Diffusion Reconstruction-based Data Likelihood Estimation for Core-Set Selection

Mingyang Chen[1,2], **Jiawei Du**[3], **Bo Huang**[1,2], **Yi Wang**[4], **Xiaobo Zhang**[5], **Wei Wang**[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]The Hong Kong University of Science and Technology, Hong Kong
[3]CFAR, A*STAR, Singapore
[4]Dongguan University of Technology, China
[5]Southwest Jiaotong University, China
mchenbt@connect.ust.hk, weiwcs@ust.hk

Abstract

Existing core-set selection methods predominantly rely on heuristic scoring signals such as training dynamics or model uncertainty, lacking explicit modeling of data likelihood. This omission may hinder the constructed subset from capturing subtle yet critical distributional structures that underpin effective model training. In this work, we propose a novel, theoretically grounded approach that leverages diffusion models to estimate data likelihood via reconstruction deviation induced by partial reverse denoising. Specifically, we establish a formal connection between reconstruction error and data likelihood, grounded in the Evidence Lower Bound (ELBO) of Markovian diffusion processes, thereby enabling a principled, distribution-aware scoring criterion for data selection. Complementarily, we introduce an efficient information-theoretic method to identify the optimal reconstruction timestep, ensuring that the deviation provides a reliable signal indicative of underlying data likelihood. Extensive experiments on ImageNet demonstrate that reconstruction deviation offers an effective scoring criterion, consistently outperforming existing baselines across selection ratios, and closely matching full-data training using only 50% of the data. Further analysis shows that the likelihood-informed nature of our score reveals informative insights in data selection, shedding light on the interplay between data distributional characteristics and model learning preferences.

Introduction

Data selection, particularly core-set selection, is crucial for enabling efficient and scalable training of deep neural networks by identifying representative subsets from massive datasets. The exponential growth of large-scale datasets across various domains (Kolesnikov et al. 2020; Brown et al. 2020; Radford et al. 2021) has rendered the direct utilization of entire datasets computationally prohibitive. Effective core-set selection can significantly accelerate model training, enhance resource efficiency and provide insights into learning preferences (Paul, Ganguli, and Dziugaite 2021a; Guo, Zhao, and Bai 2022). This is especially beneficial in resource-intensive downstream applications such as fine-tuning large-scale foundation models (Joaquin et al. 2024; Xia et al. 2024), continual learning (Yoon et al. 2022; Hao, Ji, and Liu 2023), and multi-task learning (Kung et al. 2021; Renduchintala, Bhatia, and Ramakrishnan 2024).

Existing core-set selection methods can generally be classified into score-based and optimization-based approaches (Choi, Ki, and Chung 2024). Score-based methods, popular for their scalability, typically leverage heuristic signals such as model uncertainty (Pleiss et al. 2020), training dynamics including forgetting events (Toneva et al. 2019), and loss-based metrics like EL2N (Paul, Ganguli, and Dziugaite 2021b). Despite the practical utility, these methods lack explicit modeling of data likelihood, leading such surrogate-based heuristic selections may fail to capture nuanced yet critical distributional characteristics intrinsic to model training. The comparison shown in Figure 2 further illustrates that these heuristic scores often fail to stratify samples based on their alignment with the underlying criterion, leading to selected subsets that resemble random sampling. This lack of sensitivity not only reduces their effectiveness and flexibility as selection criteria but also limits their interpretability in reflecting model learning preferences.

To address this limitation, we propose a theoretically principled and distribution-aware data scoring criterion, grounded explicitly in diffusion generative models. Specifically, we exploit the inherent generative modeling capability of Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) to directly quantify data likelihood through reconstruction deviation induced by partial reverse denoising, grounded in the Evidence Lower Bound (ELBO) of Markovian diffusion processes. Our core insight, illustrated intuitively in Figure 1, is that the reconstruction error between real data and their partially denoised reconstructions provides a precise, likelihood-sensitive signal reflecting data distributional characteristics. Lower reconstruction deviations correspond to samples closely aligned with the authentic distribution, whereas higher errors indicate samples farther from high-probability regions.

A key aspect of our methodology involves strategically selecting the reconstruction timestep, as naively chosen timesteps either degenerate into trivial reconstructions at minimal noise levels or lose discriminative sensitivity at high noise levels. Inspired by Information Bottleneck (IB) theory (Tishby, Pereira, and Bialek 2000), we formulate and solve this selection problem by maximizing the rate of decrease in mutual information between noised data and class

