arXiv:2506.22637v2 [cs.CV] 9 Jul 2025

CaO_2: Rectifying Inconsistencies in Diffusion-Based Dataset Distillation

Haoxuan Wang[1] Zhenghao Zhao[1] Junyi Wu[1]
Yuzhang Shang[2] Gaowen Liu[3] Yan Yan[1†]

[1]University of Illinois Chicago [2]University of Central Florida [3]Cisco Research

Abstract

*The recent introduction of diffusion models in dataset distillation has shown promising potential in creating compact surrogate datasets for large, high-resolution target datasets, offering improved efficiency and performance over traditional bi-level/uni-level optimization methods. However, current diffusion-based dataset distillation approaches overlook the evaluation process and exhibit two critical inconsistencies in the distillation process: (1) Objective Inconsistency, where the distillation process diverges from the evaluation objective, and (2) Condition Inconsistency, leading to mismatches between generated images and their corresponding conditions. To resolve these issues, we introduce **C**ondition-**a**ware **O**ptimization with **O**bjective-guided Sampling (**CaO**$_2$), a two-stage diffusion-based framework that aligns the distillation process with the evaluation objective. The first stage employs a probability-informed sample selection pipeline, while the second stage refines the corresponding latent representations to improve conditional likelihood. CaO$_2$ achieves state-of-the-art performance on ImageNet and its subsets, surpassing the best-performing baselines by an average of 2.3% accuracy. Code is available at* https://github.com/hatchetProject/CaO2.

1. Introduction

The rapid expansion of data scale has significantly advanced the development of machine learning, but has also placed substantial demands on the storage capacity and computational resources. To accelerate training and reduce storage requirements while maintaining comparable performance, dataset distillation (DD) [2, 38, 42, 46, 47] was introduced to construct a compact surrogate dataset that captures the most critical information from a large target dataset. Conventionally, dataset distillation was formulated as a bi-level/uni-level optimization problem, designed to match the training dynamics between a teacher model trained on the target dataset and a student model trained on the synthetic dataset [10, 24, 36, 41]. However, these matching-based methods either struggle to scale up to larger, higher-resolution datasets, or suffer from low performance.

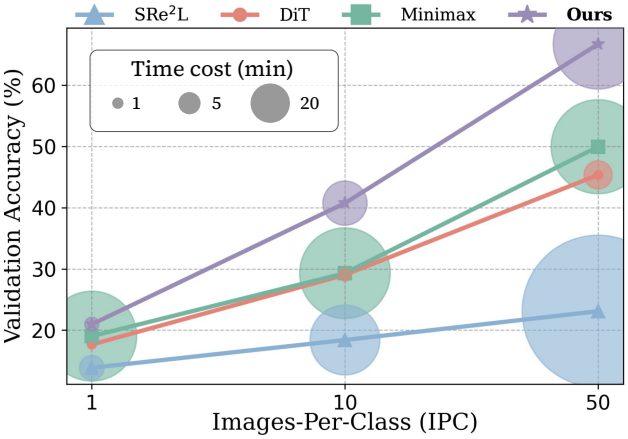


Figure 1. **Comparison of validation accuracy and distillation time for different methods under different IPCs on ImageWoof.** Our two-stage method is more efficient and can obtain better performance compared with other SOTA approaches.

Recently, diffusion-based dataset distillation [9, 33] emerged as a new paradigm for efficient data condensation. These methods utilize pre-trained diffusion models [28, 29] as strong distribution learners to filter noisy representations and retain the most important ones [3]. Therefore, their generated images are strong representations of the original target data distribution. As shown in Fig. 1, directly using randomly sampled images from the Diffusion Transformer (DiT) [28] for evaluation already outperforms the uni-level optimization method SRe^{2}L [41] by a significant margin, both in performance and distillation time. Moreover, unlike matching-based methods that rely on an additional student model and output unreadable images, diffusion-based methods are independent of the student model and produce realistic images. This enables the synthetic dataset to be seamlessly used for other tasks, such as neural architecture search [23] and continual learning [25].

These compelling advantages, nevertheless, come at the cost of achieving optimal performance. We observe that

†Corresponding author

current diffusion-based DD methods [9, 33] overlook the evaluation objective that assesses the discriminative properties of the distilled images. Instead, they focus exclusively on enhancing the fidelity of generated images, either through diffusion model training [9] or latent embedding clustering [33] (as shown in Fig. 2(a)). Such an oversight may lead to inappropriately designed distillation objectives, ultimately resulting in suboptimal evaluation accuracy.

However, incorporating the evaluation objective into the distillation process is not trivial, as diffusion models are not designed for classification tasks. Rather than directly integrating the evaluation objective, we aim to address two key inconsistencies related to it. The first is Objective Inconsistency, referring to the misalignment between the objective of the generative diffusion model and that of the evaluation model. As shown in Fig. 2(a), matching-based methods [2, 16, 44] constrain the distilled dataset to be optimized with the same objective of image classification as in the evaluation phase. In contrast, current diffusion-based methods generate the distilled dataset without any classification supervision. The second is Condition Inconsistency, stemming from the limitations of the applied conditional diffusion models themselves. In practice, diffusion models are not perfectly trained where the conditional likelihood for every generated sample is not maximized. Consequently, the obtained image-label pairs are suboptimal, which adversely impacts the training process during evaluation.

Having revealed the inconsistencies, we alleviate them by proposing a two-stage framework named **C**ondition-**a**ware **O**ptimization with **O**bjective-guided Sampling (**CaO$_2$**). As illustrated in Fig. 2(b), the first stage generates an image pool and sample from it class-wisely. By using a pre-trained lightweight classifier, we ensure that only samples confidently classified as belonging to their conditioned class are selected, thereby easing Objective Inconsistency. In the second stage, the chosen image latent is perturbed with random noise and optimized while keeping the diffusion model fixed under a task-dependent condition. The optimization objective aims to maximize the conditional likelihood of the image with respect to its condition, thereby reducing Condition Inconsistency. Our approach achieves state-of-the-art performance with better efficiency on ImageNet and its subsets. We also show that apart from pre-trained diffusion models, our method can be applied to fine-tuned models [9] and autoregressive models [19].

In summary, we make the following contributions:

- We observe that current diffusion-based dataset distillation methods overlook the evaluation process, leading to two inconsistencies in the distillation process — Objective Inconsistency and Condition Inconsistency — that hinder effective data condensation.
- To mitigate the inconsistencies, we propose CaO$_2$, a two-stage framework enabling efficient distillation with improved performance. Our framework can be applied to different model backbones, including Diffusion Transformers and autoregressive generation models such as MAR. It can also be used as a plug-and-play module for existing diffusion-based dataset distillation methods.
- Empirical results indicate that CaO$_2$ achieves state-of-the-art performance on ImageNet and its subsets, outperforming other methods by an average of 2.3% accuracy.

