

Evaluating Sample Utility for Data Selection by Mimicking Model Weights

Tzu-Heng Huang^{*†}

Manjot Bilkh^{†‡}

Frederic Sala[†]

Javier Movellan[†]

[†]University of Wisconsin-Madison

[‡]Apple Inc.

[†]{thuang273, fredsala}@wisc.edu

[‡]{mbilkh, movellan}@apple.com

January 8, 2025

Abstract

Foundation models rely on large-scale web-crawled datasets, which frequently contain noisy data, biases, and irrelevant content. Existing data selection techniques typically use human heuristics, downstream evaluation datasets, or specialized scoring models, and can overlook samples’ utility in the training process. Instead, we propose a new approach, **Mimic Score**, a data quality metric that uses a pretrained reference model as a guide to assess the usefulness of data samples for training a new model. It relies on the alignment between the gradient of the new model parameters and the vector pointing toward the reference model in weight space. Samples that misalign with this direction are considered low-value and can be filtered out. Motivated by the Mimic score, we develop **Grad-Mimic**, a data selection framework that identifies and prioritizes useful samples, automating the selection process to create effective filters. Empirically, using Mimic scores to guide model training results in consistent performance gains across six image datasets and enhances the performance of CLIP models. Moreover, Mimic scores and their associated filters improve upon existing filtering methods and offer accurate estimation of dataset quality.

1 Introduction

Large-scale web-crawled datasets are fundamental to the success of foundation models, e.g., MM1.5 [1], OpenAI CLIP [2], LLaMa-3.1 [3], and GPT-4 [4]. These datasets provide vast quantities of information but also carry noise, biases, and irrelevant content from their web sources. To mitigate these, data selection—ruling out undesirable samples—has emerged as a critical step in the model development pipeline [5; 6]. For example, the FineWeb dataset [7], containing 15 trillion tokens used for training large language models, is created through eight carefully designed filtering steps meant to refine raw web content, ensuring high model performance.

While effective, such filtering recipes face several limitations. The choices of filtering steps and rules often rely on handcrafted heuristics and require domain expertise and expensive experimentation. Such filters often fail to provide fine-grained insights into individual samples and overlook samples’ utility in training. These lead to coarse selection and suboptimal model performance.

Existing techniques often suggest selecting samples based on semantic similarity to downstream evaluation datasets [8; 9], using specialized filtering networks [10; 11; 12; 13], or training influence models to score samples [14; 15]. Yet, these approaches require access to additional datasets or involve specialized model training, adding dependencies and complexity.

We introduce the **Mimic Score**, a new data quality metric that assesses sample contribution in the weight updates using a *pretrained reference model*. We show that by leveraging a model located in a more optimal part of the weight space, we bypass the need for downstream datasets or specialized training, and can use it as a selection guide. Our approach is based on the alignment between each sample’s gradient and the direction towards reference model in weight space.

^{*} Apple Work done during an internship at Apple.

[†] Corresponding author. Email: mbilkh@apple.com & thuang273@wisc.edu.

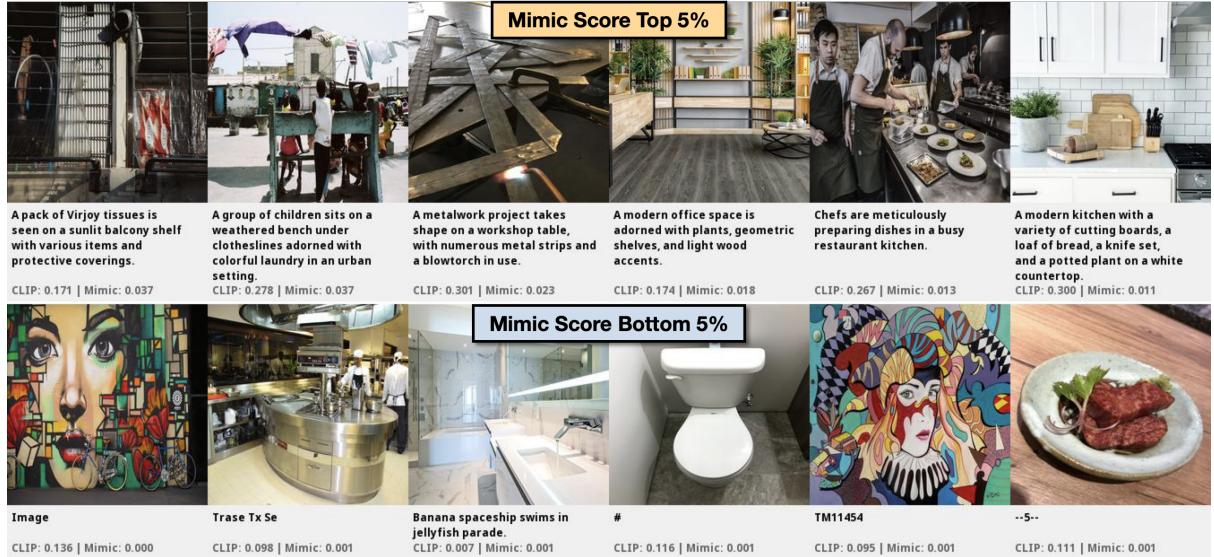


Figure 1: **High-value vs. Low-value Samples Identified by Mimic Score:** We present randomly selected samples from the top 5% (*first row*) and bottom 5% (*second row*) of web-crawled data, ranked by their mimic scores. Below each image, we show their caption and corresponding CLIP score (where a higher score suggests good quality [17]). Mimic scores closely align with their CLIP scores. High-value samples generally have detailed captions and coherent visual content, while low-value ones carry short captions and misaligned content.

Samples that potentially pull the model in undesirable directions—misguiding weight updates—are considered low-value and should be filtered out.

Building on the mimic score, we develop ***Grad-Mimic***, a data selection framework that operates in two stages:

- **Stage 1:** During training, Grad-Mimic prioritizes samples to learn by using mimic score, steering the model toward the reference model’s weight space.
- **Stage 2:** After training, Grad-Mimic identifies sample utility across training steps and aggregates these assessments for an ensemble filter, automating data selection.

We validate the effectiveness of Grad-Mimic through various experimental setups, presenting empirical results at each stage. First, we create a synthetic and controlled scenario that adds different levels of label noise. We demonstrate its capability to accurately identify mislabeled samples and provide a reliable estimation of overall dataset quality. Compared to the standard training method, Grad-Mimic successfully down-weights the contribution of undesirable samples and enhances model performance across six image datasets.

Next, we test Grad-Mimic in more challenging settings: using large-scale web-crawled multimodal data. We train CLIP models [2] from scratch on 10 million and 100 million samples from the DataComp dataset [16]. We use publicly available CLIP model weights as our reference and demonstrate that mimic scores help navigate training, resulting in performance gains across both scales. Compared to human-designed filters, our refined dataset achieves higher model performance. Furthermore, mimic scores complement CLIP score-based filters [17] by removing low-value samples, lifting model performance with less data.

