CVPR
#4003

From Reconstruction to Compression: Reinforcement Learning in Diffusion-Based Dataset Distillation

Anonymous CVPR submission

Paper ID 4003

Abstract

Dataset distillation synthesizes compact datasets that retain the training utility of much larger ones. While diffusion models are natural candidates for this task due to their generative capabilities, there are few methods that adopt them in dataset distillation compared to the matching-based approaches and label-relaxation approaches. A key reason is the fundamental mismatch between diffusion objectives and distillation goals: diffusion models are trained to reconstruct high-fidelity data, whereas distillation requires compressed, task-relevant representations. We address this gap by proposing a reinforcement learning (RL)-guided framework that steers diffusion models from reconstruction toward compression. By formulating sampling as a decision process, we optimize the generative trajectory using rewards derived from student model performance. This enables the generation of synthetic samples that maximize learning utility under strict compression budgets. Unlike prior static modifications of the diffusion process, our method dynamically adapts generation based on downstream outcomes. Experiments on standard benchmarks show that our RL-guided diffusion approach consistently improves both performance and efficiency, advancing the frontier of generative dataset distillation.

1. Introduction

Dataset distillation [3, 4, 6, 11, 16, 32, 36] emerges as a scalable alternative to coreset selection, with a critical shift in paradigm: instead of selecting a subset from the original dataset, it aims to synthesize a small number of synthetic samples that can train models to comparable performance. This synthesis-oriented nature makes generative models, particularly diffusion models [4, 21], natural candidates for distillation backbones. Given their ability to model complex data distributions and generate diverse samples, diffusion models appear well-suited to construct informative, compact datasets. As visualized in Figure 1, this shift in objective—from faithfully reconstructing individual samples to selectively synthesizing task-informative ones—calls for a fundamental rethinking of how diffusion models are em-

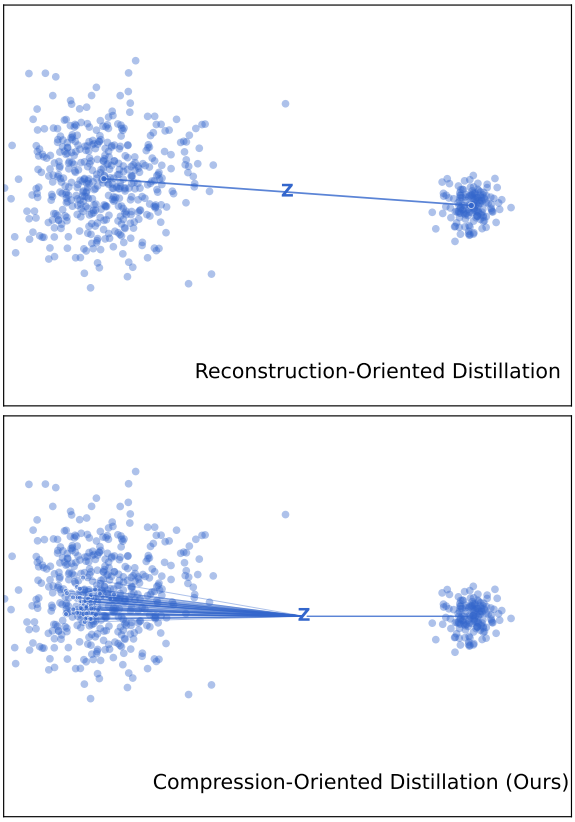

Figure 1. **Illustration of reconstruction-oriented vs. compression-oriented distillation.** Top: Existing diffusion-based distillation reconstructs high-fidelity samples by directly mapping a latent code z to individual data points, optimizing for pixel-level realism. Bottom: Our proposed Compression-Oriented Distillation introduces reward-guided sampling, enabling z to dynamically steer generation toward task-informative and compact representations, thereby capturing dense supervision with fewer samples.

ployed in this context.

However, despite the popularity of dataset distillation [17, 24] and the great success of diffusion models [15, 21], their integration remains rare. We identify a key reason behind this gap: a fundamental mismatch in objectives. Diffusion models are trained to reconstruct individual data samples with high fidelity by reversing a noise corruption process. In contrast, dataset distillation is inherently a compression task, aiming to concentrate task-relevant information into a minimal number of synthetic instances. As a result, the stronger a diffusion model becomes at reconstructing original data, the less effective it is for generating compressed data optimized for downstream learning.

Existing dataset distillation methods primarily fall into two families: (i) matching-based approaches that directly optimize synthetic samples to approximate gradients or training trajectories [3, 7, 8, 18]; and (ii) label-relaxation approaches such as SRe2L [34], which guide learning through softened targets. While both have achieved considerable progress, they suffer from scalability and generalization bottlenecks—either due to reliance on differentiable supervision or overly rigid label semantics. These limitations further motivate a flexible, model-driven distillation framework, one that can generate rather than optimize, and adapt based on downstream training outcomes.

To this end, we introduce *Compression-Oriented Distillation* (COD), a novel framework that reformulates diffusion sampling as a reinforcement learning (RL) [2, 10, 12, 19, 37] problem aimed at utility-aware data compression. Instead of statically following the reverse denoising process, we learn a policy that dynamically controls the generative trajectory to favor samples that are compact yet highly effective for downstream training. By directly optimizing this policy with task-driven reward signals, our method moves beyond heuristic guidance and enables principled generation of high-utility synthetic data under strict budget constraints.

We instantiate this framework COD using Group Relative Policy Optimization (GRPO) [12], a lightweight yet stable policy optimization method that avoids explicit value estimation. To guide sample generation, we design a reward function that combines two complementary components: (1) an entropy-based signal (R_{Ent}) that promotes informative samples by maximizing predictive uncertainty [20], and (2) a diversity-aware penalty (R_{Div}) that discourages redundancy by comparing with a memory bank of previously generated outputs. This reward-driven feedback loop steers the diffusion model beyond pixel-level fidelity, enabling it to explore and exploit regions of the data space that are optimized for learning efficiency [2]. Compared to prior methods like Minimax Diffusion that statically reshape sampling behavior, our approach offers dynamic, goal-aware control over generative processes.

