# An Adaptive Sampling Framework for Diffusion-based Dataset Distillation with High Fidelity and Diversity

**Sunbeom Jeong**[1], **Sehwan Kim**[1], **Hyeonggeun Han**[1], **Hyungjun Joo**[2], **Sangwoo Hong** [3†], **Jungwoo Lee**[1,4,5†]

[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Samsung Electronics
[3]Department of Computer Science and Engineering, Konkuk University
[4]NextQuantum, Seoul National University
[5]Hodoo AI Labs
{sb3991, sehwankim, hygnhan, junglee}@snu.ac.kr
hyjun.joo@samsung.com, swhong06@konkuk.ac.kr

**Abstract**

Dataset distillation (DD) aims to generate a compact synthetic dataset that enables efficient training of neural networks while maintaining performance comparable to that achieved with the original dataset. However, existing methods often suffer from two main limitations. They either rely on computationally intensive iterative optimization procedures or depend heavily on architecture-specific designs. These issues limit their practicality for large-scale datasets and hinder generalization across different model architectures. To overcome these challenges, recent research has explored the use of diffusion models as an architecture-agnostic approach to dataset distillation, offering improved scalability and generalization for large-scale datasets across diverse model architectures. While diffusion-based dataset distillation methods have shown considerable potential, several challenges remain. Notably, certain approaches exhibit a distributional mismatch between the pre-trained diffusion model and the target dataset, which can adversely affect the fidelity and representativeness of the generated samples. Others require substantial fine-tuning to achieve high fidelity, which negates the benefits of architectural flexibility. In this work, we propose a new diffusion-based dataset distillation framework that effectively preserves the characteristics of the original dataset without requiring any fine-tuning. Our method employs adaptive sampling and repulsion regularization to enhance both the fidelity and diversity of generated samples. As a result, the proposed approach outperforms state-of-the-art distillation methods across a wide range of datasets and model architectures.

**Code** — https://github.com/sb3991/adaptive-diffusion-dd

## 1 Introduction

Modern advancements in deep learning have enabled remarkable achievements in various fields, but these gains come at the cost of huge computational and storage requirements due to large-scale datasets and complex models. To address these challenges, Dataset Distillation (DD) has emerged as a promising solution, aiming to condense the information of large datasets into smaller, synthetic ones. These surrogate datasets enable faster training with significantly reduced memory and computation demands while maintaining remarkable performance. However, most DD methods have focused on small, low-resolution datasets, limiting their applicability to high-resolution, large-scale datasets such as ImageNet (Deng et al. 2009). Traditional approaches, including gradient matching and meta-learning-based optimization, face scalability issues due to their pixel-level optimization and costly iterative updates, making them impractical for real-world applications.

Recently, many researchers (Su et al. 2024; Gu et al. 2024; Wang et al. 2025; Zou et al. 2025) have utilized pre-trained diffusion models for DD, aiming to overcome the limitations of traditional methods. Diffusion models have demonstrated the ability to generate high-resolution images with strong generalization performance across various network architectures, highlighting their potential for DD. However, a significant challenge arises when using diffusion models for DD, primarily due to the distribution gap between diffusion models' training data and target DD datasets. This distribution gap causes generated images to deviate from the original distribution, degrading the performance of distilled datasets and limiting the applicability of the diffusion model in DD. While Minimax (Gu et al. 2024) attempted to address this issue through fine-tuning the diffusion models, this approach is computationally expensive and impractical for large-scale datasets. Foundation models, including Stable Diffusion (Rombach et al. 2022) used in methods like $D^4M$, employ Classifier-Free Guidance (CFG) (Ho and Salimans 2021) to reduce this gap by conditioning image generation on class labels or textual prompts. Despite these efforts, the distribution gap persists. Moreover, strong guidance reduces diversity, often producing repetitive or overly similar images concentrated around specific modes.

We propose a novel diffusion-based DD framework that tackles this challenge by introducing two key innovations in

†Corresponding authors: Sangwoo Hong; Jungwoo Lee

the sampling process, which are i) Adaptive Sampling with Bayesian optimization and ii) Repulsion regularization. Our approach leverages the diffusion model's capability to generate high-quality images while preserving key features of the original dataset. By controlling the influence of input images and guidance strength during the sampling stage of the diffusion model, we ensure that the generated data retains essential attributes from the source as indicated in Figure 1. At the same time, the repulsion regularization allows for controlled diversity, producing synthetic datasets that not only align with the target distribution but also introduce meaningful variations that reduce unnecessary duplication within the distilled dataset.

Our main contributions are summarized as follows:

- Leveraging pre-trained diffusion models, we propose a dataset distillation method that effectively balances the need for *fidelity* to original data and the generation of *novel, diverse* samples *without requiring any additional training* of diffusion models.
- Extensive experiments demonstrate that our proposed framework achieves *state-of-the-art performance across large-scale datasets and diverse network architectures* with significantly reduced computational requirements.
- We also provide a theoretical analysis that our proposed method effectively reduces class conditional Rényi mutual information, penalizing redundancy and enhancing the diversity of the distilled dataset.