We summarize our contributions as follows:

- **New Data Quality Metric:** We propose a novel approach to assess sample utility by computing the alignment between per-sample gradient and the vector toward the reference model’s weights.
- **Automated Data Selection Framework:** Building on the mimic score, we present Grad-Mimic, a framework that effectively identifies useful samples to improve training and automates effective data selection.
- **Dataset Quality Assessment:** The mimic score, along with its developed filters, improves existing filtering strategies and provides reliable estimations of dataset quality.

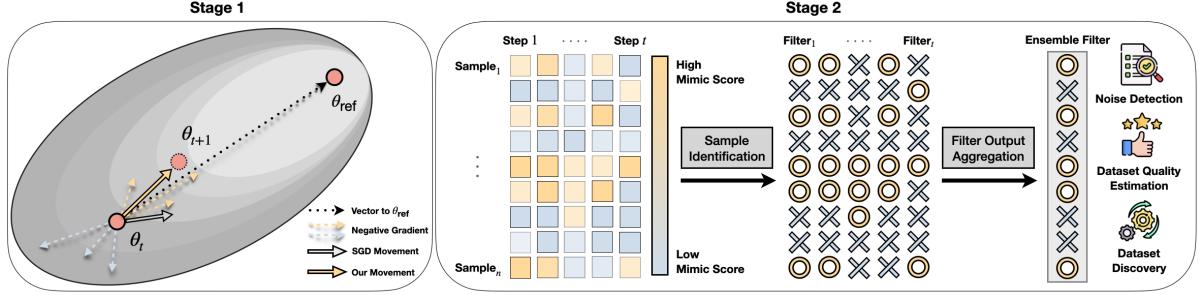


Figure 2: Grad-Mimic Two-stage Workflow: In Stage 1, Grad-Mimic uses the vector pointing toward the reference model’s weight space to measure alignments with each sample’s negative gradients. Then it adapts alignment scores to reweight gradients that contributed to the weight movements. In Stage 2, derived alignment scores, mimic scores, are used to identify low-value samples and build an ensemble filter, enabling accurate noise detection, precise dataset quality assessment, and training dataset discovery.

2 Related Work

Data Selection. Data selection techniques can generally be categorized into group- and sample-level approaches. Group-level approaches focus on optimizing the mixture of data domains to curate high-quality datasets [18; 19; 20; 21]. Sample-level methods, which are the focus of Grad-Mimic, aim to filter out individual noisy and unhelpful samples. Prior research has explored various strategies, including identifying impactful samples through gradient magnitudes [22], analyzing gradient similarities across batches [23], and selecting key samples that can capture full training update [24; 25]. These methods often rely on batch-level gradients, which can struggle with large amounts of noise in a mini-batch. Grad-Mimic addresses this challenge with the help of a reference model, acting as a reliable guide to select samples.

Data Curation for Multimodal Models. Multimodal models are typically trained on large-scale web datasets [26; 27; 16; 28], which are often noisy and require careful curation. Previous approaches have included selecting samples based on their semantic similarity to the downstream evaluation datasets [8; 9], using semantic deduplication [29], developing specialized filtering networks [10; 11; 12; 13], or training influence models [14; 15]. While effective, they require access to target datasets or introduce additional training complexities. Grad-Mimic overcomes these limitations and offers a more efficient alternative: using pretrained weights to identify useful samples.

Weak Supervision. Weak supervision is an emerging paradigm for constructing labeled datasets by combining multiple noisy label estimates [30; 31; 32; 33; 34]. These estimates typically come from sources such as heuristic labeling rules, domain knowledge, or pretrained models [35; 36], often encoded as labeling functions. The outputs of these labeling functions are modeled to assess their reliability and then aggregated to produce final labels [30]. Weak supervision has demonstrated success in various domains, including segmentation tasks [37], relation extraction [38], MRI sequence analysis [39], sensor data [40], chemical reaction extraction [41], and enhancing predictions from pretrained models [42; 43]. Unlike most existing works focusing on label aggregation for data annotation, Grad-Mimic uses weak supervision for filtering purposes. We mitigate assessment noise across training steps, aggregating filter outputs into an effective ensemble filter.

3 Evaluating Sample Utility

We quantify samples by their contributions to the learning process. *Our principle is that samples that potentially pull the model in undesirable directions, thereby misdirecting weight updates, should be considered low-value.*

We start with setup and notation, then explain how our scoring metric is derived and used in our data selection framework, Grad-Mimic.

Notation. Let $D = \{s_i\}_{i=1}^n$ denote a dataset of n samples drawn from a distribution \mathcal{D} supported on the space \mathcal{S} . At training step t , model parameters θ_t , are iteratively optimized using the dataset D . While our framework supports various training settings, we focus on supervised learning for clarity. We assume $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$, where \mathcal{X} is the input space and \mathcal{Y} is the label space. Each sample s_i can be expressed as (x_i, y_i) , where noise may be

present either in the instance x_i or in the label y_i . The empirical loss across samples is defined as $\frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i)$, and each sample's gradient with respect to the model parameters θ_t is written as: $g_{i,t} := \nabla_{\theta_t} \ell(x_i, y_i)$. A standard update to model parameters is $\theta_{t+1} := \theta_t - \eta \sum_{i=1}^b g_{i,t}$, where η is the learning rate and b is the batch size.

3.1 Mimic Score Calculation

To evaluate whether a sample steers the model in an undesirable direction, we use a pretrained model as a reference. This reference model, denoted by θ_{ref} , resides in a more optima part of the weight space, such that $\ell(\theta_{\text{ref}}) < \ell(\theta_t)$. We use the vector from the current model's weight space θ_t to θ_{ref} to measure each sample's utility in approximating a better weight configuration. These reference weights can be layer-specific, e.g., model weights in the last layer, which usually store more informative features [44; 45].

The reference model can be obtained in two ways: either by training it on the dataset to achieve a target performance, if resources permit, or more efficiently, by using publicly available pretrained models, which eliminate the need for data access and expensive training.

The vector pointing toward the reference model, at training step t , is represented by $v_t := \theta_{\text{ref}} - \theta_t$. We examine how each sample's negative gradient $-g_{i,t}$, intended for updating model weights, aligns with vector v_t . We measure the alignment degree by considering both the direction and magnitude of the negative gradient. Specifically, we compute the projection length of $-g_{i,t}$ onto v_t , yielding an alignment score $m_{i,t}$, computed as follows

$$\text{mimic_score}(s_{i,t}) := m_{i,t} = \frac{\langle -g_{i,t}, v_t \rangle}{\|v_t\|_2}. \quad (1)$$

This alignment score, named the *mimic score*, reflects how much a sample can drive the model closer to the reference model. A sample having a lower mimic score suggests it has limited utility in guiding model updates, making it a potential candidate for exclusion from future training.