In summary, this paper presents the first comprehensive study of reinforcement learning for controlling diffusion-based dataset distillation. By reinterpreting generative modeling as a compression-driven decision process, we bridge the gap between reconstruction-centric generation and training-centric distillation, setting the stage for a new class of adaptive, goal-aware synthetic data pipelines. Our contributions are summarized as follows:

- We identify a fundamental mismatch between diffusion models and dataset distillation: diffusion prioritizes reconstruction, while distillation demands compression. This insight explains the limited integration of the two paradigms.
- We propose *Compression-Oriented Distillation* (COD), a novel dataset distillation framework that formulates diffusion sampling as a reinforcement learning process guided by downstream utility.
- We instantiate COD using Group Relative Policy Optimization (GRPO) with a reward function combining entropy-based informativeness and diversity-aware regularization, enabling principled and adaptive sample generation.

2. Preliminaries

2.1. Problem Formulation: Dataset Distillation

Given a large-scale dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ are data samples drawn i.i.d. from a natural distribution $\mathcal{D}$. We denote $y_i \in \mathcal{Y} = \{1, ..., C\}$ to represent class labels. Dataset distillation aims to construct a compact synthetic dataset $\mathbb{S} = \{(s_j, \tilde{y}_j)\}_{j=1}^{M}$ with $M \ll N$ such that a model trained solely on $\mathbb{S}$ performs comparably to one trained on $\mathcal{T}$ [3, 32]:

$$\mathbb{S}^* = \arg\min_{\mathbb{S}\subset\mathbb{R}^d\times\mathcal{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(f_{\theta_{\mathbb{S}}}(x), y)\right], \quad (1)$$

where $\theta_{\mathbb{S}}$ denotes the model parameters obtained by training on $\mathbb{S}$, and $\ell(\cdot)$ is a task-specific loss function (e.g., cross-entropy). Most existing dataset distillation methods fall into two broad categories: matching-based approaches and label-relaxation approaches.

Matching-based approaches optimize synthetic data by aligning gradients or training trajectories between real and synthetic datasets [3, 6, 7, 36]. A common objective is gradient matching [36]:

$$\min_{\mathbb{S}} \sum_{(s_j,\tilde{y}_j)\in\mathbb{S}} \|\nabla_\theta \ell(f_\theta(s_j), \tilde{y}_j) - \nabla_\theta \ell(f_\theta(x_i), y_i)\|^2, \quad (2)$$

where the gradient computed on synthetic samples is forced to approximate that from real data. Trajectory-based variants extend this idea across multiple steps of optimization. While effective, such methods often require differentiability, second-order gradients, and suffer from limited scalability on larger datasets.

CVPR #4003

Label-relaxation approaches assign soft labels to synthetic samples to improve generalization [27, 29, 34]. Instead of using hard one-hot labels y_j, each synthetic sample is paired with a learnable probability vector $\tilde{y}_j \in \Delta^{C-1}$:

$$\min_{\mathbb{S}} \sum_{(s_j, \tilde{y}_j) \in \mathbb{S}} \ell(f_\theta(s_j), \tilde{y}_j), \tag{3}$$

where $\tilde{y}_j$ encodes label uncertainty or class similarity. While this approach improves performances significantly, it requires a pretained model to serve as the teacher model to generate soft labels. Xiao and He [33] reveals that removing the soft labels will cause dramatic performance drop in Label-relaxation approaches.

Both paradigms rely on direct supervision over synthetic instances. In contrast, our framework shifts the problem toward a reward-driven generative formulation, using reinforcement learning to synthesize utility-optimized training data. Unlike coreset selection [1, 5, 13, 26, 31], dataset distillation synthesizes new data instances rather than selecting from $\mathcal{T}$. This makes generative models, in particular the diffusion models, a promising approach for dataset distillation.

2.2. Diffusion Models for Generative Synthesis

Diffusion models [15, 21, 30] generate data via a two-stage process: a forward noising process and a reverse denoising process. Let $x_0 \sim \mathcal{D}$ denote a real data sample; f_θ denote the denoising network f parameterized with θ. The forward process gradually corrupts x_0 with Gaussian noise, yielding a sequence $\{x_t\}_{t=0}^{T}$. The reverse process then aims to iteratively reconstruct x_0 from pure noise $x_T \sim \mathcal{N}(0, I)$, by learning a parameterized denoising network f_θ.

Given a discretized time schedule $\{t_i\}_{i=0}^{N}$, the sampling trajectory starts from $x_0 \sim \mathcal{N}(0, b(t_{\max})^2 I)$ and proceeds via the following iterative update:

$$\mathbf{x}_{i+1} := \kappa_i \mathbf{x}_i + \eta_i f_\theta(\mathbf{x}_i \mid t_i) + \zeta_i \tilde{\epsilon}_i, \tag{4}$$

where $\tilde{\epsilon}_i \sim \mathcal{N}(0, I)$ is an optional sampling noise term (present only in SDE-based solvers), and κ_i, η_i, and ζ_i are time-dependent coefficients derived from the training-time noise schedule.

This reverse process is fundamentally designed to reconstruct a high-fidelity individual instance from Gaussian noise. The denoising network f_θ is explicitly trained to reverse the corruption applied in the forward process, which encourages the generation of samples that closely match the data distribution in pixel space or feature space. As a result, the learned generative trajectory is inherently biased toward reproducing realistic and data-faithful samples—making it highly suitable for reconstruction tasks, but potentially suboptimal for generating compressed or task-optimized representations such as those needed in dataset distillation.

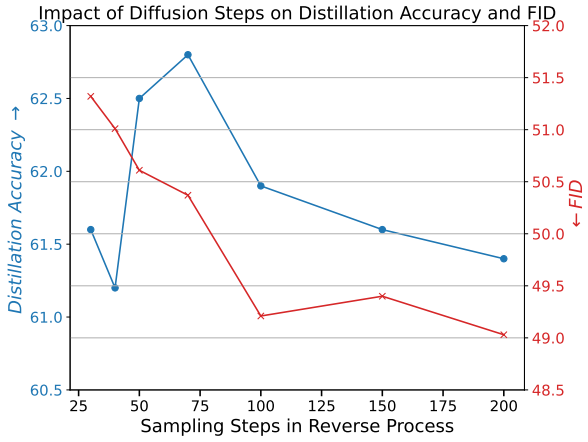

Figure 2. **Impact of Diffusion Steps on Distillation Accuracy and FID.** We vary the number of sampling steps in the reverse diffusion process and observe a trade-off between image fidelity and distillation performance. As the number of steps increases, the Fréchet Inception Distance (FID) consistently improves, indicating better reconstruction quality. However, the distillation accuracy declines, confirming that high-fidelity samples are not necessarily more informative for downstream learning. This highlights a fundamental mismatch between reconstruction and compression objectives in diffusion-based distillation.