## 2 Related Work

### 2.1 Dataset Distillation

DD focuses on compressing large-scale datasets into tiny artificial datasets while preserving their essential properties. Initially introduced as a meta-learning problem (Wang et al. 2018), DD methods have evolved to incorporate various optimization strategies. Early approaches (Nguyen et al. 2021; Nguyen, Chen, and Lee 2021) utilized kernel-based models to minimize computational overhead while preserving dataset information. Subsequently, matching-based methods emerged, focusing on aligning gradients (Zhao, Mopuri, and Bilen 2021), feature distributions (Zhao and Bilen 2023; Wang et al. 2022), and training trajectory (Cazenavette et al. 2022; Cui et al. 2023) between synthetic and real datasets. Since then, many DD methods have been proposed. DATM (Guo et al. 2024) suggests aligning early trajectories for small images-per-class (IPC) and late trajectories for large IPCs. SRe$^2$L (Yin, Xing, and Shen 2023) relabels synthetic samples using pre-trained classifiers for model updates, and RDED (Sun et al. 2024) utilizes patches with the highest confidence from real images to form synthetic data. However, many of these approaches require extensive optimization processes, and their performance is heavily dependent on the architecture.

More recently, the success of diffusion models has introduced new paradigms for DD, enabling DD on large-scale datasets with cross-architecture generalization. D$^4$M (Su et al. 2024) adopts a pretrained text-to-image diffusion model to cluster images in latent space and synthesize data.

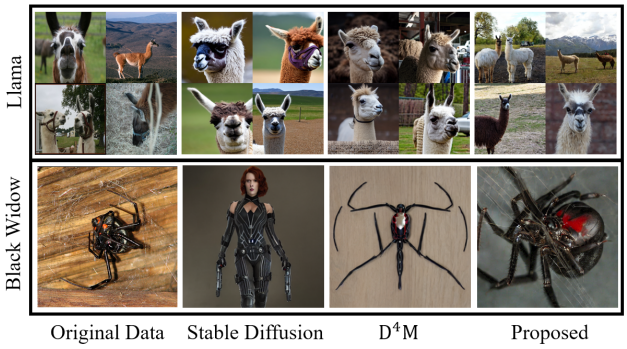


Figure 1: Comparison between generated images

However, in the absence of controls to bridge the gap between the diffusion model's pretraining data and the target dataset, the resulting images can be substantially different from the original ones, misguided by CFG that heavily depends on the label-prompt. As illustrated in Figure 1, CFG-generated images might consistently depict similar llama images, indicating limited diversity, or inaccurately generate a movie character instead of a spider, highlighting the unresolved distribution gap. D$^4$M attempts to mitigate this issue by initializing image generation with original input images, but still struggles to achieve both high fidelity and sufficient diversity simultaneously. More recently, $CaO_2$ (Wang et al. 2025) adopts high-confidence sample filtering and latent adjustment to mitigate the distribution gap, and in (Zou et al. 2025), the authors propose injecting vision–language cues to refine semantic fidelity. Nevertheless, both methods remain largely fidelity-focused, offering only limited gains in diversity and incurring extra preprocessing overhead.

On the other hand, Minimax (Gu et al. 2024) fine-tunes the DiT model (Peebles and Xie 2023) pretrained on the ImageNet dataset, applying regularization terms that encourage representativeness and diversity. Nevertheless, the computational cost of fine-tuning diffusion models can be overwhelming when the distribution of the target data for distillation is considerably different from that of the data used for training the diffusion models. Alternatively, IGD (Chen et al. 2025) and MGD$^3$ (Santiago et al. 2025) also explore training-free diffusion-based DD from diffusion models. IGD steers sampling by backpropagating an influence-based objective through the diffusion model and MGD$^3$ clusters VAE latents into discrete modes and applies mode-specific guidance and stop-guidance to push samples toward different modes. However, these methods require either gradient access to the diffusion model or additional latent clustering and architecture-specific design.

To tackle these limitations, we propose a novel DD framework that integrates adaptive sampling and repulsion regularization into the diffusion model's sampling process. Our proposed method controls the influence of the original data to improve fidelity and reduces redundancy among generated samples, generating diverse samples with high-fidelity without any computationally intensive fine-tuning.

### 2.2 Diffusion Models

Diffusion models are generative models designed to approximate the data distribution by transforming random noise into meaningful samples. The primary objective of diffusion models is to generate samples that faithfully approximate the underlying data distribution $P(x)$.

Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) introduced the concept of learning a reverse process of a fixed Markov chain to generate high-quality samples. However, due to the computational complexity involved in working with the original pixel space, optimization and evaluation in DDPMs can be resource-intensive. To address this issue, the Latent Diffusion Model (LDM) (Rombach et al. 2022) abstracts imperceptible details into a compact latent space using a VAE framework. LDMs follow a structured training process where data is first transformed into a latent space representation through an encoder and later reconstructed via a decoder. The forward process gradually adds noise from $t = 0$ to $t = 1$, expressed as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t$ is a time-dependent hyperparameter and $\epsilon \sim \mathcal{N}(0, I)$ represents the added noise. The models are trained by minimizing the mean squared error between the predicted noise and the actual noise, which can be expressed as:

$$L_t = \|\epsilon_\theta(z_t, c) - \epsilon\|_2^2. \tag{2}$$

By removing the predicted noise from $t = 1$ to $t = 0$, the reverse process efficiently generates high-quality images.