4 Grad-Mimic: Data Selection Framework

The mimic score represents each sample's usefulness in guiding the model in a better direction. Building on this metric, we propose *Grad-Mimic*, a two-stage data selection framework that first prioritizes samples to learn and then identifies useful samples to design an effective filter. A complete Grad-Mimic workflow is illustrated in Fig. 2.

4.1 Stage 1: Reweighted Gradient Updates

We first use mimic scores to aid model training by re-weighting sample gradients. Unlike standard gradient descent, which assigns equal weight to all the samples in the mini-batch, Grad-Mimic uses mimic scores to amplify helpful samples and down-weight unhelpful ones.

To achieve this, each sample's mimic score is first normalized using the softmax function with a temperature parameter, τ . The normalized score for a sample s_i in a batch of size b is computed as

$$\bar{m}_{i,t} = \frac{e^{-m_{i,t}/\tau}}{\sum_{j=1}^b e^{-m_{j,t}/\tau}}. \quad (2)$$

Then, the weight update step in Grad-Mimic is modified as

$$\theta_{t+1} := \theta_t - \eta \sum_{i=1}^b \bar{m}_{i,t} \cdot g_{i,t}. \quad (3)$$

The temperature τ controls the sensitivity of sample reweighting, allowing us to adjust how sharply the model prioritizes samples. A lower temperature results in a more aggressive focus on learning the most aligned samples, while a higher temperature encourages Grad-Mimic converges to standard gradient descent.

4.2 Stage 2: Automated Filter Design

The second stage of Grad-Mimic uses computed mimic scores to identify valuable samples, automate data selection based on these assessments, and estimate dataset quality.

4.2.1 Sample Identification

Grad-Mimic first gathers normalized mimic scores for each sample at every training step. These scores indicate the sample contributions over training, which allows us to decide whether to retain or discard a sample. Grad-Mimic supports several sample identification methods:

- *Threshold-based Selection*: A sample is chosen if its mimic score exceeds a defined threshold. This threshold could be set as $1/b$ (indicating greater than uniform weight).
- *ID Clustering*: Samples are categorized into two groups using clustering techniques such as k -means clustering [46] or Gaussian Mixture Models (GMMs), allowing more automated selection based on their assigned cluster.
- *Top- k Percent Selection*: Samples are ranked by mimic score, and the top- k percent are chosen. The value of k can be adjusted based on the available training budget.

After identification, we binarize mimic scores and categorize samples into “retain” or “discard” at each training step. We treat these assigned groups as filter outputs and then combine them for a final filter.

4.2.2 Filter Output Aggregation

While filtering at an individual training step can refine the dataset, relying on isolated filters risks overlooking training dynamics. Additionally, considering sample utility across different stages may yield complementary signals.

To address this, Grad-Mimic leverages weak supervision techniques [30; 31; 32; 33; 34], which are usually used for constructing datasets from noisy signals, to combine filter outputs across training steps. We begin with learning a model to evaluate the reliability of each step assessment [32]. Once established, this model aggregates filtering decisions into a high-quality ensemble filter. Through aggregation, Grad-Mimic mitigates assessment noise and captures training dynamics. Grad-Mimic employs the Snorkel framework [31], a widely used method in the weak supervision community.

Ultimately, the ensemble filter selects high-value samples for a refined dataset. We can estimate dataset quality by analyzing the proportion of retained samples or using samples’ mimic scores as an overall metric. We summarize steps in Grad-Mimic in Algorithm 1 placed in Appendix B.

5 Experiments

We assess the effectiveness of mimic score within Grad-Mimic using two experimental setups across datasets of varying scales and domains. We first test in a controllable setting, adding noise into sample labels (Sec. 5.1), followed by evaluations on large-scale web-crawled datasets (Sec. 5.2). Our goals are to validate the following claims in both setups:

- **Enhanced Model Performance**: The reference model acts as a reliable guide. Training a new model with samples prioritized by mimic scores enhances performance.
- **Accurate Sample Identification**: Grad-Mimic effectively detects noisy samples, automates data selection, and complements existing filtering methods.
- **Dataset Quality Assessment**: Mimic scores and their associated filters offer reliable estimates of dataset quality, highly correlated with noise levels and performance gains.

5.1 Simulating Mislabeled Samples

Setups. We begin with a controlled experiment by adding various levels of label noise to six image classification datasets. They are DTD [47], Flowers102 [48], STL10 [49], OxfordIIIT Pet [50], CIFAR10 [51], and CIFAR100 [51]. We fine-tune ViT-B/16 models [52] on these noisy datasets under two configurations: linear probing, where only the final layer is tuned, and full fine-tuning, where gradients of all model parameters are reweighted based on mimic scores to maximize the impact of useful samples. We normalize mimic scores with a temperature of 0.5 and use a batch size of 32. We simulate pretrained reference models by training ViT-B/16 models on the noise-free version of each dataset and use the last layer weights as the reference to navigate training on the noisy datasets. We detail more training configurations in Appendix C.1.

After training, we identify samples in two ways: setting one over batch size ($1/b$) as our threshold and clustering methods using k -means and GMM, then aggregate filter outputs across training steps using Snorkel framework [31].

Full Fine-tuning			DTD			Flowers102			STL10			Oxford-IIIT Pet			CIFAR10			CIFAR100			Average			
Ref. Model	65.59		91.01			93.20			82.88			94.89			81.34			84.82			84.82			
Noise Level	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6
SGD	46.91	42.02	33.24	65.54	55.90	39.84	75.35	58.57	49.98	70.43	64.62	55.63	90.31	86.69	82.01	69.42	63.77	59.28	69.66	61.93	53.33			
GraNd	29.73	8.83	15.43	10.80	2.78	27.92	78.39	52.32	64.30	56.83	45.38	45.43	87.56	83.79	80.04	18.65	8.83	7.84	46.99	33.66	40.16			
AGRA	46.28	37.29	32.39	0.62	0.93	0.82	84.35	79.94	78.17	76.97	77.11	69.39	89.23	86.79	68.09	11.96	8.44	6.88	51.57	48.42	42.62			
Grad-Match	48.88	42.66	33.24	37.92	54.94	42.62	83.86	81.16	77.18	74.35	61.73	51.16	44.48	35.69	37.59	69.41	66.21	20.09	58.82	57.07	43.65			
Grad-Mimic	49.20	42.82	33.83	68.58	56.46	44.27	72.06	71.85	83.09	81.98	78.30	73.34	90.52	89.07	66.41	73.97	74.31	24.02	72.72	68.60	54.16			