2. Related Works

2.1. Matching-based Dataset Distillation

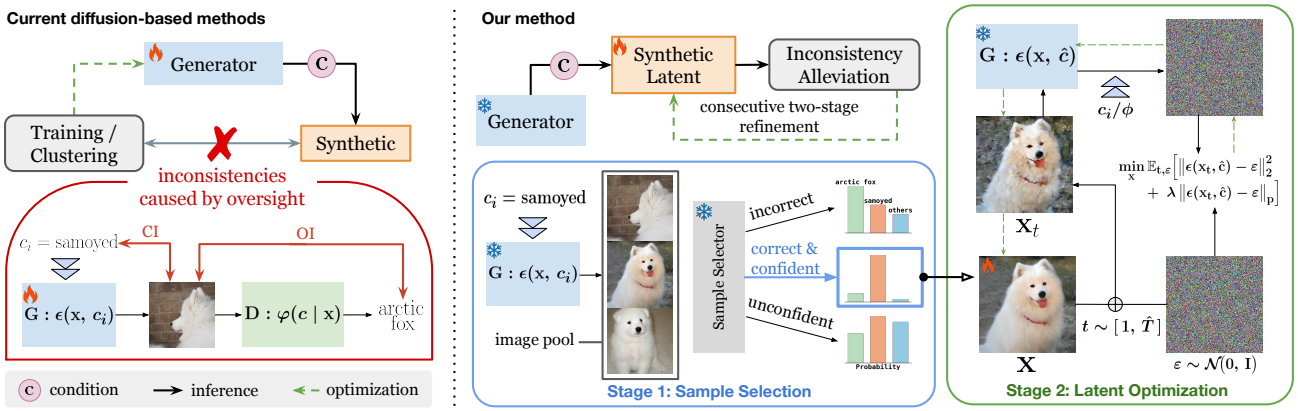
Initial works have formulated the dataset distillation problem as a bi-level optimization task, aiming to match the training characteristics of the synthetic dataset with those of the target dataset, such as gradients [16, 24, 44], feature distributions [36, 43, 45], and training trajectories [2, 6, 8, 10]. However, the bi-level distillation objective introduces large computational costs, making it impractical to generalize to large-scale datasets such as ImageNet. Even for a small dataset such as CIFAR-100, adopting the efficient matching-based method TESLA [6] to distill a subset with IPC=10 can take nearly 20 hours.

To scale efficiency, SRe^{2}L [41] and its subsequent works [31, 32, 40, 48] adopted a uni-level optimization framework. They first synthesize images by matching model outputs and Batch Norm statistics using a well-trained teacher model, then generate multiple soft labels for the distilled images. Though efficiency is improved, their performance is limited due to the inadequate matching objective and the separate processes for image and label synthesis.

2.2. Diffusion-based Dataset Distillation

An emerging line of research [5, 9, 33] utilizes pre-trained generative models for dataset distillation. GLaD [3], as a pioneering work, argues that images distilled into the latent space are less noisy and have clearer visual structures than the images in pixel space, and uses generative models for producing deep priors to integrate with matching-based methods. Minimax Diffusion [9] instead fully utilizes pre-trained diffusion models for generating realistic synthetic datasets. It finetunes the diffusion model with a minimax criterion to enhance sample representativeness and diversity. D^{4}M [33] uses pre-trained text-to-image diffusion models and adopts prototype learning to learn cluster centers for each category. These diffusion-based dataset distillation methods achieved the new state-of-the-art performance on large-scale datasets such as ImageNet, significantly outperforming the current matching-based methods in both efficiency and performance.

However, the distillation process in current diffusion-based dataset distillation methods is independent of the evaluation objective, with limited attention to the con-



(a) Previous diffusion-based methods **(b)** Illustration of our two-stage framework

Figure 2. **Comparison and illustration of previous diffusion-based methods and our approach. (a)** Current diffusion-based methods facilitate efficient distillation using a conditional diffusion model $\mathbf{G}$. However, their proposed distillation processes overlook the evaluation objective, resulting in Objective Inconsistency (OI) between distillation and evaluation, as well as Condition Inconsistency (CI) between the input condition and the generated image. **(b)** To alleviate the inconsistencies, we propose a two-stage framework: Stage 1 mitigates OI by generating an image pool and selecting more discriminative samples through a lightweight sample selector. Stage 2 refines the selected image latents using Eq. (4) and Eq. (5), reducing CI by maximizing the conditional likelihood of the generated samples. By avoiding training of the generative backbone, our method is efficient and applicable across different generative models.

straints posed by the small number of samples in the synthetic dataset. To address this discrepancy, our method introduces a two-stage framework incorporating sample selection and latent optimization, designed to achieve a more precise and consistent distillation process.

3. Methodology

3.1. Preliminary

For a conditional diffusion model with parameter θ, the process of generating a sample $\mathbf{x}_0$ adheres to a specific Markov chain structure:

$$p_\theta(\mathbf{x}_0|\mathbf{c}) = \int_{\mathbf{x}_{1:T}} p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})\, \mathrm{d}\mathbf{x}_{1:T}, \quad (1)$$

where $\mathbf{c}$ is the input condition, T is the number of denoising steps, and $p(\mathbf{x}_T)$ is the standard normal distribution. Directly maximizing the conditional likelihood $p_\theta(\mathbf{x}_0|\mathbf{c})$ is impractical, thus diffusion models instead learn to optimize the variational lower bound (ELBO) on log-likelihood:

$$\log p_\theta(\mathbf{x}_0|\mathbf{c}) \geq -\mathbb{E}_{t,\varepsilon}[\|\varepsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)\|^2]. \quad (2)$$

Given the objective of likelihood maximization, an optimal diffusion model is expected to provide accurate density estimation for each class over the entire training data distribution. This implies that diffusion models are inherently strong classifiers [4, 18], and can efficiently generate samples that represent the underlying data distribution well.

However, the effectiveness of current diffusion-based methods is limited by the misalignment between the distillation and evaluation process. In Sec. 3.2 and 3.3, we analyze two types of inconsistencies and address each to improve distillation performance. Motivated by bi-level optimization approaches, we introduce task-oriented variations in Sec. 3.4 for better method adaptation. Finally, in Sec. 3.5, we demonstrate how our method can be extended to utilize the masked autoregressive generation model.

3.2. Objective-guided Sample Selection

A conditional diffusion model takes a class condition c_i as input to generate realistic images $\mathbf{x}_0$ from sampled noises. Assume the conditional diffusion model is trained perfectly with accurate density estimations so that $p_\theta(\mathbf{x}_0|c_i)$ is maximized. Then by applying Bayes' theorem as $p_\theta(c_i|\mathbf{x}_0) \propto p_\theta(\mathbf{x}_0|c_i)p(c_i)$, we can conclude that the conditional probability $p_\theta(c_i|\mathbf{x}_0)$ is also maximized. However, this equivalence in likelihood maximization holds only for diffusion classifiers [18] with the objective in Eq. (2), and there is no guarantee that for a typical discriminative model φ with a classification objective, $p_\varphi(c_i|\mathbf{x}_0)$ can be maximized concurrently. Furthermore, [18] empirically demonstrated that, even when the discriminative model is trained on a synthetic dataset of equivalent size to the full training set, the diffusion classifier still surpasses the discriminative model by 4.8% in accuracy on ImageNet.