3. Method

3.1. The Reverse Process Performs Denoising, Not Compression

The reverse diffusion process is inherently designed as a denoising mechanism [15], not a compression pipeline. At each time step, the denoising network f_θ estimates either the original clean sample x_0 or the noise ϵ added during the forward process, conditioned on a noisy input x_t. This iterative reconstruction from Gaussian noise is shown in Equation 4, where the objective is to minimize reconstruction error between predicted and true clean samples.

This training formulation encourages the model to reproduce high-fidelity instances that resemble the original dataset distribution. However, in dataset distillation, the objective shifts: instead of reproducing all modes of the data, we seek to selectively generate samples that are maximally informative for training under tight data budgets. As shown in Figure 2, while the Fréchet Inception Distance (FID) improves with more steps the classification performance of distilled datasets degrades. This confirms that better reconstruction does not equate to better compression, and high-fidelity images are not necessarily information-dense for learning.

To address this mismatch, we aim to systematically bias the reverse process away from denoising and toward compression. The most direct solution appears to be modifying

the reverse process itself, which replaces the reconstruction-oriented dynamics with a utility-driven sampling process that prioritizes the generation of high-information-content samples.

However, this direction faces several fundamental limitations. First, shifting the diffusion objective toward compression would require defining or learning a new target distribution $\mathcal{S}$ that represents the ideal distilled dataset distribution. We denote by $\mathcal{S}$ the empirical distribution formed by all synthetic datasets produced by existing dataset distillation methods. Each method $\mathcal{A}_i$ generates a synthetic dataset $\mathcal{S}_i = \mathcal{A}_i(\mathcal{T})$, where $\mathcal{T}$ is the original training set. The distribution $\mathcal{S}$ can thus be conceptualized as:

$$\mathcal{S} = \mathbb{P}_{\mathcal{A}\sim\mathcal{M}}\left[\mathcal{A}(\mathcal{T})\right], \tag{5}$$

where $\mathcal{M}$ is the space of known distillation algorithms (e.g., DM, MTT, SRe2L, IGD).

Attempting to train a diffusion model to approximate $\mathcal{S}$ poses three major challenges. First, $\mathcal{S}$ is implicitly defined and lacks a closed-form representation—sampling from it requires exhaustively running and storing outputs from many distillation pipelines. Second, synthetic datasets from different methods are structurally inconsistent, often varying in label granularity, resolution, or supervision format, making them hard to unify under a coherent distribution. Third, even if $\mathcal{S}$ were learnable, any generative model trained to replicate it would be fundamentally limited by the diversity and quality of the existing methods. That is, the best it can do is imitate prior solutions, but never surpass them.

This motivates us to abandon the idea of statically fitting a proxy to $\mathcal{S}$ and instead adopt a reward-driven sampling mechanism that actively explores beyond it.

3.2. Reinforcement Learning for Compression-Oriented Diffusion Sampling

To explore beyond the empirical limits of $\mathcal{S}$, we propose to steer the generative trajectory using reinforcement learning (RL). Rather than statically mimicking prior synthetic datasets, our goal is to actively discover high-utility samples by assigning rewards to diffusion outcomes based on their downstream training performance. This framing naturally casts the sampling procedure as a sequential decision process, where the reverse steps of the diffusion model form a Markov chain governed by a policy π_ϕ.

The idea of prioritizing high-information-content samples is inspired by the success of dataset pruning, which shows that even within natural datasets, only a subset of examples contributes meaningfully to generalization. This observation implies that data samples are inherently unequal in the perspective of information.

Concretely, at each reverse timestep t, the policy $\pi_\phi(a_t \mid x_t)$ selects an action a_t—such as modifying the noise prediction or controlling the denoising step size—based on the current sample state x_t. The final output x_0 is evaluated via a reward function $R(x_0)$ that reflects its utility for distillation. This reward can be instantiated using downstream classification accuracy, teacher-student agreement, or information-theoretic proxies such as entropy or mutual information. The policy is then optimized to maximize expected reward:

$$\max_{\phi}, \mathbb{E}_{x_0\sim\pi_\phi}\left[R(x_0)\right]. \tag{6}$$

This formulation transforms the role of the diffusion model from a passive denoiser into an active sampler that learns to navigate toward information-rich regions of the data space. Unlike traditional guidance strategies, which rely on heuristics or task-agnostic priors (e.g., classifier gradients or class embeddings), our RL-based controller can be trained end-to-end to align sample generation directly with dataset distillation objectives.

In doing so, we depart from the conventional reconstruction pipeline and reframe dataset distillation as a compression-driven search problem over the generative trajectory space. This dynamic mechanism allows us not only to circumvent the ill-posed nature of $\mathcal{S}$ but also to transcend the limitations of existing synthetic datasets by continuously refining the sampling policy in response to feedback from training performance.

3.3. Instantiating RL-Based Sampling with GRPO and Entropy Rewards

To realize the RL formulation described above, we adopt Group Relative Policy Optimization (GRPO) as our policy learning algorithm. While standard reinforcement learning methods such as PPO [25] offer stable policy improvement, they rely heavily on value function estimation and surrogate clipping objectives, which are costly and unstable in our context. Diffusion-based sampling is inherently high-dimensional and slow, and accurate value estimation across diverse generative trajectories is impractical.

GRPO circumvents these issues by discarding value estimation altogether. Instead of modeling long-term returns, GRPO computes *relative advantages* within a group of sampled actions. Specifically, for each reverse step state x_t, we sample a group of candidate actions $\{a_t^{(i)}\}_{i=1}^{G}$, generate corresponding samples $\{x_0^{(i)}\}$, and compute their rewards $\{r_i = R(x_0^{(i)})\}$. We then normalize the rewards using z-score normalization to obtain relative advantages:

$$\bar{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}, \tag{7}$$

where $\mathbf{r} = \{r_1, \ldots, r_G\}$ is the reward vector within the group. The policy π_ϕ is updated to increase the probability

CVPR #4003

of actions with higher $\bar{r}_i$, encouraging exploration of trajectories that outperform their peers without needing explicit value estimation.

$$R_{\text{Ent}}(x_0) = \mathcal{H}(f_\theta(x_0)) = -\sum_{c=1}^{C} p_c \log p_c, \tag{8}$$

where p_c denotes the softmax probability assigned to class c by the model pretrained on original dataset. High entropy indicates model uncertainty and implies that the sample x_0 lies near the decision boundary—thus being more informative for training. Unlike label-matching losses, entropy rewards are task-agnostic, differentiable, and directly aligned with the goal of generating useful training signals.