Linear Probing			DTD			Flowers102			STL10			Oxford-IIIT Pet			CIFAR10			CIFAR100			Average				
Ref. Model	60.00			56.98			97.41			89.45			94.18			77.89			79.32						
Noise Level	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6				
SGD	51.91	47.71	44.44	28.64	21.50	15.43	96.26	95.49	93.56	87.33	85.55	83.51	92.86	92.14	91.19	75.56	74.38	73.25	72.09	69.46	66.40				
GraNd	36.22	30.59	25.32	18.23	13.90	10.49	68.70	59.55	49.68	69.47	60.02	48.13	88.02	85.61	81.48	71.60	70.51	69.27	58.71	53.36	47.40				
AGRA	41.81	36.28	31.49	41.81	36.28	31.49	96.19	95.15	93.11	85.25	82.53	78.39	92.51	92.01	90.99	72.50	71.30	69.69	71.68	68.93	65.86				
Grad-Match	51.81	47.55	41.33	27.00	20.21	14.90	96.06	95.34	93.44	86.86	85.75	83.18	92.62	90.4	91.17	75.39	74.46	73.18	71.62	69.23	66.20				
Grad-Mimic	54.68	54.10	50.43	42.75	37.10	31.71	97.16	97.00	96.90	88.80	88.25	87.14	94.15	93.92	93.80	77.24	76.82	76.05	75.80	74.53	73.01				

Table 1: **Stage 1 Results in Simulation Experiment:** Using mimic scores can effectively down-weight mislabeled samples during training, leading to improved denoising and performance gains.

	DTD			Flowers102			STL10			Oxford-IIIT Pet			CIFAR10			CIFAR100			Average						
Noise Level	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6				
Threshold (1/32)	85.13	88.91	91.82	90.93	94.9	96.92	90.31	94.11	97.21	94.41	96.86	97.93	98.37	98.19	97.19	97.66	97.57	95.80							
1D k-means	77.13	76.68	77.17	84.45	87.62	87.47	78.90	83.04	78.82	95.80	95.49	93.73	98.62	98.17	97.56	98.10	97.46	96.62							
GMM	97.85	96.72	95.68	98.39	96.96	94.21	97.96	97.16	95.69	98.04	97.07	95.86	95.70	92.89	88.62	95.86	94.31	92.55							

Table 2: **Stage 2 Results in Simulation Experiment:** We report detection results using F1-score metric. Grad-Mimic consistently achieves precise detection of mislabeled samples across datasets and identification methods. Notably, its performance remains high and robust regardless of the noise level.

Expected Results. We aim to validate that a pretrained reference model can serve as a strong guide for model training. Moreover, we expect that mimic scores obtained from Grad-Mimic can help identify low-value, mislabeled samples.

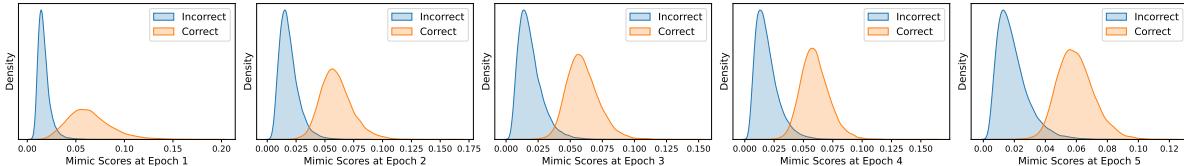


Figure 3: **Mimic Score Distribution:** Extracted from CIFAR100 dataset with a noise level of 0.5. The distributions clearly separate, aligning with the correctness of the labels.

Baselines. We compare Grad-Mimic’s training method with other approaches that use gradient information for sample prioritization. We consider: (i) *Mini-batch SGD*, the standard method, updates weights by averaging gradients within the mini-batch. (ii) *GraNd* [22] reweights samples based on their gradient norm, prioritizing data that induce greater changes. (iii) *AGRA* [23] computes the cosine similarity of each sample gradient to an average gradient of a random batch, then excludes outlier samples. (iv) *Grad-Match* [24] adapts gradients by solving a subset selection problem to identify key samples within each batch.

Stage 1 Results. We present Grad-Mimic’s two-stage results separately. The Stage 1 results, shown in Table 1, report testing accuracy across different methods and noisy datasets using full fine-tuning and linear probing. We evaluate Grad-Mimic’s robustness by averaging accuracy across datasets under different levels of label noise. Results show that Grad-Mimic effectively identifies and down-weights mislabeled samples, improving denoising and closing the performance gap to the reference model. Notably, in the full fine-tuning setup where we only target the last layer to mimic, we successfully guide the entire model weight updates, ultimately outperforming other methods. This also highlights the computational feasibility of our approach. These findings validate the idea that using the direction toward a better-performing model can guide model training and lead to enhanced performance.

Stage 2 Results. Next, we demonstrate Grad-Mimic’s capability to detect mislabeled samples. We use F1-scores to report the detection results in Table 2. Grad-Mimic accurately identifies mislabeled samples across all datasets and demonstrates robust performance when facing different noise levels.

Scale	Training Method	Mimic Layer θ_{ref}	Temperature τ	ImageNet	ImageNet dist. shifts	VTAB	Retrieval	Average over 38 datasets (\uparrow)
Small	Standard Training	—	—	0.026	0.035	0.139	0.114	0.131
				0.03	0.026	0.035	0.153	0.115 0.136
				0.05	0.026	0.035	0.147	0.114 0.135
				0.07	0.026	0.035	0.151	0.116 0.135
				0.3	0.026	0.034	0.154	0.112 0.137
	Grad-Mimic	Last MLP Layer in Text Encoder	0.5	0.027	0.035	0.152	0.114	0.139
				0.03	0.025	0.035	0.149	0.114 0.133
				0.05	0.026	0.037	0.162	0.118 0.146
				0.07	0.027	0.036	0.166	0.118 0.145
				0.3	0.026	0.035	0.161	0.114 0.144
				0.5	0.027	0.035	0.160	0.117 0.145
Medium	Standard Training	—	—	0.171	0.148	0.253	0.217	0.254
	Grad-Mimic	Last MLP Layer in Image Encoder	0.05	0.169	0.151	0.262	0.216	0.258

Table 3: **Stage 1 Results in DataComp Experiment:** On both dataset scales, with the aid of publicly available pretrained weights, Grad-Mimic consistently improves CLIP model performance.

