MGPO: MANIFOLD-GUIDED DIFFUSION ALIGNMENT FOR TASK-AWARE DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

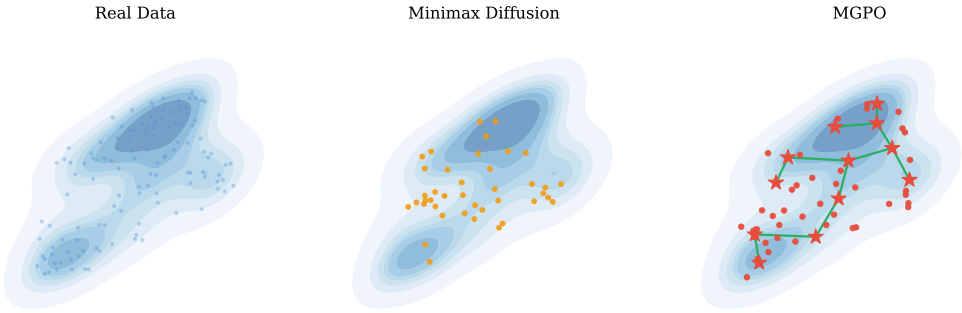

Figure 1: t-SNE of latent distributions on ImageWoof. Blue density: real data; orange dots: Minimax Diffusion; red dots: `MGPO` (ours); red stars with green edges: class-wise MST skeleton. `MGPO` covers a broader area along the MST skeleton, indicating improved manifold coverage.

ABSTRACT

Diffusion-based dataset distillation (DD) suffers from a fundamental objective mismatch: likelihood-driven diffusion models prioritize density approximation over the discriminative decision boundaries required for downstream tasks. Beyond semantic mismatch, relying solely on density also leads to geometric coverage loss, where generated samples collapse into a few high-density modes and fail to cover the manifold's structural diversity. We propose Manifold-Guided Policy Optimization (`MGPO`), which reformulates DD as a multi-objective reinforcement learning problem and achieves Dual-Space Alignment via a pixel-space discriminative reward and a latent-space geometric reward guided by a class-wise Minimum Spanning Tree (MST). The discriminative reward enforces class separability, while the MST-based geometric reward encourages generated latents to cover a sparse geometric skeleton of each class, jointly addressing both failure modes. We further provide an idealized analysis that motivates the MST-based reward, including a Hausdorff approximation bound and a subsampling bound independent of the dataset size. The reward-modular design extends to structured tasks such as object detection and segmentation by substituting the frozen task reward model. Extensive experiments show `MGPO` consistently outperforms existing methods, including a +8.0% mIoU gain on segmentation under low-budget settings.

1 INTRODUCTION

Dataset distillation (DD) (Wang et al., 2018) has emerged as a critical technique in large-scale machine learning, aiming to synthesize a small, representative subset of data from a massive original dataset. Recently, generative models, particularly diffusion models (Ho et al., 2020; Song et al., 2020), have emerged as effective backbones for this task. By leveraging powerful generative priors, diffusion models can effectively capture the complex underlying distribution of the source data to synthesize diverse and high-fidelity images. Unlike traditional gradient-matching (Zhao et al., 2020) or trajectory-matching (Cazenavette et al., 2022) approaches, which often struggle with optimization stability and limited sample diversity, these generative frameworks naturally overcome such bottlenecks. Consequently, diffusion-based methods have established state-of-the-art (SOTA) performance in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dataset distillation (Gu et al., 2024; Su et al., 2024; Wang et al., 2025; Chan-Santiago et al., 2025; Zou et al., 2025; Qi et al., 2025), offering superior results within limited computational budgets.

Despite recent advances, diffusion-based dataset distillation under extremely low-budget settings faces two distinct challenges. First, a *semantic objective mismatch*: likelihood-driven diffusion models prioritize pixel-level density approximation over the discriminative features required by downstream tasks (Biroli et al., 2024; Ho et al., 2020; Karczewski et al., 2025). Second, a *geometric coverage loss*: relying solely on density or naive classifier guidance often leads to mode collapse, failing to capture the complex geometric manifold structure and structural diversity of each class in the original data distribution. Consequently, generated samples may exhibit high perceptual fidelity yet fail to effectively represent the dataset.

These challenges show that diffusion-based distillation must transcend perceptual fidelity by resolving both issues at the same time. An effective distilled dataset requires two orthogonal properties: semantic discriminability, which pushes samples toward highly separable regions to better serve downstream tasks; and geometric preservation, which spreads samples along a sparse geometric skeleton of each class to maintain structural diversity and prevent mode collapse (Guo et al., 2024).

Following these principles, we propose Manifold-Guided Policy Optimization (`MGPO`), a framework that reformulates diffusion-based dataset distillation as a multi-objective reinforcement learning problem. Our core contribution is a Dual-Space Alignment strategy that addresses both challenges in their respective domains. We design a pixel-space discriminative reward to enforce class separability, alongside a latent-space geometric reward guided by a class-wise Minimum Spanning Tree (MST) to encourage geometric coverage. Crucially, `MGPO` utilizes a decoupled group-wise policy optimization strategy to internalize these heterogeneous, non-differentiable rewards directly into the diffusion weights. This RL formulation eliminates the need for expensive inference-time guidance, ensuring highly efficient sampling while harmonizing semantics and geometry.

We further provide an idealized analysis showing that, under smoothness and reach assumptions, a Euclidean MST δ-approximates the class manifold in Hausdorff distance with δ^2/τ tightness in the backward direction. This analysis motivates the distance term of the geometric reward; the implemented reward is evaluated empirically.

Beyond classification, the flexibility of our RL formulation unlocks a new reward-modular paradigm for dataset distillation. While the MST-based geometric prior is inherently task-agnostic, the outer framework can be adapted to complex downstream tasks, such as object detection and segmentation, by substituting the pixel-space reward model. In such scenarios, the same frozen task model provides the reward during fine-tuning and pseudo-annotates the class-conditioned generated images afterwards, without modifying the core generative architecture.

Extensive empirical evaluations across multiple benchmarks demonstrate that `MGPO` consistently outperforms existing SOTA approaches. For instance, under representative low-budget settings, our method achieves 18.3 mIoU on segmentation tasks, surpassing the strongest baseline by +8.0 mIoU.

Our contributions are summarized as follows:

- We reformulate diffusion-based dataset distillation as a multi-objective reinforcement learning problem, internalizing task-specific rewards directly into the diffusion weights and eliminating inference-time guidance overhead.
- We design a dual-space reward with a pixel-level discriminative term and a latent-level MST geometric term, addressing the semantic mismatch and geometric coverage loss.
- We provide an idealized analysis motivating the MST-based reward: under smoothness and reach assumptions, (i) a Euclidean MST approximates the class manifold in Hausdorff distance, and (ii) a subsampled MST admits a bound independent of the dataset size.
- Extensive experiments show `MGPO` achieves SOTA performance on standard classification benchmarks and extends to other complex structured tasks like object detection and semantic segmentation.

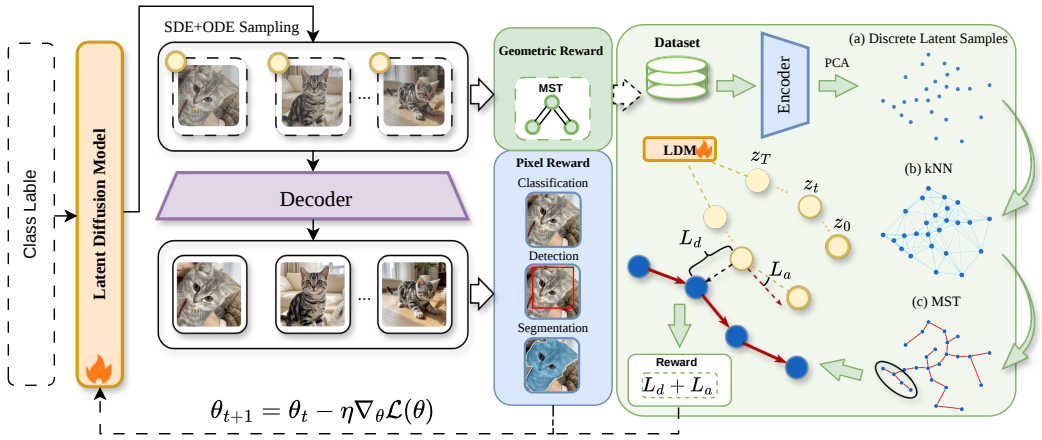

Figure 2: Overview of the proposed MGPO framework. (a) PIXEL REWARD (r_{disc}) is computed from downstream objectives such as classification, detection, and segmentation, encouraging semantic correctness and task consistency. (b) MST REWARD (r_{geo}) is constructed in latent space. Real dataset features are encoded by encoder and projected via PCA to obtain a structured latent representation. A kNN graph is built to approximate local neighborhood structure, followed by a Minimum Spanning Tree (MST) to capture the global manifold topology. The generated latent samples are aligned with this manifold structure, promoting both representativeness and structural consistency.

2 RELATED WORK

2.1 DIFFUSION MODELS FOR DATASET DISTILLATION

Recent work has explored diffusion models as powerful generative backbones for DD. Minimax (Gu et al., 2024) fine-tunes pre-trained diffusion models under a minimax objective to balance sample representativeness and diversity. D^{4}M (Su et al., 2024) incorporates text-to-image diffusion with prototype learning to improve intra-class consistency. D3HR (Zhao et al., 2025) addresses distribution mismatch via DDIM inversion (Song et al., 2020) and group-wise sampling, aligning generated and real data distributions. CaO2 (Wang et al., 2025) performs inference-time latent optimization to reduce feature level inconsistencies, while DDVLCP (Zou et al., 2025) leverages vision–language prototypes to enhance semantic coherence. Despite their success, diffusion-based approaches are inherently likelihood-centric, concentrating samples in high-density regions, which preserves perceptual fidelity but neglects boundary and hard examples essential for discriminative training. Most related to our work, ManifoldGD (Roy et al., 2026) explores manifold structure for diffusion-based distillation, but relies on inference-time guidance without formal geometric guarantees. In contrast, **MGPO** internalizes geometric constraints into the diffusion policy via RL, with an idealized Hausdorff analysis motivating the MST-based reward, while simultaneously resolving the semantic objective mismatch through dual-space alignment.

2.2 REINFORCEMENT LEARNING IN DIFFUSION MODELS

Reinforcement learning has recently been introduced to align diffusion and flow-based generative models with task-level objectives beyond likelihood-based training. A representative approach is Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which stabilizes policy updates by comparing relative rewards within groups of sampled trajectories. Subsequent works extend GRPO to generative settings. Flow-GRPO (Liu et al., 2025) adapts it to diffusion and flow matching models, while Temp-Flow-GRPO (He et al., 2025) introduces time-dependent credit assignment. Mix-GRPO (Li et al., 2026) combines multiple sampling dynamics, and structured variants such as Branch-GRPO (Li et al., 2025) and Dance-GRPO (Xue et al., 2025) improve scalability through branching and group-wise rollouts. However, existing methods focus primarily on generative quality or preference alignment. In contrast, we leverage reinforcement learning to directly optimize downstream discriminative objectives. More related works see Appendix B.1

3 PRELIMINARY

3.1 DATASET DISTILLATION

Dataset distillation aims to synthesize a small and compact dataset $\mathcal{S} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{N_S}$ that preserves the essential training information of a large-scale real dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N_T}$, where x_i denotes an input sample (e.g., an image), y_i its corresponding label, and $N_S \ll N_T$. The goal of dataset distillation is to enable models trained on $\mathcal{S}$ to achieve performance comparable to those trained on $\mathcal{T}$, while significantly reducing storage and computational costs.

3.2 DDIM SAMPLING IN LATENT SPACE

We adopt a latent diffusion formulation, where $z_0 = E(x_0)$ denotes the VAE-encoded latent, and diffusion is performed directly on the latent variables z_t. The DiT backbone parameterizes the noise prediction $\epsilon_\theta(z_t, t)$ in this latent space. We model the reverse diffusion dynamics using the generalized sampling process of Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020). The transition from state z_t to z_{t-1} is defined as:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(z_t, t) + \sigma_t \epsilon_t, \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ is standard Gaussian noise and α_t follows the standard variance schedule. This non Markovian formulation unifies deterministic and stochastic sampling through the variance parameter $\sigma_t = \eta\sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{1 - \frac{\alpha_t}{\alpha_{t-1}}}$, which is parameterized by a scalar η. The hyperparameter η explicitly controls the degree of stochasticity. When $\eta = 0$, the random noise term vanishes, reducing the sampling process to the strictly deterministic DDIM formulation. Conversely, when $\eta = 1$, the process recovers the stochastic Markovian transitions of traditional DDPM (Ho et al., 2020).

During the reward driven policy optimization phase, we maintain $\eta = 1$ to encourage stochastic trajectory exploration. Once the generative policy is fully aligned with our dual space objectives, we set $\eta = 0$ to perform highly efficient and deterministic dataset synthesis.

4 METHOD

To bridge the gap between likelihood driven diffusion generation and discriminative downstream tasks, **MGPO** internalizes dual space task objectives into the diffusion backbone via RL.

4.1 DUAL-SPACE REWARD DESIGN

To align diffusion generation with downstream objectives, we design two complementary rewards in different spaces. The discriminative reward encourages semantic separability in pixel space, while the geometric reward encourages coverage of class-level geometric structure in latent space.

Discriminative Reward (r_{disc}). To ensure the synthesized samples are informative for training downstream models, we employ a pre-trained proxy classifier f_ϕ. For a generated sample $\tilde{x}_0$ with target label c, the reward is defined as the log probability of the ground-truth class:

$$r_{disc}(\tilde{x}_0, c) = \log \left(\frac{\exp(f_\phi(\tilde{x}_0)_c)}{\sum_k \exp(f_\phi(\tilde{x}_0)_k)} \right), \tag{2}$$

where $f_\phi(\tilde{x}_0)_c$ denotes the predicted logit for class c. This continuous formulation provides a smooth reward landscape and penalizes samples that lie near ambiguous decision boundaries.