Moreover, randomness in noise initialization and sampling can produce low-quality, unrepresentative images, occasionally generating samples that are unhelpful for training or even poisonous. This issue is especially pronounced in low IPC settings, where each sample constitutes a significant portion of the dataset and is crucial for effective training. Therefore, we reveal the **Objective Inconsistency (OI)**

in diffusion-based dataset distillation, defined as:

Definition 1. *Let c_i denote the input condition, $\{\mathbf{x}_0^{i,k}\}_{k=1}^N$ represent the set of N images generated according to condition c_i, and $\{c^k\}_{k=1}^N$ denote the corresponding set of hard labels predicted by a trained classifier. Objective Inconsistency arises when there are mismatches between the input condition and the output label, i.e., $\exists k$, such that $c_i \neq c^k$.*

The Objective Inconsistency illustrates the discrepancy between the objectives of the distillation and evaluation stages in diffusion-based dataset distillation. It implies that some generated images, when used for classification, will be classified to belong to a different class than their conditioned label, impacting the correctness and fidelity of the distilled dataset. To alleviate this issue, we introduce a simple yet effective **sample selection** strategy based on classification probability to improve distillation performance. As shown in Fig. 2(b), while conventional methods directly generate the compact dataset, we adopt a post hoc approach by introducing a lightweight classifier to determine better samples for learning with the classification objective.

Specifically, to distill samples for a single class with a given IPC, we first generate an image pool of size $m \times$ IPC using a pre-trained diffusion model conditioned on that class, where m is a scaling factor. Next, we use a lightweight pre-trained classifier (e.g., ResNet-18) to obtain predictive probabilities for each generated image, selecting only those samples predicted to belong to the conditioned class. Building on insights from [20], which suggests prioritizing easier samples under lower IPC settings and harder samples for higher IPC settings, we further refine our selection by choosing the top-IPC most or least confident samples from the correct predictions. For classes lacking sufficient correct samples, we supplement the distilled dataset by randomly selecting from the remaining images. Additionally, we observe that using small scaling factors, such as $m = 2$ or $m = 4$, typically suffices to achieve strong performance, underscoring the efficiency of our strategy.

3.3. Condition-aware Latent Optimization

A good compact dataset expects each training image accurately reflecting its corresponding label. However, empirical diffusion models often fail to provide fully accurate density estimates of $p_\theta(\mathbf{x}_0|\mathbf{c})$ for all conditions, primarily due to the errors in approximating the ELBO and the presence of the non-zero diffusion training loss. Consequently, even after sample selection, the generated samples remain sub-optimal for conditional likelihood maximization when using hard labels as input conditions. We define this inherent discrepancy between the input condition and the output likelihood as the **Condition Inconsistency (CI)**:

Definition 2. *Let $\mathbf{x}_0^i$ represent the generated image latent conditioned on class c_i, $p_\theta(\mathbf{x}_0|\mathbf{c})$ be the conditional likelihood of the empirical diffusion model. Condition Inconsistency indicates that $\forall i, \exists j \neq i$ such that $p_\theta(\mathbf{x}_0^i|c_j) > 0$.*

This definition implies that the image latent generated by an empirical diffusion model is not exclusively associated with its specified condition, thereby weakening the correlation between the sample and its label in the distilled dataset. Since a diffusion model with a lower diffusion loss approximates the conditional likelihood more accurately, one straightforward approach to alleviate Condition Inconsistency is to learn a well-trained diffusion model for each class. However, this solution is computationally prohibitive and impractical given the limited amount of samples and the large number of existing conditions.

Instead of optimizing the model itself, we propose to update the generated images through **latent optimization**. By minimizing the diffusion loss with respect to the input $\mathbf{x}$, we allow the input to shift toward regions where the pre-trained diffusion model yields more accurate density estimates. Therefore, given the input Gaussian noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, input condition $\hat{\mathbf{c}}$ and the noisy input at time step t, the optimization objective is formulated as:

$$\min_{\mathbf{x}} \mathbb{E}_{t,\varepsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{c}}, t) - \varepsilon\|_2^2\right], \text{s.t. } \mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x} + \sqrt{1-\overline{\alpha}_t}\varepsilon. \tag{3}$$

$\overline{\alpha}_t$ is a constant with respect to t, and t is randomly sampled from $[1, \hat{T}]$, where $\hat{T} \ll T$. This suggests that the latent is only moderately perturbed. However, solely using the above objective may lead to optimized samples $\overline{\mathbf{x}}$ falling into regions associated with other classes [4]. To constrain the degree of update, we use the max norm as additional regularization such that $\|\overline{\mathbf{x}} - \mathbf{x}\|_\infty \leq \eta$. Thus, the final optimization objective is:

$$\min_{\mathbf{x}} \mathbb{E}_{t,\varepsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{c}}, t) - \varepsilon\|_2^2 + \lambda\|\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{c}}, t) - \varepsilon\|_\infty\right], \tag{4}$$

where λ controls the strength of regularization. Note that the above is conducted in the latent space and, for the sake of simplicity, we refrain from introducing new notations to distinguish between the image and its latent.

3.4. Task-oriented Variation

When computationally feasible, bi-level optimization methods exhibit strong performance [10, 37], primarily because aligning the learning objectives during distillation and evaluation naturally removes the need to redesign learning strategies. In contrast, non-matching-based methods enforce a strict separation between these two processes, preventing distillation-stage settings from benefiting the evaluation process. This discrepancy motivates us to refine the proposed two-stage approach with task-oriented designs.

Inspired by [1], we use the size of the distilled dataset as an evidence during the sample selection stage. Since training benefits from progressively learning more difficult samples, we propose emphasizing easier samples when training

on more challenging tasks, while prioritizing harder samples for simpler tasks. Consequently, we select correct and highly confident samples in lower IPC settings, and correct but less confident samples in higher IPC settings.

Algorithm 1: Pseudocode for CaO_2

Input: Pre-trained diffusion model $\epsilon(\mathbf{x}, \mathbf{c})$, sample selector $s(\mathbf{c}|\mathbf{x})$, target category set $\mathbb{C}$, IPC=N
Initialize: Distilled dataset $\mathcal{S} = \{\}$, scaling factor m, optimization condition $\hat{\mathbf{c}}$

for $c \in \mathbb{C}$ **do**
  // Sample Selection
  Obtain image pool $\mathcal{X}$ of size mN from $\epsilon(\mathbf{x}, c, t)$
  Calculate probabilities $\mathcal{P} = \text{softmax}(s(\mathbf{c}|\mathcal{X}))$
  Select the top-N correct and least/most confident samples $\mathcal{X}_N$ from $\mathcal{X}$ using $\mathcal{P}$
  // Latent Optimization
  Sample a random Gaussian noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$
  for $\mathbf{x} \in \mathcal{X}_N$ **do**
    for *each iteration* **do**
      Sample a random time step t
      Obtain $\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x} + \sqrt{1-\overline{\alpha}_t}\varepsilon$
      Update $\mathbf{x}$ using Eq. (4), Eq. (5)
    end
    $\mathcal{S} = \mathcal{S} \cup \{(\mathbf{x}, c)\}$
  end
end
Output: Distilled dataset $\mathcal{S}$

In the latent optimization stage, we examine the varying modeling quality across categories in diffusion models, and propose choosing different values for the input condition $\hat{\mathbf{c}}$. This approach is motivated by the fact that $\hat{\mathbf{c}}$ carries class-specific information that is inherently discriminative. For the set of classes $\mathbb{C}_e$ in an easier task and the set of classes $\mathbb{C}_h$ in a harder task, different values are considered:

$$\hat{\mathbf{c}} = c \cdot 1(c \in \mathbb{C}_e) + \phi \cdot 1(c \in \mathbb{C}_h) \tag{5}$$

where c is the true class label used for generating the latent, ϕ is the unconditional label used for classifier-free guidance and $1(\cdot)$ is the indicator function. For relatively easier tasks, we assume the input conditions can effectively provide discriminative information, and thus the true labels are used for guidance during optimization. For harder tasks, the guidance provided by the conditions may be insufficient or even detrimental, and thus classifier-free guidance is adopted. In this case, ϵ_θ is treated as a well-trained single-step denoiser. We utilize it so that, even without conditional guidance, the latent is optimized to approximate the result achieved with conditional guidance, implying that conditional information is embedded within the image latent.

