Learnability-Guided Diffusion for Dataset Distillation

Jeffrey A. Chan-Santiago
Institute of Artificial Intelligence
University of Central Florida
jeffrey.chansantiago@ucf.edu

Mubarak Shah
Institute of Artificial Intelligence
University of Central Florida
shah@crcv.ucf.edu

Abstract

Training machine learning models on massive datasets is expensive and time-consuming. Dataset distillation addresses this by creating a small synthetic dataset that achieves the same performance as the full dataset. Recent methods use diffusion models to generate distilled data, either by promoting diversity or matching training gradients. However, existing approaches produce redundant training signals, where samples convey overlapping information. Empirically, disjoint subsets of distilled datasets capture 80–90% overlapping signals. This redundancy stems from optimizing visual diversity or average training dynamics without accounting for similarity across samples, leading to datasets where multiple samples share similar information rather than complementary knowledge. We propose learnability-driven dataset distillation, which constructs synthetic datasets incrementally through successive stages. Starting from a small set, we train a model and generate new samples guided by learnability scores that identify what the current model can learn from, creating an adaptive curriculum. We introduce Learnability-Guided Diffusion (LGD), which balances training utility for the current model with validity under a reference model to generate curriculum-aligned samples. Our approach reduces redundancy by 39.1%, promotes specialization across training stages, and achieves state-of-the-art results on ImageNet-1K (60.1%), ImageNette (87.2%), and ImageWoof (72.9%). Our code is available on our project page[1].

1. Introduction

Dataset distillation has attracted broad interest for its promise to dramatically reduce training data requirements without sacrificing model performance. The goal is to synthesize a small surrogate dataset $D_{\mathcal{S}}$ from a large target dataset $D_{\mathcal{T}}$ such that models trained on $D_{\mathcal{S}}$ achieve comparable accuracy to those trained on $D_{\mathcal{T}}$. Guo et al.

[1] https://jachansantiago.github.io/learnability-guided-distillation/

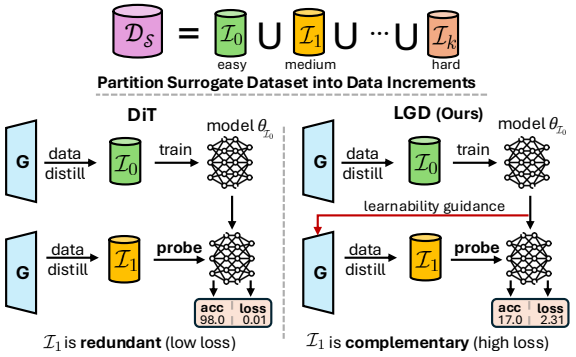


Figure 1. **Learnability-Guided Dataset Distillation.** We partition the distilled dataset $\mathcal{D}_S$ into increments $\{\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_k\}$ (**Top**). **Bottom Left (DiT):** Standard distillation generates increments independently, producing redundant samples—a model trained on $\mathcal{I}_0$ achieves 98.0% accuracy on $\mathcal{I}_1$, indicating no new information. **Bottom Right (LGD):** We condition the next increment on the model parameters $\theta_{\mathcal{I}_0}$ to guide synthesis toward samples that complement $\mathcal{I}_0$. The resulting increment $\mathcal{I}_1$ achieves only 17.0% accuracy when evaluated by the prior model, indicating it introduces substantial new learning signal.

[11] demonstrated near-lossless accuracy on small-scale benchmarks: on CIFAR-10 and CIFAR-100 [18], distilled datasets with only 100 images per class (IPC) match the full 5,000 IPC datasets—a 50× compression using $\sim 2\%$ of the data. However, scaling to larger, high-resolution datasets like ImageNet remains challenging.

Dataset distillation methods traditionally synthesize $D_{\mathcal{S}}$ through bi-level optimization that matches training trajectories [3, 6, 11, 27, 35]—optimizing synthetic data such that model parameters trained on $D_{\mathcal{S}}$ evolve similarly to those trained on $D_{\mathcal{T}}$. While effective on small datasets, this becomes computationally intractable for high-resolution images. To address scalability, generative dataset distillation [4, 5, 10, 28, 33] leverages pretrained generative models to synthesize distilled datasets at much lower cost. Recent work shows trajectory matching remains valuable: influence-guided generation [5] steers samples whose gradients align with those from training on $D_{\mathcal{T}}$.

However, this has a fundamental limitation: optimizing

all samples toward the *average* training trajectory causes convergence to similar gradient profiles rather than specialization across training phases. Model training naturally progresses through stages: early training benefits from samples with strong gradients for coarse features, while late training requires small, refined gradients for fine-grained details. A sample cannot satisfy both—optimizing for the average produces medium-strength gradients throughout, useful at no specific stage. We confirm this empirically: partitioning a 50 IPC dataset into five disjoint 10 IPC subsets, any subset captures 80–90% of the training signal from others (Figs. 1 and 2), demonstrating redundancy.

To address this, we propose learnability-driven distillation that builds the synthetic dataset incrementally. Starting from a small initial dataset (e.g., IPC = 10), we train a model to convergence, then synthesize new samples guided by learnability scores—generating samples that complement rather than replicate existing data. This incremental approach reduces redundancy by conditioning each stage on the model's evolving learning frontier. This reframes distillation as sequential learning: *given a distilled dataset and a model trained on it, generate additional samples that maximize marginal learning gains*. This approach offers a principled way to reduce redundancy and establish a foundation for methods that exploit staged learning dynamics.