While entropy-based rewards encourage the generation of uncertain and potentially informative samples, optimizing solely for entropy may lead to *mode collapse*—the repeated synthesis of ambiguous yet similar examples. To mitigate this, we introduce a diversity reward that explicitly penalizes redundancy among generated samples.

We maintain a memory bank $\mathcal{B}$ that stores the embeddings or output logits of previously generated synthetic samples. The details to implement the memory bank is clearly stated in the experiment section. For each new candidate x_0, we compute its similarity to the most similar entry in the bank and apply a penalty accordingly. The final reward function becomes:

$$R(x_0) = R_{\text{Ent}}(x_0) + R_{\text{Div}}(x_0), \quad \text{where} \tag{9}$$

$$R_{\text{Div}}(x_0) = -\lambda \cdot \max_{x' \in \mathcal{B}} \text{sim}(x_0, x'). \tag{10}$$

where $\text{sim}(x_0, x')$ measures the similarity between x_0 and a stored sample x' using cosine similarity. $R_{\text{Ent}}(x_0)$ is defined in Equation 8.The hyperparameter $\lambda > 0$ controls the trade-off between uncertainty and novelty.

This diversity-aware reward encourages the sampling policy to explore broader, less redundant regions of the generative space—promoting sample diversity without sacrificing informativeness. Empirically, we find that combining entropy and diversity signals leads to synthetic datasets that are both challenging and complementary, resulting in stronger downstream performance.

Altogether, our instantiation combines (1) a policy optimization algorithm (GRPO), (2) an information-theoretic reward signal (entropy), and (3) a diversity-aware constraint (memory bank filtering). These design choices strike a balance between sample informativeness and diversity—two pillars of effective dataset distillation.

Algorithm 1 Compression-Oriented Diffusion for Dataset Distillation

Require: Pretrained diffusion model f_θ, pretrained evaluation network h_{pt}, policy π_φ, memory bank $\mathcal{B}$, group size G, reward weight λ

for each RL iteration **do**
  Sample G initial noise vectors $\{x_T^{(i)}\}_{i=1}^{G} \sim \mathcal{N}(0, I)$
  for each $x_T^{(i)}$ **do**
    Sample actions $\{a_t^{(i)}\}$ from $\pi_\varphi(a_t \mid x_t^{(i)})$ at each timestep
    Generate sample $x_0^{(i)}$ via controlled reverse diffusion trajectory
    Compute entropy reward:

$$R_{\text{Ent}}(x_0^{(i)}) = -\sum_{c=1}^{C} p_c \log p_c \quad \text{where } \mathbf{p} = h_{\text{pt}}(x_0^{(i)})$$

    Compute diversity penalty:

$$R_{\text{Div}}(x_0^{(i)}) = -\lambda \cdot \max_{x' \in \mathcal{B}} \text{sim}(x_0^{(i)}, x')$$

    Total reward: $R^{(i)} = R_{\text{Ent}}(x_0^{(i)}) + R_{\text{Div}}(x_0^{(i)})$
  end for
  Normalize rewards: $\bar{R}^{(i)} = \frac{R^{(i)} - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}$
  Update policy π_φ using GRPO with $\bar{R}^{(i)}$
  Update memory bank: $\mathcal{B} \leftarrow \mathcal{B} \cup \{x_0^{(i)}\}_{i=1}^{G}$
end for

4. Experiments

4.1. Experimental Setup

Datasets and baselines. We evaluate our method on three benchmark datasets with increasing resolution and complexity: ImageNet-1K (224×224) and two wellknown subsets of ImageNet [23]: ImageNette, ImageWoof. For large-scale evaluation, we follow common practice and report top-1 classification accuracy under varying image-per-class (IPC) settings (e.g., 10, 50, 100). We compare with representative baselines including pixel-level methods (DM [35], IDC-1 [16]), generative methods (DiT [21]), and fine-tuned diffusion (Minimax [11]). Random and Full serve as lower and upper bounds respectively. We also compare with label-relaxation methods including Sre2L [34] and G-VBSM [27].

Evaluation protocol. Following prior work, we train standard ConvNet or ResNet architectures on the synthetic datasets for 50 to 200 epochs, depending on resolution, using SGD or Adam optimizers. We adopt consistent training schedules across baselines for fair comparison. Unless otherwise stated, evaluation is performed on the same test sets as the original datasets. For ImageNet-1K, pretrained classifiers are also used for reward calculation but not for final evaluation. While SRe2L [34] adopts an evaluation protocol using soft labels to have better performance, we only

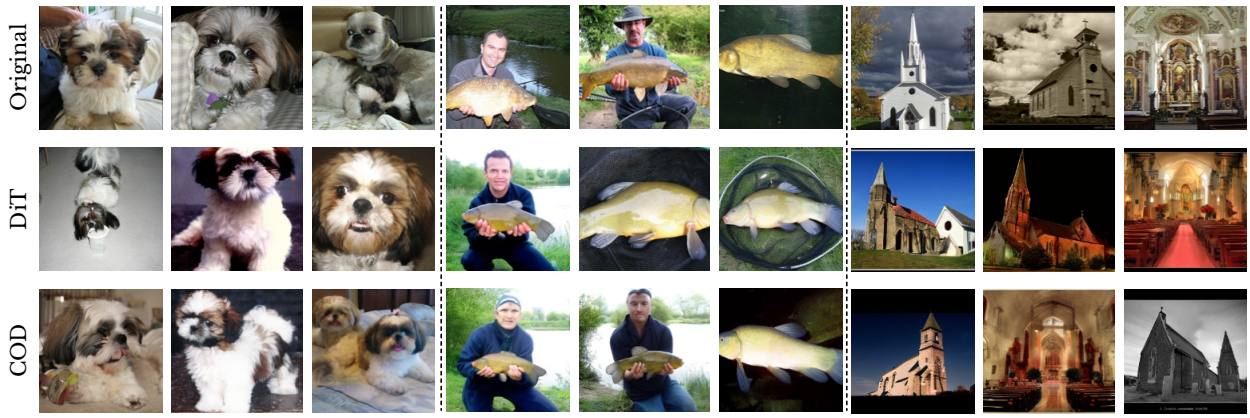

Figure 3. Visualization of random original images, images generated by baseline diffusion model (DiT) and our proposed method (COD). For each column, the generated images are based on the same random seed. Compared to DiT, COD intentionally departs from pixel-level faithfulness and produces samples that are less visually similar to the originals. This shift is consistent with our core view that high-fidelity reconstruction is misaligned with the objective of dataset distillation.