During implementation, we examine the validation accuracy of each classification task to determine the task difficulty and the specific strategy to be used. The complete procedure of our algorithm is depicted in Algorithm 1.

3.5. Extending the Generation Backbone

Apart from conventional diffusion models, the Masked Autoregressive model (MAR) [19] has recently been proposed to reinvent autoregressive image generation. MAR utilizes an autoregressive procedure to predict image patches, with a diffusion loss guiding the training process. Motivated by the similarity in the learning objective, we show that apart from conventional diffusion models, our method can also be seamlessly applied to MAR for effective dataset distillation.

While the sample selection stage remains similar, there are two major differences in the latent optimization stage compared with the diffusion model backbone. One is that instead of adding random Gaussian noises to the image latent with respect to the time step, we apply random masks on the latents with a maximum token masking ratio. The other difference is that MAR did not design a specific embedding for the classifier-free guidance, thus we replace ϕ with a self-designed zero label embedding for Eq. (5). The optimization objective can thus be formulated as:

$$\min_{\mathbf{x}} \mathbb{E}_{\mathbf{m},\varepsilon}\left[\|\epsilon_\theta(\mathbf{x}_\mathbf{m}, \hat{\mathbf{c}}) - \varepsilon\|_2^2 + \lambda\|\epsilon_\theta(\mathbf{x}_\mathbf{m}, \hat{\mathbf{c}}) - \varepsilon\|_p\right], \tag{6}$$

where $\mathbf{m} \sim \mathcal{M}(0, \mathbf{R})$ is a randomly sampled mask with a maximum masking ratio of $\mathbf{R}$, $\mathbf{x}_\mathbf{m}$ is obtained by randomly masking $|\mathbf{m}|$ tokens from $\mathbf{x}$. The workflow of our MAR-based approach is illustrated in Appendix 6.

4. Experiments

In this section, we clarify the experiment settings and the implementation details, perform comparisons with state-of-the-art methods and diffusion-based methods, and provide extensive ablation studies for discussion and analysis.

4.1. Experimental Settings

Dataset and Evaluation Settings. For the goal of more practical and realistic application, we experiment on ImageNet [7] and its subsets, including ImageWoof [15], ImageNette [14] and ImageNet-100 [30]. ImageWoof is a challenging subset containing 10 different dog breeds, while ImageNette is a simpler subset with 10 easily distinguishable categories. ImageNet-100 includes 100 randomly selected classes from the broader ImageNet. Aiming at higher compression ratios with generalizability, we adopt three images-per-class (IPC) settings of 1, 10, and 50 for each dataset.

Currently, two evaluation paradigms exist for distilled dataset assessment: one based on hard labels [9] and the other on soft labels via knowledge distillation [34]. While

Settings	IPC	ImageWoof				ImageNette			
		SRe^{2}L	Minimax	RDED	**Ours**	SRe^{2}L	Minimax	RDED	**Ours**
ResNet-18	1	13.3 ± 0.5	19.9 ± 0.2	20.8 ± 1.2	**21.1 ± 0.6**	19.1 ± 1.1	31.8 ± 0.6	35.8 ± 1.0	**40.6 ± 0.6**
	10	20.2 ± 0.2	40.1 ± 1.0	38.5 ± 2.1	**45.6 ± 1.4**	29.4 ± 3.0	61.4 ± 0.7	61.4 ± 0.4	**65.0 ± 0.7**
	50	23.3 ± 0.3	67.0 ± 1.8	68.5 ± 0.7	**68.9 ± 1.1**	40.9 ± 0.3	84.1 ± 0.2	80.4 ± 0.4	**84.5 ± 0.6**
ResNet-50	1	14.9 ± 0.6	19.5 ± 0.6	14.5 ± 1.2	**20.6 ± 1.6**	17.5 ± 3.4	27.9 ± 0.2	32.3 ± 0.8	**33.5 ± 2.1**
	10	17.3 ± 1.7	37.3 ± 1.1	29.9 ± 2.2	**40.1 ± 0.1**	49.8 ± 2.1	66.4 ± 0.4	63.9 ± 0.7	**67.5 ± 0.8**
	50	24.8 ± 0.7	64.3 ± 0.9	67.8 ± 0.3	**68.2 ± 1.1**	71.2 ± 0.3	77.1 ± 0.7	78.0 ± 0.4	**82.7 ± 0.3**
ResNet-101	1	13.4 ± 0.1	17.7 ± 0.9	19.6 ± 1.8	**21.2 ± 1.7**	15.8 ± 0.6	24.5 ± 0.1	25.1 ± 2.7	**32.7 ± 2.5**
	10	17.7 ± 0.9	34.2 ± 1.7	31.3 ± 1.3	**36.5 ± 1.4**	23.4 ± 0.8	55.4 ± 4.5	54.0 ± 0.4	**66.3 ± 1.3**
	50	21.2 ± 0.2	62.7 ± 1.6	59.1 ± 0.7	**63.1 ± 1.3**	36.5 ± 0.7	77.4 ± 0.8	75.0 ± 1.2	**81.7 ± 1.0**

Table 1. **Performance comparison (%) with the SOTA methods over ImageNet subsets.** We adopt the evaluation paradigms from [9] and [34], reporting the higher accuracy achieved between the two. Best results are marked in **bold**.

the latter often yields superior performance, it heavily depends on the expert model's accuracy and may occasionally fail (Appendix 9). *To ensure fairness and robustness, we evaluate **both** approaches and report the best result for each method.* The distilled datasets are tested across various model architectures, including ResNet-18 [11], ResNet-50, ResNet-101, EfficientNet-B0 [35], and MobileNet-V2 [13]. To ensure stability, all experiments are repeated three times.

Baselines. We primarily compare our method with other diffusion-based dataset distillation methods and SOTA baselines, including SRe^{2}L [41], Minimax Diffusion [9], D^{4}M [33] and RDED [34]. We also include comparisons with methods such as G-VBSM [31], EDC [32], IGD [5], DATM [10], and PAD [21] in Appendix 7. SRe^{2}L pioneered efficient scaling to ImageNet-1K and outperforms other matching-based methods like MTT [2] and TESLA [6]. Minimax Diffusion and D^{4}M utilized pre-trained diffusion models for efficient distillation, with [9] applying a minimax criterion to enhance sample representativeness and diversity, while [33] clusters image latents into category prototypes to incorporate label information. RDED, on the other hand, uses the original training set by selecting and grouping the most informative image crops.

Implementation Details. We adopt the pre-trained Diffusion Transformer (DiT) [28] with 256*256 resolution as the model backbone. Image sampling is performed with 50 denoising steps using a fixed random seed. For the image refinement stage, we adopt the Adam [17] optimizer with a learning rate of 0.0006, set $\lambda = 10$ for the regularization hyperparameter, and train each image for 100 iterations. During training, we fix the input noise for each image and randomly sample a time step at each iteration. All experiments can be conducted on a single RTX A6000 GPU.

4.2. Results and Discussions

We present the comparison of distillation performance in Tab. 1 and Tab. 2. Our method is independent of the evaluation model, allowing us to use the same distilled dataset for assessments across ResNets of varying depths.