Moreover, to improve the control over the output, classifier guidance (Dhariwal and Nichol 2021) and classifier-free guidance (CFG) (Ho and Salimans 2021) methods have been introduced. Classifier guidance incorporates an external classifier into the reverse process to guide the generation toward specific classes. On the other hand, CFG eliminates the dependency on an external classifier by jointly training the model on both conditional and unconditional objectives. CFG allows direct control during sampling by interpolating between the two, which can be expressed as:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, c) + \beta(\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t)), \tag{3}$$

where $\beta$ is the classifier-free guidance scale.

## 3 Background

In this section, we first introduce two key objectives of DD: *Fidelity* and *Diversity*. We also describe two core components of our method for achieving these goals: *Bayesian Optimization* (BO) and *Mutual Information* (MI).

### 3.1 Fidelity & Diversity

*Fidelity* and *Diversity* are the two central objectives of DD, steering the distilled set to preserve the semantic content of the original data and its natural variation.

*Fidelity* ensures that distilled samples remain faithful to the original distribution's semantics and structure. In diffusion-based DD, fidelity is particularly critical because pretrained diffusion models often exhibit a distribution gap from the target data, which can cause failure to preserve key characteristics of the original data.

*Diversity* captures intra-class variability and multiple modes of the target distribution. Insufficient diversity yields mode collapse, where generated samples become overly similar and fail to capture the full range of valid variations.

In DD, maintaining balance between fidelity and diversity is a critical issue. Over-emphasizing fidelity can make the distilled set rigid and less generalizable, while over-emphasizing diversity may harm fidelity.

### 3.2 Bayesian Optimization and Its Role

Bayesian Optimization (BO) is a powerful framework for optimizing costly black-box functions where gradients are unavailable. We formulate the selection of diffusion sampling parameters as a *black-box optimization problem*, and use BO to efficiently search for optimal settings that enhance the quality of the distilled dataset while minimizing costly evaluations. More details will be discussed in Section 4.2.

**Gaussian process surrogate.** BO models the objective $f(\cdot)$ with a Gaussian Process (GP) $\mathcal{GP}(m, k)$. Given observations $\mathcal{M} = \{(x_i, f_i)\}$, the GP posterior at $x$ provides a predictive mean $m(x)$ and uncertainty $\sigma(x)$, which guide sample-efficient exploration of the search space.

**Acquisition function.** In BO, candidates are chosen by maximizing an acquisition function. An acquisition function converts the GP model's predictive mean $m(x)$ and uncertainty $\sigma(x)$ into a single utility that scores where to evaluate next. We use the Upper Confidence Bound (UCB),

$$\mathrm{UCB}(x) = m(x) + \kappa\,\sigma(x),$$

where $\kappa$ balances exploration and exploitation.

### 3.3 Mutual Information

To promote diversity, we regulate the MI between generated samples during the sampling process of diffusion models. Shannon MI quantifies the dependency between two variables. If samples belonging to the same class have a high MI value, this indicates redundancy, suggesting that there exist repetitive samples within a class. To better control redundancy in DD, we employ Rényi Mutual Information (RMI), a generalization of Shannon MI, defined as:

$$I_\alpha = \frac{\log\left(\int\int p_{X,Y}(x,y)^\alpha p_X(x)^{1-\alpha} p_Y(y)^{1-\alpha}\,dx\,dy\right)}{\alpha - 1}$$

where $\alpha > 0, \alpha \neq 1$. As $\alpha \to 1$, RMI converges to Shannon MI.

In the literature (Jalali, Li, and Farnia 2023a; Friedman and Dieng 2023; Pasarkar and Dieng 2024; Jalali, Li, and Farnia 2023b; Pál, Póczos, and Szepesvári 2010), RMI has been used to measure the diversity of generative models, measuring dependencies across various distributions. Specifically, class-conditional RMI is defined as:

$$I_\alpha(X; Y \mid C) = \mathbb{E}_{P_C}[I_\alpha(X; Y \mid C = c)], \tag{4}$$

which quantifies information overlap within each class. By penalizing samples with high $I_\alpha(X; Y \mid C)$, our method promotes diversity by reducing redundancy among samples and improving coverage of the original dataset distribution.

Algorithm 1: PROPOSED ALGORITHM

**Input:** diffusion model $\mathcal{D}$, teacher $T$, real dataset $\mathcal{D}_{\text{real}}$, search grids $\mathcal{B}$ (for $\beta$), $\Gamma$ (for $\gamma$), BO budget $B$, repulsion weight $\lambda$, mini-batch size $m_{\text{seed}}$, IPC size $N$
**for each** class $c$ in the label set **do**
  $S_c \leftarrow \emptyset$ {accumulated distilled images for class $c$}
  **while** $|S_c| < N$ **do**
    Sample a mini-batch of seed images $I \subset \mathcal{D}_{\text{real}}^{(c)}$ with $|I| = m_{\text{seed}}$
    $\mathcal{M} \leftarrow \emptyset$ {BO observation set for this mini-batch}
    Initialize a GP surrogate on $\mathcal{B} \times \Gamma$
    **for** $b = 1$ **to** $B$ **do**
      Propose $(\beta, \gamma)$ via GP-UCB over $\mathcal{B} \times \Gamma$
      Generate proxy samples $G \leftarrow G_{\beta,\gamma}^{\text{proxy}}(I)$ (reduced denoising steps)
      $f \leftarrow -\mathrm{KL}\big(T(I) \,\|\, T(G)\big)$ {mini-batch fidelity}
      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\beta, \gamma, f)\}$; update GP with $\mathcal{M}$
    **end for**
    $(\beta^*, \gamma^*) \leftarrow \arg\max_{(\beta,\gamma)\in\mathcal{S}} f$
    Generate full-quality images $\hat{G} \leftarrow G_{\beta^*,\gamma^*}^{\text{full}}(I; \lambda)$ with repulsion regularization in Eq. (6)
    $S_c \leftarrow S_c \cup \hat{G}$
  **end while**
**end for**
**return** $\mathcal{S}_{\text{distilled}} = \bigcup_c S_c$