Manifold Geometric Reward (r_{geo}). Likelihood-driven diffusion models tend to concentrate samples in high-density regions, which may overlook geometrically informative regions of the class manifold. To encourage geometric coverage of each class, we introduce a geometric reward grounded in a class-wise Minimum Spanning Tree (MST), which provides a sparse geometric skeleton of the class distribution and is more discriminative than density-based regularization.

Concretely, let $\mathcal{Z}_c = \{z_i\}_{i=1}^{N_c}$ denote the VAE-encoded latent representations of class-c samples. We model the class manifold via a weighted graph $G_c = (\mathcal{Z}_c, E_c)$ with edge weights $w(u, v) = \|u - v\|_2$, and define the class-wise MST as

$$\mathcal{T}_c = \arg\min_{T \subseteq G_c} \sum_{(u,v)\in T} w(u, v), \tag{3}$$

where T spans all nodes and contains no cycles. To ensure scalability, in practice we construct $\mathcal{T}_c$ from a subsampled latent set on a sparse k-nearest-neighbor graph (see Algorithm 2).

Given a generated latent z_0 corresponding to $\tilde{x}_0$, we project it into the normalized latent space via $u_{\text{gen}} = g_c(z_0)$, and define the geometric reward as

$$\begin{aligned} r_{\text{geo}}(\tilde{x}_0, c) &= -\lambda_{\text{dist}} \min_{v\in\mathcal{V}_c} \|u_{\text{gen}} - v\|_2 - \lambda_{\text{dir}}\big(1 - |\cos(\mathbf{v}_{\text{gen}}, \mathbf{d}_{e^*})|\big), \\ e^* &= \arg\min_{(i,j)\in\mathcal{E}_c} \big\| u_{\text{mid}} - \tfrac{v_i+v_j}{2} \big\|_2. \end{aligned} \tag{4}$$

where $\mathcal{V}_c$ and $\mathcal{E}_c$ are the node and edge sets of $\mathcal{T}_c$, the first term is the Euclidean distance from the projected terminal latent to its nearest MST node, and λ_{dir} controls the strength of the directional alignment term (analysis in Appendix). The transition direction is $\mathbf{v}_{\text{gen}} = g_c(z_0) - g_c(z_{\text{mid}})$, where z_{mid} is a latent sampled from an intermediate diffusion timestep, and $\mathbf{d}_{e^*}$ is the direction of the MST edge e^* whose midpoint is nearest to the projected intermediate latent $u_{\text{mid}} = g_c(z_{\text{mid}})$. The first term encourages generated samples to remain close to the manifold skeleton, while the second encourages alignment with its local direction. By incorporating r_{geo} into policy optimization, the diffusion model internalizes the geometric prior during fine-tuning, removing the need for inference-time guidance.

Theoretical Justification. We provide two idealized results that motivate the MST-based geometric reward and its subsampled construction. Both concern a Euclidean MST built on samples from a smooth class manifold and an edge-distance formulation. They do not cover the VAE/PCA representation, the nearest-node distance used in practice, or the direction term, which we evaluate empirically. Let $\mathcal{M}_c$ be a compact, connected C^2 class submanifold with reach $\tau_c > 0$, and let the MST nodes cover $\mathcal{M}_c$ with radius $\delta < \tau_c$. Theorem 5 (Appendix A.2) shows that $d_H(\mathcal{M}_c, |\mathcal{T}_c|) \le \delta$, with the sharper backward bound $\sup_{y\in|\mathcal{T}_c|} d(y, \mathcal{M}_c) \le \delta^2/\tau_c$; that is, in this idealized setting the MST is a reach-aware geometric approximation of the class manifold. Theorem 10 (Appendix A.3) shows that an MST built on n i.i.d. samples satisfies $d_H(|\mathcal{T}(V_c)|, |\mathcal{T}(U_c)|) = \mathcal{O}\big((\log(n/\eta)/n)^{1/m_c}\big)$ with probability at least $1 - \eta$, where m_c is the intrinsic dimension; this bound is independent of the full dataset size N_c, so the per-class MST construction cost stays constant as datasets scale. These results together motivate our design of the implemented reward.

4.2 MGPO : Manifold-Guided Policy Optimization

We formulate diffusion-based dataset distillation as a sequential decision process (see Appendix A.1), where the diffusion model acts as a stochastic policy that gradually transforms Gaussian noise into a synthetic sample. At each timestep, the policy $\pi_\theta(z_{t-1} \mid z_t, c)$ predicts the reverse diffusion transition conditioned on the class label c.

Following the generalized DDIM formulation introduced in Sec. 3.2, we set $\eta = 1$ during policy optimization to introduce stochasticity into the sampling process. This injected stochasticity is vital as it enables the policy to explore a broader range within latent space, while task-driven rewards guide the trajectories toward regions that are both discriminative and geometrically structured.

Policy Optimization. To optimize the generative policy without training an explicit value network, we adopt a relative policy optimization strategy. For each class condition c, we sample a group of G trajectories $\{z_0^{(i)}\}_{i=1}^{G}$ where G denotes the rollout group size (see Sec. 5.4), and evaluate them using the dual-space rewards defined in Sec. 4.1. The diffusion policy is updated by maximizing a clipped surrogate objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{c\sim p(c),\epsilon}\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta\, \mathbb{D}_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}}\big)\right], \tag{5}$$

Algorithm 1 MGPO Training Procedure

Require: Diffusion policy π_θ, proxy reward model f_ϕ, class-wise MST skeleton $\mathcal{T}_c$ and projection function $g_c(\cdot)$

for each training iteration **do**
  Sample class label $c \sim p(c)$
  Generate G trajectories $\{\tau_i\}_{i=1}^{G}$ via SDE and ODE sampling
  for each trajectory $\tau_i = \{z_T^{(i)}, \ldots, z_0^{(i)}\}$ **do**
    Compute geometric reward $r_{\text{geo}}^{(i)}$ using latent projection $u_{\text{gen}}^{(i)} = g_c(z_0^{(i)})$
    Decode image $\tilde{x}_0^{(i)} = D(z_0^{(i)})$
    Compute discriminative reward $r_{\text{disc}}^{(i)}$ using (Eq. 2)
  end for
  Perform group-wise normalization of rewards (Eq. 6)
  Aggregate advantages across objectives (Eq. 7)
  Apply batch-level normalization (Eq. 8)
  Update diffusion policy using the MGPO objective (Eq. 5)
end for

where $\rho_i = \frac{\pi_\theta(z_{t-1}^{(i)}|z_t^{(i)},c)}{\pi_{\theta_{\text{old}}}(z_{t-1}^{(i)}|z_t^{(i)},c)}$ denotes the importance sampling ratio and $\hat{A}_i$ is the aggregated advantage of the i-th trajectory.

The MST skeleton and the projection function $g_c(\cdot)$ are pre-computed following Algorithm 2. MGPO training procedure is summarized in Algorithm 1.

Multi-Reward Aggregation. A key challenge in dual-space optimization is that semantic and geometric rewards may lie on different numerical scales. Directly combining them can lead to unstable training. Following GDPO (Liu et al., 2026), we adopt a three-stage advantage normalization strategy.

First, for each reward type $k \in \{\text{disc}, \text{geo}\}$, we perform group-wise normalization across the G rollouts:

$$A_{i,k} = \frac{r_k^{(i)} - \text{mean}(\{r_k^{(j)}\}_{j=1}^{G})}{\text{std}(\{r_k^{(j)}\}_{j=1}^{G}) + \epsilon}. \tag{6}$$

This step removes scale differences while preserving the relative ranking of samples within the group.

Second, the normalized advantages are combined through a weighted summation:

$$A_i^{\text{sum}} = w_{\text{disc}} A_{i,\text{disc}} + w_{\text{geo}} A_{i,\text{geo}}. \tag{7}$$

where w_{disc} and w_{geo} weight the discriminative and geometric rewards (Sec. 5.4).

Finally, we apply batch-level normalization across all samples in the training batch:

$$\hat{A}_i = \frac{A_i^{\text{sum}} - \text{mean}(\{A_j^{\text{sum}}\})}{\text{std}(\{A_j^{\text{sum}}\}) + \epsilon}. \tag{8}$$

This normalization stabilizes policy gradients and prevents numerical explosion when fine-tuning large diffusion backbones.

4.3 Reward-Modular Extension to Structured Tasks

While the geometric reward r_{geo} in Section 4.1 is task-agnostic, structured tasks such as object detection and segmentation require pixel-space rewards that capture spatial, in addition to semantic, correctness. The modular design of MGPO allows substituting r_{disc} with such task-specific rewards while keeping the policy optimization, geometric reward, and overall training pipeline unchanged. We illustrate this with two examples; each task still requires a *task-appropriate* reward design, but the burden is reduced from modifying the framework to specifying a scalar reward signal.

Object Detection. To encourage clear spatial boundaries without composite losses, we derive the reward from a pretrained detector. For an image $\tilde{x}_0$ conditioned on class c, we apply Non-Maximum Suppression to filter redundant predictions, and define the reward as the maximum remaining confidence score for class c:

$$r_{\text{det}}(\tilde{x}_0, c) = \max\big(\mathcal{S}_c(\tilde{x}_0)\big), \tag{9}$$

where $\mathcal{S}_c(\tilde{x}_0)$ denotes the set of post-NMS confidence scores.

Image Segmentation. Instead of relying on predefined mask annotations, we extract the spatial probability map from a pre-trained segmentation network. To prevent reward hacking via isolated high-confidence pixels, we average the top-K pixel confidences for class c:

$$r_{\text{seg}}(\tilde{x}_0, c) = \frac{1}{K} \sum_{k=1}^{K} \text{TopK}\big(\mathcal{P}_c(\tilde{x}_0)\big)_k, \tag{10}$$

where $\mathcal{P}_c(\tilde{x}_0)$ is the flattened probability map. The hyperparameter K enforces a minimum expected pixel area (analysis in Appendix).

In both cases, the reward depends only on the prompted class c. Spatial annotations are obtained by class-conditioned image generation followed by pseudo-annotation with the same frozen task model. For detection, we keep the post-NMS predictions of class c (score threshold 0.05, IoU threshold 0.5) as boxes. For segmentation, we keep the pixels whose argmax prediction is c as the mask; the Top-K operation in Eq. (10) only computes the scalar reward. Images without a valid class-c prediction are kept with empty annotations (detection) or an all-background mask (segmentation). The current formulation therefore does not model multi-class co-occurrence or inter-class relations (see Appendix F). This reward-modular design allows MGPO to extend to new structured tasks through reward specification rather than architectural modification, though designing an effective reward for each task remains a non-trivial step (see Appendix C for further design considerations).

5 EXPERIMENTS

Table 1: Comparison of SOTA dataset distillation methods on ImageWoof under various IPC and backbone settings. The best results are marked as bold, and the second are underlined.

IPC (Ratio)	Model	Random	K-Center	Herding	DiT	Minimax	DM	IDC-1	D^4M	DDVLCP	ManifoldGD	MGPO (Ours)	Full
10 (0.8%)	ConvNet-6	24.3±1.1	19.4±0.9	26.7±0.5	34.2±1.1	37.0±1.0	26.9±1.2	33.3±1.1	29.4±0.9	34.8±2.4	36.9±0.6	**37.2±0.8**	86.4±0.2
	ResNetAP-10	29.4±0.8	22.1±0.1	32.0±0.3	34.7±0.5	39.2±1.3	30.3±1.2	39.1±0.5	33.2±2.1	39.5±1.5	38.3±0.4	**40.9±1.1**	87.5±0.5
	ResNet-18	27.7±0.9	21.1±0.4	30.2±1.2	34.7±0.4	37.6±0.9	33.4±0.7	37.3±0.2	32.3±1.2	39.9±2.6	39.2±0.7	**40.1±0.3**	89.3±1.2
20 (1.6%)	ConvNet-6	29.1±0.7	21.5±0.8	29.5±0.3	36.1±0.8	37.6±0.2	29.9±1.0	35.5±0.8	34.0±2.3	37.9±1.9	37.7±0.6	**39.6±1.3**	86.4±0.2
	ResNetAP-10	32.7±0.4	25.1±0.7	34.9±0.1	41.1±0.8	45.8±0.5	35.2±0.6	43.4±0.3	40.1±1.6	44.5±2.2	45.6±0.5	**46.1±0.3**	87.5±0.5
	ResNet-18	29.7±0.5	23.6±0.3	32.2±0.6	40.5±0.5	42.5±0.6	29.8±1.7	38.6±0.2	38.4±1.1	44.5±2.0	42.4±0.4	**45.9±0.2**	89.3±1.2
50 (3.8%)	ConvNet-6	41.3±0.6	36.5±1.0	40.3±0.7	46.5±0.8	53.9±0.6	44.4±1.0	43.9±1.2	47.4±0.9	54.5±0.6	51.3±0.2	**55.7±0.4**	86.4±0.2
	ResNetAP-10	47.2±1.3	40.6±0.4	49.1±0.7	49.3±0.2	56.3±1.0	47.1±1.1	48.3±1.0	51.7±3.2	57.3±0.5	57.9±0.6	**58.1±0.2**	87.5±0.5
	ResNet-18	47.9±1.8	39.6±1.0	48.3±1.2	50.1±0.5	57.1±0.6	46.2±0.6	48.3±0.8	53.7±2.2	58.9±1.5	58.2±0.6	**59.7±0.4**	89.3±1.2
70 (5.4%)	ConvNet-6	46.3±0.6	38.6±0.7	46.2±0.6	50.1±1.2	55.7±0.9	47.5±0.8	48.9±0.7	50.5±0.4	55.8±1.7	55.3±0.6	**56.4±1.3**	86.4±0.2
	ResNetAP-10	50.8±0.6	45.9±1.5	53.4±1.4	54.3±0.9	58.3±0.2	51.7±0.8	52.8±1.8	54.7±1.6	60.6±0.3	60.6±0.1	**60.8±0.4**	87.5±0.5
	ResNet-18	52.1±1.0	44.6±1.1	49.7±0.8	51.5±1.0	58.8±0.7	51.9±0.8	51.1±1.7	56.3±1.8	60.3±0.3	61.0±0.6	**61.2±0.9**	89.3±1.2
100 (7.7%)	ConvNet-6	52.2±0.4	45.1±0.5	54.4±1.1	53.4±0.3	61.1±0.7	55.0±1.3	53.2±0.9	57.9±1.5	**62.7±1.4**	60.6±0.2	61.3±1.1	86.4±0.2
	ResNetAP-10	59.4±1.0	54.8±0.2	61.7±0.9	58.3±0.8	64.5±0.2	56.4±0.8	56.1±0.9	59.5±1.8	65.7±0.5	65.2±0.3	**65.8±0.6**	87.5±0.5
	ResNet-18	61.5±1.3	50.4±0.4	59.3±0.7	58.9±1.3	65.7±0.4	60.2±1.0	58.3±1.2	63.8±1.3	68.3±0.4	66.3±0.6	**68.4±0.2**	89.3±1.2

5.1 EXPERIMENTAL SETUP

Generative Backbone. For classification experiments, we adopt a pretrained DiT-XL/2 (Peebles & Xie, 2022) as the generative backbone and fine-tune it using DiffFit (Xie et al., 2023), all synthetic images are generated at a resolution of 256×256. For detection and segmentation, we follow the experimental protocol of UniDD (Qi et al., 2025) and employ Stable Diffusion v1.4 (Rombach et al., 2022) as the generative backbone to synthesize labeled images with structured annotations.