ImageWoof and ImageNette: As two smaller subsets of ImageNet, these tasks capture different aspects of the ImageNet distribution. ImageWoof includes various classes within a single species, testing the representativeness of the distilled dataset. In contrast, ImageNette, an easier dataset with distinct categories, assesses the diversity of the distilled dataset. Tab. 1 demonstrates that on these two datasets, SRe^{2}L performs sub-optimally, while Minimax Diffusion achieves results comparable to RDED. Our method surpasses the best baselines, achieving an average accuracy improvement of 1.6% across all settings for ImageWoof and 4.3% for ImageNette.

ImageNet-100 and ImageNet-1K: We further scale up to evaluations on 100 and 1,000 classes, with results shown in Tab. 2. For ImageNet-100, our method achieves an average accuracy improvement of 1.8% across all settings. On ImageNet-1K, we include D^{4}M [33] for comparison, using the reported accuracies and assuming zero deviation, as repeated results were not provided in the original paper. Our method consistently outperforms other methods in all settings, yielding an average accuracy improvement of 1.5%.

4.3. Ablation Studies

We conduct extensive ablation studies to evaluate the effectiveness of the different components in our method, and explore the impact of different hyperparameter choices.

Component Analysis. We analyze the effectiveness of each component in our method using ResNet-18 as the model for evaluation. Tab. 3 demonstrates that both components individually improve upon the baseline, and their combination yields even larger performance gains. Furthermore, the comparable accuracy improvements provided by each component indicate their equal importance.

Selection Protocols. We first examine the effects of selection pool size and selection strategy. As shown in Fig. 3(b), an optimal pool size for sample selection typically ranges around 2×IPC or 4×IPC. Interestingly, selecting from a

Settings	IPC	ImageNet-100				ImageNet-1K				
		SRe^{2}L	Minimax	RDED	**Ours**	SRe^{2}L	Minimax	D^{4}M	RDED	**Ours**
ResNet-18	1	3.0±0.3	7.3±0.1	8.1±0.3	**8.8±0.4**	0.1±0.1	5.9±0.2	-	6.6±0.2	**7.1±0.1**
	10	9.5±0.4	32.0±1.0	36.0±0.3	**36.6±0.2**	21.3±0.6	44.3±0.5	27.9±0.0	42.0±0.1	**46.1±0.2**
	50	27.0±0.4	63.9±0.1	61.6±0.1	**68.0±0.5**	46.8±0.2	58.6±0.3	55.2±0.0	56.5±0.1	**60.0±0.0**
ResNet-50	1	1.5±0.0	6.8±0.5	6.5±0.3	**7.3±0.4**	1.0±0.0	5.1±0.5	-	5.4±0.2	**7.0±0.4**
	10	3.4±0.1	30.8±0.4	29.5±0.7	**35.0±0.6**	28.4±0.1	49.7±0.8	33.5±0.0	43.6±0.5	**53.0±0.2**
	50	10.8±0.3	67.4±0.3	68.8±0.2	**70.1±0.1**	55.6±0.3	64.8±0.1	62.4±0.0	64.6±0.1	**65.5±0.1**
ResNet-101	1	2.1±0.1	5.4±0.6	6.1±0.8	**6.6±0.4**	0.6±0.1	4.0±0.5	-	5.9±0.4	**6.0±0.4**
	10	6.4±0.1	29.2±1.0	33.9±0.1	**34.5±0.4**	30.9±0.1	46.9±1.3	34.2±0.0	48.3±1.0	**52.2±1.1**
	50	25.7±0.3	67.4±0.6	66.0±0.6	**70.8±0.2**	60.8±0.5	65.6±0.1	63.4±0.0	61.2±0.4	**66.2±0.1**

Table 2. **Performance comparison (%) over ImageNet-100 and ImageNet-1K.** We adopt the evaluation paradigms from [9] and [34], reporting the higher accuracy achieved. '-' indicates missing results from the original paper. Best results are marked in **bold**.

OSS	CLO	Woof IPC=10	Woof IPC=50	Nette IPC=10	Nette IPC=50
-	-	38.7±1.1	66.1±0.8	61.7±1.7	82.0±1.1
✔	-	42.6±1.1	68.5±0.9	63.7±0.6	83.5±1.1
-	✔	42.1±1.2	67.9±0.8	64.2±2.1	83.9±0.7
✔	✔	45.6±1.4	68.7±0.5	65.0±0.7	84.5±0.6

Table 3. Accuracy (%) comparison of different components.

larger set of generated samples does not necessarily improve performance. We suspect this is due to two factors: first, while a larger pool provides more diversity, selecting from it may reduce representativeness [9]; second, although our empirical results (as will be discussed later) support the effectiveness of the current selection criterion, it is not guaranteed to consistently identify the optimal samples. Developing more advanced and complex selection protocols may address these issues, but is beyond the scope of this study.

For the selection criterion, we evaluate several strategies: random sampling, EL2N score [27] based selection, selecting only correct samples, selecting correct and more confident samples, and selecting correct but least confident samples. The EL2N score measures sample difficulty via probability output, with lower scores indicating samples that are easier to discriminate. As shown in Fig. 3(c), selecting correct and more confident samples generally yields the best performance across most cases. However, random sampling performs best on ImageNette with 50 IPC, likely because the diffusion model is well-trained on ImageNette classes, preserving both representativeness and diversity. Random selection effectively covers the modeled distribution, whereas probability-based selection protocols may introduce unintended bias into the distilled dataset.

Input condition comparison. We further analyze the choice of $\hat{\mathbf{c}}$ in Eq. (5). In Fig. 3(d), we present the performance difference, calculated as $Acc(\text{true}) - Acc(\text{NULL})$, when varying only the conditions used for latent optimization. The results suggest that for more challenging tasks such as ImageWoof and ImageNet-1K, unconditional labels yield better performance, while for easier tasks, such as ImageNette, using true class labels is preferable.

L_p norm	None	L_1	L_2	L_∞
ImageWoof (IPC=10)	42.7±0.7	43.4±2.2	44.1±0.7	**44.4±0.2**
ImageNette (IPC=10)	62.2±1.4	61.3±0.3	62.5±1.0	**63.5±0.8**

Table 4. Accuracy (%) comparison of different regularization.

Effect of regularization. Lastly, we examine the effect of different regularization forms in Eq. (4). We compare no regularization, L_1, L_2, and L_∞ norms, tuning the hyperparameter λ from $(10, 1, 0.1, 0.01)$ to achieve the best performance. Tab. 4 presents the results, where using regularization most of the time improves performance and the L_∞ norm performs the best. We attribute this to the L_∞ norm's ability to enhance the robustness of generated latent representations, aiding to prevent the learning process from producing outliers or extreme values.

