Mitigating The Distribution Shift of Diffusion-based Dataset Distillation

Supplementary Material

1. Theoretical and Intuitive Justification

Our framework is grounded in both statistical learning theory and intuitive geometric interpretations of the distillation process. Here, we provide a rigorous justification for our two core design choices.

1.1. Semantic Sparsity (RSM)

Theoretical Perspective: Dataset Distillation as Lossy Compression.

We ground our approach in the information bottleneck (IB) principle. Dataset distillation is fundamentally a lossy compression task: maximizing the mutual information $I(X_{\text{syn}}; Y)$ under a severe rate constraint $|X_{\text{syn}}| = N$. The standard diffusion objective ($\mathcal{L}_{\text{diff}}$) approximates minimizing the KL-divergence between the model and real data distributions. This effectively maximizes the likelihood of the *entire* data distribution, forcing the model to allocate capacity to model aleatoric noise and task-irrelevant high-frequency details (*e.g.*, background texture). For DD, this is inefficient "lossless" compression. By imposing an L_1 penalty on the predicted clean latent $\hat{z}_0$, we enforce a sparse prior. This acts as an *information bottleneck*, filtering out non-robust, high-frequency variations while retaining the core semantic structures (low-frequency manifolds) essential for classification.

To verify that this mechanism is *semantic simplification* rather than *magnitude reduction*, we performed a spectrum analysis in Fig. 1. The results show that RSM significantly steepens the decay of the spectrum (blue vs. red). This indicates a reduction in the effective intrinsic dimension. In deep feature spaces, leading singular values encode dominant semantic modes, while the tail encodes stochastic noise. By suppressing the tail while preserving the head, RSM functions as a semantic filter, selectively discarding redundancy.

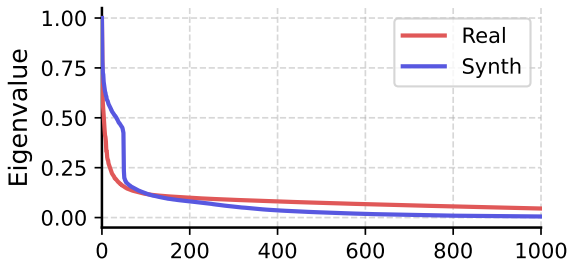

Figure 1. Spectral analysis of real and synthetic datasets.

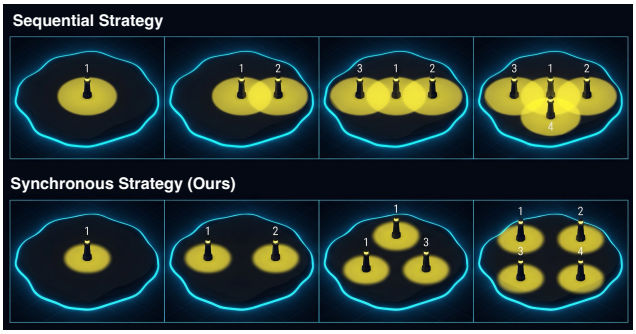

Figure 2. **The Lighthouse Analogy for Joint Sampling.** By optimizing the set collaboratively, CGS ensures that the limited samples (lighthouses) maximize the coverage of the data manifold (coastline) without redundancy (overlap).

1.2. Synchronous Sampling (CGS)

Theoretical Perspective: Global Set Optimization. The utility of a distilled dataset is a set function $U(S)$, where $S = \{z^{(1)}, \ldots, z^{(N)}\}$. This function is typically submodular (exhibiting diminishing returns). Standard guided sampling such as MiniMax [2] or DD-IGD [1] treats this as N independent optimization problems: $\max_z U(\{z\} \cup S_{\text{fixed}})$. This is a *sequential greedy algorithm*, which is known to converge only to a local optimum with a $(1 - \frac{1}{e})$ approximation bound [3]. Our Joint Sampling paradigm lifts the optimization to the product space $\mathcal{Z}^N$. By coupling the denoising paths, we perform gradient ascent directly on the joint utility $U(S)$. The DPP loss specifically maximizes the volume of the parallelotope spanned by the feature vectors in kernel space, which is a differentiable proxy for the set's diversity and coverage, ensuring a solution closer to the global optimum.

We can also interpret our motivation intuitively by an analogy of lighthouse optimization. Imagine the task of placing lighthouses to illuminate a dark island (the data manifold).

- **Sequential Strategy:** Place the first lighthouse at the center point, without knowing where the future ones will go. And the second one could only go to one side because that spot is "optimal" locally. You end up with 10 lighthouses clumped together, leaving the rest of the coast dark. This is diversity collapse.
- **Synchronousx Strategy (Ours):** Control all lighthouses simultaneously and feel a "repulsive force" (DPP) pushing them apart so they don't overlap, and a "gravitational pull" (mean matching) keeping them centered on the coastline. They naturally spread out to cover the en-

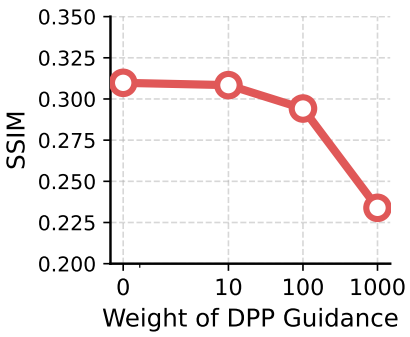

Figure 3. Comparison of image diversity with different DPP weight.

tire shape of the coast evenly.

This collaborative behavior, illustrated in Fig. 2, is intrinsic to our joint framework and unattainable by sequential methods.

2. Implementation Details

2.1. Hyperparameter Settings

Key hyperparameters for our main experimental configurations are listed in Tab. 1.

Table 1. Key hyperparameter settings for different experimental configurations.

Dataset	IPC	RSM Stage		CGS Stage	
		λ	lr	η_{DPP}	μ_{DM}
ImageNette	10	3e-3	1e-3	100	1000
	50	5e-3	1e-3	100	10000
	100	5e-3	1e-3	100	3000
ImageWoof	10	1e-2	1e-3	300	3000
	50	1e-2	1e-3	100	1000
	100	1e-2	1e-3	100	1000

3. Image Quality Analysis

Image FID. We computed FID on ImageWoof (IPC=100) and our method achieves **48.86**, surpassing Minimax (56.53), which confirms that our techniques enhance the distribution fidelity of the synthetic dataset.

Image diversity. We show a quantitative comparison for DPP in the right subfigure above. The curve of SSIM versus different DPP weights demonstrates that DPP effectively enhances diversity.

References

[1] 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

[2] 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

[3] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1): 265–294, 1978. 1