Implementation Details. For classification, student networks are trained from scratch for 300 epochs using SGD with cosine learning-rate decay. For downstream tasks, we use Faster R-CNN (Ren et al., 2016) with a ResNet-50 backbone for object detection and LR-ASPP with a MobileNetV3 (Howard et al., 2019) backbone for semantic segmentation, with standard pretrained backbone weights. All experiments are conducted on a single NVIDIA A100 GPU (80GB). To ensure statistical reliability, each synthetic dataset is generated three times and each downstream model is trained with three random seeds; we report mean $\pm$ standard deviation. More details see Appendix C.

Baselines. For classification, we compare against representative optimization-based and diffusion-based dataset distillation methods, including DM (Zhao & Bilen, 2023), IDC (Kim et al., 2022), SRe^{2}L (Yin et al., 2023), RDED (Sun et al., 2024), Minimax (Gu et al., 2024), D^{4}M (Su et al., 2024), DDVLCP (Zou et al., 2025), CaO2 (Wang et al., 2025), and ManifoldGD (Roy et al., 2026). For segmentation and object detection, we primarily compare with UniDD (Qi et al., 2025), a recent universal dataset distillation framework. Since diffusion-based distillation for structured outputs remains relatively underexplored, we additionally include coreset selection baselines (Random, Uniform, K-Center (Sener & Savarese, 2018), and Herding (Welling, 2009)).

Evaluation Protocol. For classification, we evaluate ConvNet-6, ResNetAP-10, and ResNet-18 trained from scratch. For segmentation and object detection, we follow standard training recipes with pretrained backbones as described above. Following prior works, we report the better performance between hard-label and soft-label supervision.

5.2 Results on Classification

We first evaluate our approach on classification benchmarks, providing comprehensive comparisons with multiple state-of-the-art dataset distillation methods across four benchmark datasets: ImageWoof (Howard, 2019b), ImageNette (Howard, 2019a), ImageIDC (Kim et al., 2022) and ImageNet-1K (Russakovsky et al., 2015). The results are summarized in Tables 1, 2 and 3. Additional comparison experiment results with CaO2 (Wang et al., 2025) are provided in Appendix D, Table 7.

Table 2: Comparison on ImageNette and ImageIDC under different IPC settings. Results are obtained on ResNet-18. The best results are marked as bold and the second are underlined.

	IPC	Random	DiT	Minimax	D^{4}M	DDVLCP	MGPO (Ours)
Nette	10	54.2±1.6	59.1±0.7	58.6±0.4	60.9±1.7	<u>64.8±3.6</u>	**66.9±2.2**
	20	63.5±0.5	64.8±1.2	70.6±0.3	66.3±1.3	<u>71.4±0.5</u>	**73.9±0.7**
	50	76.1±1.1	73.3±0.9	<u>83.7±0.4</u>	77.7±1.1	81.2±0.8	**85.4±1.1**
IDC	10	48.1±0.8	54.1±0.4	51.9±1.4	52.8±0.5	**57.0±1.4**	<u>54.6±1.7</u>
	20	52.5±0.9	58.9±0.2	59.1±3.7	58.5±0.4	<u>63.3±1.2</u>	**64.8±0.4**
	50	68.1±0.7	64.3±0.6	69.4±1.4	69.1±0.8	<u>71.9±0.4</u>	**77.8±0.9**

Table 3: Comparison of dataset distillation methods on ImageNet-1K under different IPC settings.

IPC	SRe^{2}L	RDED	DiT	Minimax	MGPO (Ours)
10	21.3±0.6	42.0±0.1	39.6±0.4	44.3±0.5	**45.1±0.6**
50	46.8±0.2	56.5±0.1	52.9±0.6	58.6±0.3	**59.1±0.4**

5.3 Result on Detection and Segmentation

We further investigate whether the distilled datasets preserve spatial structures that generalize to downstream tasks, particularly semantic segmentation and object detection.

Object Detection. Table 4 reports object detection results on Pascal VOC 2012 (Everingham et al.) and MS-COCO (Lin et al., 2015). Under aggressive compression, our method substantially outperforms coreset selection baselines, which indicates that the distilled images retain object-level localization cues rather than merely category-level appearance statistics.

Segmentation. Table 5 summarizes segmentation performance on Pascal VOC (Everingham et al.). Our method achieves consistently higher mIoU across compression ratios, suggesting that pixel-level structural information is preserved during distillation and remains useful for downstream learning.

Table 4: Performance comparison with coreset selection methods. Metrics are reported as mAP and $AP_{50}(\%)$ across different selection ratios.

Method	Pascal VOC						MS COCO					
	0.5%		1%		2%		0.25%		0.5%		1%	
	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
Random	0.8±0.2	3.1±0.4	4.2±0.5	13.7±0.6	12.4±0.4	34.3±0.5	0.5±0.1	1.7±0.3	3.7±0.2	10.1±0.3	7.2±0.8	17.3±0.9
Uniform	0.9±0.1	3.4±0.3	5.7±0.2	17.7±0.4	13.8±0.3	36.2±0.4	0.8±0.2	2.4±0.5	3.4±0.4	9.5±0.6	7.4±0.5	17.6±0.5
K-Center	0.5±0.1	2.1±0.3	3.6±0.6	12.3±0.3	10.9±0.6	29.3±0.6	0.4±0.1	1.5±0.2	3.2±0.5	9.5±0.5	6.1±0.3	15.4±0.6
Herding	0.6±0.2	2.4±0.2	3.5±0.5	11.9±0.5	10.4±0.4	28.7±0.7	0.5±0.1	1.8±0.4	3.5±0.3	9.7±0.3	6.7±0.4	16.3±0.7
UniDD	8.5±0.4	22.3±0.6	16.8±0.5	38.9±0.7	23.9±0.5	48.5±0.6	4.5±0.3	10.3±0.4	7.1±0.4	16.9±0.3	10.8±0.4	22.5±0.5
MGPO(Ours)	**13.4**±0.2	**26.3**±0.5	**21.1**±0.4	**41.8**±0.3	**26.9**±0.9	**53.1**±1.3	**5.8**±0.2	**13.9**±0.3	**8.3**±0.5	**19.9**±0.2	**11.7**±0.8	**25.4**±0.8
Full Data	51.4±0.8 / 80.3±0.4						32.6±0.7 / 51.4±0.8					

Table 5: Semantic Segmentation results.

Methods	Semantic Segmentation (mIoU)			
	3.5%	7%	14%	Full
Random	6.7±0.4	9.6±0.5	11.3±0.4	32.9±0.6
Uniform	6.5±0.6	9.3±0.4	12.1±0.5	
K-Center	6.1±0.4	8.5±0.6	10.4±0.7	
Herding	6.2±0.7	8.7±0.4	10.6±0.3	
UniDD	10.3±0.5	12.3±0.3	14.9±0.7	
MGPO (Ours)	**18.3±0.5**	**20.4±0.4**	**23.1±0.6**	

Table 6: Ablation study of rewards.

Method	ImageWoof		ImageNette	
	IPC = 10	IPC = 50	IPC = 10	IPC = 50
DiT	34.7±0.5	49.3±0.2	58.9±1.2	82.9±0.1
Pixel Reward	37.0 ±0.3	50.5±0.5	64.1±0.9	84.9±1.1
Latent Reward	36.8±0.3	50.0±0.2	65.9±0.2	84.9±1.3
Both (Ours)	**40.9±1.1**	**57.9±0.2**	**66.9±2.2**	**85.4±1.1**

5.4 Ablation Study

We conduct ablation studies to understand how each reward component contributes to both semantic fidelity and spatial consistency. Specifically, we progressively enable the baseline DiT objective, Pixel Reward, MST Reward, and their combination, with results reported in Table 6. In general, when both rewards are applied, best performance is observed across different IPC settings.

We further analyze robustness across training hyperparameters (Figure 3) and finetuning design choices (Figure 4). MGPO remains stable across all settings, yielding the best trade-off.

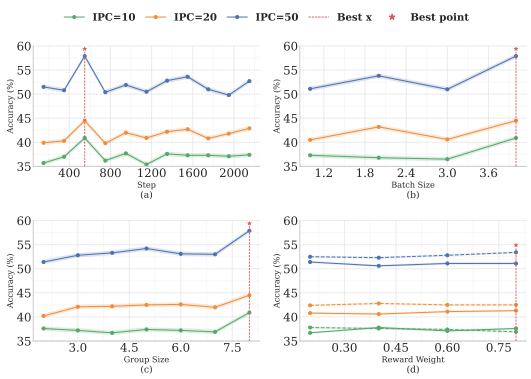

Figure 3: Hyperparameter analysis of MGPO.

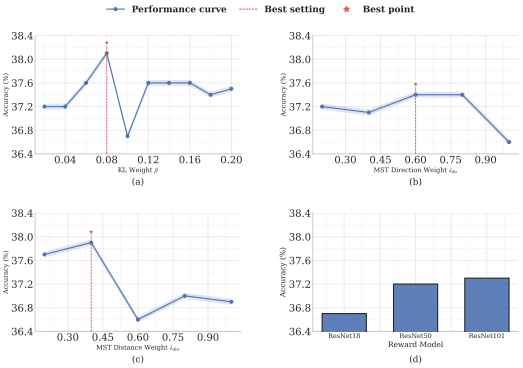

Figure 4: Ablation studies on key design choices of MGPO finetuning process.

6 Conclusion

In this paper, we proposed MGPO , a reward driven RL framework that resolves the fundamental objective mismatch in diffusion-based dataset distillation. By integrating pixel level discriminative and latent level geometric rewards, MGPO effectively aligns the generative process with downstream tasks. Furthermore, our plug-and-play reward design enables the framework to extend beyond standard classification to complex downstream tasks like object detection and semantic segmentation. Extensive experiments show that MGPO achieves state of the art performance across multiple classification benchmarks and successfully reaches state of the art on detection and segmentation tasks.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for the following tasks with required disclosure: *proof assistance*, namely helping to derive and write the proofs of the theoretical results in Appendix A.2 and Appendix A.3 (Lemmas 1–11, Theorems 5 and 10, Proposition 8, and the associated corollaries and remarks), and *drafting* parts of the manuscript, primarily the related-work discussion (Section 2 and Appendix B.1) and the implementation descriptions in Appendix C. We have not used generative AI tools for research ideation, hypothesis refinement, methodology design, implementation of the method, running experiments, data cleaning, or result interpretation; the remaining required disclosure tasks (e.g., translation and qualitative analysis) are not applicable to this work. Note that the diffusion models used to synthesize distilled datasets are the object of study and are not AI assistance in preparing this paper. Additionally, we used generative AI tools to aid and polish the writing (grammar, wording, and clarity) throughout the paper.

We have reviewed all AI-assisted work. Every AI-assisted proof was checked step by step by at least two authors, who verified all assumptions, constants, and cited results against the original sources (e.g., Berenfeld et al. (2022), Niyogi et al. (2008)). All AI-drafted or AI-polished text was revised by the authors, and every citation was checked manually against the referenced paper. All numerical results were produced by the authors' own code and were not generated or edited by AI tools. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, user studies, or the collection of new data. All experiments use publicly available benchmarks (ImageNet-1K and its subsets ImageWoof, ImageNette, and ImageIDC; PASCAL VOC 2012; MS-COCO), used under their respective licenses and terms of use; we do not redistribute the original images. Some of these datasets contain images of people collected from the web, and the synthesized distilled datasets may therefore reflect content and biases present in the source data. In addition, because `MGPO` optimizes the generator with frozen pretrained reward models (classifier, detector, or segmentation network), any biases in these reward models may be internalized and amplified in the distilled data; we discuss this risk and recommend auditing reward models before deployment in Appendix F. We do not foresee direct harmful applications beyond those common to generative models and dataset compression. The authors declare no conflicts of interest or sponsorship that could have influenced this work.

REPRODUCIBILITY STATEMENT

We have made the following efforts to ensure reproducibility. The full method is described in Section 4, including the DDIM sampling formulation (Section 3.2), the reward definitions (Section 4.1), and the complete training procedure (Algorithm 1). The MST construction is given in Algorithm 2, with all preprocessing steps and hyperparameters (e.g., k-NN size, PCA dimension, number of sampled latents per class) detailed in Appendix C.1; the latent geometric reward and the pixel-space rewards for classification, detection, and segmentation are specified in Appendices C.2 and C.3 (Algorithms 3 and 4). Experimental settings, including backbones, student architectures, training schedules, evaluation protocols, and the number of runs and random seeds, are given in Section 5.1, and hyperparameter sensitivity analyses are reported in Section 5.4 and Appendix D. For the theoretical results, all assumptions are stated explicitly and complete proofs are provided in Appendices A.1, A.2, and A.3. All datasets are publicly available. The source code will be released upon acceptance.