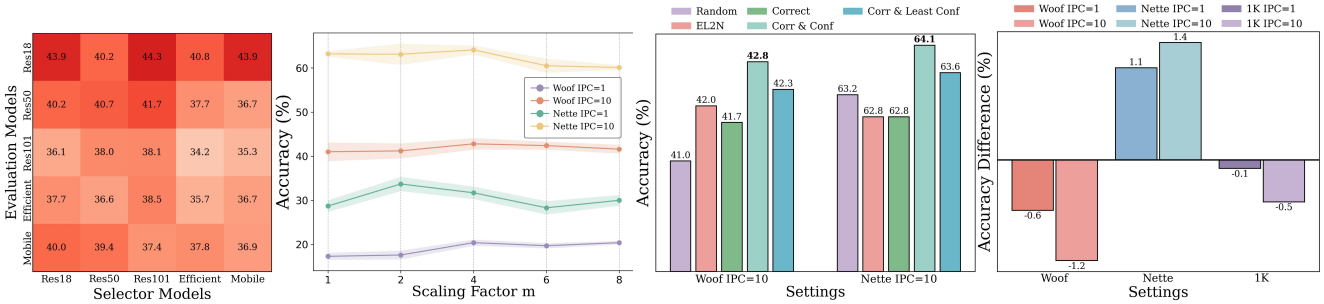
4.4. Cross-Architecture Performance

We evaluate and analyze two types of cross-architecture performances. The first is the conventional comparison between different evaluation models when trained on the same synthetic dataset. Since diffusion-based methods do not rely on the evaluation model during distillation, Tab. 1 and Tab. 2 themselves already show these comparison results, demonstrating the strong cross-architecture ability.

The second type of evaluation involves comparing performance across different combinations of selectors and evaluation models. As shown in Fig. 3(a), the models included from left to right, top to bottom, are ResNet-18, ResNet-50, ResNet-101, EfficientNet-B0, MobileNet-V2, respectively. We observe that using stronger selectors does not consistently improve performance, and the lightweight ResNet-18 is sufficient for achieving good results efficiently. Therefore, we use the pre-trained ResNet-18 as the sample selector in all our experiments.

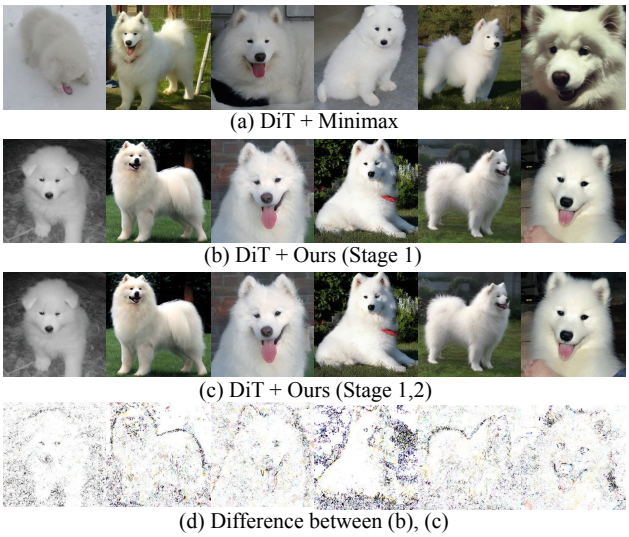
4.5. Visualization Comparison

We visualize obtained samples using the same random seeds in Fig. 4, qualitatively revealing that our method produces more discriminative and visually refined images.



(a) Cross-architecture performance **(b)** Effect of image pool size **(c)** Selection strategy comparison **(d)** Condition choice comparison

Figure 3. **Ablation studies on different components. (a)** illustrates the cross-architecture performance for selector and evaluation models, showing that a lightweight ResNet-18 generally leads to better performance in most cases. **(b)** examines the effect of the scaling factor, showing that a larger pool size is not always optimal. As a result, we adopt a pool size of 2×IPC or 4×IPC. **(c)** compares different selection strategies, concluding that prioritizing correct and more confident samples generally leads to better performance. **(d)** calculates the accuracy difference when applying various conditions for latent optimization, suggesting that true labels should be used for easier class samples, while unconditional labels are more suitable for harder class samples.



(a) DiT + Minimax

(b) DiT + Ours (Stage 1)

(c) DiT + Ours (Stage 1,2)

(d) Difference between (b), (c)

Figure 4. **Visualization of distilled images.** We compare the distilled images from Minimax Diffusion with the two stages of our method, also highlighting the differences introduced by latent optimization, where white color denotes unchanged regions.

Interestingly, latent optimization introduces imperceptible changes to the human eye. By examining the absolute differences between images before and after latent optimization, we find that the optimization primarily refines object boundaries and background details, suggesting that images might be enhanced by emphasized semantic features and improved robustness against adversarial pixels [4]. Appendix 10 provides more visualizations and analysis.

4.6. Extending to Different Backbones

Integration with MAR. As discussed in Sec. 3.5, we extend our method to masked autoregressive models [19], using MAR-Base with 32 autoregressive steps and 100 denoising steps per autoregressive step. From a 2×IPC image pool, we select the most confident correct samples. For latent optimization, we apply a 0.25 max masking ratio, train for 100 steps with a 0.0001 learning rate, and use L_∞ regularization. Tab. 5 shows that MAR alone is a strong baseline on ImageNet subsets, and combining it with our method boosts accuracy by 3.3% on ImageWoof and 2.0% on ImageNette. Visualizations are in Appendix 10.

Method	Woof IPC=10	Woof IPC=50	Nette IPC=10	Nette IPC=50
MAR [19]	38.6±0.4	69.1±0.2	68.8±1.8	86.3±0.4
+ CaO_2	**43.9±0.8**	**70.3±0.1**	**71.8±1.6**	**87.3±0.1**
Minimax [9]	40.1±1.0	67.0±1.8	61.4±0.7	83.9±0.2
+ CaO_2	**45.7±1.5**	**70.0±1.4**	**64.1±1.8**	**85.1±1.0**

Table 5. Performance (%) of integration with different backbones.

Integration with Minimax Diffusion. We further integrate our method with Minimax Diffusion [9], a finetuned diffusion transformer that improves sample quality without altering the model architecture. Using a 2×IPC image pool, L_∞ regularization, and true label guidance (except zero embeddings for ImageWoof at IPC=10), Tab. 5 shows that we further improve the performance with an average improvement of 4.3% on ImageWoof and 2.0% on ImageNette. Visualizations are available in Appendix 10. In summary, CaO_2 serves as an effective plug-and-play module for both diffusion models with diverse training objectives and masked autoregressive models with distinct generation schemes.

5. Conclusion

We propose CaO_2, an efficient two-stage framework for large-scale dataset distillation, addressing two key inconsistencies in diffusion-based dataset distillation: Objective Inconsistency and Condition Inconsistency. Our method achieves state-of-the-art performance on ImageNet and its subsets across various evaluation settings. Additionally, CaO_2 can be seamlessly integrated with different model backbones to enhance distillation performance, showcasing its versatility as a highly effective plug-and-play approach.