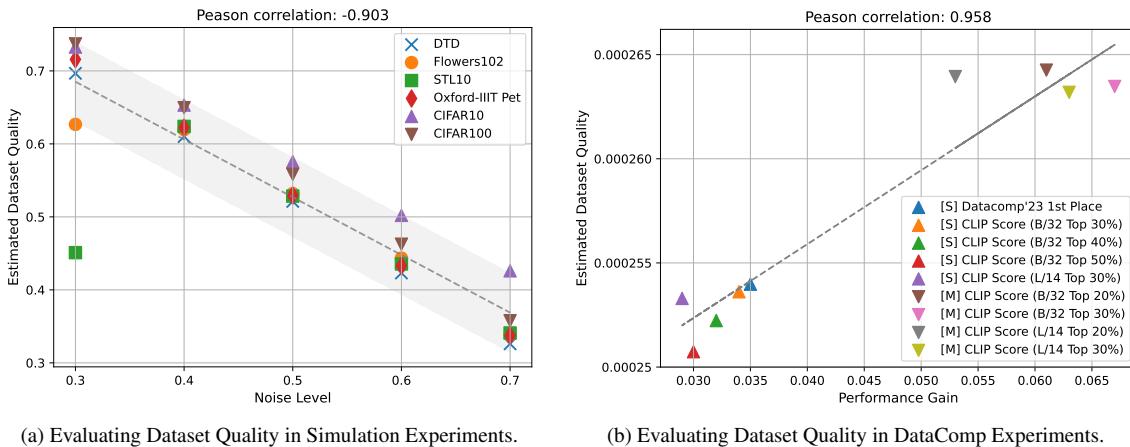


Figure 4: **Mimic scores offer precise dataset quality estimation:** Figure 4a computes the portion of retained samples to evaluate dataset quality in simulation experiments. Figure 4b takes the average on mimic scores in each filtered dataset for estimation. [S/M] denotes the scale of DataComp dataset—small [S] or medium [M]—that the filter is designed on.

We use CIFAR100 dataset as an example and visualize mimic score distribution at each training step in Figure 3. In this dataset, half of the labels are flipped. We see their distributions clearly separate into two groups, confirming that mimic scores can serve as informative signals to identify samples. Despite variations over training, these dynamics can be effectively captured in our aggregation step.

We use an ensemble filter from each dataset created by GMM clustering to estimate dataset quality using the proportion of retained data. The results, displayed in Figure 4a, highly correspond to the presence of label noise, with a Pearson correlation of **-0.903**. This demonstrates mimic score’s effectiveness for dataset quality estimation.

5.2 Data Selection in Large-scale Web Datasets

Setups. We evaluate Grad-Mimic in a more challenging setting using million-scale web-crawled datasets. We use the small- and medium-scale DataComp datasets [16], which contain approximately 10 million and 100 million image-caption pairs¹, respectively, for CLIP model pretraining [2]. We follow the training setup from DataComp [53; 16] and use publicly available pretrained CLIP weights as our reference ². This pretrained model,

¹We identified that some URLs provided by DataComp dataset are now broken. This means that our results might not be comparable to previous approaches on the DataComp Leaderboard. See [here](#) for details.

²<https://huggingface.co/laion/CLIP-ViT-B-32-DataComp.XL-s13B-b90K>

Scale	Filtering Strategy	Dataset Size	ImageNet	ImageNet dis. shifts	VTAB	Retrieval	Average over 38 datasets (\uparrow)
Small	No Filtering	10.7M	0.026	0.035	0.139	0.114	0.131
	Basic Filtering	3M	0.031	0.040	0.159	0.115	0.141
	Mimic Score Top 30%	3M	0.028	0.037	0.151	0.107	0.135
	Mimic Score Top 35%	3.8M	0.026	0.037	0.158	0.107	0.143
	Mimic Score Top 40%	4.2M	0.028	0.035	0.151	0.110	0.139
	CLIP Score (B/32 Top 30%)	3M	0.051	0.054	0.183	0.118	0.165
	CLIP Score \ Mimic Score Bottom 15%	2.7M	0.047	0.053	0.174	0.106	0.163
Medium	CLIP Score \ Mimic Score Bottom 20%	2.6M	0.047	0.052	0.173	0.113	0.164
	No Filtering	101.9M	0.026	0.035	0.139	0.114	0.254
	Basic Filtering	30M	0.209	0.176	0.264	0.232	0.267
	Mimic Score Top 30%	29M	0.192	0.163	0.269	0.218	0.268
	CLIP Score (B/32 Top 30%)	30M	0.275	0.229	0.332	0.246	0.321
	CLIP Score \ Mimic Score Bottom 5%	28.1M	0.270	0.227	0.339	0.246	0.323
	CLIP Score \ Mimic Score Bottom 10%	27.7M	0.274	0.226	0.334	0.243	0.323
Large	CLIP Score \ Mimic Score Bottom 30%	25.3M	0.264	0.223	0.330	0.244	0.322

Table 4: **Stage 2 Results in DataComp Experiment:** Mimic score-based filters perform better than basic filtering and complement CLIP score-based filters by removing low-value samples. Symbol “\” denotes the exclusion of curated datasets.

trained on the DataComp-1B dataset (1.4 billion samples), represents the best-performing model accessible to us. It serves as a proxy to the ideal reference point for scenarios where training such powerful models is infeasible. We target to mimic the final MLP layer in the text and image encoders respectively. We evaluate model performance using DataComp benchmark, which includes 38 diverse image classification and retrieval tasks. Additional training details can be found in Appendix C.2.

In Stage 2, we use top- k percent selection method to create the final filter, subsequently training a CLIP model on the curated dataset to validate filter’s effectiveness.

Expected Results. We anticipate mimic scores help large-scale training focus on high-value samples, while the derived scores will be instrumental in automating effective data curation.

Baselines. In the first stage, we compare Grad-Mimic to vanilla training, where each sample contributes in the weight update equally. After training, we compare our mimic score-based filter against the following methods: (1) *No Filtering*: using the entire training pool, (2) *Basic Filtering* [16]: selecting samples based on predefined criteria like caption length, image size, and caption language—representing a human-designed filter, (3) *CLIP Score* [17]: selecting top- k percent samples based on embedding alignment between images and captions. Scores are computed by OpenAI’s CLIP ViT-B/32 model pretrained on over 400 million samples [2]. We see CLIP score as a strong baseline, which is prebuilt through heavy training using hundreds of millions of samples.

Stage 1 Results. Table 3 presents pretraining results for CLIP models on both dataset scales. Grad-Mimic consistently outperforms standard training across all the temperature settings. Interestingly, we find that mimicking the last layer of the image encoder yields more performance gains compared to targeting the text encoder. These results highlight Grad-Mimic’s success in using pretrained model weights as a guide to prioritize high-value samples to learn.

Stage 2 Results. We use the derived mimic scores from our best-performing model in Stage 1 to design filters and evaluate their effectiveness. Results are presented in Table 4. Grad-Mimic successfully automates the selection process, outperforming the basic filtering strategy at both dataset scales. Additionally, we remove less useful samples identified by Grad-Mimic to improve CLIP score-based filter (B/32 Top 30%). This complementary approach not only enhances model performance but also improves data efficiency by reducing the training set size, specifically cutting down **4.7 million** samples in the medium-scale dataset.