REFERENCES

Clément Berenfeld, John Harvey, Marc Hoffmann, and Krishnan Shankar. Estimating the reach of a manifold via its convexity defect function. *Discrete & Computational Geometry*, 67(2):403–438, 2022.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

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*, pp. 4750–4759, 2022.

Jeffrey A. Chan-Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. Mgd^3: Mode-guided dataset distillation using diffusion models, 2025. URL https://arxiv.org/abs/2505.18963.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

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*, pp. 15793–15803, 2024.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. URL https://arxiv.org/abs/1905.02244.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019a. URL https://github.com/fastai/imagenette.

Jeremy Howard. Imagewoof: a subset of 10 classes from imagenet that aren't so easy to classify, March 2019b. URL https://github.com/fastai/imagenette#imagewoof.

Rafał Karczewski, Markus Heinonen, and Vikas Garg. Diffusion models as cartoonists: The curious case of high density regions, 2025. URL https://arxiv.org/abs/2411.01293.

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, 2022. URL https://arxiv.org/abs/2205.14959.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2026. URL https://arxiv.org/abs/2507.21802.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models, 2025. URL https://arxiv.org/abs/2509.06040.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di ZHANG, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oCBKGw5HNf.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1–3):419–441, 2008.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Ding Qi, Jian Li, Junyao Gao, Shuguang Dou, Ying Tai, Jianlong Hu, Bo Zhao, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Towards universal dataset distillation via task-driven diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10557–10566, 2025. doi: 10.1109/CVPR52734.2025.00987.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. URL https://arxiv.org/abs/1506.01497.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Ayush Roy, Wei-Yang Alex Lee, Rudrasis Chakraborty, and Vishnu Suresh Lokhande. Manifoldgd: Training-free hierarchical manifold guidance for diffusion-based dataset distillation. *arXiv preprint arXiv:2602.23295*, 2026.

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. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach, 2018. URL https://arxiv.org/abs/1708.00489.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5809–5818, 2024.

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.

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Cao2: Rectifying inconsistencies in diffusion-based dataset distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4722–4731, October 2025.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1121–1128, 2009.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning, 2023. URL https://arxiv.org/abs/2304.06648.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36: 73582–73603, 2023.

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

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Lin Zhao, Yushu Wu, Xinru Jiang, Jianyang Gu, Yanzhi Wang, Xiaolin Xu, Pu Zhao, and Xue Lin. Taming diffusion for dataset distillation with high representativeness. *arXiv preprint arXiv:2505.18399*, 2025.

Yawen Zou, Guang Li, Duo Su, Zi Wang, Jun Yu, and Chao Zhang. Dataset distillation via vision-language category prototype. *arXiv preprint arXiv:2506.23580*, 2025.

APPENDIX

CONTENTS OF APPENDIX

A Theoretical Analysis **14**

A.1 Reward-Aligned Diffusion as Policy Optimization 14

A.2 Theoretical Justification of MST as a Manifold Skeleton 15

A.3 Subsample-MST Approximation and Scalability 18

B Related Work **22**

B.1 Optimization-Based Dataset Distillation 22

B.2 Diffusion Models for Dataset Distillation 22

B.3 Reinforcement Learning in Diffusion models 22

C Implementation Details **23**

C.1 MST Construction Details 23

C.2 Latent Geometric Reward Details 24

C.3 Pixel Classifier Reward Details 25

D Additional Experiments **25**

D.1 Experiment Results on Different Student Model 25

D.2 Hyperparameter Analysis 26

E Visualization **27**

E.1 Results on Classification 27

E.2 Results on Detection 27

E.3 Results on Segmentation 27

F Limitations and Broader Impacts **27**

A THEORETICAL ANALYSIS

A.1 REWARD-ALIGNED DIFFUSION AS POLICY OPTIMIZATION

Let $p_\theta(x_0)$ denote the data distribution induced by a the current diffusion model, and let $p_{\text{ref}}(x_0)$ denote the distribution of the frozen pretrained model Standard diffusion training maximizes likelihood and therefore prioritizes perceptual fidelity. However, in dataset distillation, our objective is not merely realism, but the generation of samples that are informative for downstream tasks.

To incorporate task objectives, we consider a reward-reweighted target distribution

$$p_\beta(x_0) \;\propto\; p_{\text{ref}}(x_0)\exp\big(\beta R(x_0)\big), \tag{11}$$

where $R(x_0)$ denotes a task-specific reward and $\beta > 0$ controls alignment strength. This formulation shifts probability mass toward task-relevant regions of the data manifold.

We view diffusion sampling process as a trajectory $\tau = \{x_T, x_{T-1}, \ldots, x_0\}$ induced by a parameterized policy $p_\theta(\tau)$.

Under this perspective, aligning diffusion generation with task objectives amounts to optimizing the expected terminal reward over trajectories, regularized towards the pretrained model:

$$\mathcal{J}_\beta(\theta) = \mathbb{E}_{\tau\sim p_\theta}\big[R(x_0)\big] \;-\; \tfrac{1}{\beta}\,\mathrm{KL}\big(p_\theta(x_0)\,\|\,p_{\mathrm{ref}}(x_0)\big). \tag{12}$$

Over all distributions of x_0, the maximizer of Eq. (12) is exactly the reward-reweighted density in Eq. (11), since $\mathcal{J}_\beta = \frac{1}{\beta}\log Z_\beta - \frac{1}{\beta}\mathrm{KL}(p_\theta \,\|\, p_\beta)$ with $Z_\beta = \mathbb{E}_{p_{\mathrm{ref}}}[\exp(\beta R)]$. Policy optimization over the diffusion parameters therefore approximates reward-aligned sampling within the model class, without modifying the inference procedure.

This policy view naturally enables the incorporation of heterogeneous reward signals. In our setting, the reward function will consist of (i) a discriminative term defined in pixel space and (ii) a geometric term defined in latent space, allowing alignment of both semantic and structural properties.

A.2 Theoretical Justification of MST as a Manifold Skeleton

Setup. Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact, connected, C^2-embedded submanifold with reach

$$\tau \coloneqq \mathrm{reach}(\mathcal{M}) > 0.$$

Let $U = \{u_1, \ldots, u_N\} \subset \mathcal{M}$ be a finite point set on the manifold, and define its *covering radius*

$$\delta \coloneqq \sup_{x\in\mathcal{M}} \min_{1\le i\le N} \|x - u_i\|_2. \tag{13}$$

Let $\mathcal{T}_{\mathrm{MST}} = (U, E_{\mathrm{MST}})$ denote the Euclidean minimum spanning tree on U, and write

$$|\mathcal{T}_{\mathrm{MST}}| \coloneqq \bigcup_{(u_i,u_j)\in E_{\mathrm{MST}}} [u_i, u_j] \;\subset\; \mathbb{R}^d$$

for its geometric realization, where $[u_i, u_j]$ denotes the closed line segment in $\mathbb{R}^d$ between u_i and u_j.

Scope. The analysis in this subsection and in Appendix A.3 is idealized: it concerns a Euclidean MST built on points sampled from a smooth class manifold, with the distance to the tree realization $|\mathcal{T}_{\mathrm{MST}}|$ (i.e., to the nearest edge segment). It motivates the geometric reward but does not by itself cover the implemented reward, which uses the nearest-node distance in the VAE/PCA representation, a kNN-restricted MST, and a direction term. proposition 8 below states what carries over to the implementation.

The goal of this subsection is to establish that, under the regime $\delta < \tau$, the MST realization $|\mathcal{T}_{\mathrm{MST}}|$ approximates the manifold $\mathcal{M}$ in the Hausdorff metric at rate δ in the forward direction and rate δ^2/τ in the backward direction. The argument proceeds via four lemmas: (*i*) a one-sided Hausdorff bound from the manifold to any spanning tree (lemma 1); (*ii*) the existence of a short-edged spanning tree on U (lemma 2); (*iii*) a bottleneck property of the MST (lemma 3); (*iv*) a chord-to-manifold distance bound based on convexity defect (lemma 4). These ingredients are then combined in theorem 5.

Lemma 1 (One-sided Hausdorff bound)**.** *For any spanning tree $\mathcal{T}$ on the vertex set U,*

$$\sup_{x\in\mathcal{M}} d(x, |\mathcal{T}|) \;\le\; \delta.$$

Proof. Fix $x \in \mathcal{M}$. By the definition of δ in eq. (13), there exists $u_i \in U$ such that

$$\|x - u_i\|_2 \le \delta.$$

Since u_i is a vertex of $\mathcal{T}$, we have $u_i \in |\mathcal{T}|$. Therefore

$$d(x, |\mathcal{T}|) \;\le\; \|x - u_i\|_2 \;\le\; \delta.$$

Taking the supremum over $x \in \mathcal{M}$ proves the claim. □

Lemma 2 (Existence of a short-edged spanning tree)**.** *There exists a spanning tree on U whose every edge has Euclidean length at most 2δ.*

Proof. For each $i \in \{1, \dots, N\}$, define the closed manifold cap

$$A_i \coloneqq \overline{B}(u_i, \delta) \cap \mathcal{M}.$$

By the definition of δ, every $x \in \mathcal{M}$ lies within distance δ of some u_i, hence

$$\mathcal{M} \;=\; \bigcup_{i=1}^{N} A_i.$$

Claim: the intersection graph of $\{A_i\}_{i=1}^N$ is connected. Suppose, for contradiction, that the intersection graph is disconnected, so there is a partition

$$\{1, \dots, N\} = I \sqcup J, \qquad I, J \neq \varnothing,$$

satisfying

$$A_i \cap A_j = \varnothing \qquad \text{for all } i \in I,\ j \in J.$$

Define

$$X \coloneqq \bigcup_{i \in I} A_i, \qquad Y \coloneqq \bigcup_{j \in J} A_j.$$

Each A_i is closed in $\mathcal{M}$ (as the intersection of a closed Euclidean ball with $\mathcal{M}$), so X and Y are closed in $\mathcal{M}$. They are nonempty, disjoint, and satisfy $\mathcal{M} = X \cup Y$. In particular, $X = \mathcal{M} \setminus Y$ is also open in $\mathcal{M}$, so X is a nonempty proper clopen subset of $\mathcal{M}$. This contradicts the connectedness of $\mathcal{M}$, establishing the claim.

Now construct a graph G on vertex set U by joining u_i and u_j whenever $A_i \cap A_j \neq \varnothing$. By the claim, G is connected. For any such edge (u_i, u_j), choose $x \in A_i \cap A_j$; the triangle inequality yields

$$\|u_i - u_j\|_2 \;\le\; \|u_i - x\|_2 + \|x - u_j\|_2 \;\le\; 2\delta.$$

Thus every edge of G has length at most 2δ, and any spanning tree of G inherits this property. □

Lemma 3 (Bottleneck bound on MST edge length)**.** *Let*

$$\ell_{\mathrm{MST}} \coloneqq \max_{(u_i, u_j) \in E_{\mathrm{MST}}} \|u_i - u_j\|_2.$$

Then $\ell_{\mathrm{MST}} \le 2\delta$.

Proof. By Lemma 2, there exists a spanning tree on U whose maximum edge length is at most 2δ. Every Euclidean minimum spanning tree is also a minimum bottleneck spanning tree, i.e., it minimizes the maximum edge length among all spanning trees. Hence

$$\ell_{\mathrm{MST}} \;\le\; 2\delta.$$ □

Lemma 4 (Chord distance to manifold via convexity defect)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ have reach $\tau > 0$. Suppose $a, b \in \mathcal{M}$ satisfy*

$$\|a - b\|_2 = \ell < 2\tau.$$

Then for every $y \in [a, b]$,

$$d(y, \mathcal{M}) \;\le\; \tau - \sqrt{\tau^2 - \tfrac{\ell^2}{4}} \;\le\; \frac{\ell^2}{4\tau}. \tag{14}$$

Proof. Let $t \coloneqq \ell/2 < \tau$. The smallest closed ball in $\mathbb{R}^d$ containing $\{a, b\}$ has radius t, namely the ball centered at the midpoint of $[a, b]$. Since

$$[a, b] = \mathrm{conv}\{a, b\},$$

every $y \in [a, b]$ belongs to the t-local convex hull of $\mathcal{M}$ in the sense of (Berenfeld et al., 2022, Definition 4.1). Therefore

$$d(y, \mathcal{M}) \;\le\; h_{\mathcal{M}}(t),$$

where $h_{\mathcal{M}}$ denotes the convexity defect function of $\mathcal{M}$.

For a set of reach τ, the quarter-circle bound of (Berenfeld et al., 2022, Definition 4.1 and the bound stated immediately thereafter) states

$$h_{\mathcal{M}}(t) \;\le\; \tau - \sqrt{\tau^2 - t^2}, \qquad 0 \le t < \tau.$$

Substituting $t = \ell/2$ yields the first inequality of eq. (14).

For the second inequality, observe that for any $0 \le s \le \tau^2$,

$$\tau - \sqrt{\tau^2 - s} \;=\; \frac{s}{\tau + \sqrt{\tau^2 - s}} \;\le\; \frac{s}{\tau},$$

where the equality follows from rationalizing the numerator and the inequality from $\sqrt{\tau^2 - s} \ge 0$. Setting $s = \ell^2/4$ gives

$$\tau - \sqrt{\tau^2 - \tfrac{\ell^2}{4}} \;\le\; \frac{\ell^2}{4\tau},$$

which completes the proof. □

We are now ready to state and prove the main result of this subsection.

Theorem 5 (MST–manifold Hausdorff bound)**.** *Suppose $\delta < \tau$. Then*

$$d_H\big(\mathcal{M},\, |\mathcal{T}_{\mathrm{MST}}|\big) \;\le\; \delta. \tag{15}$$

More precisely, the two one-sided bounds

$$\sup_{x \in \mathcal{M}} d\big(x,\, |\mathcal{T}_{\mathrm{MST}}|\big) \;\le\; \delta, \tag{16}$$

$$\sup_{y \in |\mathcal{T}_{\mathrm{MST}}|} d\big(y,\, \mathcal{M}\big) \;\le\; \tau - \sqrt{\tau^2 - \delta^2} \;\le\; \frac{\delta^2}{\tau} \tag{17}$$

hold simultaneously.