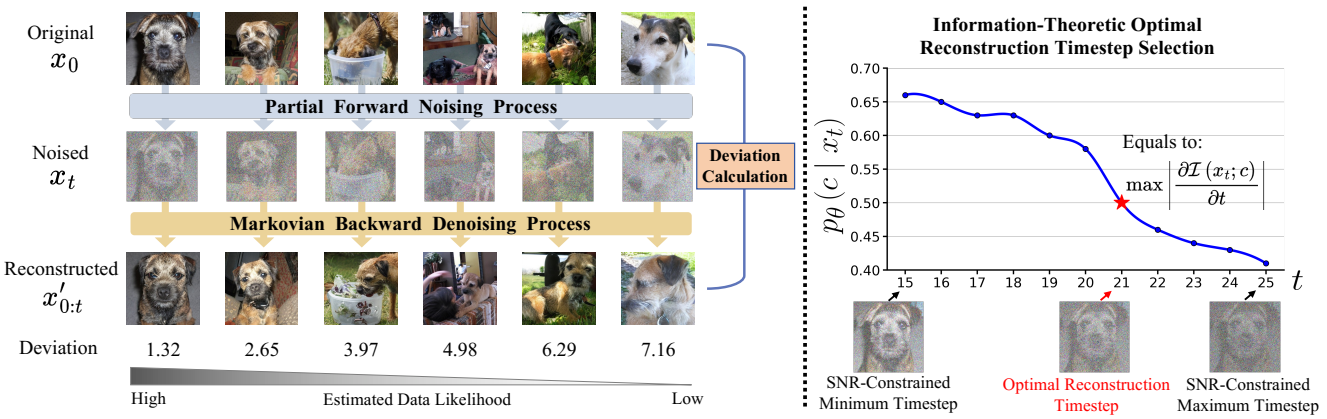


Figure 1: Intuitive illustration of our likelihood-informed scoring and optimal reconstruction timestep selection. **Left:** The deviation between real data point x_0 and the reconstructed $x'_{0:t}$ serves as a likelihood-sensitive signal, with lower deviation indicating higher estimated data likelihood. The visualized examples reveal a semantic illustration: *high-likelihood samples typically feature class-relevant objects that are spatially prominent and well-formed*; *moderate-likelihood samples often contain target objects that are less visually salient*, e.g., occupying smaller regions or blended with irrelevant elements; and *low-likelihood samples exhibit apparent out-of-distribution characteristics*, leading to significant semantic shifts after reconstruction. **Right:** We select the optimal reconstruction timestep ($0 < t < T$) by maximizing the drop rate $|\partial \mathcal{I}(x_t; c)/\partial t|$. Following Lemma 1, we equivalently maximize it by the time derivative of $\log p_\theta(c \mid x_t)$ predicted by a *diffusion classifier* (Li et al. 2023). The search is constrained to $\text{SNR}(t) \in [\gamma_{\min}, \gamma_{\max}]$ to avoid degenerate regions of timesteps.

labels, thus ensuring the most informative and discriminative timestep selection. This principled selection strategy avoids empirical grid search inefficiencies and theoretically grounds our approach in information-theoretic optimization.

In summary, our contributions are threefold:

- We present a novel, distribution-aware scoring criterion for core-set selection, grounded theoretically in the connection between diffusion reconstruction error and data likelihood.
- We introduce an efficient information-theoretic approach to select reconstruction timesteps, ensuring that the deviation provides a reliable signal indicative of underlying data likelihood.
- Extensive experiments on ImageNet demonstrate that using our proposed reconstruction deviation as a scoring criterion consistently yields superior performance. Further analysis of the likelihood-informed nature of our method provides new insights into the relationship between model learning preferences and the distributional characteristics of training data.

Background

Diffusion Model Preliminaries. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) define a generative process by learning to reverse a fixed forward noising process. Let $x_0 \sim q^c(x)$ denote a data sample drawn from the real distribution conditioned on the class label $c \in \{1, \ldots, C\}$. The forward process gradually adds Gaussian noise over T discrete timesteps using a fixed variance schedule $\{\beta_t\}_{t=1}^T$. The marginal distribution at step t can be expressed in closed form as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ denotes the noise retention factor. To generate data, diffusion models learn a reverse Markov process $p_\theta(x_{t-1} \mid x_t, c)$ that approximates the time-reversed dynamics of the forward process. The model is trained by maximizing a variational lower bound (ELBO) on the conditional log-likelihood:

$$\begin{aligned} \log p_\theta(x_0 \mid c) &\geq E_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T}, c)}{q(x_{1:T} \mid x_0)} \right] \\ &\approx -E_{t,\epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2 \right] + \mathcal{C}, \end{aligned} \tag{2}$$

where $\mathcal{C}$ is a constant independent of model parameters. In practice, $p_\theta(x_{t-1} \mid x_t, c)$ is modeled as a Gaussian with its mean predicted by a neural network $\epsilon_\theta(x_t, t, c)$ trained to reconstruct the injected noise ϵ. At inference, one samples $x_T \sim \mathcal{N}(0, I)$ and applies the learned reverse transitions:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c) \right) + \sigma_t z, \tag{3}$$

where $z \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, and σ_t^2 denotes the posterior variance.

Core-Set Selection. Recent advances in core-set selection can be broadly categorized into score-based and optimization-based approaches. Score-based methods evaluate the importance or utility of individual data via surrogate metrics. These include instance-wise influence estimation methods such as data shapley (Ghorbani and Zou 2019; Kwon, Rivas, and Zou 2021; Kwon and Zou 2022), influence functions (Koh and Liang 2017; Pruthi et al. 2020); and training dynamics based metrics such as forgetting events (Toneva et al. 2019), EL2N (Paul, Ganguli, and Dziugaite

2021b), memorization (Feldman and Zhang 2020), and CG-score (Ki, Choi, and Chung 2023). In contrast, optimization-based methods formulate subset selection as an optimization problem, aiming to approximate the diverse characteristics of the full dataset (Mirzasoleiman, Bilmes, and Leskovec 2020; Yang, Kang, and Mirzasoleiman 2023; Pooladzandi, Davini, and Mirzasoleiman 2022). However, such methods are often computationally intensive and underperform compared to score-based ones (Choi, Ki, and Chung 2024). Our method also falls under the score-based category but differs fundamentally in both principle and implementation. Unlike prior approaches relying on certain surrogate behaviours, our method derives selection signals from a theoretically grounded connection between diffusion reconstruction deviation and data likelihood, providing a distribution-aware criterion. To the best of our knowledge, this is the first work to leverage data likelihood as a scoring signal for core-set selection, offering a novel and interpretable perspective on the relationship between data distributional characteristics and model performance.

