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 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.

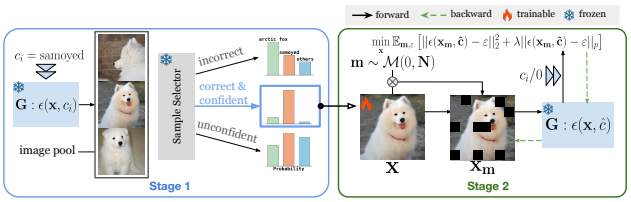

Figure 5. Pipeline of our method when applied to the Masked Autoregressive model.

7. More Baseline Comparisons

7.1. Quantitative Comparison with DD Methods

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.

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. 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

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. 9.

$\hat{T}$	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 9. 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

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

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. 10, 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 10. 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. 11 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 11. 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. 12, 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 12. 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.

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.

Figure 8. Examples of our distilled images on ImageNette.

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 boundaries* 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.