minibatch gradient on b_t using parameters θ^t
  $\theta^{t+1} \leftarrow \theta^t - \eta g_t$
end for

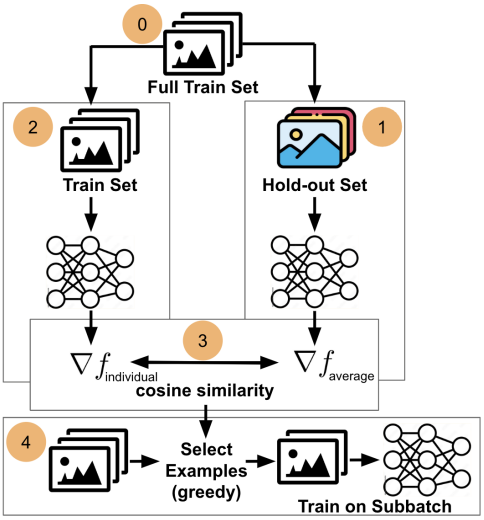

Figure 3. Overview of GIST.

- **Coherence:** A gradient-based ablation of GIST that removes the holdout set and instead aligns individual gradients with the average gradient of the current superbatch. This method selects examples whose gradients have the highest cosine similarity to the superbatch gradient, thereby prioritizing coherence within the current training batch rather than generalization to an external validation set.

4.1. Experiment Results

We define the **selection fraction** as the proportion of data selected from the superbatch to form the subbatch used for

training. We evaluate all selection fractions from 0.10 to 1.0, in increments of 0.10. Results are shown in Figure 1 for a selection fraction of 0.60. We verify that accuracy gains outweigh additional computational overhead from GIST in Figure 4. We also report the accuracies and speedup for each of these methods at all selection fractions in Figure 5. Full experiment details can be found in Appendix A.

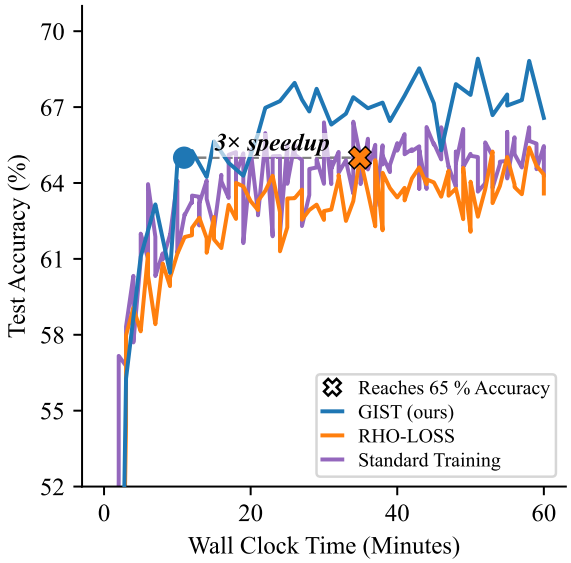

Figure 4. **Comparison of classification accuracy versus training compute cost (clock time) on Clothing-1M at a 0.60 selection fraction.** GIST achieves higher accuracy than baseline methods RHO-LOSS and Standard Training (batch size = selected subbatch size of other methods) when computational cost is held constant across methods.

GIST outperforms all baselines across all selection fractions greater than 0.3, demonstrating its effectiveness in identifying high-value training examples. Meanwhile, RHO-LOSS only outperforms at selection fractions 0.1 and 0.2. Our method surpasses all baselines by up to 4% with a training speedup of up to 6x.

5. Conclusion

In this work, we introduced Gradient Informed Selection Training (GIST), a method for active data selection that leverages a small holdout set to guide training via gradient alignment. We demonstrated that aligning per-example gradients from each training superbatch with the holdout set gradient leads to strong selection of training samples, outperforming baselines such as RHO-LOSS, Coherence, and Random selection. Our method shows strong results on both noisy datasets like Clothing1M and does so without requiring teacher models, auxiliary objectives or heuristics.

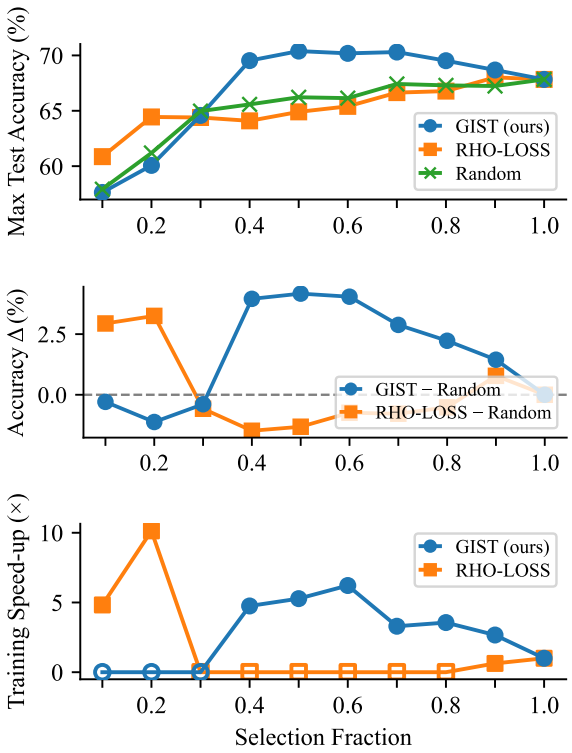

Figure 5. **Top:** Max test accuracy achieved across selection fractions; **Middle:** Difference in accuracy between GIST/RHO-LOSS and random; **Bottom:** Training speedup across selection fractions. Computed by dividing the step at which random selection achieves max accuracy by the step at which the alternative method exceeds that accuracy. Speedup is 0 if it never achieves a greater max accuracy (marked by no-fill points).