adopt this protocol for ImageNet-1K experiments; all subset results are reported under standard hard-label evaluation for comparability.

Diffusion backbone. We adopt latent DiT [21] as our diffusion backbone, using a pretrained VAE encoder-decoder to map between image and latent space. All experiments use DDIM [28] sampling with 50 steps. The policy network π_φ operates over the noise prediction module of the reverse process and is trained using GRPO. For reward computation, we use a pretrained ImageNet-1K classifier f_{pt} to evaluate entropy.

Memory Bank Implementation. To support the diversity-aware reward R_{div}, we maintain a dynamic memory bank that stores previously generated synthetic samples. At the beginning of policy training, the memory bank is cold-started by populating it with a fixed number of synthetic samples generated unconditionally from the pretrained diffusion model. The total number of stored samples is set equal to the target dataset size to avoid memory growth.

During training, each newly generated sample is compared against existing entries in the memory bank. If a sample is found to be highly similar to any stored instance (based on cosine similarity in a pretrained feature space), it is discarded from reward calculation and excluded from memory bank updates. Otherwise, the sample is appended to the memory bank, and the most similar existing item is removed to maintain a fixed memory size. This design ensures continual refresh of diverse representations without allowing the memory bank to grow, enabling efficient and scalable diversity estimation.

Figure 4. **Reward components over training steps.** The total reward (blue) increases steadily throughout training, driven by the entropy-based component R_{ent} (red) and the diversity-based penalty R_{div} (yellow). Notably, R_{ent} saturates early, while R_{div} continues to rise, indicating a shift in policy focus from informativeness to sample diversity as training progresses.

Training details. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Each GRPO update uses group size $G = 4$, and the memory bank retains up to 512 embeddings per class. We adopt cosine annealing for the policy learning rate and freeze the pretrained diffusion and classifier networks throughout the process. Further details (e.g., entropy temperature, policy depth) are detailed in Appendix A.

4.2. Experimental Results

Main Results on ImageNet. We report the top-1 classification accuracy on Nette and Woof subsets under varying architectures (ConvNet-6, ResNetAP-10, ResNet-18) [14] and image-per-class (IPC) budgets (10, 50, 100). The re-

CVPR #4003

Table 1. **Comparison of distillation performance across multiple methods, architectures, and datasets.** We report top-1 classification accuracy (%) on Nette and Woof subsets of the ImageNet-1K dataset under varying architectures (ConvNet-6, ResNetAP-10, ResNet-18) and image-per-class (IPC) budgets (10, 50, 100). COD (Ours) consistently achieves competitive or superior performance across settings, particularly under low IPC (e.g., 10), demonstrating its advantage in generating informative and compressed synthetic datasets. *Full* denotes training on the complete original dataset and serves as an upper bound.

Subset	**Nette**									**Woof**								
Architecture	ConvNet-6			ResNetAP-10			ResNet-18			ConvNet-6			ResNetAP-10			ResNet-18		
IPC	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100
Random	46.0 ±0.5	71.8 ±1.2	79.9 ±0.8	54.2 ±1.2	77.3 ±1.0	81.1 ±0.6	55.8 ±1.0	75.8 ±1.1	82.0 ±0.4	24.3 ±1.1	41.3 ±0.6	52.2 ±0.4	29.4 ±0.8	47.2 ±1.3	59.2 ±0.9	27.7 ±0.9	47.9 ±1.8	61.5 ±1.3
DM	49.8 ±1.1	70.3 ±0.8	78.5 ±0.8	60.2 ±0.7	76.7 ±1.0	80.9 ±0.7	60.9 ±0.7	75.0 ±1.0	81.5 ±0.4	26.9 ±1.2	43.8 ±1.1	50.1 ±0.9	29.8 ±1.0	47.1 ±1.1	56.4 ±0.8	30.2 ±0.6	46.2 ±0.6	60.2 ±1.0
IDC-1	48.2 ±1.2	72.4 ±0.7	80.6 ±1.1	60.4 ±0.6	77.4 ±0.7	81.5 ±1.2	61.0 ±0.8	77.8 ±0.7	81.7 ±0.8	33.3 ±1.1	42.6 ±0.9	51.0 ±1.1	38.5 ±0.7	48.3 ±1.0	56.1 ±0.9	36.7 ±0.8	48.3 ±0.8	57.7 ±0.8
DiT	56.2 ±1.3	73.3 ±0.9	78.2 ±0.3	62.8 ±0.8	76.9 ±0.5	80.1 ±1.1	62.5 ±0.9	75.2 ±0.9	77.8 ±0.6	32.3 ±0.8	46.5 ±0.8	53.4 ±0.3	34.7 ±0.5	49.3 ±0.2	58.3 ±0.8	34.7 ±0.4	50.1 ±0.5	58.9 ±1.3
Minimax	58.2 ±0.9	76.6 ±0.2	**81.1 ±0.3**	63.2 ±1.0	78.2 ±0.7	**81.3 ±0.9**	**64.9 ±0.6**	78.1 ±0.6	81.3 ±0.7	33.5 ±1.4	50.7 ±1.8	**57.1 ±1.9**	39.2 ±1.3	56.3 ±1.0	64.5 ±0.2	37.6 ±0.9	57.1 ±0.6	**65.7 ±0.4**
COD (Ours)	**59.2 ±0.9**	**74.8 ±0.5**	78.8 ±1.6	**64.3 ±0.3**	**78.5 ±0.8**	81.0 ±1.0	63.8 ±1.2	**78.7 ±0.7**	**81.6 ±1.3**	**36.0 ±0.9**	**51.3 ±1.0**	55.2 ±1.5	**41.6 ±1.3**	**58.2 ±0.8**	**65.6 ±0.6**	**44.0 ±1.8**	**59.2 ±1.0**	65.4 ±0.9
Full	94.3 ±0.5	94.3 ±0.5	94.3 ±0.5	94.6 ±0.5	94.6 ±0.5	94.6 ±0.5	95.3 ±0.6	95.3 ±0.6	95.3 ±0.6	85.9 ±0.4	85.9 ±0.4	85.9 ±0.4	87.2 ±0.6	87.2 ±0.6	87.2 ±0.6	89.0 ±0.6	89.0 ±0.6	89.0 ±0.6