Lastly, we gather mimic scores from datasets that are curated by various CLIP score-based filters. Besides, we include the top-ranked approach during our time of experimentation [54], which uses an orthogonal approach based on object detection models. By calculating the average mimic score in each curated dataset, we compare with their

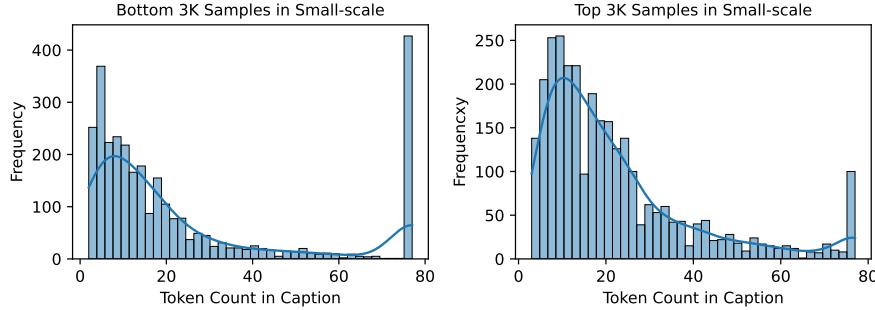


Figure 5: The filtering logic behind mimic score aligns with handcrafted heuristics, particularly in caption length. Low-value samples (in the left figure) tend to have captions that are either too short or exceed the maximum token length, i.e., 77.

corresponding performance gains. The results, shown in Figure 4b, reveal a strong alignment between performance gains and average mimic scores, with a Pearson correlation of **0.958**, demonstrating that mimic scores can serve as a reliable metric for estimating dataset quality.

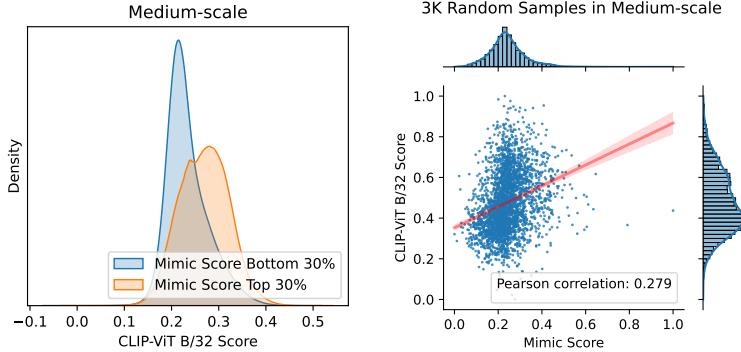


Figure 6: Mimic score distribution and its relationship with CLIP score.

Sample Analysis. We analyze high- and low-value samples ranked by mimic score. In Figure 5, we compare the token usage distribution in captions. We find that bottom-ranked samples often carry captions that are either too short (one or two words) or excessively long (exceeding the maximum token limit in CLIP’s tokenizer), unlike top-ranked samples. Moreover, mimic score-based filter effectively identifies low-value samples, such as Captcha-like images in the small-scale dataset and empty or visually incoherent images in the medium-scale dataset (see examples shown in Figure 1). *These filtering patterns align with human intuition but are automatically captured by Grad-Mimic framework.*

We select the top 30% and bottom 30% of samples ranked by their mimic scores and visualize their distribution along with corresponding CLIP scores. In Figure 6, we observe high-value samples distributed towards higher CLIP scores, while low-value samples are more concentrated in lower CLIP score ranges (half of the samples are scored less than 0.2). Furthermore, we calculate their correlation using randomly selected samples. For a better visualization, we normalize both scores into the range 0 to 1. We see a positive correlation between them, supporting the reliability of mimic scores in identifying high-quality samples.

Predicting Pretraining Dataset We explore an interesting application of mimic score: predicting the pretraining dataset. Specifically, we ask: *how accurately can Grad-Mimic check whether a given sample was used to train a reference model?* We test this hypothesis on the small-scale DataComp dataset. We first apply various pre-built filters to curate datasets and train CLIP models on each. Then, we use their final layer weights as the reference and apply Grad-Mimic to train on the entire data pool. Our goal is to see whether the top-ranked samples (matched in size to the curated datasets) identified by mimic scores appear in their curated dataset.

We evaluate our approach against random selection using Jaccard similarity and percentage of overlap. As shown in

Filtering Strategy	Random Selection		Mimic Score Selection	
	Jaccard Similarity	Percentage of Overlap	Jaccard Similarity	Percentage of Overlap
CLIP-ViT B/32 Top 30%	0.149	0.258	0.232	0.376
CLIP-ViT L/14 Top 30%	0.150	0.260	0.236	0.382
DataComp’23 Top-ranked	0.166	0.284	0.271	0.426

Table 5: Mimic score can help in predicting whether a sample was used to train the reference model.

Table 5, mimic score-based selection identifies more samples used to train the reference model compared to random sampling. We achieve a **42.6%** overlap with DataComp’23 best-performing filtered dataset *by simply mimicking the final layer weights without direct access to their filtering steps*.

6 Conclusion

We introduce a new data quality metric called Mimic Score, designed to assess samples’ utility in the training process. We quantify samples by measuring the alignment of per-sample gradient and the vector toward a pretrained reference model. Building on Mimic Score, we develop Grad-Mimic to automate data selection process. Empirically, we demonstrate that Grad-Mimic effectively prioritizes identified samples to learn, resulting in enhanced model performance. Moreover, Mimic Score can serve as a reliable metric for creating effective filters and estimating dataset quality. Developed filters outperform human-designed ones and exhibit a strong correlation with performance gains.

7 Acknowledgments

We thank Kwei-Herng Lai, Prasad Gabbur, Meng Cao, Ryan Jia, Russ Webb, Vedant Joshi, Albert Ge, and John Cooper for their helpful feedback and valuable discussion.

References

- [1] Zhang, H.; Gao, M.; Gan, Z.; Dufter, P.; Wenzel, N.; Huang, F.; Shah, D.; Du, X.; Zhang, B.; Li, Y.; others MM1. 5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning. *arXiv preprint arXiv:2409.20566 2024*,
- [2] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others Learning transferable visual models from natural language supervision. International conference on machine learning. 2021; pp 8748–8763.
- [3] Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; others The llama 3 herd of models. *arXiv preprint arXiv:2407.21783 2024*,
- [4] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; others Gpt-4 technical report. *arXiv preprint arXiv:2303.08774 2023*,
- [5] Albalak, A.; Elazar, Y.; Xie, S. M.; Longpre, S.; Lambert, N.; Wang, X.; Muennighoff, N.; Hou, B.; Pan, L.; Jeong, H.; others A survey on data selection for language models. *arXiv preprint arXiv:2402.16827 2024*,
- [6] Bai, T.; Liang, H.; Wan, B.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; He, C.; Yuan, B.; Zhang, W. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv preprint arXiv:2405.16640 2024*,
- [7] Penedo, G.; Kydlíček, H.; Lozhkov, A.; Mitchell, M.; Raffel, C.; Von Werra, L.; Wolf, T.; others The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557 2024*,
- [8] Wang, Y.; Chen, Y.; Yan, W.; Fang, A.; Zhou, W.; Jamieson, K.; Du, S. S. CLIPLoss and Norm-Based Data Selection Methods for Multimodal Contrastive Learning. *arXiv preprint arXiv:2405.19547 2024*,
- [9] Thrush, T.; Potts, C.; Hashimoto, T. Improving pretraining data using perplexity correlations. *arXiv preprint arXiv:2409.05816 2024*,