Reconstruction-based OOD Detection. Several recent works explored using diffusion models for out-of-distribution (OOD) detection by exploiting their reconstruction behaviour under partial noising or masking (Graham et al. 2023; Liu et al. 2023; Bellier and Audebert 2024). These methods measure reconstruction errors at fixed or multiple noise levels as heuristics for identifying OOD data. While effective for OOD detection, they lack theoretical grounding in the data likelihood of in-distribution samples and provide no analysis for data selection. In particular, the choice of reconstruction timestep is typically made via empirical grid search. As there exists no principled metric to assess the discriminative utility for data selection, effectiveness can only be validated via downstream training, which is computationally expensive. In contrast, our method establishes a formal connection between reconstruction deviation and data likelihood and proposes an efficient, information-theoretic criterion for selecting reconstruction timesteps aligned with reliable likelihood estimation.

Method

Estimating Data Likelihood with Diffusion Reconstruction Deviation

In this work, we propose to leverage diffusion generative models, which are explicitly trained to learn the authentic data distribution, as a natural lens through which data likelihood can be assessed. Specifically, we show that the deviation between a sample and its reconstruction via Markovian diffusion provides an effective, likelihood-sensitive signal for identifying samples that are most aligned with the selection criteria.

To formalize this idea, we consider the reconstruction deviation between a real data point x_0 and its denoised estimate $x'_{0:t}$, obtained by reversing a partial forward process from an intermediate noised x_t. This deviation is defined as $\Delta x_0(t) := |x'_{0:t} - x_0|$. The following theorem formalizes the inverse relationship between the reconstruction deviation and the log-likelihood of a data point.

Theorem 1 (Inverse Dependence of Reconstruction Deviation on Log-Likelihood)**.** *Let $x_0 \in R^d$, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$, and $x'_{0:t}$ be the reconstructed data obtained by DDPM denoising from x_t, the expected squared reconstruction deviation satisfies:*

$$E_\epsilon \left[\|\Delta x_0(t)\|^2\right] \geq -\kappa(t) \log q(x_0) + \mathcal{C}_{noise}(t),$$

where $\kappa(t) = \frac{1}{t}\sum_{s=1}^{t} \frac{1}{\sigma_s^2}$, and $\mathcal{C}_{noise}(t)$ is a constant independent of x_0.

The detailed proof is provided in Appendix A.1. Intuitively, Theorem 1 suggest that when x_0 resides in a high-density region, the reverse process should reconstruct it with relatively small distortion. In contrast, low-density or off-manifold samples tend to incur larger reconstruction errors. While the theorem provides a lower bound rather than an exact equivalence, it establishes a principled, theoretically grounded link between reconstruction deviation and data likelihood, justifying the use of deviation as a distribution-aware scoring signal for core-set selection.

In practice, we adopt the Deterministic Denoising Implicit Model (DDIM) (Song, Meng, and Ermon 2021) for reconstructing $x'_{0:t}$, which preserves the same marginal distributions as DDPM under the given noise schedule while enabling more efficient denoising from intermediate steps x_t. The DDIM update is given by:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon. \tag{4}$$

To robustly measure reconstruction deviation, we use the Learned Perceptual Image Patch Similarity (LPIPS), which computes perceptual similarity via the ℓ_2 distance between deep features from a pretrained network (e.g., ResNet (He et al. 2016)), and is known to correlate well with human judgments (Zhang et al. 2018).

Information-Theoretic Optimal Reconstruction Timestep Selection

The effectiveness of reconstruction deviation as a proxy for data likelihood critically depends on the choice of the target reconstruction timestep t. At extreme timesteps, the approach degenerates: when $t \to T$, the noised input x_t approaches pure Gaussian noise, making the reconstruction independent of x_0; conversely, when $t \to 0$, minimal noise injection yields $x'_0 \approx x_0$, rendering the deviation uninformative. Consequently, an effective timestep t should be selected to distinguish between high and low-likelihood data. However, identifying such an optimal t is challenging due to the absence of a principled metric for discriminative utility and the high cost of downstream validation.

Intuitively, an optimal timestep t should balance information preservation about the original data and discriminative sensitivity regarding label information. This criterion naturally aligns with minimizing an information bottleneck (IB) formulation (Tishby, Pereira, and Bialek 2000) involving the original data x_0, intermediate noised data x_t, and associated class labels c:

$$\min_t E_{x_0} \left[\mathcal{I}(x_0; x_t) - \beta \mathcal{I}(x_t; c)\right], \tag{5}$$

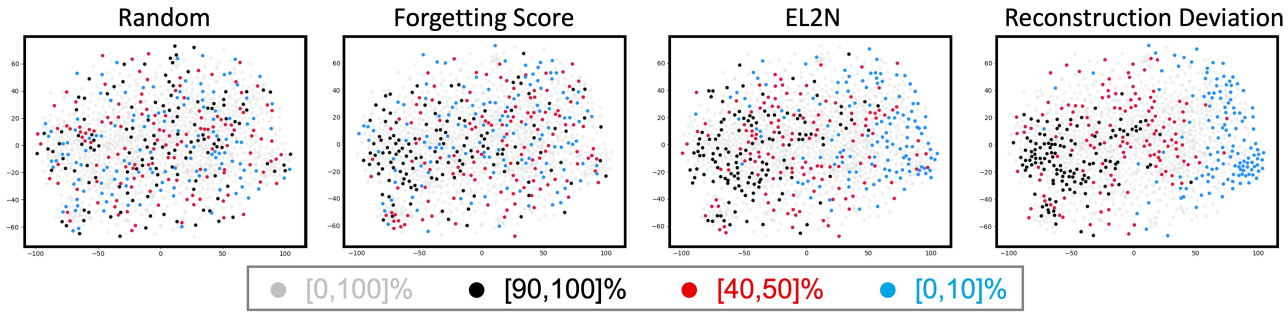


Figure 2: t-SNE visualization of stratified samples from the ImageWoof dataset. Samples are grouped by ascending score ranges for Forgetting Score (Toneva et al. 2019), EL2N (Paul, Ganguli, and Dziugaite 2021b), and our proposed Reconstruction Deviation, with Random representing equally sized random groups for reference. The distributions show that Forgetting Score and EL2N produce stratifications that resemble random sampling, whereas Reconstruction Deviation yields more distinct and semantically coherent groupings.