Proof. The forward bound eq. (16) is exactly lemma 1.

For the backward bound eq. (17), by lemma 3,

$$\ell_{\mathrm{MST}} \;\le\; 2\delta \;<\; 2\tau,$$

so lemma 4 applies to every edge $[u_i, u_j] \in E_{\mathrm{MST}}$. Each point y on such an edge therefore satisfies

$$d(y, \mathcal{M}) \;\le\; \tau - \sqrt{\tau^2 - \tfrac{\|u_i - u_j\|_2^2}{4}}.$$

Since the function $\ell \mapsto \tau - \sqrt{\tau^2 - \ell^2/4}$ is monotone non-decreasing on $[0, 2\tau)$, and $\|u_i - u_j\|_2 \le \ell_{\mathrm{MST}} \le 2\delta$, we have

$$d(y, \mathcal{M}) \;\le\; \tau - \sqrt{\tau^2 - \tfrac{\ell_{\mathrm{MST}}^2}{4}} \;\le\; \tau - \sqrt{\tau^2 - \delta^2} \;\le\; \frac{\delta^2}{\tau},$$

where the last inequality follows from lemma 4 applied with $s = \delta^2$. Taking the supremum over $y \in |\mathcal{T}_{\mathrm{MST}}|$ yields eq. (17).

To deduce the combined Hausdorff bound eq. (15), note that

$$d_H\big(\mathcal{M},\, |\mathcal{T}_{\mathrm{MST}}|\big) \;=\; \max\left\{ \sup_{x \in \mathcal{M}} d\big(x, |\mathcal{T}_{\mathrm{MST}}|\big),\; \sup_{y \in |\mathcal{T}_{\mathrm{MST}}|} d(y, \mathcal{M}) \right\}.$$

By eqs. (16) and (17), both terms inside the maximum are at most

$$\max\Big\{ \delta,\; \tau - \sqrt{\tau^2 - \delta^2} \Big\}.$$

It remains to show $\tau - \sqrt{\tau^2 - \delta^2} \le \delta$ whenever $\delta < \tau$. Equivalently, $\tau - \delta \le \sqrt{\tau^2 - \delta^2}$, which after squaring becomes

$$(\tau - \delta)^2 \;\le\; \tau^2 - \delta^2,$$

i.e., $-2\tau\delta + 2\delta^2 \le 0$, i.e., $\delta \le \tau$, which holds by assumption. Hence

$$d_H\big(\mathcal{M},\, |\mathcal{T}_{\mathrm{MST}}|\big) \;\le\; \delta.$$

□

Remark 6. The bounds in theorem 5 are *asymmetric*: the forward direction (manifold to MST) decays linearly in δ, while the backward direction (MST to manifold) decays quadratically as δ^2/τ. The asymmetry arises because the MST realization includes line segments that may bow off the manifold, but the bowing distance is controlled by the manifold's reach via lemma 4. This quadratic backward rate is what enables lemma 3 to feed into a tight bound: a generic spanning tree on U would only give a linear $\mathcal{O}(\delta)$ bound in both directions.

Remark 7 (Geometric approximation, not topology preservation). Theorem 5 is a statement about Hausdorff closeness only. The MST is a one-dimensional acyclic graph, so $|\mathcal{T}_{\mathrm{MST}}|$ has trivial first homology and cannot, in general, preserve cycles, higher-dimensional homology, or the intrinsic dimension of $\mathcal{M}$. For example, for points sampled on a circle the MST is a path, which breaks the cycle. We therefore use the MST as a sparse geometric skeleton that encourages coverage of the class support, not as a topology-preserving representation.

What carries over to the implemented reward. The proofs of lemmas 1 to 3 use only that $\mathcal{M}$ is compact and connected; smoothness and reach enter only through lemma 4, i.e., the backward bound eq. (17). This allows us to state a bound for the nearest-node distance used in practice, directly in the deployed representation.

Proposition 8 (Nearest-node distance in the deployed representation)**.** *Let $K \subset \mathbb{R}^D$ be compact and connected, let $U = \{u_1, \dots, u_N\} \subset K$ have covering radius $\delta_K := \sup_{x\in K} \min_i \|x - u_i\|_2$, and let $\mathcal{T}$ be the Euclidean MST on U. Define the node distance $D_{\mathrm{node}}(y) := \min_i \|y - u_i\|_2$ and the edge distance $D_{\mathrm{edge}}(y) := d(y, |\mathcal{T}|)$. Then:*

(i) $\sup_{x\in K} D_{\mathrm{node}}(x) \le \delta_K$*;*

(ii) for every $y \in \mathbb{R}^D$, $D_{\mathrm{edge}}(y) \le D_{\mathrm{node}}(y) \le D_{\mathrm{edge}}(y) + \delta_K$*;*

(iii) if $V \subseteq U$*, then for every* $y \in \mathbb{R}^D$, $0 \le D^V_{\mathrm{node}}(y) - D^U_{\mathrm{node}}(y) \le \max_{u\in U} \min_{v\in V} \|u - v\|_2 \le \delta_K(V)$*, where $\delta_K(V)$ is the covering radius of V over K.*

In particular, with $K = g_c(\mathcal{X}_c)$ the image of a compact, connected class support $\mathcal{X}_c$ under the continuous encoder-and-whitening map g_c (Appendix C.1), (i)–(iii) hold in the VAE/PCA space where the reward is computed, without requiring g_c to be isometric or invertible.

Proof. (i) is the definition of δ_K. For (ii), the left inequality holds because every node lies on $|\mathcal{T}|$. For the right inequality, let $p \in [u_i, u_j]$ be a point of $|\mathcal{T}|$ nearest to y. The nearer endpoint of this edge is within $\|u_i - u_j\|_2/2 \le \ell_{\mathrm{MST}}/2$ of p, and the arguments of lemmas 2 and 3, which use only compactness and connectedness of K, give $\ell_{\mathrm{MST}} \le 2\delta_K$. Hence $D_{\mathrm{node}}(y) \le \|y - p\|_2 + \delta_K = D_{\mathrm{edge}}(y) + \delta_K$. For (iii), the left inequality holds since $V \subseteq U$. For the right one, let $u^* \in U$ be nearest to y and $v^* \in V$ nearest to u^*; then $D^V_{\mathrm{node}}(y) \le \|y - u^*\|_2 + \|u^* - v^*\|_2 \le D^U_{\mathrm{node}}(y) + \max_{u\in U} \min_{v\in V} \|u - v\|_2$, and the last quantity is at most $\delta_K(V)$ because $U \subset K$. Finally, the continuous image of a compact connected set is compact and connected. □

Remark 9 (What remains empirical). Proposition 8 does not assume smoothness or positive reach of K, so the backward bound eq. (17) is not claimed in the deployed space. It also does not cover the direction term, which is evaluated empirically, or the kNN restriction used to build the MST: if the symmetrized kNN graph contains a Euclidean MST of U, then its MST is a Euclidean MST of U, and otherwise the two trees can differ. Finally, the relation between distances in $g_c(\mathcal{X}_c)$ and downstream utility is not established theoretically and is assessed through the experiments.

A.3 Subsample-MST Approximation and Scalability

Setup. We adopt the notation of section A.2. Let $\mathcal{M}_c \subset \mathbb{R}^d$ denote the class-c data manifold, assumed to be a compact, connected, C^2-embedded submanifold with reach $\tau_c := \mathrm{reach}(\mathcal{M}_c) > 0$ and intrinsic dimension m_c. Let μ_c be a probability measure on $\mathcal{M}_c$ with density (with respect to the volume form) bounded below by $\rho_{\min} > 0$.

For any finite point set $W \subset \mathcal{M}_c$, define its covering radius

$$\delta(W) := \sup_{x\in\mathcal{M}_c} \min_{w\in W} \|x - w\|_2,$$

and let $\mathcal{T}(W) = (W, E_{\mathrm{MST}}(W))$ be the Euclidean minimum spanning tree on W, with geometric realization

$$|\mathcal{T}(W)| := \bigcup_{(u,v)\in E_{\mathrm{MST}}(W)} [u, v] \subset \mathbb{R}^d.$$

The Hausdorff distance between two compact sets $A, B \subset \mathbb{R}^d$ is

$$d_H(A, B) := \max\left\{ \sup_{a\in A} \inf_{b\in B} \|a - b\|_2,\ \sup_{b\in B} \inf_{a\in A} \|a - b\|_2 \right\}.$$

Let $U_c = \{u_1, \ldots, u_{N_c}\} \subset \mathcal{M}_c$ denote the full real sample set for class c, drawn i.i.d. from μ_c, and let $V_c \subseteq U_c$ with $|V_c| = n$ be a uniform random subset. By the exchangeability of i.i.d. samples, V_c is itself distributed as an i.i.d. sample of size n from μ_c, a characterization we adopt in the analysis below.

The aim of this subsection is to show that, with high probability, the subsample MST $|\mathcal{T}(V_c)|$ approximates the full MST $|\mathcal{T}(U_c)|$ in Hausdorff distance at a rate determined only by the intrinsic dimension m_c, justifying the practical use of small random subsamples in the geometric reward computation (Algorithm 1 of the main paper).

Theorem 10 (Subsample-MST approximation)**.** *Let $\mathcal{M}_c$, μ_c, U_c, and V_c be as above. Then there exist constants $C_{\mathcal{M}_c} > 0$ (depending only on $\mathcal{M}_c$ and $\rho_{\min}$) and $n_0 = n_0(\mathcal{M}_c, \tau_c)$ such that, for every $n \geq n_0$ and every $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\delta(V_c) \ \leq\ C_{\mathcal{M}_c} \left(\frac{\log(n/\eta)}{n} \right)^{1/m_c} \ <\ \tau_c, \tag{18}$$

and consequently

$$d_H\big(|\mathcal{T}(V_c)|,\ |\mathcal{T}(U_c)|\big) \ \leq\ 2\, C_{\mathcal{M}_c} \left(\frac{\log(n/\eta)}{n} \right)^{1/m_c}. \tag{19}$$

Moreover, the geometric reward r_{geo} defined in Eq. (4) of the main paper satisfies

$$\left| r_{\mathrm{geo}}^{V_c}(\tilde{x}_0, c) - r_{\mathrm{geo}}^{U_c}(\tilde{x}_0, c) \right| \ \leq\ 2\, C_{\mathcal{M}_c} \left(\frac{\log(n/\eta)}{n} \right)^{1/m_c} + \lambda_{\mathrm{dir}}, \tag{20}$$

for every generated latent $\tilde{x}_0$, where r_{geo}^{W} denotes the geometric reward computed using the MST on point set W. The first term captures the contribution of the distance component of r_{geo} and vanishes as $n \to \infty$; the second term is a worst-case bound on the direction component (see remark 12).

The proof proceeds in three steps: (*i*) a probabilistic bound on the covering radius of the random subsample (lemma 11); (*ii*) a deterministic Hausdorff bound between the two MSTs via the triangle inequality, leveraging theorem 5; (*iii*) a Lipschitz argument transferring the geometric bound to the distance component of the geometric reward, combined with a worst-case bound on the direction component.

Step 1: Covering radius of a random subsample. We first invoke a classical result on the convergence of empirical covering radius on a manifold.

Lemma 11 (Covering radius rate, (Niyogi et al., 2008; Penrose, 2003))**.** *Under the assumptions of theorem 10, there exists a constant $C_{\mathcal{M}_c} > 0$ depending only on $\mathcal{M}_c$ and $\rho_{\min}$ such that, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\delta(V_c) \ \leq\ C_{\mathcal{M}_c} \left(\frac{\log(n/\eta)}{n} \right)^{1/m_c}. \tag{21}$$

Proof. Cover $\mathcal{M}_c$ by $\mathcal{N}(\rho)$ geodesic balls of radius ρ, where

$$\mathcal{N}(\rho) \leq c_1\, \rho^{-m_c}\, \mathrm{vol}(\mathcal{M}_c)$$

for some constant $c_1 = c_1(\mathcal{M}_c)$, by standard volume comparison on bounded-reach submanifolds. For $\rho < \tau_c/2$, each such ball has μ_c-mass at least $\rho_{\min}\, c_2\, \rho^{m_c}$ by the lower density assumption combined with the small-ball volume bound on $\mathcal{M}_c$.

Setting

$$\rho := \left(\frac{\log(n/\eta)}{c_3\, n} \right)^{1/m_c}$$

for an appropriate constant c_3, and applying a union bound over the cover, the probability that some ball contains no point of V_c is at most

$$\mathcal{N}(\rho) \left(1 - \rho_{\min}\, c_2\, \rho^{m_c}\right)^n \;\le\; \eta.$$

On this event, every $x \in \mathcal{M}_c$ lies within ρ of some $v_i \in V_c$, yielding eq. (21) with $C_{\mathcal{M}_c} := c_3^{-1/m_c}$. □

Choosing n_0 large enough that the right-hand side of eq. (21) is strictly less than τ_c, we obtain eq. (18) for all $n \ge n_0$.

Step 2: Hausdorff distance between the two MST realizations. Apply theorem 5 to the point set $W = U_c$ and to $W = V_c$. Since $V_c \subseteq U_c$, the covering radius is monotone non-increasing under enlargement of the point set, i.e., $\delta(U_c) \le \delta(V_c)$. Hence, on the event of lemma 11,

$$\delta(U_c) \;\le\; \delta(V_c) \;<\; \tau_c,$$

so theorem 5 applies to both sets and gives

$$d_H\big(\mathcal{M}_c,\, |\mathcal{T}(U_c)|\big) \le \delta(U_c), \qquad d_H\big(\mathcal{M}_c,\, |\mathcal{T}(V_c)|\big) \le \delta(V_c).$$

The triangle inequality for Hausdorff distance yields

$$\begin{aligned} d_H\big(|\mathcal{T}(V_c)|,\, |\mathcal{T}(U_c)|\big) \;&\le\; d_H\big(|\mathcal{T}(V_c)|,\, \mathcal{M}_c\big) + d_H\big(\mathcal{M}_c,\, |\mathcal{T}(U_c)|\big) \\ &\le\; \delta(V_c) + \delta(U_c) \\ &\le\; 2\,\delta(V_c). \end{aligned}$$

Combined with eq. (18), this establishes eq. (19).