- [10] Fang, A.; Jose, A. M.; Jain, A.; Schmidt, L.; Toshev, A.; Shankar, V. Data filtering networks. *arXiv preprint arXiv:2309.17425* **2023**,
- [11] Chen, J.; Mueller, J. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776* **2024**,
- [12] Wettig, A.; Gupta, A.; Malik, S.; Chen, D. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739* **2024**,
- [13] Thakkar, M.; Bolukbasi, T.; Ganapathy, S.; Vashishth, S.; Chandar, S.; Talukdar, P. Self-influence guided data reweighting for language model pre-training. *arXiv preprint arXiv:2311.00913* **2023**,
- [14] Lin, X.; Wang, W.; Li, Y.; Yang, S.; Feng, F.; Wei, Y.; Chua, T.-S. Data-efficient Fine-tuning for LLM-based Recommendation. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024; pp 365–374.
- [15] Yu, Z.; Das, S.; Xiong, C. MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models. *arXiv preprint arXiv:2406.06046* **2024**,
- [16] Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; others Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* **2024**, 36.
- [17] Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* **2021**,
- [18] Fan, S.; Pagliardini, M.; Jaggi, M. Doge: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393* **2023**,
- [19] Xie, S. M.; Pham, H.; Dong, X.; Du, N.; Liu, H.; Lu, Y.; Liang, P. S.; Le, Q. V.; Ma, T.; Yu, A. W. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems* **2024**, 36.
- [20] Chen, M. F.; Hu, M. Y.; Lourie, N.; Cho, K.; Ré, C. Aioli: A Unified Optimization Framework for Language Model Data Mixing. *arXiv preprint arXiv:2411.05735* **2024**,
- [21] Chen, M.; Roberts, N.; Bhatia, K.; Wang, J.; Zhang, C.; Sala, F.; Ré, C. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems* **2024**, 36.
- [22] Paul, S.; Raffel, C.; Schoenholz, S. S. Deep Learning on a Data Diet: Finding Important Examples Early in Training. *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.
- [23] Sedova, A.; Zellinger, L.; Roth, B. Learning with noisy labels by adaptive gradient-based outlier removal. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2023; pp 237–253.
- [24] Killamsetty, R.; Pervaje, A.; Mirzaoleiman, B. Grad-Match: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.
- [25] Mirzaoleiman, B.; Bilmes, J.; Leskovec, J. Coresets for data-efficient training of machine learning models. *International Conference on Machine Learning*. 2020; pp 6950–6960.
- [26] Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; others Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **2022**, 35, 25278–25294.
- [27] Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* **2021**,
- [28] Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM* **2016**, 59, 64–73.
- [29] Abbas, A.; Tirumala, K.; Simig, D.; Ganguli, S.; Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* **2023**,
- [30] Ratner, A. J.; De Sa, C. M.; Wu, S.; Selsam, D.; Ré, C. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems* **2016**, 29.

- [31] Ratner, A.; Bach, S. H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid training data creation with weak supervision. Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases. 2017; p 269.
- [32] Ratner, A.; Hancock, B.; Dunnmon, J.; Sala, F.; Pandey, S.; Ré, C. Training complex models with multi-task weak supervision. Proceedings of the AAAI Conference on Artificial Intelligence. 2019; pp 4763–4771.
- [33] Fu, D.; Chen, M.; Sala, F.; Hooper, S.; Fatahalian, K.; Ré, C. Fast and three-rious: Speeding up weak supervision with triplet methods. International Conference on Machine Learning. 2020; pp 3280–3291.
- [34] Roberts, N.; Li, X.; Huang, T.-H.; Adila, D.; Schoenberg, S.; Liu, C.-Y.; Pick, L.; Ma, H.; Albargouthi, A.; Sala, F. Autows-bench-101: Benchmarking automated weak supervision with 100 labels. *Advances in Neural Information Processing Systems* **2022**, *35*, 8912–8925.
- [35] Huang, T.-H.; Cao, C.; Bhargava, V.; Sala, F. The ALCHEmist: Automated Labeling 500x CHEaper Than LLM Data Annotators. *arXiv preprint arXiv:2407.11004* **2024**,
- [36] Huang, T.-H.; Cao, C.; Schoenberg, S.; Vishwakarma, H.; Roberts, N.; Sala, F. ScriptoriumWS: A Code Generation Assistant For Weak Supervision. *ICLR Deep Learning for Code Workshop* **2023**,
- [37] Hooper, S.; Wornow, M.; Seah, Y. H.; Kellman, P.; Xue, H.; Sala, F.; Langlotz, C.; Re, C. Cut out the annotator, keep the cutout: better segmentation with weak supervision. International Conference on Learning Representations. 2020.
- [38] Liu, L.; Ren, X.; Zhu, Q.; Zhi, S.; Gui, H.; Ji, H.; Han, J. Heterogeneous supervision for relation extraction: A representation learning approach. *arXiv preprint arXiv:1707.00166* **2017**,
- [39] Fries, J. A.; Varma, P.; Chen, V. S.; Xiao, K.; Tejeda, H.; Saha, P.; Dunnmon, J.; Chubb, H.; Maskatia, S.; Fiterau, M.; others Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature communications* **2019**, *10*, 3111.
- [40] Khattar, S.; O’Day, H.; Varma, P.; Fries, J.; Hicks, J.; Delp, S.; Bronte-Stewart, H.; Re, C. Multi-frame weak supervision to label wearable sensor data. ICML Time Series Workshop. 2019.
- [41] Mallory, E. K.; de Rochemonteix, M.; Ratner, A.; Acharya, A.; Re, C.; Bright, R. A.; Altman, R. B. Extracting chemical reactions from text using Snorkel. *BMC bioinformatics* **2020**, *21*, 1–15.
- [42] Smith, R.; Fries, J. A.; Hancock, B.; Bach, S. H. Language models in the loop: Incorporating prompting into weak supervision. *ACM/JMS Journal of Data Science* **2024**, *1*, 1–30.
- [43] Arora, S.; Narayan, A.; Chen, M. F.; Orr, L.; Guha, N.; Bhatia, K.; Chami, I.; Sala, F.; Ré, C. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441* **2022**,
- [44] Ghiasi, A.; Kazemi, H.; Borgnia, E.; Reich, S.; Shu, M.; Goldblum, M.; Wilson, A. G.; Goldstein, T. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727* **2022**,
- [45] Gandselman, Y.; Efros, A. A.; Steinhhardt, J. Interpreting CLIP’s Image Representation via Text-Based Decomposition. *arXiv preprint arXiv:2310.05916* **2023**,
- [46] Wu, X. Optimal quantization by matrix searching. *Journal of algorithms* **1991**, *12*, 663–673.
- [47] Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; Vedaldi, A. Describing Textures in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014; pp 3606–3613.
- [48] Nilsback, M.-E.; Zisserman, A. Automated flower classification over a large number of classes. *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing* **2008**, 722–729.
- [49] Coates, A.; Ng, A. Y.; Lee, H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). 2011; pp 215–223.
- [50] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; Jawahar, C. Cats and Dogs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012; pp 3498–3505.
- [51] Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; 2009.
- [52] Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**,