where $\mathcal{I}(\cdot;\cdot)$ denotes mutual information, and $\beta > 0$ trades off between compression of the original information and preservation of discriminative power. In this IB context, x_t is viewed as a compressed representation of x_0, regulated by the timestep t, and $\mathcal{I}(x_0; x_t)$ quantifies the retained information content. Thus, this quantity can be approximately characterized using the signal-to-noise ratio $\text{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$ (Luo 2022). To circumvent the instability and computational overhead associated with the selection of the hyperparameter β, we propose using $\mathcal{I}(x_0; x_t)$ as a regularization criterion. Specifically, we constrain the candidate range of timesteps t by restricting the signal-to-noise ratio $\text{SNR}(t)$ to lie within the empirical interval $[\gamma_{\min}, \gamma_{\max}] = [0.05, 1]$.

Conversely, $\mathcal{I}(x_t; c)$ quantifies the mutual information between the intermediate noised data x_t and its corresponding class c. As t increases, this term decreases monotonically and asymptotically approaches zero. Consequently, directly maximizing $\mathcal{I}(x_t; c)$ leads to the selection of t at the upper bound $\text{SNR}(t) = \gamma_{\max}$, degenerating the selection strategy into a naive grid search. Intuitively, the optimal timestep t should correspond to the boundary at which the distribution $q(x_t|x_0)$ transitions significantly from authentic data towards noise. At this boundary, the rate of decrease of mutual information between x_t and labels c should be maximized. Based on this insight, we formally define the optimal timestep selection criterion as:

$$t^* = \arg\max_t \left| \frac{\partial \mathcal{I}(x_t; c)}{\partial t} \right| \text{ s. t. } \text{SNR}(t) \in [\gamma_{\min}, \gamma_{\max}]. \tag{6}$$

Lemma 1. *Let* $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, *where* $\epsilon \sim \mathcal{N}(0, I)$, *and assume a uniform prior over classes* $p(c) = \frac{1}{C}$. *For all* $t \in [0, T]$,

$$\left| \frac{\partial \mathcal{I}(x_t; c)}{\partial t} \right| = \left| E_{x_0, c}\left[\frac{\partial}{\partial t} \log p(c \mid x_t) \right] \right|.$$

The detailed proof of Lemma 1 is provided in Appendix A.2. According to this lemma, the time derivative of mutual information can alternatively be computed through the expectation of the time derivative of $\log p(c \mid x_t)$. Based on Bayes' theorem and the uniform prior assumption of $p(c)$, we employ the diffusion classifier (Li et al. 2023) to approximate class probabilities based on the ELBO formulation of the data log-likelihood shown in Equation (2):

$$p_\theta(c \mid x_t) = \frac{\exp\left\{-E_\epsilon[\|\epsilon - \epsilon_\theta(x_t, c)\|^2]\right\}}{\sum_{c' \in C} \exp\left\{-E_\epsilon[\|\epsilon - \epsilon_\theta(x_t, c')\|^2]\right\}}. \tag{7}$$

Note that we omit the expectation over t as discussed in (Li et al. 2023) for testing on each candidate timestep. Furthermore, since the input timestep t is inherently discrete within practical diffusion formulation, the partial derivative $\frac{\partial}{\partial t} \log p_\theta(c \mid x_t)$ cannot be directly calculated using automatic differentiation tools (e.g., PyTorch's autograd). To address this limitation, we propose using a finite difference method to approximate the partial derivative:

$$\frac{\partial}{\partial t} \log p_\theta(c \mid x_t) \approx \frac{\log p_\theta(c \mid x_{t+\Delta t}) - \log p_\theta(c \mid x_{t-\Delta t})}{2\Delta t}, \tag{8}$$

where $\Delta t = 1$ in our practical implementation. Finally, we perform a Monte Carlo estimate of the expectation over $x_0 \sim q^c(x)$. Specifically, for each class in the target dataset, we randomly sample B data points from its class-conditional subset. The class-wise optimal reconstruction timestep is determined by solving the following objective:

$$t_c^* = \arg\max_t \sum_{i=1}^{B} \left| \frac{\log p_\theta(c \mid x_{t+\Delta t}^{(i)}) - \log p_\theta(c \mid x_{t-\Delta t}^{(i)})}{2\Delta t} \right| \quad \text{subject to} \quad \text{SNR}(t) \in [\gamma_{\min}, \gamma_{\max}]. \tag{9}$$

Analysis of Likelihood-Informed Core-Set Construction

Better Stratification and Interpretability of Distribution-Aware Likelihood Score. The foundation of data selection lies in the assumption that the score function quantifies how well each sample aligns with a target property, such as forgetting events, model uncertainty, or, in our case, distributional fidelity. Thus, an informative score should induce