Step 3: Stability of the geometric reward. Fix a generated latent $\tilde{x}_0$ with class label c, and let $u_{\text{gen}} = g_c(z_0)$ denote its projected representation as in Sec. 4.1 of the main paper. The geometric reward (Eq. (4) of the main paper) admits the decomposition

$$r^W_{\text{geo}}(\tilde{x}_0, c) = -\, D_W \;-\; \lambda_{\text{dir}}\, A_W,$$

where

$$D_W := \min_{e \in E_{\text{MST}}(W)} \text{dist}(u_{\text{gen}}, e), \qquad A_W := 1 - |\cos(\mathbf{v}_{\text{gen}}, \mathbf{v}^*_{e_W})|,$$

and $e^*_W \in E_{\text{MST}}(W)$ is the MST edge nearest to u_{gen} in W.

Distance term. Since $|\mathcal{T}(V_c)|$ and $|\mathcal{T}(U_c)|$ are compact subsets of $\mathbb{R}^d$, and the map $w \mapsto \text{dist}(u_{\text{gen}}, w)$ is 1-Lipschitz,

$$\big|D_{V_c} - D_{U_c}\big| \;=\; \Big| \inf_{w \in |\mathcal{T}(V_c)|} \|u_{\text{gen}} - w\| - \inf_{w \in |\mathcal{T}(U_c)|} \|u_{\text{gen}} - w\| \Big| \;\le\; d_H\big(|\mathcal{T}(V_c)|,\, |\mathcal{T}(U_c)|\big). \quad (22)$$

Direction term. Since $|\cos(\cdot,\cdot)| \in [0,1]$, we have the worst-case bound

$$\big|A_{V_c} - A_{U_c}\big| \;\le\; 1. \quad (23)$$

This bound cannot in general be improved through Hausdorff-distance arguments alone: for class manifolds of intrinsic dimension $m_c \ge 2$, two MSTs whose geometric realizations are close as point sets may nonetheless contain edges whose directions span different subspaces of the local tangent plane, so the nearest-edge direction $\mathbf{v}^*_{e_W}$ is not a Hausdorff-continuous functional of $|\mathcal{T}(W)|$ (see remark 12).

Combining the two,

$$\big|r^{V_c}_{\text{geo}} - r^{U_c}_{\text{geo}}\big| \;\le\; |D_{V_c} - D_{U_c}| + \lambda_{\text{dir}}\, |A_{V_c} - A_{U_c}| \;\le\; d_H\big(|\mathcal{T}(V_c)|,\, |\mathcal{T}(U_c)|\big) + \lambda_{\text{dir}}.$$

Substituting eq. (19) yields eq. (20), which completes the proof of theorem 10. □

Implemented distance term. For the nearest-node distance used in practice, proposition 8(iii) gives, deterministically and in the deployed representation,

$$0 \;\le\; D^{V_c}_{\mathrm{node}}(u_{\mathrm{gen}}) - D^{U_c}_{\mathrm{node}}(u_{\mathrm{gen}}) \;\le\; \delta(V_c),$$

so the distance component of the implemented reward changes by at most $\lambda_{\mathrm{dist}}\,\delta(V_c)$ under subsampling. Where the assumptions of lemma 11 hold, $\delta(V_c) \le C_{\mathcal{M}_c}(\log(n/\eta)/n)^{1/m_c}$ with probability at least $1-\eta$. This bound does not depend on the MST edges, so it is unaffected by the kNN restriction; the direction term remains uncontrolled, as discussed next.

Remark 12 (On the direction component). The bound eq. (20) retains a constant term λ_{dir} that does not vanish with $\delta(V_c) \to 0$. This reflects an intrinsic limitation: while the distance component D_W depends only on $|\mathcal{T}(W)|$ as a point set and is therefore controlled by Hausdorff distance, the direction component A_W depends on the orientation of the nearest MST edge, which is not preserved under Hausdorff convergence on manifolds of intrinsic dimension ≥ 2. In our experiments we set $\lambda_{\mathrm{dir}} = 0.6$ (chosen via ablation; cf. Fig. 3(b) of the main paper), so the absolute size of the direction contribution remains small relative to the dynamic range of the reward, preserving training stability. A finer analysis of the direction component, e.g. in expectation over u_{gen}, is left to future work.

The stability of the geometric reward guarantees that the excess-risk bound of the main paper carries over with δ replaced by $\delta(V_c)$, up to the additive direction-component constant.

Corollary 13 (Excess risk under subsampled MST)**.** *Under the assumptions of theorem 10 together with those of the main-paper excess-risk bound, the classifier $\hat{h}_S$ trained on the MGPO-distilled set generated using the subsample-MST satisfies, with probability at least $1-\eta-\delta_0$,*

$$R_\mu(\hat{h}_S) - \min_{h^*\in\mathcal{H}} R_\mu(h^*) \;\le\; 2\,L_\ell\left(\varepsilon + \frac{\delta(V_c)^2}{\tau_c} + \lambda_{\mathrm{dir}}\right) + 4\,\mathfrak{R}_{N_S}(\ell\circ\mathcal{H}) + 2\,L_\ell\,C_{\mathcal{M}}\,N_T^{-1/m} + \mathcal{O}\left(B\sqrt{\tfrac{\log(1/\delta_0)}{N_S}}\right).$$

where $\delta(V_c) \le C_{\mathcal{M}_c}\big(\log(n/\eta)/n\big)^{1/m_c}$.

Proof. Substitute the reward stability bound from eq. (20) into the excess-risk decomposition: the δ-dependent term is replaced by $\delta(V_c)$, and the direction-component constant λ_{dir} contributes additively through the Lipschitz constant L_ℓ. The result follows from a union bound over the high-probability events of theorem 10 and the main-paper excess-risk bound. □

Corollary 14 (Sample complexity for fixed precision, independent of N_c)**.** *Fix any target covering radius $\delta_{\mathrm{target}} \in (0, \tau_c)$ and any $\eta \in (0,1)$. Then with probability at least $1-\eta$,*

$$n \;=\; \Theta\left(\frac{\log(1/\eta) + m_c\log(1/\delta_{\mathrm{target}})}{\delta_{\mathrm{target}}^{m_c}}\right)$$

*random samples suffice to ensure $\delta(V_c) \le \delta_{\mathrm{target}}$. In particular, the required n depends only on the manifold parameters $(\tau_c, m_c, \rho_{\min})$ and the target precision δ_{target}, and is **independent of the full dataset size** N_c.*

Proof. Set $\delta(V_c) = \delta_{\mathrm{target}}$ in eq. (21) and solve for n:

$$\delta_{\mathrm{target}} = C_{\mathcal{M}_c}\left(\tfrac{\log(n/\eta)}{n}\right)^{1/m_c} \;\Longrightarrow\; \tfrac{n}{\log(n/\eta)} = \Theta\big(\delta_{\mathrm{target}}^{-m_c}\big).$$

Inverting gives the stated scaling. None of the constants involved depend on N_c. □

Remark 15 (Practical implications for ImageNet-scale data). The empirical intrinsic dimension of natural image manifolds, estimated by maximum-likelihood methods, lies in the range $m_c \in [26, 43]$ for ImageNet-scale data (Pope et al., 2021), with class-conditional manifolds typically exhibiting smaller intrinsic dimension due to reduced intra-class variation. By corollary 14, the subsample size n required to reach a fixed geometric precision δ_{target} is independent of N_c, so a single choice of n generalizes across datasets of different scale. In our implementation, taking $n = 200$ random

samples per class gives a ratio of the two covering-radius rates of $(N_c/200)^{1/m_c}$ (ignoring logarithmic factors), which evaluates to ≈ 1.07–1.13 on ImageNet-1K ($N_c \approx 1300$, $m_c \in [26, 43]$). This is a heuristic comparison of upper-bound rates rather than a bound on the realized covering radii, and the intrinsic-dimension estimates refer to image space rather than the VAE/PCA space. It is consistent with, but does not by itself establish, the empirical scalability of MGPO to ImageNet-1K (Table 3).

B Related Work

B.1 Optimization-Based Dataset Distillation

Dataset distillation was originally introduced as a meta-learning problem that directly optimizes a small set of synthetic examples such that models trained on them achieve comparable performance to training on the full dataset (Wang et al., 2018). A dominant line of follow-up work formulates distillation as bi-level optimization, where the outer loop updates synthetic data and the inner loop trains a network on these synthetic samples(Zhao et al., 2020). Subsequent studies reduce accumulative error in optimization process by matching training trajectory. Later, SRe^{2}L (Yin et al., 2023) series of work further unleashed the ability of decoupling the distillation process by squeeze and recover approach. Besides, an alternative line of works matches higher-level statistics such as distribution or feature in representation space to avoid expensive unrolling, improving efficiency while sometimes sacrificing fidelity of training dynamics. Nevertheless, these optimization-based methods often suffer from heavy compute or memory overhead due to unrolling, sensitivity to architecture choices, and difficulties in scaling to higher resolutions or large numbers of classes.

B.2 Diffusion Models for Dataset Distillation

Recent advances have explored diffusion models as powerful generative backbones for dataset distillation, enabling the synthesis of compact yet highly representative datasets. As a pioneering work, Minimax Diffusion (Gu et al., 2024) leverages pre-trained diffusion models and further fine-tunes them under a minimax criteria to jointly enhance sample representativeness and diversity. Building upon this idea, D^{4}M (Su et al., 2024) incorporates pre-trained text-to-image diffusion models with prototype learning, clustering latent features to learn category prototypes and thereby improving intra-class consistency and visual fidelity. To mitigate distribution mismatch and stochasticity in latent sampling, D3HR (Zhao et al., 2025) proposes a domain mapping framework that applies DDIM inversion (Song et al., 2020) to project VAE latents into a Gaussian space with improved normality, combined with a group-based sampling strategy to better align the generated data distribution with the real one. More recently, CaO2 (Wang et al., 2025) rectifies feature and class level inconsistencies by inference time latent optimization. DDVLCP (Zou et al., 2025) extends diffusion-based distillation to the vision–language setting by constructing paired image–text prototypes using large language models, which effectively enhances semantic coherence and reduces artifacts such as missing or illogical objects. Despite their success, diffusion-based approaches are inherently likelihood-centric, concentrating samples in high-density regions, which preserves perceptual fidelity but neglects boundary and hard examples essential for discriminative training.

B.3 Reinforcement Learning in Diffusion models

Reinforcement learning has recently been adopted to align diffusion and flow based generative models with task level objectives that are difficult to encode through likelihood based training alone. A representative line of work builds upon Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which updates generative policies by comparing relative rewards within groups of sampled trajectories, enabling stable optimization without explicit value function estimation.

Following the introduction of GRPO, several variants have been proposed to adapt it to different generative settings. Flow-GRPO (Liu et al., 2025) extends group relative optimization to diffusion and flow matching models by introducing controlled stochasticity into the sampling process. Temporal variants such as Temp-Flow-GRPO (He et al., 2025) further incorporate time dependent credit assignment to emphasize critical denoising stages. Mix-GRPO (Li et al., 2026) combines multiple sampling dynamics within a unified optimization framework to balance exploration and efficiency. More recently, structured rollout strategies such as Branch-GRPO (Li et al., 2025) and Dance-

Algorithm 2 Per-class MST Construction in Latent Space

Require: Real dataset $\mathcal{T}$, VAE encoder $E(\cdot)$, max samples per class N_c, PCA dimension d, kNN size k
Ensure: Class-wise MST skeleton $\{\mathcal{T}_c\}_c$ and PCA statistics (μ_c, W_c)
1: **for** each class c **do**
2: Sample up to N_c real images $\{x_i\}$ from class c
3: Encode latents $z_i = E(x_i)$ and flatten to vectors
4: Stack all latent vectors into matrix Z
5: Compute centered features $Z_c = Z - \mu_c$
6: Perform PCA projection to obtain normalized embeddings $u_i = W_c(z_i - \mu_c)$
7: Construct a k-nearest-neighbor graph over $\{u_i\}$
8: Compute the minimum spanning tree $\mathcal{T}_c$ from the sparse adjacency graph
9: **end for**

GRPO (Xue et al., 2025) have been proposed to improve scalability and stability through branching trajectories and group wise reward aggregation.

Despite these advances, existing GRPO-based methods primarily focus on improving generative quality or preference alignment. In contrast, our work leverages reinforcement learning to explicitly incorporate downstream discriminative objectives into the diffusion generation process, addressing the objective mismatch between likelihood-driven diffusion models and dataset distillation for classification.

C Implementation Details

C.1 MST Construction Details

The class-wise MST skeleton is pre-computed once before `MGPO` fine-tuning. For each class c, we sample up to N_c real images from the training set. All images are resized to 256×256, center-cropped, converted to tensors, and normalized to the range $[-1, 1]$. Each image is then encoded by a frozen Stable Diffusion VAE encoder. We use the mean of the encoder posterior as the latent representation and divide it by the standard latent scaling factor 0.18215. The resulting latent tensor is flattened into a vector. Stacking all vectors from class c produces a latent matrix

$$Z \in \mathbb{R}^{N \times D},$$

where N is the number of collected samples and D is the flattened latent dimension.

Class-wise PCA normalization. To obtain a compact and numerically stable embedding space, we perform PCA on the centered latent matrix. Let

$$\mu_c = \frac{1}{N} \sum_{i=1}^{N} z_i$$

be the class mean and define the centered matrix

$$Z_c = Z - \mu_c.$$

We compute a low-rank PCA decomposition of Z_c and retain the top d principal components. Let $V \in \mathbb{R}^{D \times d}$ denote the projection matrix and $S \in \mathbb{R}^d$ the singular values. To avoid numerical instability, singular values are lower-bounded by a small constant ε. Each latent vector z is then projected into a normalized embedding space by

$$u = g_c(z) = (z - \mu_c) V \operatorname{diag}(S)^{-1}.$$

This transformation simultaneously performs dimensionality reduction and whitening, making distances more comparable across directions. In our implementation, the PCA dimension is set to $d = 128$.