sults are demonstrated in Table 1. COD achieves consistent improvements over prior diffusion-based methods (DM [35], DiT [21]) and optimization-based methods (IDC-1 [16]), especially under low-data regimes such as IPC=10. Compared to the strongest baseline, Minimax, our method exhibits competitive performance across nearly all settings. However, the performance gap between COD and Minimax remains small. This is expected, as both approaches share a similar underlying philosophy: Minimax explicitly modifies the denoising network during training to favor discriminative gradients, while COD fine-tunes the sampling trajectory via reinforcement learning. Despite differing in implementation (training vs. inference), both methods achieve comparable expressivity in guiding generation away from pixel-level fidelity and toward task-relevant content.

We report top-1 accuracy on ImageNet-1K with IPC = 10 and 50 in Table 2. COD achieves the highest accuracy at IPC = 50, surpassing both optimization-based (SRe^{2}L [34], G-VBSM [27], RDED [29]) and generative (DiT [21], Minimax [11]) methods. The consistent improvement demonstrates the effectiveness of our reward-driven policy in scaling to large-scale distillation.

Trade-off Between Accuracy and Fidelity. We investigate how the number of reverse diffusion steps affects the trade-off between sample fidelity and distillation performance. As shown in Figure 2, increasing steps leads to lower Fréchet Inception Distance (FID), indicating improved visual quality. However, distillation accuracy peaks at 75 steps and declines thereafter. This confirms a key insight: higher-fidelity samples are not necessarily more informative for training, and optimizing for visual realism can hurt task-specific compression.

Reward Dynamics Analysis. To understand how our reward function evolves during training, we track the total reward and its two components (R_{ent}, R_{div}) across policy updates. As shown in Figure 4, the total reward increases consistently, indicating effective policy learning. The informativeness term R_{ent} rises rapidly in early stages and then saturates, reflecting that informative sample selection is quickly optimized. In contrast, the diversity term R_{div} grows more gradually, highlighting a shift in focus from informativeness to diversity as training progresses. This dynamic illustrates the complementary nature of the reward design, encouraging both discriminative and varied sample generation over time.

Additional Results in Supplementary. Due to space constraints, we include several extended experiments in the supplementary material. These include (1) cross-architecture evaluation on ImageNet-1K, which demonstrates the robustness of our method across different backbone networks; (2) an ablation study isolating the effects of the entropy-based reward (R_{ent}) and the diversity-based reward (R_{div}), showing that both components contribute positively to performance, though their combination yields diminishing returns due to partial redundancy; and (3) visual-

CVPR #4003

Table 2. Top-1 accuracy (%) on ImageNet-1K under different distillation methods with IPC = 10 and 50. COD achieves the highest accuracy when IPC=50, outperforming optimization-based (SRe^{2}L, G-VBSM, RDED) and generative (DiT, Minimax) baselines. This demonstrates the effectiveness of reward-driven sampling in scaling dataset distillation to challenging large-scale benchmarks.

Dataset	IPC	SRe^{2}L	G-VBSM	RDED	DiT	Minimax	COD(Ours)
ImageNet-1K	10	21.3 ±0.6	31.4 ±0.5	42.0 ±0.1	39.6 ±0.4	44.3 ±0.5	**45.0 ±0.3**
	50	46.8 ±0.2	51.8 ±0.4	56.5 ±0.1	52.9 ±0.6	58.6 ±0.3	**59.4** ±0.4

izations of generated samples that qualitatively reflect the trade-off between fidelity and informativeness. All code and implementation details are also provided in the supplement for reproducibility.

4.3. Discussion

This work takes a first step toward bridging generative modeling and dataset distillation by introducing a reward-driven formulation over the diffusion sampling process. While prior approaches often rely on handcrafted objectives or direct optimization of synthetic data, our method shows that reinforcement learning can provide a principled mechanism for exploring informative regions of the sample space.

Visualization. The qualitative comparison in Figure 3 further illustrates the core shift enabled by our framework—from reconstruction to compression. While the DiT backbone tends to replicate the dominant visual modes of the original dataset, COD deliberately deviates from pixel-level fidelity and instead synthesizes samples that emphasize class-discriminative structures, pose variation, and decision-boundary cues. Notably, COD images often appear less realistic or less similar to their original counterparts; however, this deviation is not a defect but a direct consequence of optimizing for learning utility rather than appearance. This trend mirrors our quantitative findings in Figure 2, where higher visual fidelity (lower FID) correlates with worse distillation accuracy.

Limitations. However, our framework also presents several limitations. First, incorporating reinforcement learning—though conceptually appealing—introduces training instability. Although GRPO offers a lightweight and gradient-regularized alternative to value-based methods, it still requires careful tuning of sampling frequency, reward scaling, and update schedules to achieve consistent convergence. Second, while our policy successfully shifts the generative behavior from reconstruction toward compression, it does so by modifying the sampling trajectory rather than the underlying diffusion model itself. The denoising backbone remains trained to match the natural data distribution, and therefore retains an inherent bias toward data fidelity. As a result, the full potential of compression-oriented generation is still constrained by the original training objective of the generative model.

These limitations point to promising future directions, such as integrating downstream utility signals into the training of the generative model itself, or developing more stable and expressive learning frameworks beyond policy optimization to further improve the quality and utility of distilled samples.

Future Work. More broadly, we view *Compression-Oriented Distillation* as a paradigm shift for the dataset distillation community. Rather than treating generative models as static decoders of the original dataset, we advocate for a dynamic, policy-guided generation process in which synthetic data is optimized for task-specific utility. Our framework—based on reinforcement learning and built upon a transformer-based diffusion backbone—demonstrates that modern generative architectures can be harnessed not just for realism, but for strategic, goal-aware data construction.