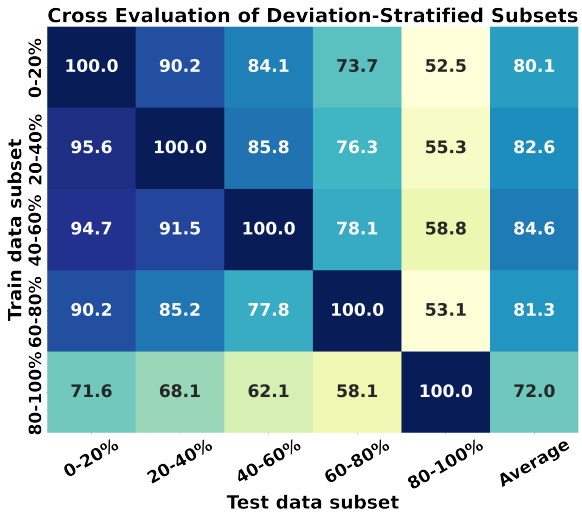


Figure 3: Cross evaluation of ResNet-18 models trained on deviation-stratified subsets. Each model is trained on one subset and tested across all five. Lower indices (e.g., $0-20\%$) indicate higher estimated likelihoods.

a meaningful stratification of the dataset: samples across different score quantiles ought to exhibit distinct structural distributions. In contrast, ineffective scores tend to produce subsets whose distributions resemble random sampling. In Figure 2, we visualize t-SNE embeddings of samples selected by different scoring methods. Compared to EL2N and Forgetting Score, which yield noticeably entangled quantile regions resembling quasi-random partitions, our reconstruction deviation score produces clearer separation across quantiles. Interestingly, we also observe that the dense cluster of black points in the left-central region of the Forgetting plot, corresponding to the most frequently forgotten samples, aligns most closely with our stratification. The likelihood-informed nature of our score suggests that the Forgetting Score tends to favor low-likelihood data while lacking discriminative ability among less-forgotten examples.

Moderate-Likelihood Data Benefits Core-Set Construction. A natural follow-up question to our earlier findings is: *which regions of the likelihood spectrum yield the most informative subset for model training?* To explore this, we adapt the "training set split" experiment from (Choi, Ki, and Chung 2024). Specifically, we sort the ImageWoof training set by reconstruction deviation and divide it into five equal-sized subsets, from the top 20% (highest estimated likelihood) to the bottom 20%. We train a ResNet-18 model on each subset and evaluate its accuracy across all five. As shown in Figure 3, models trained on moderate-likelihood subsets consistently generalize better, whereas those trained solely on the lowest or highest extremes perform poorly. To deepen understanding, we evaluate two contrasting selection strategies for leveraging our reconstruction-based score. **(1) Coverage-Centric Selection (CCS)** (Zheng et al. 2023) divides per-class samples—sorted by reconstruction deviation—into $\mathcal{B}$ strata and uniformly samples from each until the budget is met. **(2) Best Window Selection (BWS)** (Choi, Ki, and Chung 2024) scans fixed-width windows over the sorted list using a step size of 5%, with starting points in the range $[0, \min(50\%, 100\% - budget)]$, and selects the one achieving the highest validation accuracy on the full set. As shown in Table 1, BWS consistently outperforms CCS across different selection ratios. Notably, we observe that, except at extremely low or high ratios, the optimal window start point generally falls within the $[20\%, 40\%]$ range. We attribute this to BWS selecting moderate-likelihood subsets that better align with the curriculum-like preference in model training (Hacohen and Weinshall 2019), whereas CCS may introduce overly heterogeneous features, especially under tight budgets. Moreover, the model's preference for moderate-likelihood data may arise from a similar effect as data augmentation techniques such as CutMix (Yun et al. 2019), which improve robustness by introducing mixed or ambiguous class signals. As illustrated in Figure 1 and 7-8 in Appendix E, moderate-likelihood samples often contain less salient features, thereby prompting the model to generalize under partial or uncertain semantic cues.