Contributions. Our main contributions are:

- *Learnability-driven incremental framework.* (Sec. 4.1) We construct distilled datasets stage-by-stage, conditioning each increment on the current model's learnability to generate complementary rather than redundant training signals.
- *Learnability-guided synthesis.* (Sec. 4.3) We propose Learnability-Guided Diffusion (LGD) that conditions generation on the current model state to synthesize samples that complement existing data, integrated into diffusion sampling to automatically generate informative increments.
- *Redundancy analysis.* (Sec. 5.2) Our incremental framework enables quantifying information overlap in distilled datasets, revealing 80–90% redundancy in existing methods.
- *Improved sample efficiency.* (Secs. 5.1 and 5.2) We achieve state-of-the-art or competitive results on ImageNet-1K (60.1%), ImageNette (82.6–87.2%), and ImageWoof (53.9–72.9%), while reducing redundancy by 39.1%.

2. Related Work

Dataset Distillation. Early methods synthesize compact datasets through optimization that matches training dynamics using gradient matching [27, 35], distribution matching [24, 31, 37, 38], or trajectory matching [3, 6, 11]. While effective on small benchmarks, these pixel-optimization approaches struggle with high-resolution datasets. Decoupled methods [26, 32] improve scalability through sequential stages, enhanced by multi-architecture training [25], patch-based composition [29], and curriculum strategies [20]. However, pixel-space constraints limit sample diversity, motivating generative approaches.

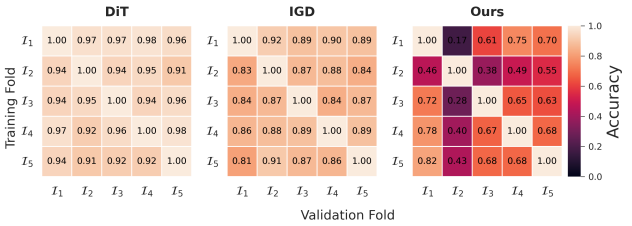


Figure 2. **Cross-validation across distilled data increments ($\mathcal{I}_1 - \mathcal{I}_5$) for IPC 50 on ImageNette.** Each heatmap shows accuracy when training on one increment (rows) and evaluating on another (columns). DiT[23] and IGD [5] exhibit high cross-increment accuracy due to overlapping information, while LGD yields lower off-diagonal scores, indicating more **complementary and diverse** increments.

Generative Dataset Distillation. Recent works [4, 5, 10, 28, 33, 36] leverage pretrained generative models to overcome pixel-space limitations. Diffusion-based methods offer realistic generation: Minimax Diffusion [10] balances diversity and representativeness through diffusion fine-tuning, MGD^3 [4] and D^{4}M [28] conditions on feature-space modes for broader coverage, and Influence-Guided Diffusion (IGD) [5] guides generation to match training gradients from the full dataset. However, these methods synthesize all samples uniformly, causing convergence to similar gradient profiles. Yet these approaches suffer from inherent redundancy because they optimize for visual diversity or average training trajectories without accounting for training signal similarity across samples. This produces datasets where multiple samples teach the model similar information rather than providing complementary knowledge across training stages. Our work addresses this through learnability-driven synthesis: conditioning each increment on the current model state to generate complementary rather than redundant samples.

Curriculum Learning and Data Curation. Our approach connects to curriculum learning, where training difficulty evolves with model competence [2, 22]. Recent methods [9, 12, 34, 39] identify learnable samples by tracking training signals, filtering samples that are already mastered or too difficult. Feedback-driven synthesis [1, 13] generates harder examples tailored to model weaknesses. We extend this to dataset distillation by iteratively synthesizing data increments that maximize marginal learning gains, transforming distillation into a learnability-driven curriculum that adapts to evolving training needs.

3. Background

Dataset Distillation. Dataset distillation aims to create a small synthetic dataset $\mathcal{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{M}$ that captures the essential information from a large training dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N_\mathcal{T}}$, where $M \ll N_\mathcal{T}$. The key requirement is that models trained on the small dataset $\mathcal{D}$ should perform nearly as well as models trained on the full dataset $\mathcal{T}$. We denote these models as $\theta_\mathcal{D}$ and $\theta_\mathcal{T}$ respectively, with the goal that $\mathcal{A}(\theta_\mathcal{D}) \approx \mathcal{A}(\theta_\mathcal{T})$, where $\mathcal{A}(\cdot)$ measures test accuracy. The budget for distillation is specified in images per class (IPC). Modern approaches [4, 5, 10, 28, 33, 36] use generative models to synthesize $\mathcal{D}$ by matching the learning behavior on real data:

$$\min_{\mathcal{D}} \left\| \mathbb{E}_{x\sim P_{\text{data}}}\left[\ell(\theta_\mathcal{T}(x), y)\right] - \mathbb{E}_{x\sim P_{\text{data}}}\left[\ell(\theta_\mathcal{D}(x), y)\right] \right\| \tag{1}$$

where ℓ is a loss function and P_{data} is the data distribution. This generative approach can create entirely new samples rather than just selecting from existing data, enabling greater flexibility and diversity.

Diffusion Models. Diffusion models [14] generate images through a two-stage process. First, the *forward process* gradually adds noise to real images over T steps. Starting from a clean image x_0, each step adds a small amount of Gaussian noise: $q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where β_t controls how much noise to add at step t. We can jump directly to any noisy version using $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

Second, the *reverse process* learns to remove noise step-by-step. A neural network $\epsilon_\phi(x_t, t)$ predicts the noise at each step, and we use this to compute the mean of the reverse distribution:

$$\mu_\phi(x_t) = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\phi(x_t, t)\right) \tag{2}$$

The denoised sample is then drawn as $x_{t-1} = \mu_\phi(x_t) + \sigma_t\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ adds controlled randomness. Starting from pure noise, we repeatedly apply this denoising step to generate new images. For class-conditional generation, we provide the class label as input: $\epsilon_\phi(x_t, t, c)$.