We believe this direction opens up a rich avenue for future research: leveraging increasingly powerful generative models, especially diffusion and transformer-based architectures, not merely as sample generators, but as active agents in data compression, selection, and synthesis. As foundation models continue to scale in capacity and generality, coupling them with task-aware decision-making mechanisms may fundamentally redefine how we construct and optimize training datasets across domains.

5. Conclusion

We introduced *Compression-Oriented Distillation* (COD), a reinforcement learning framework that guides diffusion models to generate informative and compact synthetic data for dataset distillation. By shifting the objective from reconstruction to compression, our method departs from static denoising and instead learns a dynamic sampling policy optimized for downstream utility. Through entropy-driven and diversity-driven rewards, our approach enables principled control over generative trajectories without modifying the diffusion training objective. This work bridges generative modeling and data distillation, paving the way for future research that further integrates task-aware objectives with advanced generative architectures.

References

[1] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017. 3

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 2, 11

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10708–10717, 2022. 1, 2, 11

[4] Mingyang Chen, Jiawei Du, Bo Huang, Yi Wang, Xiaobo Zhang, and Wei Wang. Influence-guided diffusion for dataset distillation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 11

[5] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012. 3

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 6565–6590, 2023. 1, 2

[7] Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3749–3758, 2023. 2

[8] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. 2

[9] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *Advances in Neural Information Processing Systems*, 2023. 11

[10] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 2, 11

[11] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2024. 1, 5, 7, 11, 12

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 12, 13

[13] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 126–134, 2005. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[16] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pages 11102–11118. PMLR, 2022. 1, 5, 7

[17] Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):17–32, 2024. 2

[18] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. 2

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[20] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 20596–20607, 2021. 2

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 2, 3, 5, 6, 7

[22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. 12, 13

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 5

[24] Noveen Sachdeva and Julian J. McAuley. Data distillation: A survey. *Trans. Mach. Learn. Res.*, 2023. 2

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4, 12

[26] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 3

[27] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16709–16718, 2024. 3, 5, 7

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 6

[29] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. 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, pages 9390–9399, 2024. 3, 7, 11

[30] Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? *arXiv preprint arXiv:2502.13129*, 2025. 3

[31] Ivor W Tsang, James T Kwok, Pak-Ming Cheung, and Nello Cristianini. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(4), 2005. 3

[32] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 2, 11

[33] Lingao Xiao and Yang He. Are large-scale soft labels necessary for large-scale dataset distillation? *arXiv preprint arXiv:2410.15919*, 2024. 3

[34] Zeyuan Yin, Eric P. Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from A new perspective. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. 2, 3, 5, 7

[35] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 6503–6512, 2023. 5, 7

[36] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. 1, 2

[37] Muxin Zhou, Zeyuan Yin, Shitong Shao, and Zhiqiang Shen. Self-supervised dataset distillation: A good compression is all you need. *arXiv preprint arXiv:2404.07976*, 2024. 2

CVPR #4003

A. More Related Work

Trajectory Matching Approaches Dataset distillation was first cast as a bi-level optimisation that matches gradients between real and synthetic data to preserve learning signals while using orders-of-magnitude fewer images [32]. Subsequent trajectory matching methods align full optimisation paths rather than single-step gradients, capturing richer learning dynamics and markedly boosting cross-architecture transferability [3]. However, early variants were memory-intensive and limited to small datasets. Recent work addresses these bottlenecks compress back-prop storage to enable ImageNet-1K distillation with constant memory while Du et al. introduce sequential matching that partitions long trajectories into manageable blocks [9]. Together, these advances push trajectory matching to large-scale settings, narrowing the test-accuracy gap to within a few points of full-data training.

Diffusion-Based Distillation Diffusion models provide a powerful generative prior for synthesising realistic yet compact datasets. Minimax Diffusion fine-tunes a DiT backbone adversarially, balancing sample fidelity and discriminative utility to outperform pixel-level baselines on ImageNet subsets [11]. Influence-Guided Diffusion (IGD) further removes heavy retraining by steering the sampling trajectory with mutual-information rewards, producing diverse, class-informative images at scale [4]. Complementary to these tuning-heavy approaches, RDED composes high-quality image patches without gradient updates, achieving strong performance under extreme budgets of 10 images per class [29]. Collectively, these studies demonstrate that diffusion priors can retain visual fidelity, diversity and class coverage even when the synthetic dataset is compressed by two orders of magnitude.

Reinforcement Learning for Diffusion Control Viewing denoising as a sequential decision process opens the door to policy-gradient fine-tuning. DDPO learns to adjust reverse-time steps with task-level rewards, improving alignment, aesthetics and even compressibility of generated images [2]. DPOK augments this framework with KL-regularised updates for greater stability and sample quality [10]. Building on these ideas, our Compression-Oriented Distillation uses utility-driven rewards—combining uncertainty and diversity—to bias diffusion trajectories toward samples that maximise downstream accuracy under strict image-per-class budgets, bridging reconstruction-centric generation and training-centric distillation.

B. Cross–Architecture Comparisons on ImageNet-1K

Table 3 reports top-1 accuracy of our Compression-Oriented Distillation (COD) against the recent RDED baseline on four unseen backbones of varying capacity—ResNet101, MobileNet-V2, EfficientNet-B0 and Swin Transformer—under two compression budgets.

Overall, COD delivers consistent gains at IPC50, exceeding RDED by an average of **3.7** percentage points across architectures. Improvements are particularly pronounced on ResNet101 (+4.9) and Swin Transformer (+4.0). At the stricter IPC10 setting, COD still outperforms RDED on two of four backbones and yields an average uplift of **4.3** points, driven largely by a sizeable margin on EfficientNet-B0 (+15.4). These results demonstrate that reward-driven diffusion sampling scales effectively across diverse network families while maintaining strong performance under aggressive data budgets.

C. Ablation study

To validate the effectiveness of our reward function design, we conduct comprehensive ablation studies on the ImageWoof and ImageNette datasets using ResNetAP-10 architecture. The experiments systematically evaluate the contribution of each reward component to the overall distillation performance.