# 4 Proposed Method

We propose a diffusion-based DD method that enhances both fidelity and diversity through two key components:

1. Adaptive Sampling with BO: Automatically selects sample-specific sampling parameters by optimizing a fidelity objective, without finetuning the diffusion model.
2. Repulsion Regularization: Encourages diversity by reducing redundancy. Our theoretical analysis demonstrates that this regularization mechanism reduces the cross-conditional RMI.

## 4.1 Overall Distillation Process

We begin by describing our overall distillation process. In this process, random samples from the original dataset are selected and used as initialization points. For each class $c$, we sample a mini-batch $I \subset \mathcal{D}_{\text{real}}^{(c)}$ and run BO to select $(\beta, \gamma)$ that maximize a teacher-guided fidelity score on proxy generations with fewer denoising steps. We then regenerate the batch with the full sampler at $(\beta^*, \gamma^*)$ while applying repulsion regularization (Eq. (6)) to reduce redundancy, and accumulate the results until the target image per-class (IPC) size is satisfied. Algorithm 1 summarizes the overall process.

## 4.2 Adaptive Sampling with BO

We now provide more details about adaptive sampling.

**Problem Setting** In diffusion-based DD, the quality of generated samples is mainly influenced by two key sampling parameters[1], which are **CFG scale** $\beta$ and **denoising strength** $\gamma$. The optimal values of $(\beta, \gamma)$ vary depending on the class, prompt alignment, and initialization, resulting in a data-dependent and non-stationary objective. This motivates *adaptive* parameter selection rather than fixed settings. To enable adaptive parameter selection, we employ *teacher-guided fidelity* as the BO objective, where a higher consistency between teacher predictions on real and generated samples results in a higher score $f$.

**Adaptive Sampling** Unlike conventional methods (e.g., D$^4$M, Minimax) with fixed sampling strategies, we *dynamically* adjust the diffusion process using latent representations of real seeds to balance fidelity and diversity by modulating:

- **CFG scale** ($\beta$): controls the strength of adherence to the label prompt.
- **Denoising strength** ($\gamma$): adjusts how far the sample deviates from the initial image.

A higher CFG scale $\beta$ can enhance perceptual quality by exploiting the pretrained prior, but it can also lead to misalignment with the original image. The denoising strength $\gamma$ sets the reverse-process start step ($t_0{=}\gamma$) rather than a fixed $t_0{=}1$, thereby controlling how much of the original image is preserved. Using a smaller $\gamma$ retains dataset-specific details, while a larger $\gamma$ promotes diversity.

Since optimal values for $(\beta, \gamma)$ differ across samples and prompts, a fixed sampling strategy often fails to maintain both fidelity and diversity. Therefore, we evaluate candidate settings using the prediction agreement with a teacher network $T(x)$ as a proxy for fidelity. Moreover, since $T(x)$ is a black-box function of $(\beta, \gamma)$, we employ BO to adapt these parameters for each sample.

**Detailed Bayesian Optimization Process**

**Fidelity objective** For a batch $I$ and generated images $G = G_{\beta,\gamma}(I)$, we define the fidelity score using soft labels from the teacher as follows:

$$f(\beta, \gamma) = -\mathrm{KL}\big(P_{\text{orig}} \,\|\, P_{\text{gen}}\big), \tag{5}$$

where $P_{\text{orig}} = T(I)$ and $P_{\text{gen}} = T(G)$. Larger fidelity scores indicate stronger teacher agreement to the generated images.

**BO loop** Using the GP surrogate and UCB acquisition described in Section 3.2, BO proposes $(\beta, \gamma)$ candidates. To reduce the computational cost, each candidate is scored with a *proxy sampler* that uses *only half denoising steps* of the full sampler.

**Selection and regeneration** After the BO budget is exhausted, we select the best observed $(\beta^*, \gamma^*)$ and regenerate images with the full sampler at these settings.

## 4.3 Repulsion Regularization

We also introduce repulsion regularization to prevent the over-representation of dominant modes in distilled datasets. Repulsion regularization penalizes redundant structures in the latent space by incorporating a redundancy-aware kernel into the diffusion process.

[1]This will be further analyzed in Section 6

Specifically, in each diffusion step for batch latent variable $\{\mathbf{z}_i\}_{i=1}^N$, we modify the update rule to penalize samples with redundant information, which can be expressed as

$$\mathbf{z}_{i,t-1} = \frac{1}{\sqrt{\alpha_t}}\Big[\mathbf{z}_{i,t} - \eta_t\, \tilde{\epsilon}_\theta(\mathbf{z}_{i,t}, c) - \eta_t \lambda \sum_{j \neq i} \nabla_{\mathbf{z}_i} \mathrm{k}\big(\mathbf{z}_{i,t}, \mathbf{z}_{j,t}\big)\Big] + \sigma_t\, \boldsymbol{\xi}_i. \quad (6)$$

where:

- $\mathrm{k}(\cdot,\cdot)$ is a kernel that measures the redundancy between samples.
- $\lambda$ controls how strongly we push the latent vectors apart.
- $\eta_t = \dfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}$ is the effective DDPM step.
- $\sigma_t$ is the noise scale and $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \mathbf{I})$.
- $\tilde{\epsilon}_\theta(\mathbf{z}_t, c)$ is the noise prediction.