Sparse kNN graph construction. After obtaining the normalized embeddings $\{u_i\}_{i=1}^N$, we construct a sparse k-nearest-neighbor graph in the embedding space. For each node u_i, we identify its k nearest neighbors according to Euclidean distance. An adjacency matrix A is then constructed where

$$A_{ij} = \begin{cases} \|u_i - u_j\|_2, & \text{if } u_j \text{ is among the } k \text{ nearest neighbors of } u_i, \\ +\infty, & \text{otherwise.} \end{cases}$$

The graph is symmetrized by assigning the same edge weight to both A_{ij} and A_{ji} once either direction is observed. This sparse construction significantly reduces graph density and improves scalability. In practice, we use $k = 10$.

Minimum spanning tree extraction. Given the sparse adjacency matrix, we compute the minimum spanning tree using a standard graph solver. The resulting tree is represented as an edge set

$$\mathcal{E}_c = \{(i, j)\},$$

which connects all nodes while minimizing the total edge weight. Ideally, a spanning tree over N nodes contains $N - 1$ edges. The extracted tree therefore provides a sparse skeleton of the class manifold that preserves global connectivity while avoiding redundant edges.

Saved statistics. For each class, we store the PCA mean μ_c, the projection matrix V, the singular values S, the normalized node embeddings, and the MST edge list. These quantities define both the class-wise tree $\mathcal{T}_c$ and the projection function $g_c(\cdot)$ used during `MGPO` training. When a generated latent z_0 is obtained at the end of the diffusion trajectory, it is projected using the same transformation $g_c(\cdot)$ and evaluated according to its geometric relation to the nearest edge of the corresponding MST skeleton.

C.2 Latent Geometric Reward Details

We provide additional implementation details for the latent geometric reward r_{geo} used in `MGPO`. The reward is evaluated on a two-stage diffusion trajectory that combines a stochastic SDE transition with a deterministic ODE refinement. This design enables both stable policy optimization and geometry-aware trajectory supervision.

Latent projection. To measure geometric consistency with the class manifold, both z_{t_m} and z_0 are projected into the class-specific PCA space used for MST construction.

Let $g_c(\cdot)$ denote the projection function defined in Section C.1. For a flattened latent vector $\text{vec}(z)$, the projection is

$$u = g_c(\text{vec}(z)).$$

Thus we obtain

$$u_{t_m} = g_c(\text{vec}(z_{t_m})), \qquad u_0 = g_c(\text{vec}(z_0)).$$

Let $\mathcal{T}_c = (\mathcal{V}_c, \mathcal{E}_c)$ denote the class-wise MST skeleton constructed from real data.

Distance-to-manifold term. The first component of the latent reward encourages the terminal latent to stay close to the class manifold represented by the MST nodes.

Let $\{v_i\}$ denote the nodes in $\mathcal{V}_c$. The distance penalty is defined as

$$\mathcal{L}_{\text{dist}} = \min_i \|u_0 - v_i\|_2.$$

This term measures how far the generated latent deviates from the learned manifold skeleton.

Direction alignment term. To additionally constrain the diffusion trajectory, we measure whether the local latent transition direction aligns with the direction of nearby MST edges.

The sample transition direction in the projected space is

$$d_{\text{sample}} = \frac{u_0 - u_{t_m}}{\|u_0 - u_{t_m}\|_2}.$$

For each MST edge $(i, j) \in \mathcal{E}_c$, we define

$$m_{ij} = \frac{v_i + v_j}{2}, \qquad d_{ij} = \frac{v_j - v_i}{\|v_j - v_i\|_2}.$$

We first find the nearest edge midpoint to the intermediate latent

$$(i^*, j^*) = \arg \min_{(i,j)\in\mathcal{E}_c} \|u_{t_m} - m_{ij}\|_2.$$

The direction penalty is then defined as

$$\mathcal{L}_{\text{dir}} = 1 - \langle d_{\text{sample}}, d_{i^*j^*} \rangle.$$

This term becomes small when the local diffusion update follows the tangent direction of the class manifold.

Final latent reward. The geometric reward is defined as the negative weighted sum of the two penalties:

$$r_{\text{geo}} = -\lambda_{\text{dist}}\mathcal{L}_{\text{dist}} - \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}.$$

Therefore, higher reward values correspond to smaller manifold deviation and better alignment with the manifold geometry.

C.3 Pixel Classifier Reward Details

We provide additional implementation details for the pixel-space discriminative reward r_{disc} used in `MGPO`. The pixel reward evaluates the semantic consistency of generated samples in pixel space and is computed from the decoded image $\tilde{x}_0 = \mathcal{D}(z_0)$ corresponding to the terminal latent produced by the diffusion trajectory.

Classification. For standard image classification, the discriminative reward is derived from a pretrained proxy classifier f_ϕ. Given a generated image $\tilde{x}_0$ and target class c, the reward is defined as the log-probability of the ground-truth class:

$$r_{\text{disc}}(\tilde{x}_0, c) = \log \left(\frac{\exp(f_\phi(\tilde{x}_0)_c)}{\sum_k \exp(f_\phi(\tilde{x}_0)_k)} \right), \tag{24}$$

which provides a smooth reward landscape and penalizes samples near ambiguous decision boundaries.

Object Detection. For object detection, the discriminative reward is derived from a pretrained Faster R-CNN (Ren et al., 2016) with a ResNet-50 FPN backbone. Given a generated image $\tilde{x}_0$ conditioned on class c, the reward is the maximum post-NMS confidence score for class c, as defined in eq. (9) of the main paper. The full computation is described in Algorithm 3.

Semantic Segmentation. For semantic segmentation, the discriminative reward is derived from a pretrained LR-ASPP (Howard et al., 2019) with a MobileNetV3-Large backbone. Given a generated image $\tilde{x}_0$ conditioned on class c, the reward is the mean confidence over the top-K activated pixels of the spatial probability map for class c, as defined in eq. (10) of the main paper. The TopK aggregation prevents reward hacking by enforcing a minimum expected object area, discouraging the policy from exploiting isolated high-confidence pixels. The full computation is described in Algorithm 4.

D Additional Experiments

D.1 Experiment Results on Different Student Model

As shown in Table 7, we report the performance of different distillation methods under the soft-label based training protocol. Experiments are conducted on ImageWoof across multiple backbone

Algorithm 3 Detection Reward Computation

Require: Generated image $\tilde{x}_0$, target class c, pretrained detector f_ϕ (Faster R-CNN)
Ensure: Scalar reward r_{det}

proposals $\leftarrow f_\phi$.region_proposals($\tilde{x}_0$)
detections $\leftarrow$ NMS(proposals, τ_{iou}=0.5, τ_{score}=0.05)
$\mathcal{S}_c \leftarrow \{\text{score}(d) \mid d \in \text{detections},\ \text{class}(d) = c\}$
if $\mathcal{S}_c = \emptyset$ **then**
    $r_{\text{det}} \leftarrow -1.0$ ▷ no valid detection for class c
else
    $r_{\text{det}} \leftarrow \max(\mathcal{S}_c)$ ▷ eq. (9) of main paper
end if
return r_{det}

Algorithm 4 Segmentation Reward Computation

Require: Generated image $\tilde{x}_0$, target class c, pretrained segmentation network f_ϕ (LR-ASPP), TopK fraction α
Ensure: Scalar reward r_{seg}

$\mathcal{P}_c \leftarrow f_\phi(\tilde{x}_0)[c] \in \mathbb{R}^{H \times W}$ ▷ softmax probability map for class c
$\mathbf{p} \leftarrow \text{flatten}(\mathcal{P}_c) \in \mathbb{R}^{H \cdot W}$
$K \leftarrow \lfloor \alpha \cdot H \cdot W \rfloor$ ▷ number of top pixels
if $\max(\mathbf{p}) < \epsilon$ **then**
    $r_{\text{seg}} \leftarrow -1.0$ ▷ no activation for class c
else
    $r_{\text{seg}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \text{TopK}(\mathbf{p})_k$ ▷ eq. (10) of main paper
end if
return r_{seg}

architectures, including ResNet-18, ResNet-50, and ResNet-101, under various IPC settings. Across all settings, MGPO consistently achieves the best performance. Notably, the performance gains remain stable across different backbone capacities, indicating that the distilled datasets produced by MGPO generalize well to diverse student architectures. In particular, under low IPC setting (IPC = 10), MGPO boosts ResNet-18 performance from 45.6% (CaO^2) to 47.3%, and ResNet-50 accuracy from 40.1% to 41.3%. Under higher IPC (IPC = 50), MGPO continues to outperform strong baselines such as Minimax (Gu et al., 2024), RDED (Sun et al., 2024), and CaO^2 (Wang et al., 2025). This suggests that our reward-guided distillation framework is able to capture class-discriminative information that transfers effectively across models of varying depth and capacity.

D.2 Hyperparameter Analysis

We further study the sensitivity of our method to several key hyperparameters. All experiments are conducted on ImageWoof with 10 IPC distilled datasets. Specifically, we analyze the parameters involved in the construction of the MST that models the latent geometric structure. By varying these parameters while keeping other settings fixed, we evaluate the robustness of our framework and examine how they influence the quality of the generated synthetic data and the downstream performance (Figure 5).

Table 7: Performance comparison with soft-label-based baselines on ImageWoof under different IPC and backbone settings. The best results in each setting are highlighted in bold.

Backbone	IPC	SRe^{2}L (Yin et al., 2023)	Minimax (Gu et al., 2024)	RDED (Sun et al., 2024)	CaO^2 (Wang et al., 2025)	MGPO (Ours)
ResNet-18	10	20.2$_{\pm 0.2}$	40.1$_{\pm 1.0}$	38.5$_{\pm 2.1}$	45.6$_{\pm 1.4}$	**47.3**$_{\pm 0.8}$
	50	23.3$_{\pm 0.3}$	67.0$_{\pm 1.8}$	68.5$_{\pm 0.7}$	68.9$_{\pm 1.1}$	**69.7**$_{\pm 1.6}$
ResNet-50	10	17.3$_{\pm 1.7}$	37.3$_{\pm 1.1}$	29.9$_{\pm 2.2}$	40.1$_{\pm 0.1}$	**41.3**$_{\pm 1.1}$
	50	24.8$_{\pm 0.7}$	64.3$_{\pm 0.9}$	67.8$_{\pm 0.3}$	68.2$_{\pm 1.1}$	**69.0**$_{\pm 0.4}$
ResNet-101	10	17.7$_{\pm 0.9}$	34.2$_{\pm 1.7}$	31.3$_{\pm 1.3}$	36.5$_{\pm 1.4}$	**37.8**$_{\pm 1.6}$
	50	21.2$_{\pm 0.2}$	62.7$_{\pm 1.6}$	59.1$_{\pm 0.7}$	63.1$_{\pm 1.3}$	**63.5**$_{\pm 1.5}$

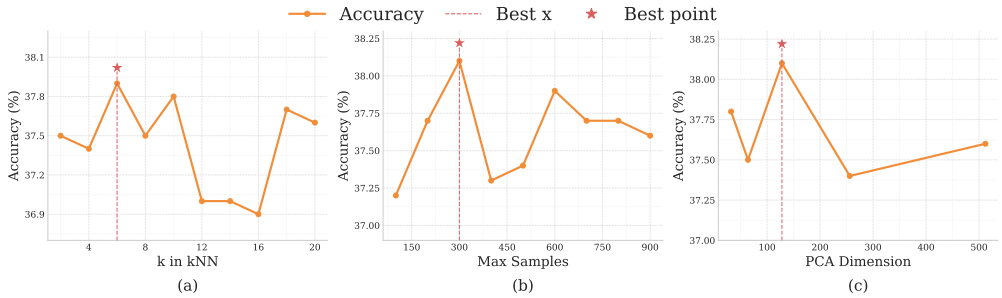

Figure 5: Ablation studies on the hyperparameters used in MST construction, including the number of neighbors k in the kNN graph, the maximum number of sampled latent features per class, and the PCA projection dimension.

Overall, MGPO exhibits stable performance across a broad range of hyperparameter settings. The results suggest that the method is not overly sensitive to individual parameters, and moderate values generally provide the best balance between optimization stability and reward effectiveness.

E Visualization

E.1 Results on Classification

Figure 6, 7, and 8 demonstrate the visualization results on ImageWoof, ImageNette and ImageIDC, respectively.

E.2 Results on Detection

Figure 9 demonstrates the visualization results on detection results.

E.3 Results on Segmentation

Figure 10 demonstrates the visualization results on segmentation results.

F Limitations and Broader Impacts

Limitations. While MGPO establishes an efficient and adaptable paradigm for task aware dataset distillation, it still possesses certain limitations. First, from a theoretical perspective, as discussed in remark 12, the direction component of the geometric reward relies on local tangent approximations which may exhibit fluctuations on high intrinsic dimension manifolds. Second, although MGPO eliminates the computational bottleneck of inference time guidance, the reinforcement learning policy optimization phase introduces additional offline training overhead compared to purely training free selection methods. Finally, scaling the class wise MST construction to datasets with millions of categories might require more advanced hierarchical clustering algorithms to maintain strict scalability.



Figure 6: Comparison between samples generated by Minimax Diffusion (left) and generated by the proposed MGPO method (right) on ImageWoof (Classification).

Broader Societal Impacts. Our research presents significant positive societal impacts by heavily compressing the data required to train deep learning models. This inherently reduces the carbon footprint and energy consumption associated with large scale model training, democratizing artificial intelligence by allowing researchers with limited computational resources to train state of the art models. However, as a potential negative impact, if the pre-trained reward models (such as the proxy classifier or detector) contain inherent biases regarding demographic attributes, the policy optimization process might internalize and amplify these biases within the distilled dataset. Therefore, practitioners must ensure that the reward models used for distillation are thoroughly audited for fairness and equity prior to deployment in sensitive applications.