C.1. Reward Component Analysis

The ablation study examines three configurations: (1) baseline without any reward components, (2) entropy reward only ($R_{\texttt{Ent}}$), and (3) diversity reward only ($R_{\texttt{Div}}$). The entropy reward promotes samples with high predictive uncertainty, while the diversity reward encourages exploration of different regions in the data space.

C.2. Performance Impact

Results demonstrate that both reward components contribute positively to distillation performance, with the diversity reward showing particularly strong improvements on the ImageWoof dataset. The entropy reward provides consistent gains across both datasets, indicating its effectiveness in generating informative samples. The combination of both components yields optimal performance, validating our multi-component reward design.

Table 5 presents the ablation study results, demonstrating that both reward components are essential for optimal performance. The diversity reward shows particularly strong improvements on ImageWoof (+5.5% at 10-IPC), while the entropy reward provides consistent gains across both datasets. These results validate our reward function design and highlight the importance of balancing informativeness and diversity in dataset distillation.

D. Hyperparameters Setup

Our Compression-Oriented Distillation (COD) framework employs a carefully tuned set of hyperparameters to balance

Table 3. Top-1 accuracy (%) on ImageNet-1K with 10 and 50 images per class (IPC10 / IPC50). Numbers are mean ± standard deviation over three runs.

Method	ResNet101		MobileNet-V2		EfficientNet-B0		Swin Transformer	
	IPC10	IPC50	IPC10	IPC50	IPC10	IPC50	IPC10	IPC50
RDED	48.3±1.0	61.2±0.4	40.4±0.1	53.3±0.2	31.0±0.1	58.5±0.4	42.3±0.6	53.2±0.8
COD	50.8±0.3	66.1±0.4	40.1±0.4	56.3±0.5	46.4±0.2	61.4±0.2	42.1±0.6	57.2±0.6

Table 4. Hyperparameters Setup for Compression-Oriented Distillation (COD). Key parameters include reward weights for entropy (W_{Ent}) and diversity (W_{Div}), training configuration (batch size, epochs, learning rate, α, β,), and GRPO clip (ϵ).

	W_{Ent}	W_{Div}	batch size	epochs	lr	α	β	ϵ
args	1.0	0.5	16	40	1e-4	0.4	0.6	0.2

Table 5. Ablation study of reward components on ImageWoof and ImageNette datasets using ResNetAP-10. R_{Ent} denotes the Entropy Reward and R_{Div} denotes the Diversity Reward. Results show mean ± standard deviation over multiple runs.

R_{Ent}	R_{Div}	ImageWoof		ImageNette	
		10-IPC	50-IPC	10-IPC	50-IPC
-	-	$34.9_{\pm 0.9}$	$50.8_{\pm 1.1}$	$62.8_{\pm 0.8}$	$76.9_{\pm 0.5}$
✓	-	$38.2_{\pm 1.1}$	$54.6_{\pm 0.7}$	$61.4_{\pm 0.7}$	$77.1_{\pm 0.9}$
-	✓	$40.4_{\pm 0.8}$	$56.7_{\pm 0.9}$	$62.3_{\pm 0.4}$	$77.3_{\pm 0.8}$

exploration and exploitation in the reinforcement learning process. The configuration is designed to maximize the informativeness and diversity of generated samples while maintaining stable training dynamics.

D.1. Reward Function Configuration

The reward function in our GRPO-based framework combines multiple components with empirically tuned weights. The entropy reward weight $W_{\text{Ent}} = 1.0$ encourages the generation of samples that maximize predictive uncertainty, thereby promoting informative samples that lie near decision boundaries. The diversity reward weight $W_{\text{Div}} = 0.5$ penalizes redundancy by comparing generated samples against a memory bank of previously synthesized outputs, ensuring sample diversity without sacrificing informativeness.

D.2. Training Parameters

The training process is configured with a batch size of 16 and runs for 40 epochs to ensure sufficient exploration of the generative space. We employ a learning rate of 1×10^{-4} with AdamW optimizer, which provides stable convergence for the policy optimization process. The epsilon parameter $\epsilon = 0.2$ controls the clipping range for the GRPO algorithm, ensuring policy updates remain within reasonable bounds.

E. Training Efficiency Analysis

The introduction of a reinforcement learning (RL) framework in our work inevitably incurs additional computational overhead. To quantify this, we compared the training efficiency quantitatively of our proposed Compression-Oriented Distillation (COD) to the state-of-the-art (SOTA) baseline, Minimax [11], in producing distilled datasets, using a single NVIDIA RTX 4090 GPU. We discovered that COD is significantly more efficient than Minimax. Specifically, COD converges in approximately 5.5 hours, whereas Minimax requires over 7 hours—a notable efficiency gain of around 30%. We attribute this significant advantage primarily to the lightweight nature of the Group Relative Policy Optimization (GRPO) [12] algorithm we employ. While the ultimate performance of COD is currently bottlenecked by the reward function used to evaluate synthetic sample compression, its substantial efficiency advantage, coupled with the strong potential of RL in controlling generative models, makes it a highly promising direction for dataset distillation research.

F. Training algorithm selection

Our choice of Group Relative Policy Optimization (GRPO) [12] over other widely-used policy optimization algorithms, such as Proximal Policy Optimization (PPO) [25] or Direct Preference Optimization (DPO) [22], was a deliberate decision based on the specific challenges inherent to the task of dataset distillation.

Algorithms like PPO [25] typically necessitate an additional critic, or value model, to estimate state values. This architecture presents two primary obstacles in the context of our task. First, introducing this additional critic model significantly increases the computational overhead (often by 3–5×), as it requires separate training and inference. Second, and more critically, this approach suffers from a strong reward function dependency. The accuracy of the

critic model is highly dependent on a precisely defined reward function—a significant challenge, as defining a reward that accurately evaluates the compression effectiveness of a synthetic sample is itself an inherently difficult and unsolved research problem in dataset distillation. While DPO [22] reformulates the objective to avoid a separate reward model, it is designed for preference data, which does not naturally align with our task of evaluating the utility of synthetic images.

GRPO [12] circumvents these issues by eschewing a critic model entirely. This critic-free nature allows us to directly use a relatively accurate reward metric—namely our designed R_{Ent} and R_{Div}—to optimize the policy, thereby making the application of RL to dataset distillation computationally feasible and efficient.