- [53] Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023; pp 2818–2829.
- [54] Huang, T.-H.; Shin, C.; Tay, S. J.; Adila, D.; Sala, F. Multimodal Data Curation via Object Detection and Filter Ensembles. *arXiv preprint arXiv:2401.12225* **2024**,
- [55] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; others Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.

The appendix is structured as follows. First, we discuss the broader impacts of Grad-Mimic and its limitations in Appendix A. Next, we organize the steps involved in Grad-Mimic Stage 1 in Appendix B. Then, we provide experimental details, including training configurations, and computation resources in Appendix C.1 and in Appendix C.2. Lastly, we present ablation studies in Appendix D.

A Discussion

A.1 Broader Impacts and Limitations.

The importance of data selection has grown in building foundation models. We demonstrate that Grad-Mimic can effectively identify and exclude noisy or unhelpful samples. This leads to reduced computational and resource costs. Besides, Grad-Mimic provides the second benefit by enabling models to focus their training on high-value samples, which contributes to performance improvements.

We do not foresee explicit negative impacts arising from Grad-Mimic. However, if the reference model used for sample selection is unreliable, there remains a risk of refining suboptimal datasets. To address this, incorporating human expert-designed filtering rules may be necessary to ensure the quality of curated datasets.

B Grad-Mimic Algorithm

We summarize the algorithm steps in Grad-Mimic for computing mimic scores and updating model weights. Our algorithm supports various training settings, such as supervised learning and self-supervised learning. We have validated both training scenarios through simulation experiments (Sec. 5.1) and by training on web-crawled datasets with CLIP models (Sec. 5.2). Moreover, these experiments evaluate Grad-Mimic performance across three different training configurations: full-parameter fine-tuning, linear probing, and training models from scratch.

Algorithm 1 Grad-Mimic (Stage 1)

Input: reference model θ_{ref} , new model θ , dataset D , number of training steps T , batch size b , learning rate η , temperature τ

Output: samples' mimic scores $\bar{m}_{i,t}$, trained model θ_T

```

Initialize new weights  $\theta_0$ 
for  $t \in [T]$  do
    uniformly sample a batch  $\{s_i\}_{i=1}^b$  from  $D$ 
     $v_t \leftarrow \theta_{\text{ref}} - \theta_t$  ▷ compute the vector as guide
    for  $i \in [b]$  do
         $g_{i,t} \leftarrow \nabla_{\theta_t} \ell(s_i)$  ▷ compute per-sample gradients
         $m_{i,t} \leftarrow \frac{\langle -g_{i,t}, v_t \rangle}{\|v_t\|_2}$  ▷ compute mimic scores
    end for
     $\bar{m}_{:,t} = \frac{e^{-m_{:,t}/\tau}}{\sum_{j=1}^b e^{-m_{j,t}/\tau}}$  ▷ normalize mimic scores
     $\theta_{t+1} := \theta_t - \eta \sum_{i=1}^b \bar{m}_{i,t} \cdot g_{i,t}$  ▷ reweight gradients
end for

```

C Experimental Details

We provide more details about our training setups, and computational resources.

C.1 Simulation Experiments

In the experiments of simulating mislabeled samples (Sec. 5.1), we fine-tune ViT-B/16 models pretrained on the ImageNet-21k dataset [55]. Each dataset is trained with a batch size of 32, a learning rate of 1e-4, the AdamW optimizer, and for 5 epochs. We evaluate Grad-Mimic using various temperature values τ (1.0, 0.9, 0.8, 0.7, 0.6, and 0.5) and report the performance at $\tau = 0.5$ in the main paper. Results for other temperature values are presented in Table 6. To simulate the reference model, we train ViT-B/16 models on noise-free versions of the datasets using identical configurations but with different random seeds for weight initialization. These experiments were performed on a Nvidia Tesla A100.

	Temperature τ	DTD	Flower102	Oxford-IIIT Pet	CIFAR10	CIFAR100
SGD	—	52.8	35.6	87.9	93.2	76.2
Grad-Match	—	53.1	32.1	87.4	93.0	76.1
Grad-Mimic	1.0	56.5	44.9	89.0	94.1	77.0
	0.9	56.4	45.3	89.1	94.1	77.1
	0.8	56.3	45.8	89.1	94.1	77.2
	0.7	56.3	46.1	89.2	94.2	77.2
	0.6	56.0	46.4	89.3	94.2	77.3

Table 6: **Stage 1 Results in Simulation Experiment with Different Temperature:** Lower temperature makes the model focus more on training with high-value samples, resulting in a higher testing accuracy.

C.2 DataComp Experiments

We adopt the training setup from DataComp [53; 16]. CLIP models are trained from scratch using a contrastive objective over image-caption pairs. Each model is trained for 5 epochs with a batch size of 4096. The total number of seen samples is 12.8M for the small-scale dataset and 128M for the medium-scale dataset. These experiments ran on 8 Nvidia Tesla A100s.

We test Grad-Mimic using different parts of weights in the reference model, specifically the final MLP layer in the image and text encoders. The layers used are “visual.transformer.resblocks.11.mlp.c_fc.weight” and “text.transformer.resblocks.11.mlp.c_fc.weight”. Grad-Mimic’s performance is evaluated using 38 diverse downstream tasks [16] with various temperature values τ (0.03, 0.05, 0.07, 0.3, and 0.5). Results are presented in Table 3.

D Ablation Studies

The Choice of Temperature. We present Grad-Mimic results with different temperature values and compare them to baseline methods (SGD and Grad-Match [24]). The results in the simulation experiments are presented in Table 6. We set the noise level to 0.3 and fine-tune ViT model under linear probing configuration. Grad-Mimic outperforms the baseline methods across all temperature settings. Lower temperatures yield better testing accuracy, as they normalize mimic scores in the way that encourages the model to focus more on high-value samples during training.