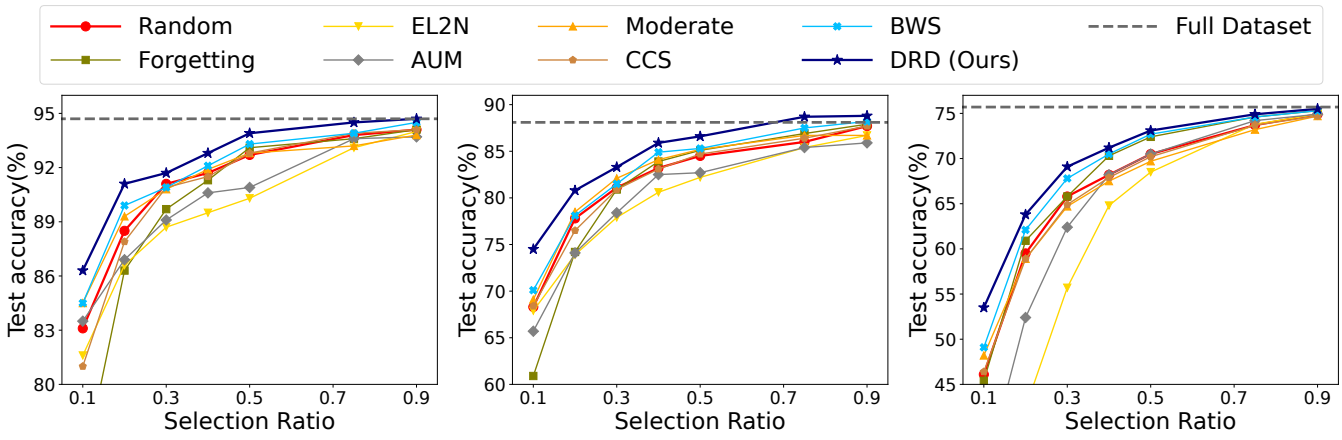
Experiments

Experimental Setup

Evaluation Protocol. We conduct core-set selection experiments on three datasets with increasing levels of difficulty: ImageNette, ImageWoof, and ImageNet-1K (Russakovsky et al. 2015). ImageNette contains 10 classes with low intra-class similarity, while ImageWoof comprises 10 visually similar dog breeds, posing a more challenging classification task. We use ResNet-18 for ImageNette and ImageWoof, and ResNet-50 for ImageNet-1K. We compare our **D**iffusion **R**econstruction **D**eviation (**DRD**) method against seven baselines, including: (1) **Random**, (2) **Forgetting** (Toneva et al. 2019), (3) **EL2N** (Paul, Ganguli, and Dziugaite 2021b), (4) **AUM** (Pleiss et al. 2020), (5) **Moderate** (Xia et al. 2023), and two score-oriented selection strategies—(6) **CCS** (Zheng et al. 2023) and (7) **BWS** (Choi, Ki, and Chung 2024). Forgetting score is used as the underlying criterion for both CCS and BWS, following their default setup. More details about the baselines and experiments are provided in Appendix B.

Implementation Details. We use the pretrained Diffusion Transformer (DiT) model from the official PyTorch implementation[1], trained with a full diffusion schedule of $T = 1000$ steps. During inference, we adopt the standard DDIM sampler with $T = 50$ and perform partial-step reconstruction from a selected $t < 50$, determined by our IB-based timestep selection strategy. To approximate $p_\theta(c \mid x_t)$ in Equation (7), we sample 20 groups of random noise. We set $B = 20$ for the Monte Carlo estimate in Equation (9). For data selection, we use the BWS strategy by default, scanning fixed-width windows over the samples sorted by DRD score and selecting the one yielding the highest validation accuracy on the full training set. Detailed window start points are reported in Appendix B.2. All reported results of our method can be obtained using a single RTX 4090 GPU.

[1]https://github.com/facebookresearch/DiT



(a) ImageNette (b) ImageWoof (c) ImageNet-1K

Figure 4: **(a, b, c) Core-set selection results.** Test accuracy of models trained on data subsets selected by different methods at varying selection ratios on ImageNette, ImageWoof, and ImageNet-1K. Our DRD method outperforms all baselines across selection budgets and approaches full-data performance with only 50% of the data. See Appendix C for detailed results.

Benchmark Evaluations

In Figure 4, we report the test accuracy of models trained on data subsets selected by various methods across a range of selection ratios on ImageNette, ImageWoof, and ImageNet-1K. Across all benchmarks and selection budgets, our proposed DRD score consistently achieves superior performance. Notably, DRD achieves near full-data performance using only 50% of the data across all datasets. A particularly informative comparison lies in contrasting our DRD method with BWS under a shared selection strategy. Both methods adopt an identical sliding-window mechanism guided by full-data validation accuracy to identify optimal windows. The only difference stemming from the scoring criterion is that BWS defaults to using the Forgetting score, whereas DRD relies on our likelihood-sensitive reconstruction deviation. Despite operating under the same selection protocol, DRD achieves significantly better results, underscoring the effectiveness of our score. Specifically, under selection ratios $\leq 40\%$, DRD surpasses BWS (with Forgetting score) by average margins of 1.3%, 2.5%, and 2.0% on ImageNette, ImageWoof, and ImageNet-1K, respectively.

Furthermore, we provide a cross-architecture evaluation in Appendix D, where different data selection methods are evaluated on EfficientNet-B0 and ViT across two datasets. Overall, DRD generally achieves the highest test accuracy under different selection ratios, with particularly notable gains on the more challenging ImageWoof dataset, demonstrating its strong cross-architecture generalizability.

Ablation Study and Analysis

Score-Oriented Selection Strategy Comparison. To better understand the impact of selection strategy on overall performance, we compare two score-oriented strategies CCS and BWS across three scoring functions: Forgetting, EL2N, and our DRD. Experiments are conducted on ImageWoof under varying selection ratios. As shown in Table 1, BWS consistently outperforms CCS across all scoring functions and selection ratios, highlighting the advantage of window-based selection over bin-based coverage. Among the three scores, DRD basically achieves the best performance under both strategies, confirming its strength as a likelihood-sensitive and distribution-aware signal. Notably, BWS with DRD yields the highest accuracy across all selection ratios. These results suggest that DRD not only provides a more effective ranking signal but also synergizes particularly well with curriculum-aligned strategies such as BWS.

Effectiveness of Information-Theoretic Timestep Selection. Figure 5 compares the performance of our information-bottleneck-informed (IB-informed) timestep selection method against fixed timesteps chosen via naive grid search on ImageNette and ImageWoof. Across all selection ratios, our method generally yields superior test accuracy, demonstrating that the class-wise deviations computed using our selected timesteps lead to more discriminative likelihood estimation. Empirically, the selected timesteps tend to cluster around $t = 20$, explaining why fixed $t = 20$ performs comparably well. This suggests that the noise level at this range strikes a balance between retaining class-relevant features and enabling informative reconstruction. However, naive grid search requires computing reconstruction deviations for the entire dataset at multiple candidate timesteps, followed by data selection and full training for each setting, making it computationally prohibitive. In contrast, we identify effective timesteps using a lightweight Monte Carlo estimate over a small subset of samples, typically within a few minutes. This efficiency, together with its strong empirical performance, makes our approach both practical and principled for the core-set construction.