Guided Sampling. We can steer diffusion models to generate images with specific properties by adding guidance signals during sampling. Classifier guidance [8] adjusts the noise prediction to favor a target class c:

$$\tilde{\epsilon}_\phi(x_t, t, c) = \epsilon_\phi(x_t, t, c) + \lambda\nabla_{x_t}\log p(c|x_t) \tag{3}$$

where λ controls the guidance strength. The gradient term $\nabla_{x_t}\log p(c|x_t)$ steers generation toward class c. Recent work in distillation [4, 5] uses similar guidance based on proxies for training utility—guiding the model to generate samples that will be most useful for learning. We build on this idea by making the guidance adaptive to the current training state.

4. Method

Existing dataset distillation methods typically optimize synthetic images to mimic the gradient trajectories or training dynamics obtained from the full dataset. In contrast, we propose an *incremental* formulation that builds the distilled dataset stage by stage, where each newly generated increment is optimized to maximize the model's learning signal given its current knowledge. Our key insight is that by aligning the synthesis process with the learner's evolving state, each synthetic sample can contribute complementary, non-redundant information, leading to a more efficient and adaptive curriculum.

4.1. Incremental Distillation Formulation

We formulate learnability-driven distillation as an incremental synthesis problem. Let $\mathcal{D}$ denote the final distilled dataset of size M, partitioned into K disjoint increments $\mathcal{I}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}$ such that $\sum_{i=1}^{K} N_i = M$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for $i \neq j$. At stage i, the model with parameters θ_{i-1} is trained on the cumulative dataset

$$\mathcal{D}_i = \bigcup_{k=1}^{i} \mathcal{I}_k, \tag{4}$$

resulting in updated parameters θ_i. This process repeats until all K increments have been synthesized and incorporated into $\mathcal{D}$.

Given the current model state θ_{i-1}, we seek the next increment $\mathcal{I}_i$ that maximizes its contribution to learning:

$$\mathcal{I}_i^* = \arg\max_{\mathcal{I}} \mathcal{L}(\theta_{i-1}, \mathcal{I}), \tag{5}$$

where $\mathcal{L}(\theta_{i-1}, \mathcal{I})$ measures how much the model can learn from $\mathcal{I}$. However, without constraints, this process can produce degenerate samples that drift away from meaningful semantics or correct labels. We therefore regularize the synthesis with a reference model θ^* trained on the full dataset:

$$\mathcal{I}_i^* = \arg\max_{\mathcal{I}} \left[\mathcal{L}(\theta_{i-1}, \mathcal{I}) - \mathcal{L}(\theta^*, \mathcal{I})\right]. \tag{6}$$

The second term acts as a regularizer, penalizing samples that are difficult or misclassified by the reference model. This objective naturally promotes the generation of examples that are hard for θ_{i-1} yet still learnable under θ^*, ensuring each increment targets learnable knowledge gaps.

Incremental distillation as synthesis and analysis framework. Beyond synthesis, our incremental formulation provides an *analysis framework* for diagnosing sample redundancy. By partitioning any dataset $\mathcal{D}$ into increments and evaluating cross-increment learning dynamics, we can quantify information overlap across different distillation methods. For synthesis, each increment is conditioned on the evolving model state to maximize complementary information. For evaluation, the final dataset