Acknowledgments: This research is supported by NSF IIS-2525840, CNS-2432534, ECCS-2514574, NIH 1RF1MH133764-01 and Cisco Research unrestricted gift. This article solely reflects opinions and conclusions of authors and not funding agencies.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009. 4

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 6

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[4] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023. 3, 4, 8

[5] 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. 2, 6

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 2, 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3758, 2023. 2

[9] 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 (CVPR)*, 2024. 1, 2, 5, 6, 7, 8

[10] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 4, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6

[12] Yang He, Lingao Xiao, and Joey Tianyi Zhou. You only condense once: Two rules for pruning condensed datasets, 2023. 1

[13] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 6

[14] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. 5

[15] Jeremy Howard. Imagewoof: a subset of 10 classes from imagenet that aren't so easy to classify, 2019. 5

[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 (ICML)*, 2022. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[18] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, 2023. 3

[19] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. 2, 5, 8, 1

[20] Zekai Li, Ziyao Guo, Wangbo Zhao, Tianle Zhang, Zhi-Qi Cheng, Samir Khaki, Kaipeng Zhang, Ahmad Sajedi, Konstantinos N Plataniotis, Kai Wang, and Yang You. Prioritize alignment in dataset distillation, 2024. 4

[21] Zekai Li, Ziyao Guo, Wangbo Zhao, Tianle Zhang, Zhi-Qi Cheng, Samir Khaki, Kaipeng Zhang, Ahmad Sajedi, Konstantinos N Plataniotis, Kai Wang, and Yang You. Prioritize alignment in dataset distillation, 2024. 6

[22] Zekai Li, Xinhao Zhong, Zhiyuan Liang, Yuhao Zhou, Mingjia Shi, Ziqiao Wang, Wangbo Zhao, Xuanlei Zhao, Haonan Wang, Ziheng Qin, Dai Liu, Kaipeng Zhang, Tianyi Zhou, Zheng Zhu, Kun Wang, Guang Li, Junhao Zhang, Jiawei Liu, Yiran Huang, Lingjuan Lyu, Jiancheng Lv, Yaochu Jin, Zeynep Akata, Jindong Gu, Rama Vedantam, Mike Shou, Zhiwei Deng, Yan Yan, Yuzhang Shang, George Cazenavette, Xindi Wu, Justin Cui, Tianlong Chen, Angela Yao, Manolis Kellis, Konstantinos N. Plataniotis, Bo Zhao, Zhangyang Wang, Yang You, and Kai Wang. Dd-ranking: Rethinking the evaluation of dataset distillation. GitHub repository, 2024. 3

[23] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search, 2019. 1

[24] Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching, 2023. 1, 2

[25] Wojciech Masarczyk and Ivona Tautkute. Reducing catastrophic forgetting with learning on synthetic data, 2020. 1

[26] Brian B. Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models, 2024. 1

[27] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training, 2023. 7

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1, 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[30] 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

[31] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching, 2024. 2, 6

[32] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation, 2025. 2, 6

[33] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ4m: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5809–5818, 2024. 1, 2, 6

[34] 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 (CVPR)*, 2024. 5, 6, 7, 2

[35] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 6

[36] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022. 1, 2

[37] Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective, 2025. 4

[38] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020. 1

[39] Yue Xu, Yong-Lu Li, Kaitong Cui, Ziyu Wang, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Distill gold from massive ores: Bi-level data pruning towards efficient dataset distillation, 2024. 1

[40] Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. *arXiv preprint arXiv:2311.18838*, 2023. 2

[41] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 6

[42] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[43] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 2

[44] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. 2

[45] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 2

[46] Zhenghao Zhao, Yuzhang Shang, Junyi Wu, and Yan Yan. Dataset quantization with active learning based adaptive sampling. *arXiv preprint arXiv:2407.07268*, 2024. 1

[47] Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets, 2024. 1

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

CaO_2: Rectifying Inconsistencies in Diffusion-Based Dataset Distillation

Supplementary Material

The supplementary material is organized as follows: Sec. 6 presents the process of integrating our method with MAR; Sec. 7 includes more baseline comparisons and discussions; Sec. 8 provides more ablation studies; Sec. 9 provide a more in-depth analysis of different evaluation paradigms; Sec. 10 shows more examples of distilled images across different datasets; and Sec. 11 discusses the limitations and broader impacts.

6. Generalizing to MAR

Fig. 5 shows the pipeline of how we utilize the MAR [19] backbone for our framework. The process differs from the DiT-based pipeline in two aspects: (1) Instead of perturbing the input latent using Gaussian noise w.r.t. random time steps, we perturb the input by randomly masking patches w.r.t. a maximum masking ratio; (2) The unconditional guidance is not available in MAR, thus we use a zero label embedding obtained by reformulating the `Embedding()` layer as a linear layer. The first stage of sample selection is the same as that of Fig. 2.

We find that MAR exhibits stronger distillation performance than DiT, and is more efficient in both distillation time and GPU memory cost. We utilize the MAR-Base model, but observe that using larger versions such as MAR-Large and MAR-Huge does not lead to better performance.

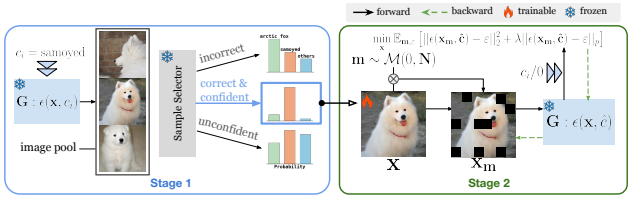


Figure 5. Pipeline of our method when applied to the Masked Autoregressive model.

7. More Baseline Comparisons

7.1. Quantitative Comparison with DD Methods

IPC	10	50
G-VBSM	31.4±0.5	51.8±0.4
EDC	48.6±0.3	58.0±0.2
DiT-IGD	45.5±0.5	59.8±0.3
Ours	46.1±0.2	60.0±0.0

Table 6. Baseline comparison on ImageNet-1K.

Woof	IPC=10	IPC=50
DiT-IGD	67.7±0.3	81.0±0.7
Ours (w/o SS)	65.0±0.7	84.5±0.6
Nette	IPC=10	IPC=50
DiT-IGD	44.8±0.8	62.0±1.1
Ours (w/o SS)	45.6±1.4	68.9±1.1

Table 7. Comparison w/o SS.

In Tab. 6, we report performance on ImageNet-1K using ResNet-18. We adopt IGD's DiT version for fair comparison. In Tab. 7, we further compare with IGD without using sample selection (SS), showing the standalone effectiveness of single-stage CaO_2.

In Tab. 8, we present CIFAR-10 results. DATM and PAD are strong trajectory matching methods but less efficient and scalable, representing a different paradigm from us.

IPC	SRe^{2}L	Ours	DATM	PAD
10	29.3±0.5	39.0±1.5	66.8±0.2	76.1±0.3
50	45.0±0.7	64.0±0.9	67.4±0.3	77.0±0.5

Table 8. Comparison on CIFAR10.

7.2. Discussion on More Related Works

LD3M [26] is similar to GLaD but replaces the GAN backbone with a diffusion model, combining matching-based approaches (e.g., MTT) with objectives that align latents to real datasets. In contrast, our method is orthogonal, as we avoid dataset matching and instead focus on fully leveraging the diffusion model, leading to improved efficiency and scalability. YOCO [12] and BiLP [39] use sample selection as preprocessing for matching-based DD to improve efficiency and denoise source data, while our method acts as a post-processing step tailored to diffusion-based DD. We also tested their protocols (EL2N, LBPE) and observed up to a 3% performance drop compared to our design.

8. More Ablations

Hyperparameter Recommendations. Though we ablated on the various hyperparameters, only a few need to be tuned. It is recommended to always use L_∞ with $\lambda = 10$, and pool size of $2/4\times$ IPC.

Time Cost Comparison. We provide the quantitative results for Fig. 1 here:

IPC	SRe^{2}L	DiT	Minimax	Ours
1	299	12	3967	99
10	2392	81	4036	960
50	11338	388	4343	3958

Table 9. Time Cost (s) Comparison on ImageWoof.