Hyperparameter Sensitivity of Reconstruction Timestep Selection. We conduct a sensitivity analysis on two key hyperparameters of our information-theoretic timestep selection strategy: the number of class samples B used for

Selection	Score	10%	30%	75%
CCS	Forgetting	68.4±0.9	80.9±0.3	86.4±0.2
	EL2N	64.1±0.5	75.8±0.6	83.6±0.8
	DRD	68.9±0.1	80.1±0.8	86.9±0.5
BWS	Forgetting	70.1±0.7	*81.5±0.5*	87.3±0.8
	EL2N	68.9±0.5	80.9±0.5	86.5±0.3
	DRD	**74.5±0.5**	**83.3±0.4**	**88.7±0.8**

Table 1: **Score-oriented selection strategy comparison.** Test accuracy on ImageWoof under varying selection ratios using two selection strategies: CCS (bin-based coverage-centric sampling) and BWS (validation-guided window search), each applied to three scores.

B	$\#\epsilon$	10%	30%	75%
20	5	67.8±0.4	81.1±0.6	86.3±0.2
	20	74.5±0.5	83.3±0.4	88.5±0.8
	40	74.6±0.4	83.1±0.4	88.6±0.7
40	5	68.3±0.4	82.1±0.5	87.1±0.4
	20	**74.7±0.3**	83.3±0.4	88.5±0.6
	40	74.6±0.3	**83.4±0.4**	**88.7±0.7**

Table 2: **Hyperparameter sensitivity of timestep selection.** Test accuracy on ImageWoof under varying selection ratios with different numbers of class-conditional samples B and noise samples $\#\epsilon$ used in our IB-based reconstruction timestep selection strategy.

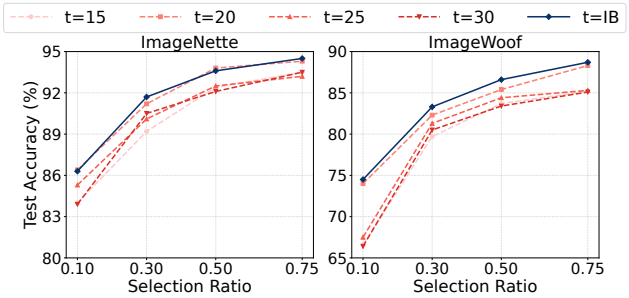


Figure 5: **Comparison of different reconstruction timesteps.** Test results of fixed timesteps and timesteps selected by our IB-informed method on ImageNette and ImageWoof. While grid search requires full reconstruction and evaluation over the entire dataset, our method identifies effective, class-wise timesteps using lightweight Monte Carlo estimates, *enabling timestep selection within minutes*.

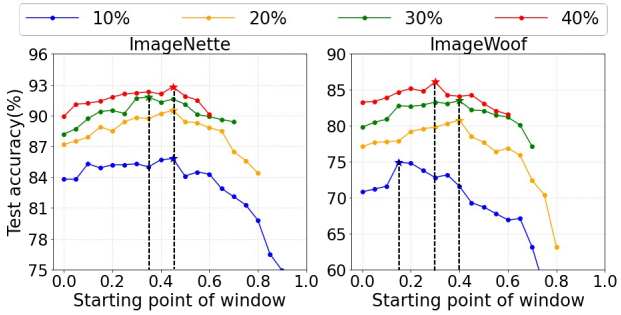


Figure 6: **Comparison of selected window start points based on DRD score.** Test accuracy on ImageNette and ImageWoof when sliding a fixed-width selection window across the DRD-sorted score list. High-performing subsets consistently arise from windows starting between 20% and 40% of the ranked list, reflecting a moderate-likelihood preference in model learning.

Monte Carlo estimation in Equation (9), and the number of noise samples $\#\epsilon$ used to approximate $p_\theta(c \mid x_t)$ of the diffusion classifier as defined in Equation (7). Results on ImageWoof are summarized in Table 2. We observe that using an extremely small number of noise samples (e.g., $\#\epsilon = 5$) leads to noticeably degraded performance, particularly under low selection ratios. This is likely due to unreliable estimation of $p_\theta(c \mid x_t)$, which in turn causes instability in the estimated timestep derivative and, ultimately, degrades the reliability of the reconstruction deviation. However, once $\#\epsilon$ is set to 20 or higher, performance stabilizes and becomes largely insensitive to further increases. A similar trend is observed with respect to B. Based on these findings, we set $B = 20$ and $\#\epsilon = 20$ as default values.

Analysis of Selection Window Start Points. To investigate how the structure induced by our DRD score relates to core-set quality, we perform a sliding-window analysis over the sorted score list. For each selection ratio, we fix the window width and vary its starting point, then evaluate the test accuracy of models trained on the resulting subsets. The results, shown in Figure 6, reveal consistent and interpretable patterns across both ImageNette and ImageWoof. In particular, we find that the highest-performing subsets consistently emerge from a mid-range window—typically starting between the 20% and 40% mark in the DRD-sorted list. Performance degrades sharply when selecting from either extreme, suggesting that samples with the highest or lowest DRD scores contribute less effectively to learning. This observation reinforces our earlier hypothesis: the most informative and discriminative subsets often lie within the moderate-likelihood region, where examples are neither trivially easy nor overly noisy.

Conclusion

In this work, we propose a novel and principled approach to core-set selection by leveraging diffusion-based reconstruction deviation as a likelihood-informed scoring criterion. By formally linking reconstruction error to data likelihood through the ELBO of diffusion models and guided by an information-theoretic reconstruction timestep selection strategy, our method offers both theoretical interpretability and practical effectiveness. Experiments on ImageNet demonstrate consistent performance gains over existing score-based methods. Besides empirical improvements, our analysis sheds light on the relationship between data distributional characteristics and model learning preferences.