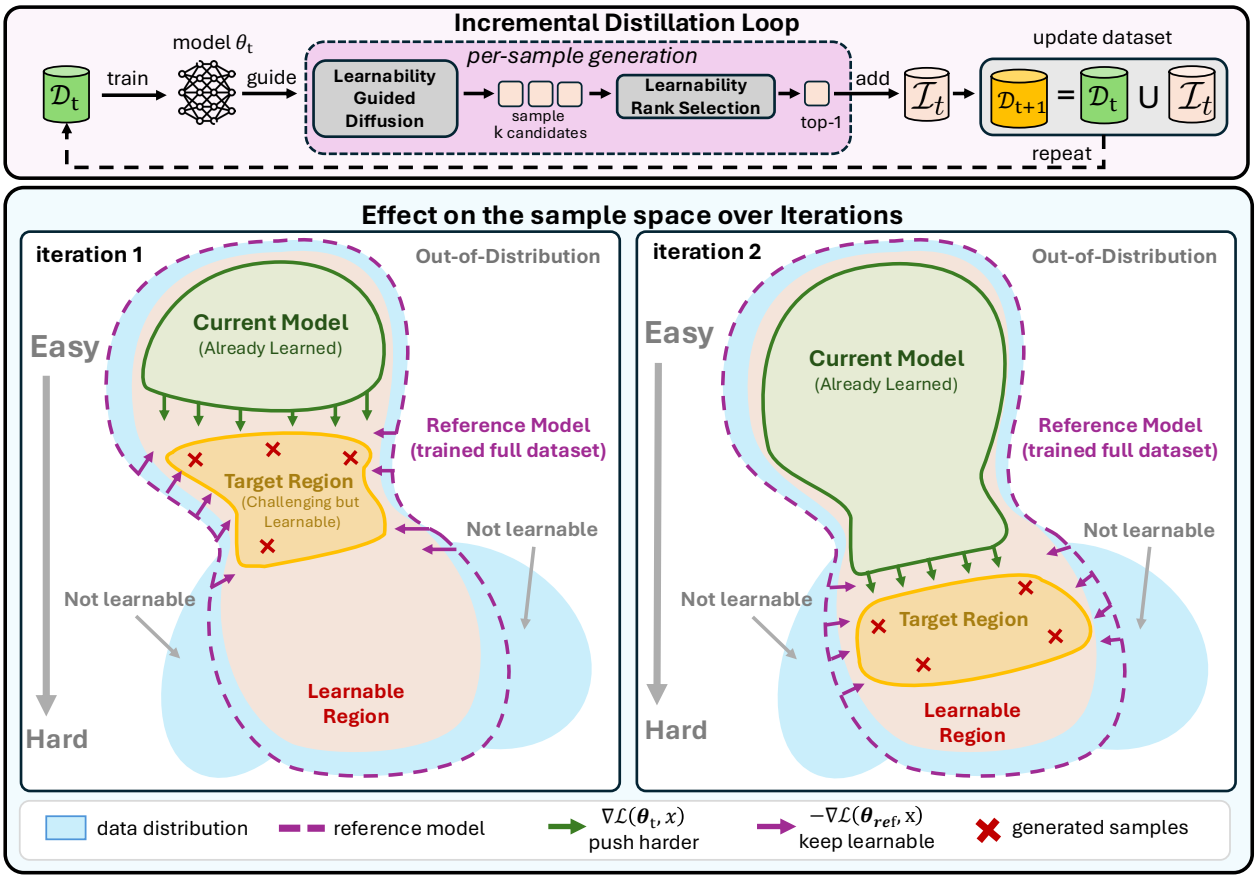


Figure 3. **Overview of our learnability-guided iterative generation framework.** (*Top*) Incremental distillation loop: we iteratively train model θ_t on cumulative dataset $\mathcal{D}_t$, generate samples using our learnability guidance, select high-quality samples via learnability ranking, and augment the dataset. (*Bottom*) Effect on sample space: The current model θ_t (green) expands over iterations, while the reference model θ^* (purple, fixed) defines the learnable region. Generated samples (red $\times$) land in the learnable gap between boundaries, automatically synthesizing samples that complement the current model's learned distribution.

$\mathcal{D}$ is compatible with both incremental training and standard static protocols, enabling direct comparison with prior work. Fig. 3 (top) illustrates this incremental loop: at each stage, we train model θ_i on the cumulative dataset, generate candidates via learnability-guided diffusion, select high-quality samples, and augment the dataset for the next iteration.

4.2. Data Seed Initialization

Our learnability-driven synthesis requires an initial seed dataset $\mathcal{D}_1$. By default, we use IGD [5] distilled images at 10 IPC per class, though $\mathcal{D}_1$ can also be sampled from pretrained diffusion models or other distilled datasets. This seeded initialization accelerates convergence by starting from a distilled dataset rather than random samples. Ablations on seed choices are in the supplementary material.

4.3. Learnability-Guided Diffusion Sampling

During synthesis, we incorporate a learnability criterion derived from Eq. (6) into the diffusion sampling process, guiding the denoising trajectory toward regions with high learnability scores.

Learnability Score. For a candidate sample (x, y), we define its learnability [9, 21] as

$$\mathcal{S}(x, y) = \mathcal{L}(\theta_{i-1}, x, y) - \omega \cdot \mathcal{L}(\theta^*, x, y), \tag{7}$$

where $\mathcal{L}(\theta, x, y)$ is a prediction loss (e.g., cross-entropy) and ω controls the reference model regularization strength. High $\mathcal{S}$ indicates the current model struggles while the reference model does not—the sample is both informative and semantically valid.

Learnability Guidance. We modulate the diffusion model's predicted noise $\epsilon_\phi(x_t, t, y)$ with the gradient of the learnability score:

$$\tilde{\epsilon}_\phi(x_t, t, y) = \epsilon_\phi(x_t, t, y) + \lambda \cdot \rho_t \cdot \nabla_{x_t}\mathcal{S}(x_t, y), \tag{8}$$

where λ controls guidance strength and $\rho_t = \sqrt{1-\bar{\alpha}_t}\frac{\|\epsilon_\phi(x_t,t,y)\|}{\|\nabla_{x_t}\mathcal{S}(x_t,y)\|}$ is a timestep-dependent scaling factor [5] that normalizes guidance magnitude relative to

Table 1. **Comparison across distilled IPC budgets on Nette and Woof evaluated on different network architectures**. Mean±std accuracy; best per row in **bold**. IGD, MGD^3, and LGD used a pretrained DiT as the diffusion backbone.

Dataset	Model	IPC	Random	DM [37]	IDC-1 [17]	DiT [23]	Minimax [10]	IGD [5]	MGD^3 [4]	**LGD (Ours)**	Full
Nette	ConvNet-6	50	71.8±1.2	70.3±0.8	72.4±0.7	74.1±0.6	76.9±0.9	80.9±0.9	80.9±2.3	**82.6±0.7**	94.3±0.5
		100	79.9±0.8	78.5±0.8	80.6±1.1	78.2±0.3	81.1±0.3	84.5±0.7	86.5±0.9	**87.2±0.7**	
	ResNetAP-10	50	77.3±1.0	76.7±1.0	77.4±0.7	76.9±0.5	78.2±0.7	81.0±1.2	81.2±1.0	**84.3±0.5**	94.6±0.5
		100	81.1±0.6	80.9±0.7	81.5±1.2	80.1±1.1	81.3±0.9	85.2±0.5	85.5±1.0	**87.2±0.9**	
	ResNet-18	50	75.8±1.1	75.0±1.0	77.8±0.7	75.2±0.9	78.1±0.6	81.0±0.7	81.5±3.4	**85.0±0.9**	95.3±0.6
		100	82.0±0.4	81.5±0.4	81.7±0.8	77.8±0.6	81.3±0.7	84.4±0.8	85.6±0.2	**86.9±0.6**	
Woof	ConvNet-6	50	41.9±1.4	43.8±1.1	42.6±0.9	48.5±1.3	50.7±1.8	**54.2±0.7**	53.4±0.4	53.9±2.2	85.9±0.4
		100	52.3±1.5	50.1±0.9	51.0±1.1	54.2±1.5	57.1±1.9	61.1±1.0	59.2±0.9	**61.9±2.0**	
	ResNetAP-10	50	50.1±1.6	47.8±1.2	49.3±0.9	55.8±1.1	59.8±0.8	**62.7±1.2**	59.4±0.2	62.5±0.7	87.2±0.6
		100	59.2±0.9	59.8±1.3	56.4±0.5	62.5±0.9	66.8±1.2	69.7±0.9	66.1±0.8	**71.1±0.8**	
	ResNet-18	50	54.0±0.8	53.9±0.7	54.5±1.0	57.4±0.7	60.5±0.5	62.0±1.1	63.9±0.3	**65.1±0.7**	89.0±0.6
		100	63.6±0.5	64.9±0.7	57.7±0.8	62.3±0.5	67.4±0.7	70.6±1.8	71.3±0.5	**72.9±0.6**	