When $\lambda = 0$, the update rule reduces to the standard diffusion step without repulsion. By incorporating this repulsion term directly into each diffusion iteration, the latent diverges more effectively, mitigating the generation of duplicate samples. Consequently, the model yields a set of images that naturally cover a wider range of appearances and styles, boosting diversity. In our experiments, we use Radial Basis Function (RBF) kernel, $k(x, y) = \exp(-\|x - y\|_2^2/(2\sigma^2))$, with bandwidth $\sigma > 0$ which is widely recognized for its ability to handle complex, nonlinear relationships in data.

Our repulsion regularization is supported by Theorem 1, which shows that our regularization effectively reduces class conditional RMI and enhances diversity.

**Theorem 1.** *Let the $k(x, y) = \exp(-\|x - y\|_2^2/(2\sigma^2))$ be a positive-definite RBF kernel with bandwidth $\sigma > 0$. Then minimizing $\sum_{i,j\in V} \mathrm{k}\Big(\mathbf{z}_i^{(t)}, \mathbf{z}_j^{(t)}\Big)$ decreases $\widehat{I}_\alpha\left(\mathbf{X}_{1:n}|C\right)$, where $\widehat{I}_\alpha\left(\mathbf{X}_{1:n}|C\right)$ denotes the estimated class-conditional RMI between generated images.*

We provide the proof in the extended paper. By using a repulsion regularization, our proposed method reduces class-conditional RMI, penalizing unnecessary repetition between generated samples and promoting diversity.

### 4.4 Training with Soft-Labels

After synthesizing the distilled dataset using the diffusion model, we assess its effectiveness by training networks $S_\theta$. It has been known (Qin, Deng, and Alvarez-Melis 2024) that using soft labels from a teacher model for training large-scale synthetic datasets can significantly enhance the accuracy and generalizability of the student model. Consequently, many researchers have utilized this soft-label approach (Shen and Xing 2022), including TESLA (Cui et al. 2023), RDED (Sun et al. 2024), and $\mathrm{D}^4\mathrm{M}$ (Su et al. 2024). Therefore, we also implement the soft labels during the training process, following (Shen and Xing 2022).

| Method | Scalability | Train-Free | Dist. Match | Novelty |
|---|---|---|---|---|
| DATM | - | ✓ | ✓ | ✓ |
| Minimax | ✓ | - | ✓ | ✓ |
| $\mathrm{D}^4\mathrm{M}$ | ✓ | ✓ | - | ✓ |
| RDED | ✓ | ✓ | ✓ | - |
| Proposed | ✓ | ✓ | ✓ | ✓ |

Table 1: Properties and performance of various SOTA DD methods. 'Train-Free' indicates that fine-tuning of diffusion models is not required for new datasets, while 'Distribution match' means that the generated data follows the original data's distribution. Unlike other DD methods, our approach is the only scalable DD method that can generate diverse and novel samples with high fidelity, without requiring training.

## 5 Comparison with the SOTA Methods

### 5.1 Property Comparison

We first analyze the efficiency and generalization capability of our proposed method by comparing it with the state-of-the-art DD methods. Comparison results are detailed below and summarized in Table 1.

- **DATM:** DATM is an optimization-based DD method capable of achieving performance comparable to the original dataset. However, the excessive optimization process of DATM limits its scalability to high-resolution datasets.
- **$\mathrm{D}^4$M:** $\mathrm{D}^4\mathrm{M}$ is a diffusion-based DD method that uses the prototype samples generated by the original dataset as initialization points for the reverse process of diffusion. As a training-free approach, it offers computational efficiency, but uncontrolled guidance may sometimes produce samples with low fidelity.
- **Minimax:** Minimax fine-tunes a diffusion transformer (DiT) to generate diverse and representative samples. Minmax requires much higher computational and memory overhead compared to training-free approaches.
- **RDED:** RDED distills the datasets through cropping and selecting patches from the original dataset. Since it only reuses existing patches, it cannot generate novel samples absent in the original dataset. This limitation also raises concerns about privacy protection when direct use of original data is not permissible.

Contrary to previous methods, our proposed method effectively scales to large datasets, requires no additional training, maintains fidelity, and is capable of generating novel samples that were not present in the original dataset.

### 5.2 Experimental Results

We now evaluate the performance of the proposed method across various datasets and networks. The detailed experimental settings are provided in the extended paper.