6. Discussion and Future Work

One promising application of GIST is learning in privacy-sensitive settings. In scenarios where a confidential or copyrighted dataset cannot be used directly for training (due to legal, ethical, or security constraints) GIST enables indirect training by selecting subbatches of general training data that approximate the gradient directions of the confidential set. Since the model never directly optimizes over the private data nor stores any of its examples or labels, this approach could satisfy certain compliance requirements while enabling more generalizable learning. Potential applications include medical diagnosis, governmental applications, and federated learning, where privacy-preserving training is crucial.

Another avenue for future research is improving the quality of the holdout set. While we currently use a small random subset, future work could explore construction of the holdout set using dataset distillation or optimized processes. Additionally, adaptive holdout sets that evolve during training may allow GIST to better capture shifting model uncertainties over time.

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. In *Knowledge-Based Systems*, page 106771. Elsevier BV, 2021. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 1, 2

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. 1

[4] Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. 2

[5] Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning, 2022. 2

[6] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *ICLR*, 2020. 2

[7] Zhijie Deng, Peng Cui, and Jun Zhu. Towards accelerated model training via bayesian data selection. In *Advances in Neural Information Processing Systems*, 2023. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 7

[9] Guy Hacohen and Daphna Weinshall. How to select which active learning strategy is best suited for your specific problem and budget. In *Advances in Neural Information Processing Systems*, 2023. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2015. 7

[11] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, I. Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *ArXiv*, abs/2211.01866, 2022. 1

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*. 7

[13] Eeshaan Jain, Tushar Nandy, Gaurav Aggarwal, Ashish Tendulkar, Rishabh Iyer, and Abir De. Efficient data subset selection to generalize training across models: Transductive and inductive networks. In *Advances in Neural Information Processing Systems*, pages 4716–4740. Curran Associates, Inc., 2023. 2

[14] Ayrton San Joaquin, Bin Wang, Zhengyuan Liu, Nicholas Asher, Brian Lim, Philippe Muller, and Nancy F. Chen. In2core: Leveraging influence functions for coreset selection in instruction finetuning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. 2

[15] Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 2

[16] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*. 2

[17] 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*, 2021. 2

[18] Krishnateja Killamsetty, Guttu Sai Abhishek, Aakriti, Alexandre V. Evfimievski, Lucian Popa, Ganesh Ramakrishnan, and Rishabh Iyer. Automata: Gradient based data subset selection for compute-efficient hyper-parameter tuning. In *Advances in Neural Information Processing Systems*, 2022. 2

[19] Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C. Knoll. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. In *IEEE Transactions on Intelligent Vehicles*, 2024. 1

[20] Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 2, 3, 7

[21] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning (ICML)*. 2

[22] Hiep Nguyen, Lynn Yip, and Justin DeBenedetto. Automatic quality estimation for data selection and curriculum learning. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 212–220, Miami, FL, USA, 2024. Association for Computational Linguistics. 2

[23] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Con-

fident learning: Estimating uncertainty in dataset labels. In *Journal of Artificial Intelligence Research*, 2022. 2

[24] Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[26] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 2

[27] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. In *International Journal of Computer Vision*, 2022. 1

[28] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. 2

[29] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2

[30] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes. Submodular subset selection for large-scale speech training data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3311–3315, 2014. 2

[31] Biao Wu, Fang Meng, and Ling Chen. Curriculum learning with quality-driven data selection, 2024. 2

[32] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015. 2

[33] Yichen Xie, Mingyu Ding, Masayoshi TOMIZUKA, and Wei Zhan. Towards free data selection with general-purpose models. In *Advances in Neural Information Processing Systems*, pages 1309–1325. Curran Associates, Inc., 2023. 2

[34] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021. 1

[35] Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data, 2024. 2

[36] Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Shengnan Liu, and Wenwu Zhu. CurBench: Curriculum learning benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62088–62107. PMLR, 2024. 2

A. Appendix

A.1. Clothing-1M Experiment Details

All models are pretrained on ImageNet, following the setup in RHO-LOSS [20], and then fine-tuned on Clothing-1M for a total of 6 effective training epochs. We define one effective epoch as processing a number of samples equal to the full dataset size. For example, with a subset selection fraction of 0.10, each training pass sees (updates weights on) only 10% of the data, so 60 such passes are needed to match the sample count of 6 full epochs. For the holdout set, we randomly split out 200 images per class (2800 total), which is 0.28% of the entire training set size.

A.2. Models

We use the ViT-Small architecture [8] for all experiments, as it offers a balance between performance and efficiency. Moreover, unlike architectures that include BatchNorm layers [12] (e.g., ResNets [10]), ViTs enable gradient averaging without batch-dependent side effects, which is important for our selection strategy. For the RHO-LOSS experiments, we use a smaller ViT-Tiny model as the teacher.

A.3. Hyperparameters

All models are trained using the AdamW optimizer with β_1 = 0.9, β_2 = 0.999, weight decay = 0.05, eps = 1e-8, learning rate = 0.001. For all experiments, we use a batch size of 320, following the RHO-LOSS setup for reproducubility. We finetuned all weights, not just the final layer.