the noise level. This transforms diffusion synthesis into an *active learning mechanism*: each increment $\mathcal{I}_i$ maximizes training value for the current learner state, forming a curriculum of increasing difficulty with non-redundant supervision. Fig. 3 (bottom) shows generated samples landing in the learnable gap between current model θ_i and reference model θ^*.

Guidance Schedule and Diversity. Following [4, 5], we apply guidance only during timesteps $t \in [10, 45]$ (out of 50 steps), as full-trajectory guidance degrades performance. Within this window, we apply deviation guidance [5]: for each sample x_t, we subtract $\gamma \nabla_{x_t} \mathcal{G}_D(x_t)$ from the predicted noise. Let $\mathcal{M}^c$ denote the memory buffer of previously generated samples for class c. The guidance objective $\mathcal{G}_D(x) = \frac{x \cdot \tilde{x}^*}{\|x\| \|\tilde{x}^*\|}$ measures cosine similarity to the nearest sample $\tilde{x}^* \in \mathcal{M}^c$, and γ controls repulsion strength. This enhances intra-class diversity by steering generation away from existing samples.

4.4. Learnability Sample Selection

While learnability guidance steers diffusion trajectories toward informative regions during synthesis, the stochastic nature of sampling can still produce a mix of highly and weakly learnable examples. To refine the distilled dataset, we introduce a per-sample learnability selection step applied after generation: for each sample to be added to the dataset, we generate κ candidate versions via learnability-guided diffusion sampling, score each using Eq. (7), and retain only the highest-scoring candidate. This selected sample is then added to the memory buffer $\mathcal{M}^c$, and the process repeats sequentially for subsequent samples, ensuring each new addition is selected in the context of the already-constructed dataset.

Candidate Generation and Selection. At each incremental stage i, we fill N_i sample positions per class by generating κ candidates $\{(x^c_{j,k}, y^c_{j,k})\}_{k=1}^{\kappa}$ for each position $j \in 1, \ldots, N_i$ via learnability-guided diffusion sampling. Each candidate is scored using Eq. (7), and we retain the top-scoring sample:

$$(x^c_j, y^c_j) = \text{Top}\,1\big((xj, k^c, yj, k^c, \mathcal{S}(x^c_{j,k}, y^c_{j,k}))_{k=1}^{\kappa}\big) \quad (9)$$

The selected samples are appended to the memory buffer $\mathcal{M}^c$ and repeated for all N_i positions to form $\mathcal{I}_i$, encouraging diversity while ensuring high-learnability increments aligned with Eq. (6).

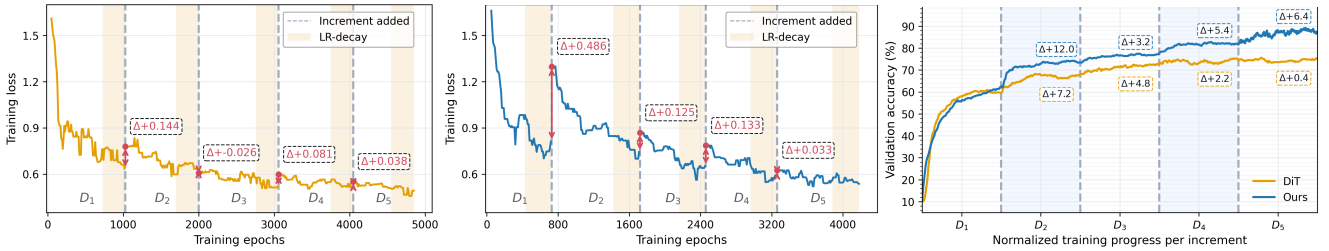
5. Experiments

Datasets. We evaluate on three 256×256 datasets: ImageNette [15], ImageWoof [16], and ImageNet-1K [7], in various IPC from 20 to 100.

Protocols. For ImageNette/ImageWoof, we use the hard-label protocol [10]: train target networks (ConvNet-6, ResNet-AP-10, ResNet-18) from scratch on synthetic data with ground-truth labels. For ImageNet-1K, we use the soft-label protocol [10, 29]: each distilled image x_i is divided into M regions, and each region $x_{i,m}$ is paired with a soft label $y_{i,m} = \ell(\phi_{\theta^\tau}(x_{i,m}))$ generated by a ResNet-18 teacher. All experiments use standard augmentation (random crop, CutMix) and report mean±std of top-1 accuracy over multiple runs following [4, 5, 10].

Hyperparameters. We set the learnability guidance strength $\lambda = 15$, reference model weight $\omega = 0.5$, and over-generation factor $\kappa = 3$ for selection. Guidance is applied during timesteps $t \in [10, 45]$ (out of 50 diffusion steps). For deviation guidance, we use $\gamma = 50$. The initial seed dataset uses 10 IPC per class. Ablation studies and further details are in the supplementary material.