**Visualization of distilled dataset** First, we present images generated by our method in Figure 2, along with those created by $\mathrm{D}^4\mathrm{M}$ and the original images. Our method notably improves the fidelity and diversity of the synthesized

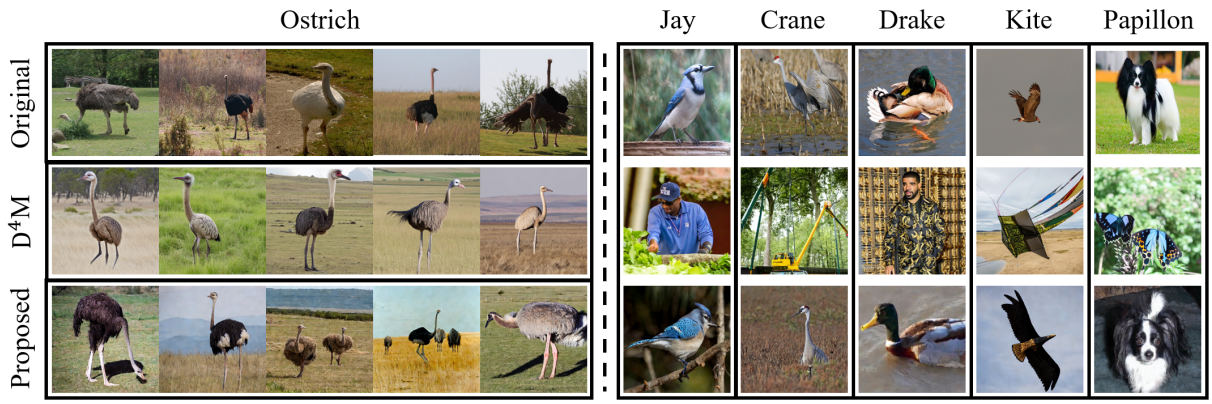


Figure 2: Visualization of random original images, images generated by D$^4$M (Su et al. 2024), and our proposed method. The first row comes from the original dataset and the second row comes from D$^4$M. In comparison to previous methods, our method significantly enhances the fidelity to the original data distribution and the diversity of the synthesized dataset.

| | | ResNet-18 | | | | | | ResNet-101 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dataset | IPC | SRe$^2$L | RDED | D$^4$M | Minimax | CaO$_2$ | Proposed | SRe$^2$L | RDED | D$^4$M | Minimax | CaO$_2$ | Proposed |
| ImageNette | 10 | 29.4 | 61.4 | 67.4 | 57.6 | 65.0 | **74.6** | 23.4 | 54.0 | 57.2 | 34.4 | **66.3** | 60.4 |
| | 50 | 40.9 | 80.4 | 81.6 | 73.0 | 84.5 | **87.8** | 36.5 | 75.0 | 73.6 | 64.8 | 81.7 | **86.0** |
| | 100 | – | 86.2 | 87.0 | 76.0 | – | **92.1** | – | 85.6 | 87.0 | 75.8 | – | **91.3** |
| ImageWoof | 10 | 20.2 | 38.5 | 43.0 | 37.6 | 45.6 | **57.2** | 17.7 | 31.3 | 32.0 | 38.3 | 36.5 | **39.6** |
| | 50 | 23.3 | 68.5 | 71.6 | 57.1 | 68.9 | **72.6** | 21.2 | 59.1 | 54.2 | 54.9 | 63.1 | **69.4** |
| | 100 | – | 69.4 | 73.0 | 63.4 | – | **78.7** | – | 66.6 | 74.8 | 61.6 | – | **76.7** |
| ImageNet-100 | 10 | 9.5 | 36.0 | 30.2 | 22.5 | 36.6 | **39.0** | 6.4 | 33.9 | 29.4 | 19.6 | 34.5 | **37.1** |
| | 50 | 27.0 | 61.6 | 67.2 | 40.9 | 68.0 | **70.7** | 25.7 | 66.0 | 71.2 | 37.1 | 70.8 | **72.7** |
| | 100 | – | 68.0 | 75.4 | 61.8 | – | **78.6** | – | 74.2 | 76.8 | 58.9 | – | **81.0** |
| ImageNet-1K | 10 | 21.3 | 42.0 | 27.9 | 17.4 | 46.1 | **46.6** | 30.9 | 48.3 | 34.2 | 17.2 | 52.2 | **56.3** |
| | 50 | 46.8 | 56.5 | 55.2 | 32.4 | 60.0 | **61.4** | 60.8 | 61.2 | 63.4 | 38.2 | 66.2 | **67.8** |
| | 100 | 52.8 | 60.5 | 59.3 | 52.7 | – | **63.2** | 62.8 | 64.1 | 66.5 | 53.2 | – | **68.8** |

Table 2: Experimental results on high-resolution dataset.

images. To be more specific, in the '*Ostrich*' class, images generated by D$^4$M exhibit high-quality appearances but tend to feature similar poses and only highlight the most prominent aspects of the objects. In contrast, our proposed method produces images with diverse poses, backgrounds, and appearances. Additionally, when the CFG is misaligned with the original distribution, such as in the '*Drake*' and '*Kite*' classes, D$^4$M generates completely inaccurate images. Conversely, our method successfully creates images that retain features of the original dataset. This suggests that the adaptive sampling and repulsion regularization implemented in our approach significantly enhance coverage of the original distribution, ensuring high fidelity and increased diversity.

**Performance in high-resolution Datasets** Following (Su et al. 2024; Gu et al. 2024; Sun et al. 2024), we have experimented on ImageNet and its three subsets: ImageNette, ImageWoof, and ImageNet-100. We have used ResNet-18 and ResNet-101 network architectures, and the experimental results are indicated in Table 2. Our proposed method demonstrates the highest performance across various datasets and networks, outperforming the baseline by a large margin.