Level of noise perturbation. Beyond selection strategy and condition choice, we also investigate the impact of varying noise perturbation levels in the latent optimization process. Greater perturbation severity introduces noisier image input during latent optimization, thereby increasing denoising difficulty and accentuating key semantic features. The degree of perturbation is determined by the maximum time step $\hat{T}$, where we randomly sample $t \sim [1, \hat{T}]$. A larger $\hat{T}$ increases the amount of noisier inputs during latent optimization. Let T represent the total number of time steps; the impact of noise level is detailed in Tab. 10.

Figure 6. More examples of our distilled images on ImageWoof.

$\hat{T}$	ImageWoof IPC=10	ImageNette IPC=10
T/12	42.7±0.8	61.9±1.6
T/8	**44.4±0.2**	61.9±1.6
T/4	42.3±1.0	62.9±1.0
T/2	42.3±1.6	**63.5±0.8**
T	42.7±0.7	62.3±0.7

Table 10. Effect of the noise perturbation level.

From the table, we observe that for challenging tasks like ImageWoof, a lower level of noise perturbation is more advantageous, while for easier tasks like ImageNette, a relatively higher noise level is beneficial. Additionally, an extremely low noise level yields sub-optimal performance, as does using all time steps. We speculate that this is because the latent optimization process requires a minimum noise level to improve image robustness. For harder tasks, optimization should be conservative to avoid shifting images toward the region of another class, while for easier tasks, a more aggressive approach enhances discriminative features.

Effect of stage ordering. We analyze the ordering the of the current stage designs. As shown in Tab. 11, reversing the stages reduces performance and increases distillation time due to the additional latents requiring optimization.

Acc (%) / Time (min)	Woof IPC=10	Woof IPC=50	Nette IPC=10	Nette IPC=50
CaO$_2$	45.6 / 15	68.9 / 64	65.0 / 15	84.5 / 64
Reverse	37.3 / 46	67.7 / 115	61.9 / 46	83.0 / 115

Table 11. Effect of stage ordering.

Superiority of using generated images. We justify when generated synthetic images may be a better solution than randomly sampled real images. Tab. 12 shows that diffusion-generated images perform better than carefully selected real ones, especially under lower IPC settings. A similar phenomenon is also observed on ImageNette.

Acc (%)	IPC=1 R18	IPC=1 R50	IPC=1 R101	IPC=10 R18	IPC=10 R50	IPC=10 R101
Real	13.1±0.8	13.8±0.6	14.4±1.2	39.1±0.9	36.9±0.5	31.8±0.9
Gen	19.5±0.8	19.9±0.5	20.0±0.9	42.6±1.1	38.5±0.3	36.4±1.1

Table 12. Comparison on ImageWoof (same selection settings).

Comparison with classifier-guided models. Fig. 7 compares the performance of using classifier-guided models with classifier-free counterparts. The reasons we do not use classifier-guided models are threefolds: (1) From the table, we see that guided-diffusion empirically provides limited discriminative information, performing similarly to its classifier-free counterpart. (2) They also require additional classifiers, increasing parameters and being slower than a simple ResNet. (3) Most diffusion models are trained with CFG, thus we focus on this family of models to be more generalizable.

	IPC=10	IPC=50
CG	42.6±0.7	67.2±0.6
CFG	43.3±1.9	66.8±1.5

Figure 7. Comparison of using classifier-guidance or not.

9. Influence of different evaluation paradigms

We compare the popularly used hard-label [9] and soft-label [34] evaluation metrics in Tab. 13, using distilled images from Minimax Diffusion as an example. From the table, we show that neither of the two approaches can always obtain better performance.

Setting	Woof IPC=1	Woof IPC=10	Woof IPC=50	Nette IPC=1	Nette IPC=10	Nette IPC=50
Hard-label [9]	**19.9±0.2**	36.2±0.2	57.6±0.9	**31.8±0.6**	54.9±0.1	74.2±1.3
Soft-label [34]	18.2±1.1	**40.1±1.0**	**67.0±1.8**	22.6±1.2	**61.4±0.7**	**83.9±0.2**

Table 13. Comparison on Minimax images using ResNet18.

We also observe other cases where using hard-labels outperform soft-labels:

- For ResNet50 training on ImageNet-100 with Minimax images, using the ResNet18 model to generated soft-labels leads to only 1.0% accuracy. This indicates that a good expert is critical for successful guidance.
- For ResNet101 training on ImageNet-1K (IPC=1) with our method, using hard-labels leads to 6.0 ± 0.4 accuracy while using soft-labels leads to 5.8 ± 0.7 accuracy. We induce that the prior knowledge from the expert may be insufficient when the IPC is low.

Figure 8. Examples of our distilled images on ImageNette.

From the above results, we conclude that there is currently no unified evaluation paradigm that is being simultaneously effective, stable, and does not require external prior knowledge. Relevant works such as DD-Ranking [22] were developed, but yet (March 2025) does not support ImageNet-level datasets. Benchmarking and unifying the distilled datasets remains an open question and is of vital importance.

10. Additional Visualizations

We provide more visualization results here for a comprehensive analysis of our method.

Distilled images of ImageWoof and ImageNette. Fig. 6 and 8 show examples of distilled images under IPC=10 for ImageNette and ImageWoof. Three samples are shown for each category. From the distilled images, we see that our method effectively covers the class distribution and produces high-fidelity images. One thing we noticed is that although the classification performance on ImageNette is significantly higher than that of ImageWoof, the sample quality of both tasks is similar. The reason is straightforward: the categories in ImageNette are distinct, and therefore, easily distinguishable. This observation indicates that the class composition of a task matters, suggesting that more attention should be paid to the tasks than to the individual classes during distillation, supporting the design of our approach.

Distilled images with Minimax Diffusion backbone. We further provide examples of the images generated via the Minimax backbone. Fig. 9 shows examples of the distilled images in ImageWoof. Compared to the DiT backbone, the use of the Minimax Diffusion backbone further enhances the diversity of the distilled images. This phenomenon also suggests the extensibility of our proposed method, indicating its applicability as a plug-and-play module for existing and future work.

Distilled images with MAR backbone. Fig. 10 presents example distilled images generated using MAR as model backbone. Interestingly, although MAR-distilled images achieve higher classification performance compared to those distilled with DiT, we observe that their image quality is generally lower. In fact, the images shown are those selected for their best visual quality. We conjecture that the reason might be: although the overall image quality is low, the essential features related to the corresponding category are emphasized, while background and irrelevant features are de-emphasized. As a result, even if the images appear visually poor to human observers, they possess strong discriminative capabilities.

Figure 9. Examples of our distilled images when using the Minimax Diffusion model as backbone.

Figure 10. Examples of our distilled images when using MAR as backbone.

More analysis on Fig. 4. The optimization objective improves image-label consistency, refining *category bound-*

aries to enhance class characteristics. *Background* adjustments may occur because diffusion models, trained with a noise prediction objective, only fully denoise as $t \to \infty$. Under limited NFE, generated latents remain partially denoised, and the changes likely result from removing residual noise.

11. Limitations and Potential Improvements

Although diffusion-based methods demonstrate strong performance, their applicability is constrained by the limited conditions these models can handle (e.g. DiTs can only deal with ImageNet classes). Employing text-to-image models such as Stable Diffusion can help mitigate this issue, but the large model size and absence of classification constraints may hinder practical application.Therefore, developing efficient and task-adaptive approaches based on text-to-image models might be a way to enable effective handling of arbitrary classes. Moreover, the two inconsistencies we observe arise from the fundamental difference between generation and discrimination. Thus, developing a unified framework for both generation and classification may also significantly advance the field of diffusion-based dataset distillation.