Evaluation Settings. We evaluate under two complementary protocols: *Static Evaluation* trains models from scratch on the complete distilled dataset (standard evaluation). *Incremental Distillation* constructs datasets incre-



(a) DiT Incremental Training (b) Ours Incremental Training (c) Validation Accuracy Progression Comparison

Figure 4. **Incremental training dynamics of DiT and our method.** (a-b) show the training loss across successive data increments ($D_1 \rightarrow D_5$), where each increment adds new samples followed by a 300-epoch learning-rate decay (light beige). Our method yields stronger loss spikes (Δ) after each increment, suggesting the added data is harder and complementary. (c) compares normalized validation accuracy per increment between DiT and our method, highlighting consistent accuracy gains and faster convergence for ours.

Table 2. **ImageNet-1K:** Performance comparison over ResNet-18 with state-of-the-art dataset distillation methods.

SRe^{2}L [32]	G-VBSM [19]	RDED [29]	DiT [23]	Minimax [10]	DiT-IGD [5]	MGD^3[4]	DiT+LGD (Ours)
46.8±0.2	51.8±0.4	56.5±0.1	52.9±0.6	58.6±0.3	59.8±0.3	60.2±0.1	60.1±0.1

Table 3. **Cross-Architecture Evaluation on ImageNette.** Accuracy (%) when models trained on surrogate architectures are evaluated on different target architectures. Bold indicates the best within each IPC.

IPC	Surrogate	Evaluated on		
		ConvNet-6	ResNetAP-10	ResNet-18
50	ConvNet-6	**83.5±0.4**	83.5±0.8	83.3±1.0
	ResNetAP-10	82.6±0.7	**84.7±1.2**	**85.0±0.9**
	ResNet-18	81.1±1.1	82.3±0.3	83.6±1.0
100	ConvNet-6	86.0±0.6	86.7±0.9	86.6±0.6
	ResNetAP-10	**87.2±0.7**	**87.7±0.6**	86.9±0.6
	ResNet-18	86.7±1.6	87.3±1.2	**87.7±0.7**

Table 4. **Progression across data increments on ImageNette.** Accuracy at each IPC level as additional increments are sequentially added (IPC10 $\rightarrow$ IPC100).

Method	IPC10	IPC20	IPC30	IPC40	IPC50	IPC100
DiT	61.0±1.5	68.2±1.1	73.0±1.7	75.2±1.8	75.6±1.1	78.2±0.7
IGD	**64.5±0.9**	71.6±1.3	74.8±1.7	78.1±1.2	79.9±0.3	84.3±0.1
LGD (Ours)	64.1±1.3	**75.9±0.3**	**79.3±1.6**	**83.1±0.8**	**85.2±0.8**	**89.1±0.7**

mentally (10 IPC per stage), training models on cumulative data after each increment to analyze curriculum effects. For our method, incremental construction yields the final dataset; baselines distill each increment independently.

5.1. Main Results: Static Evaluation

We compare our incrementally constructed datasets with current SOTA methods under standard static training protocols. Models train from scratch on the complete distilled dataset across three benchmarks, enabling direct comparison with prior work.

ImageNette and ImageWoof. On the ImageNette dataset, our method achieves notable performance gains of

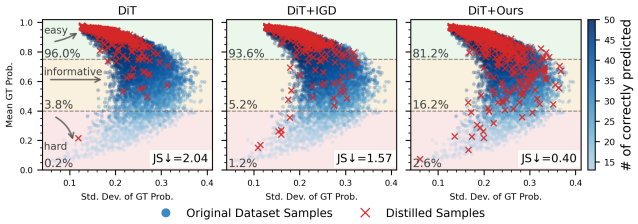


Figure 5. **Learning-dynamics visualization of original and distilled samples.** Each point shows a sample's mean and standard deviation of ground-truth class probability across training (50 epochs). Top-left points are easy (high μ, low σ^2), bottom-left are hard, and mid-right indicate informative samples. Our method yields distilled samples that form a more informative (16.2%) and harder dataset (2.6%), roughly 3$\times$ and 2$\times$ over IGD, respectively, aligning more closely with the original training dynamics distribution, as shown by the lowest Jensen–Shannon divergence (JS $\downarrow$).

1.7%, 2.0%, 2.7%, and 4.0% across various IPC values and architectures, surpassing DiT+IGD (see Tab. 1). At 50 IPC, we reach 82.6-85.0% across ConvNet-6, ResNet-AP-10, and ResNet-18, while at 100 IPC, we achieve 86.9-87.2%. Similarly, on the ImageWoof dataset, our method demonstrates competitive performance at 50 IPC (53.9-65.1%) and achieves the best results at 100 IPC (61.9-72.9%), consistently outperforming DiT+IGD and DiT+MGD^3.

ImageNet-1K. Tab. 2 shows comparison to SOTA in ImageNet-1K at 50 IPC with ResNet-18. Our method achieves 60.1%, matching state-of-the-art performance (MGD^3: 60.2%) while substantially outperforming the base diffusion approach (DiT: 52.9%, improvement of 7.2%) and decouple methods. This demonstrates effective scaling to complex, large-scale datasets despite incremental construction.