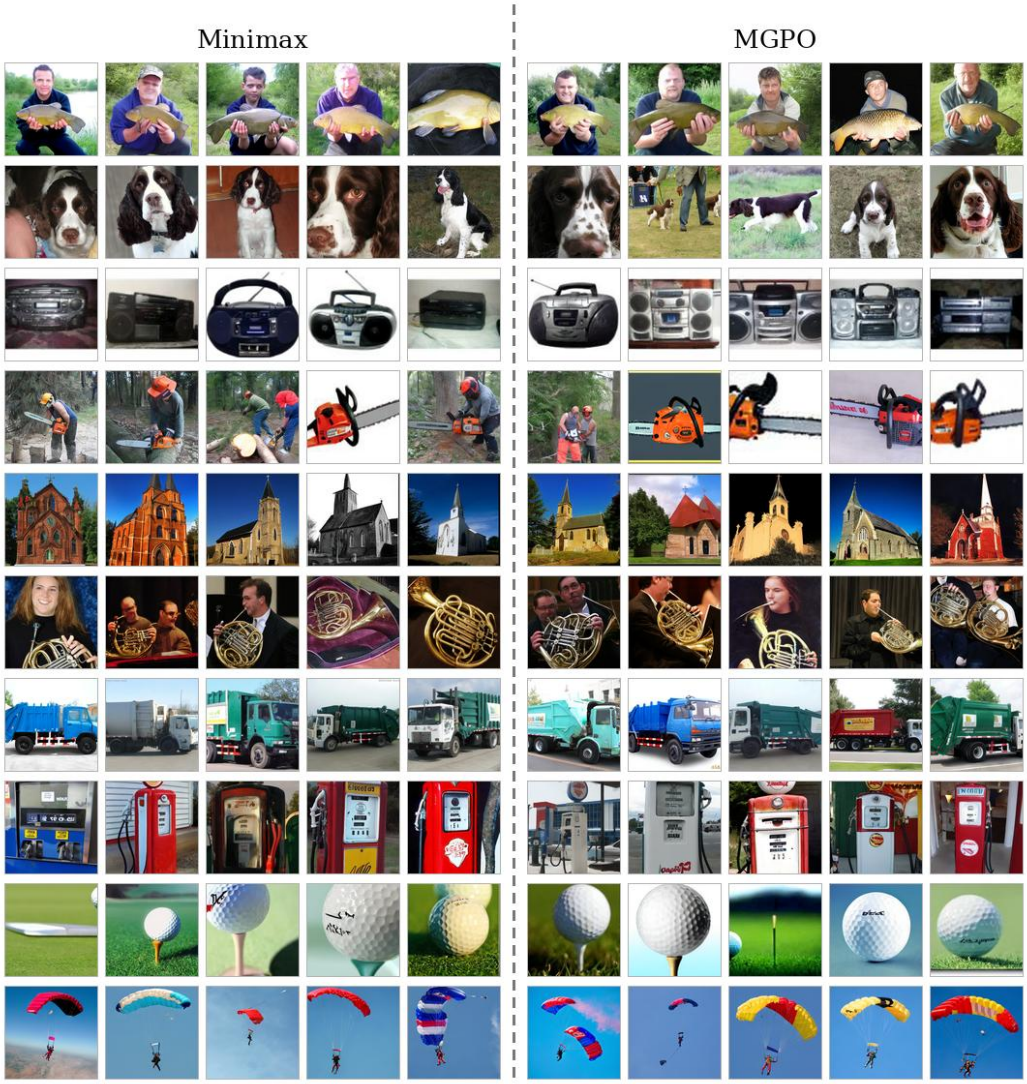

Figure 7: Comparison between samples generated by Minimax Diffusion (left) and generated by the proposed MGPO method (right) on ImageNette (Classification).

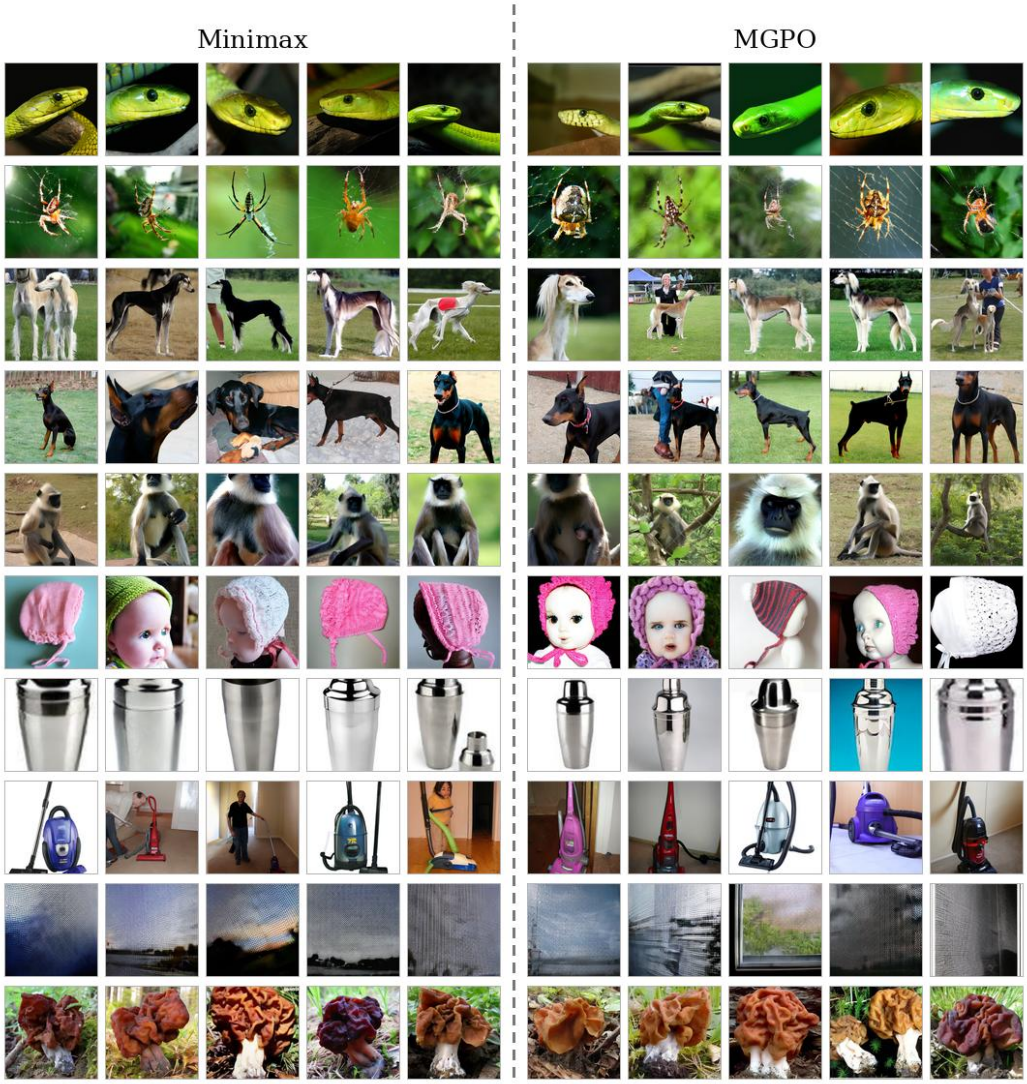

Figure 8: Comparison between samples generated by Minimax Diffusion (left) and generated by the proposed MGPO method (right) on ImageIDC (Classification).

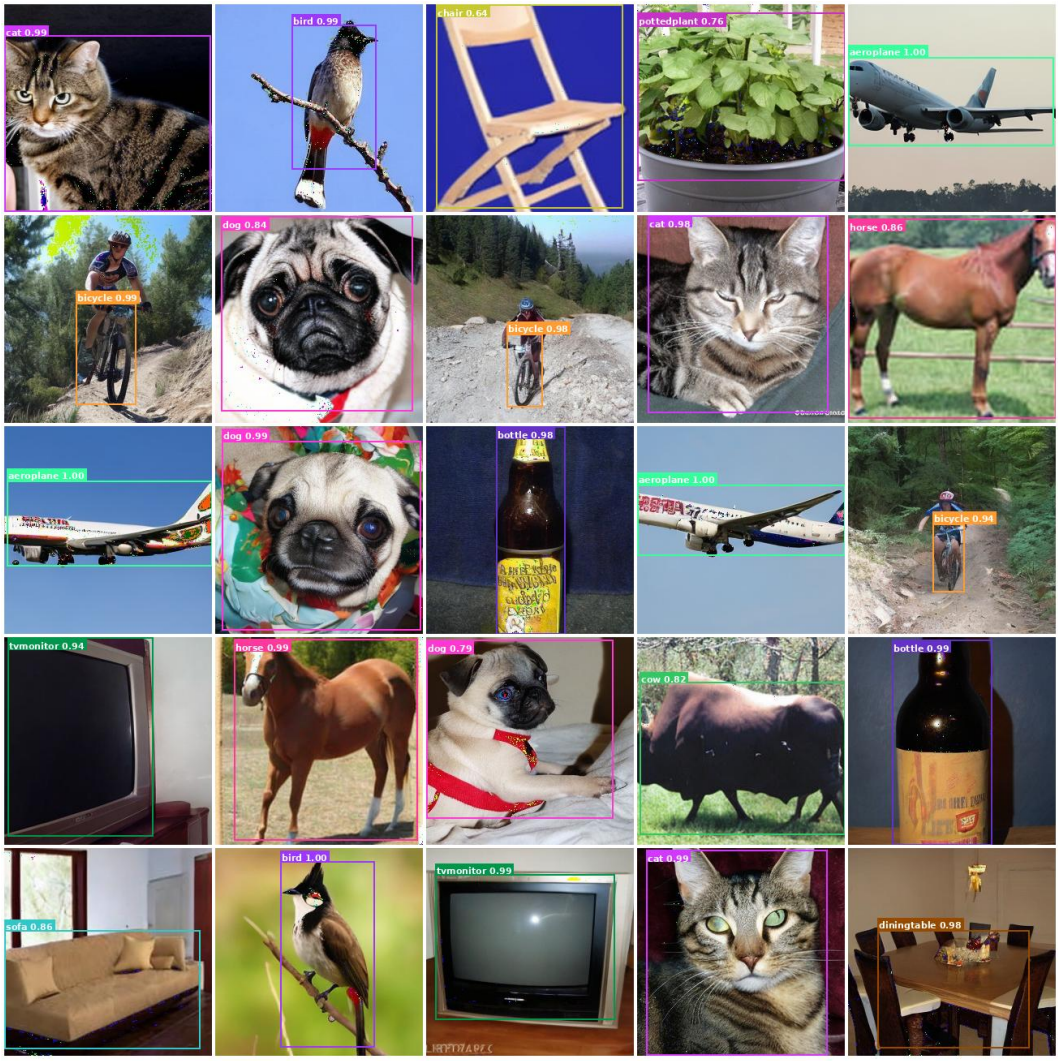

Figure 9: Samples generated by the proposed MGPO (Object Detection).

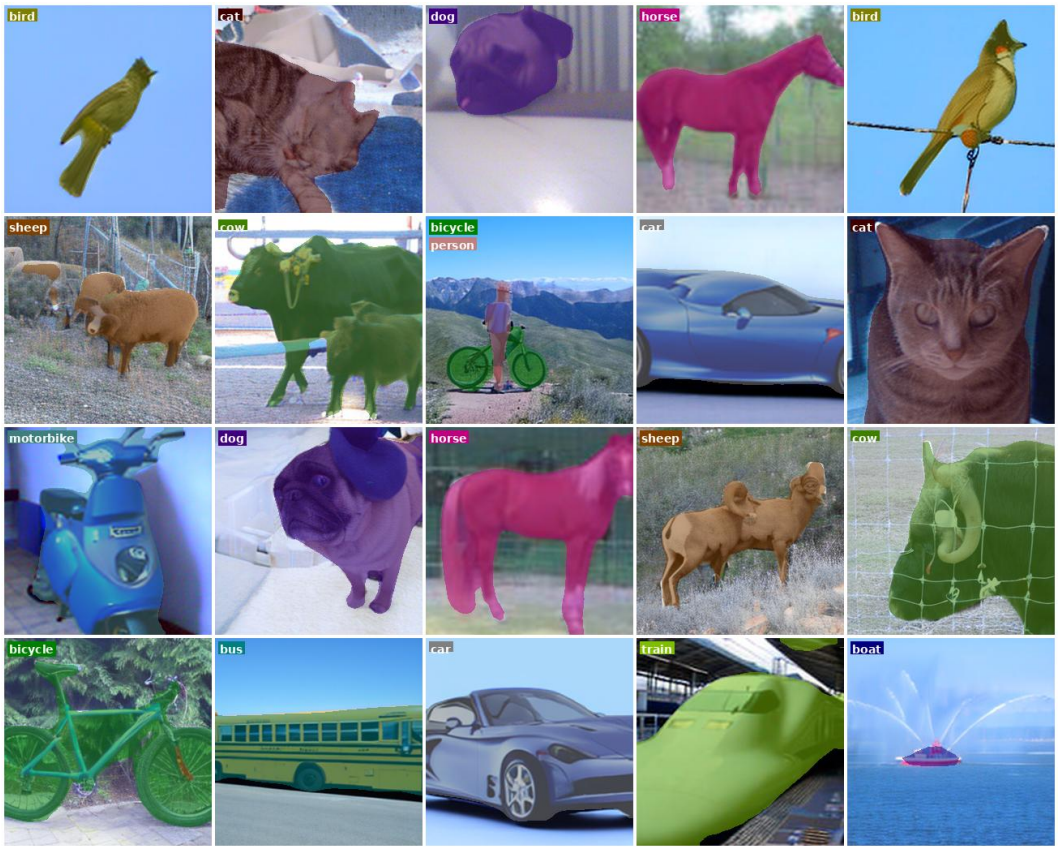

Figure 10: Samples generated by the proposed MGPO (Segmentation).