**Comparison with recent training-free diffusion-based distillation methods** We also compare our method with MGD$^3$ and IGD in Table 3. Our core novelty is that we address a challenging and practical scenario where the backbone's pre-training and target data are unaligned. By utilizing BO and repulsion regularization, we enable DD in the unaligned setting, which is critical for generalizability, as it's impractical to have a pretrained diffusion model for every new dataset. Our method achieves the highest accuracy in difficult unaligned cases, even surpassing the performance of MGD$^3$ and IGD from the favorable aligned setting.

## 5.3 Image Quality Analysis

We now provide a quantitative analysis of the generation quality of synthetic datasets using diffusion models. To eval-

| ImageNet, IPC=50 | Ours | MGD$^3$ | IGD |
|---|---|---|---|
| Align. assumption | X | △ | O |
| Acc ( %, aligned case, DiT) | – | 60.2 | 59.8 |
| Acc ( %, unaligned case, LDM) | **61.4** | 58.5 | – |

Table 3: Property and performance comparison with recent diffusion-based DD methods. Align assumption indicates whether a method assumes the pre-training dataset for diffusion backbone is aligned with the target dataset for DD.

| Metrics | RDED | D$^4$M | Minimax | Proposed |
|---|---|---|---|---|
| FID Score ↓ | 56.4 | 61.7 | 57.9 | **48.7** |
| KID Score (1e-2) ↓ | 1.73 | 0.92 | 1.57 | **0.72** |
| Precision ↑ | 0.61 | 0.67 | **0.88** | 0.84 |
| Recall ↑ | 0.01 | 0.14 | 0.21 | **0.45** |
| Density ↑ | 1.63 | 1.35 | 1.84 | **1.86** |
| Coverage ↑ | 0.25 | 0.24 | 0.31 | **0.33** |

Table 4: Diversity and Fidelity analysis based on FID score, KID score, Precision, Recall, Density, and Coverage.

uate the diversity and fidelity of the generated data, we use Fréchet Inception Distance (FID) (Heusel et al. 2017), Kernel Inception Distance (KID) (Bińkowski et al. 2018), Precision, Recall (Kynkäänniemi et al. 2019), Density, and Coverage (Naeem et al. 2020), which are widely used metrics for assessing the synthetic image quality. FID, KID, and Precision are metrics closely related to the fidelity of the dataset. Recall, Density, and Coverage are closely associated with the diversity of the dataset.

As shown in Table 4, our proposed methods effectively improve all the metrics. Notably, our method significantly outperforms D$^4$M and even surpasses Minimax on all metrics except precision. This indicates that our method maintains fidelity by generating samples that closely resemble the original dataset, while enhancing diversity at the same time by covering a broader range of the original data distribution.

### 5.4 Distillation Cost Analysis

We also analyze distillation time and GPU memory usage, with results summarized in Table 5. MTT and TESLA, which are optimization-based DD methods, require high GPU memory and generation time. Diffusion-based methods such as D$^4$M and Minimax show much lower memory and generation time. Notably, our method is training-free, achieving lower generation time and memory than Minimax. Moreover, unlike D$^4$M, our method does not require latent prototype generation, resulting in $40.6\%$ memory usage of D$^4$M while achieving similar generation time. We also note that of this 2.7s per-image generation time of our method, the BO overhead accounts for 0.7s (26%).

## 6 Effect of Adaptive Sampling

Moreover, to analyze the effect of $\beta$, $\gamma$ on the quality of the generated image, we visualize the images generated with various $\beta$, $\gamma$ in Figure 3. As demonstrated in the figure, employing adaptive sampling that utilizes an adequate $(\beta, \gamma)$ pair is essential for generating high-fidelity samples. However, the appropriate values vary across different datasets and classes, and testing all possible combinations is resource-intensive. Therefore, we employ BO to identify the optimal values, effectively determining the settings for producing high-fidelity samples.

| | MTT | TESLA | Minimax | D$^4$M | Proposed |
|---|---|---|---|---|---|
| Time(s) | 45.0 | 46.0 | 10.1 | **2.6** | 2.7 |
| GPU(GB) | 79.9 | 13.9 | 10.0 | 6.4 | **2.6** |

Table 5: Comparison of diffusion-based distillation methods in terms of time and GPU memory usage.

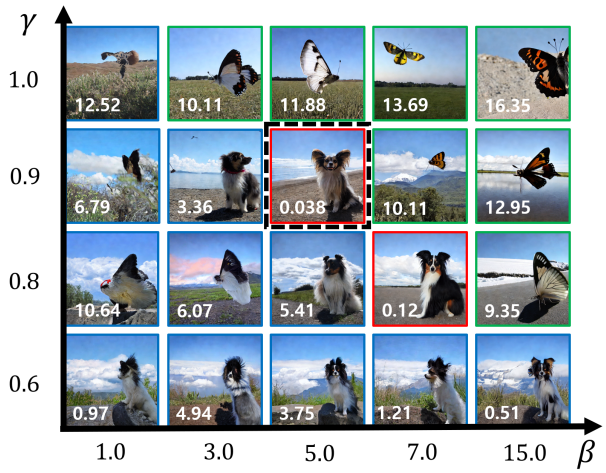


Figure 3: The effect of $\beta$ and $\gamma$ on the generated samples "papillon". The **blue** line represents samples with insufficient detail due to a low CFG scale, while the **green** line represents samples with low fidelity due to incorrect guidance. Conversely, the **red** line indicates high-quality samples with sufficient detail. We also indicate the fidelity score from Equation (5) in the figure and indicate the sample whose parameters are selected by BO using **black** dotted lines.