Cross-Architecture Generalization. Tab. 3 shows comparison of cross-architecture transfer on ImageNette. Datasets distilled with ResNet-AP-10 achieve 85.0% when evaluated on ResNet-18 at 50 IPC, demonstrating strong architecture-agnostic performance. At 100 IPC, cross-architecture accuracy remains high (86.0-87.7%) with minimal degradation across all architecture pairings, confirming our method learns transferable representations rather than

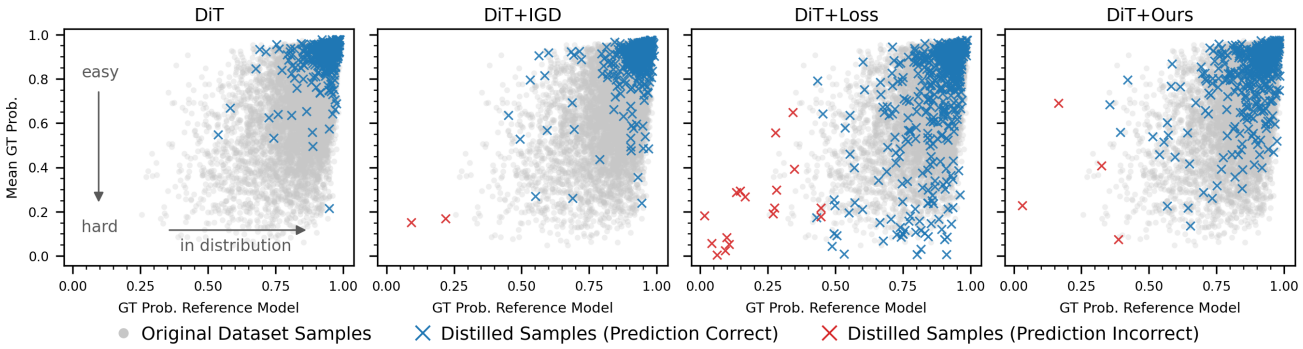


Figure 6. **In-distribution and learning-dynamics analysis of distilled datasets.** Each point represents a sample described by its ground-truth (GT) class probability from the reference model (x-axis, measuring in-distribution likelihood $p(y|x)$) and its mean GT probability across training epochs (y-axis, reflecting sample difficulty). Original samples (gray) delineate the in-distribution region, where higher y indicates easier examples. **DiT** concentrates on easy, high-confidence samples; **DiT+IGD** extends slightly toward mid and hard areas; **DiT+Loss** covers a broader difficulty range but introduces several non-representative (low-x) samples; **DiT+Ours** achieves a balanced spread—capturing informative mid and hard examples while remaining closely aligned with the in-distribution region.

architecture-specific patterns.

Across all three datasets, our method achieves state-of-the-art or competitive performance (ImageNet-1K: 60.1%, ImageNette: 82.6-87.2%, ImageWoof: 53.9-72.9%), with consistent improvements over uniform guidance methods. These results validate that our learnability-driven construction yields high-quality datasets. We next analyze why this approach succeeds by examining sample redundancy and training dynamics.

5.2. Diagnosing Sample Redundancy

To understand why learnability-driven synthesis works, we analyze sample redundancy in distilled datasets. We partition each 50 IPC dataset into $K = 5$ disjoint 10 IPC increments and measure cross-increment learning: training on one increment and evaluating on another. High cross-increment accuracy indicates redundancy (overlapping information); low accuracy indicates complementary information.

Fig. 2 shows cross-validation heatmaps on ImageNette. DiT exhibits severe redundancy with 91-98% off-diagonal accuracy (average: 94.7%)—models trained on any increment already capture most information from others. IGD improves slightly to 81-92% (average: 87.1%) but substantial overlap remains. Our method achieves only 17-82% cross-increment accuracy (average: 57.65%), confirming that increments contain complementary information. Our method reduces redundancy by 39.1% relative to DiT, validating that learnability-guided synthesis conditioned on model state effectively diversifies training signals.

Beyond synthesis, this demonstrates our incremental formulation as a diagnostic tool for analyzing any distillation method. By partitioning existing datasets and measuring cross-increment dynamics, researchers can systematically evaluate information distribution and identify redundancy.

5.3. Incremental Training Dynamics

We now examine how reduced redundancy translates to improvement. We train models incrementally on distilled datasets, adding one 10 IPC increment at a time (5 total). For each increment, we train the model until the training loss no longer improves, then apply a learning rate decay for 300 more epochs before adding a new increment.

Loss Spikes and Information Content. Figs. 4a and 4b shows training loss across increments. DiT exhibits small, uniform loss spikes (avg. $\Delta = 0.06$) after each increment, consistent with redundant samples. Our method shows larger loss spikes when adding new data ($\Delta_1 = 0.486 \rightarrow \Delta_5 = 0.033$, avg. $\Delta = 0.20$), confirming increments of increasing difficulty. Despite larger disruptions, our method converges faster to higher accuracy.

Sustained Marginal Gains. Fig. 4c and Tab. 4 show accuracy progression. Our method achieves 85.2% vs. 79.9% (IGD) and 75.6% (DiT) at 50 IPC. As shown in Fig. 4c, our method maintain sustained gains across stages (+12.0%, +3.2%, +5.4%, +6.4%) while DiT shows declining returns (+7.2%, +4.8%, +2.2%, +0.4%) . Our final increment (40→50 IPC) alone provides more gain than DiT's last two combined.

These dynamics validate our curriculum hypothesis: conditioning synthesis on model state automatically generates samples at the frontier of current capability—hard enough for new information, learnable enough for efficient absorption.

5.4. Sample Quality and Difficulty Analysis

We validate that our samples align with natural difficulty distributions while maintaining semantic validity.