Acknowledgments

Jiawei Du was supported by the A*STAR Career Development Fund (Grant No. C233312004) and by the National Research Foundation, Singapore under its Digital Trust Centre Innovation Grant (DTC Award No: DTC-IGC-02). Yi Wang was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120058). Wei Wang was supported by Advanced Materials-National Science and Technology Major Project (Grant No. 2025ZD0620100), Guangdong Provincial Key Laboratory of Integrated Communication, Sensing, and Computation for Ubiquitous Internet of Things (Grant No. 2023B1212010007), the Guangzhou Municipal Science and Technology Project (Grant Nos. 2023A03J0003, 2023A03J0013, and 2024A03J0621), and the Institute of Education Innovation and Practice Project (Grant No. HKUST(GZ)-ROP2025015).

References

Bellier, G. L.; and Audebert, N. 2024. Detecting Out-Of-Distribution Earth Observation Images with Diffusion Models. In *CVPR Workshops*, 481–491. IEEE.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Choi, H.; Ki, N.; and Chung, H. W. 2024. BWS: Best Window Selection Based on Sample Scores for Data Pruning across Broad Ranges. In *ICML*. OpenReview.net.

Feldman, V.; and Zhang, C. 2020. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. In *NeurIPS*.

Ghorbani, A.; and Zou, J. Y. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2242–2251. PMLR.

Graham, M. S.; Pinaya, W. H.; Tudosiu, P.-D.; Nachev, P.; Ourselin, S.; and Cardoso, J. 2023. Denoising Diffusion Models for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2947–2956.

Guo, C.; Zhao, B.; and Bai, Y. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, 181–195. Springer.

Hacohen, G.; and Weinshall, D. 2019. On The Power of Curriculum Learning in Training Deep Networks. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2535–2544. PMLR.

Hao, J.; Ji, K.; and Liu, M. 2023. Bilevel Coreset Selection in Continual Learning: A New Formulation and Algorithm. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 51026–51049. Curran Associates, Inc.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Joaquin, A. S.; Wang, B.; Liu, Z.; Asher, N.; Lim, B.; Muller, P.; and Chen, N. F. 2024. In2Core: Leveraging Influence Functions for Coreset Selection in Instruction Finetuning of Large Language Models. *arXiv preprint arXiv:2408.03560*.

Ki, N.; Choi, H.; and Chung, H. W. 2023. Data Valuation Without Training of a Model. In *ICLR*. OpenReview.net.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 491–507. Springer.

Kung, P.; Yin, S.; Chen, Y.; Yang, T.; and Chen, Y. 2021. Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity. In *EMNLP (1)*, 416–428. Association for Computational Linguistics.

Kwon, Y.; Rivas, M. A.; and Zou, J. 2021. Efficient Computation and Analysis of Distributional Shapley Values. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, 793–801. PMLR.

Kwon, Y.; and Zou, J. 2022. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, 8780–8802. PMLR.

Li, A. C.; Prabhudesai, M.; Duggal, S.; Brown, E.; and Pathak, D. 2023. Your Diffusion Model is Secretly a Zero-Shot Classifier. In *ICCV*, 2206–2217. IEEE.

Liu, Z.; Zhou, J. P.; Wang, Y.; and Weinberger, K. Q. 2023. Unsupervised Out-of-Distribution Detection with Diffusion Inpainting. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 22528–22538. PMLR.

Luo, C. 2022. Understanding Diffusion Models: A Unified Perspective. *CoRR*, abs/2208.11970.

Mirzasoleiman, B.; Bilmes, J. A.; and Leskovec, J. 2020. Coresets for Data-efficient Training of Machine Learning Models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 6950–6960. PMLR.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021a. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021b. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *NeurIPS*, 20596–20607.

Pleiss, G.; Zhang, T.; Elenberg, E. R.; and Weinberger, K. Q. 2020. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *NeurIPS*.

Pooladzandi, O.; Davini, D.; and Mirzasoleiman, B. 2022. Adaptive Second Order Coresets for Data-efficient Machine Learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 17848–17869. PMLR.

Pruthi, G.; Liu, F.; Kale, S.; and Sundararajan, M. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Renduchintala, H. S. V. N. S. K.; Bhatia, S.; and Ramakrishnan, G. 2024. SMART: Submodular Data Mixture Strategy for Instruction Tuning. In *ACL (Findings)*, 12916–12934. Association for Computational Linguistics.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3): 211–252.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Toneva, M.; Sordoni, A.; des Combes, R. T.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2019. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *ICLR (Poster)*. OpenReview.net.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *International Conference on Machine Learning (ICML)*.

Xia, X.; Liu, J.; Yu, J.; Shen, X.; Han, B.; and Liu, T. 2023. Moderate Coreset: A Universal Method of Data Selection for Real-world Data-efficient Deep Learning. In *ICLR*. OpenReview.net.

Yang, Y.; Kang, H.; and Mirzasoleiman, B. 2023. Towards Sustainable Learning: Coresets for Data-efficient Deep Learning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 39314–39330. PMLR.

Yoon, J.; Madaan, D.; Yang, E.; and Hwang, S. J. 2022. Online Coreset Selection for Rehearsal-based Continual Learning. In *ICLR*. OpenReview.net.

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*, 6022–6031. IEEE.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 586–595. Computer Vision Foundation / IEEE Computer Society.

Zheng, H.; Liu, R.; Lai, F.; and Prakash, A. 2023. Coverage-centric Coreset Selection for High Pruning Rates. In *The Eleventh International Conference on Learning Representations*.