## 7 Conclusion

In this work, we proposed a novel diffusion-based DD framework that effectively generates diverse and realistic samples. By leveraging adaptive sampling and repulsion regularization, our approach enhances the diversity of the distilled dataset while maintaining fidelity to the original data distribution. Unlike previous methods that rely on architecture-dependent optimization or expensive fine-tuning, our framework does not require any fine-tuning process and generalizes well across different architectures, significantly reducing both memory requirements and distillation time. Experiments demonstrate that our method outperforms state-of-the-art DD approaches across various datasets and architectures.

## Acknowledgments

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2025-16072709(25%), RS-2024-00451435(20%), RS-2024-00413957(15%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(10%), RS-2025-02273157(10%), RS-2025-25442149(10%) RS-2021-II211343(10%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

## References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset Distillation by Matching Training Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 4750–4759.

Chen, M.; Du, J.; Huang, B.; Wang, Y.; Zhang, X.; and Wang, W. 2025. Influence-Guided Diffusion for Dataset Distillation. In *The Thirteenth International Conference on Learning Representations*.

Cui, J.; Wang, R.; Si, S.; and Hsieh, C.-J. 2023. Scaling Up Dataset Distillation to ImageNet-1K with Constant Memory. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 6565–6590. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. 248–255.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 8780–8794. Curran Associates, Inc.

Friedman, D.; and Dieng, A. B. 2023. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*.

Gu, J.; Vahidian, S.; Kungurtsev, V.; Wang, H.; Jiang, W.; You, Y.; and Chen, Y. 2024. Efficient Dataset Distillation via Minimax Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15793–15803.

Guo, Z.; Wang, K.; Cazenavette, G.; LI, H.; Zhang, K.; and You, Y. 2024. Towards Lossless Dataset Distillation via Difficulty-Aligned Trajectory Matching. In *The Twelfth International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Jalali, M.; Li, C. T.; and Farnia, F. 2023a. An Information-Theoretic Evaluation of Generative Models in Learning Multi-modal Distributions. In *Advances in Neural Information Processing Systems*, volume 36, 9931–9943. Curran Associates, Inc.

Jalali, M.; Li, C. T.; and Farnia, F. 2023b. An Information-Theoretic Evaluation of Generative Models in Learning Multi-modal Distributions. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 9931–9943. Curran Associates, Inc.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7176–7185. PMLR.

Nguyen, T.; Chen, Z.; and Lee, J. 2021. Dataset Meta-Learning from Kernel Ridge-Regression. In *International Conference on Learning Representations*.

Nguyen, T.; Novak, R.; Xiao, L.; and Lee, J. 2021. Dataset Distillation with Infinitely Wide Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 34, 5186–5198. Curran Associates, Inc.

Pál, D.; Póczos, B.; and Szepesvári, C. 2010. Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs. In Lafferty, J.; Williams, C.; Shawe-Taylor, J.; Zemel, R.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Pasarkar, A. P.; and Dieng, A. B. 2024. Cousins Of The Vendi Score: A Family Of Similarity-Based Diversity Metrics For Science And Machine Learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, 3808–3816. PMLR.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4195–4205.

Qin, T.; Deng, Z.; and Alvarez-Melis, D. 2024. A Label is Worth A Thousand Images in Dataset Distillation. In *Advances in Neural Information Processing Systems*, volume 37, 131946–131971.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent

Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Santiago, J. A. C.; praveen tirupattur; Nayak, G. K.; Liu, G.; and Shah, M. 2025. MGD$^3$ : Mode-Guided Dataset Distillation using Diffusion Models. In *Forty-second International Conference on Machine Learning*.

Shen, Z.; and Xing, E. 2022. A Fast Knowledge Distillation Framework for Visual Recognition. In *Computer Vision – ECCV 2022*, 673–690. Springer Nature Switzerland.

Su, D.; Hou, J.; Gao, W.; Tian, Y.; and Tang, B. 2024. Dˆ4: Dataset Distillation via Disentangled Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5809–5818.

Sun, P.; Shi, B.; Yu, D.; and Lin, T. 2024. On the Diversity and Realism of Distilled Dataset: An Efficient Dataset Distillation Paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9390–9399.

Wang, H.; Zhao, Z.; Wu, J.; Shang, Y.; Liu, G.; and Yan, Y. 2025. CaO2: Rectifying Inconsistencies in Diffusion-Based Dataset Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, K.; Zhao, B.; Peng, X.; Zhu, Z.; Yang, S.; Wang, S.; Huang, G.; Bilen, H.; Wang, X.; and You, Y. 2022. CAFE: Learning To Condense Dataset by Aligning Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12196–12205.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959*.

Yin, Z.; Xing, E.; and Shen, Z. 2023. Squeeze, Recover and Relabel: Dataset Condensation at ImageNet Scale From A New Perspective. In *Advances in Neural Information Processing Systems*, volume 36, 73582–73603. Curran Associates, Inc.

Zhao, B.; and Bilen, H. 2023. Dataset Condensation With Distribution Matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6514–6523.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2021. Dataset Condensation with Gradient Matching. In *International Conference on Learning Representations*.

Zou, Y.; Li, G.; Su, D.; Wang, Z.; Yu, J.; and Zhang, C. 2025. Dataset Distillation via Vision-Language Category Prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.