Difficulty Distribution. Following [12, 30], we characterize samples by mean (μ) and variance (σ^2) of GT probability across 50 training epochs. Easy samples have high

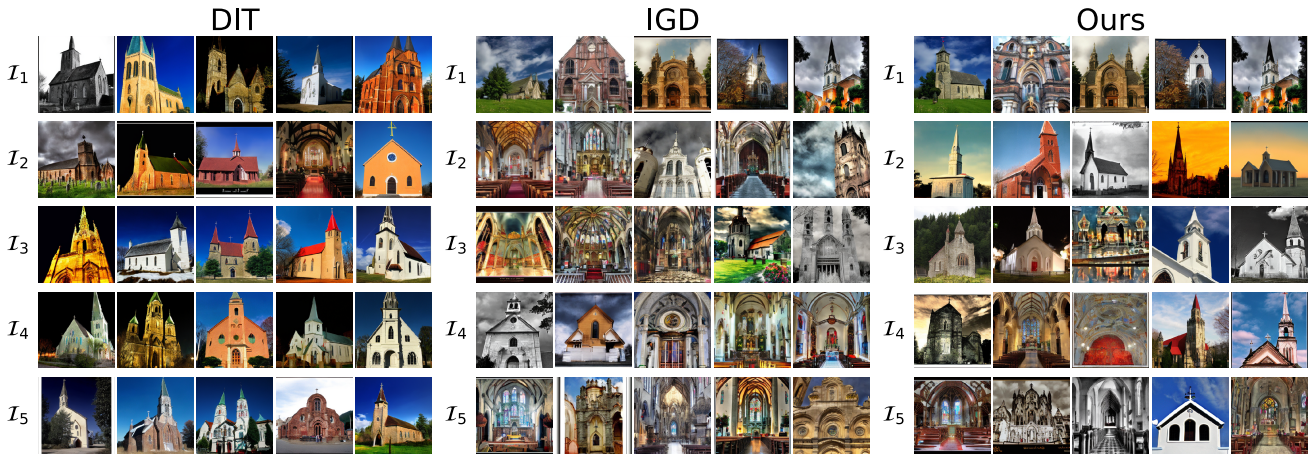


Figure 7. **Visual diversity in incrementally distilled datasets.** Samples from increments $\mathcal{I}_1 - \mathcal{I}_5$ (50 IPC total) on the Church class. **DiT** generates repetitive samples with similar architectural styles and lighting. **IGD** improves slightly but exhibits clustering around Gothic exteriors. **LGD (Ours)** produces diverse samples spanning multiple architectural styles (traditional, modern, ornate), perspectives (exterior and interior views), and lighting conditions (day, night, golden hour). This diversity reflects our curriculum-based synthesis: early increments capture simpler structures while later increments progressively introduce architectural complexity and challenging interior scenes, reducing redundancy and maximizing sample utility across the distilled dataset.

μ, low σ^2; hard samples have low μ, low σ^2; informative samples have mid-to-high μ and high σ^2 (model improves during training).

Fig. 5 shows (μ, σ^2) distributions. DiT concentrates in easy regions (>80% with $\mu > 0.8$), IGD improves (60% easy) but retains clustering, while our method distributes broadly across easy, informative (high σ^2), and hard regions. Jensen-Shannon divergence quantifies alignment: our method achieves **JS = 0.40** vs. DiT (2.04) and IGD (1.57)—5× and 4× improvements, confirming better alignment with natural data difficulty characteristics.

In-Distribution Analysis and Semantic Consistency Fig. 6 characterizes samples by reference model confidence $p(y|x)$ (x-axis, in-distribution likelihood) and mean GT probability at training (y-axis, difficulty). DiT clusters in easy regions (top-right). We also evaluate our method without the $\mathcal{L}(\theta^*, x, y)$ regularization term (denoted as DiT+Loss). While this covers a broader difficulty range, it introduces out-of-distribution samples that are misclassified by the reference model (red × markers at low x-values). Our method achieves balanced spread across difficulty levels ($x > 0.6$ for most samples), with red × appearing in genuinely hard regions (bottom-right) rather than out-of-distribution areas. This validates our learnability regularization (Eq. (6)): the $\mathcal{L}(\theta^*, x, y)$ term prevents out-of-distribution drift while enabling hard in-distribution samples. These analyses confirm our method spans natural difficulty ranges without semantic drift, explaining the sustained improvements in Sec. 5.3.

Qualitative Results. To demonstrate how our curriculum-based approach introduces diversity progressively, we visualize samples from each increment $\mathcal{I}_1 - \mathcal{I}_5$ for the Church class in Fig. 7. Compared to DiT's repetitive samples and IGD's clustering around similar structures, our method progressively introduces diversity across increments. Early increments capture simpler, easily learned structures that establish foundational features, while later increments introduce architectural complexity and challenging scenes. This curriculum design ensures that each increment contributes unique visual information rather than duplicating patterns from previous increments, maximizing sample utility across the distilled dataset.

6. Conclusion

Existing dataset distillation methods suffer from substantial redundancy, with disjoint subsets capturing 80–90% overlapping training signals. This arises because methods optimize for visual diversity or average trajectories without considering training signal complementarity across samples. We propose learnability-driven dataset distillation that constructs synthetic datasets incrementally, conditioning each stage on the current model state to generate complementary rather than redundant samples. Our learnability-guided diffusion automatically produces curriculum-aligned samples by balancing current-model informativeness with reference-model validity. This reduces redundancy by 39.1%, enables specialization across training phases, and achieves state-of-the-art results: ImageNet-1K (60.1%), ImageNette (87.2%), and ImageWoof (72.9%). Our framework establishes foundations for adaptive distillation methods that exploit staged learning dynamics to maximize sample efficiency.

References

[1] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. In *Forty-second International Conference on Machine Learning*, 2025. 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 2

[3] 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 (CVPR)*, pages 10718–10727, 2022. 1, 2

[4] Jeffrey A. Chan-Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. MGD3: Mode-guided dataset distillation using diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. 1, 2, 3, 5, 6

[5] Mingyang Chen, Jiawei Du, Bo Huang, Yi Wang, Xiaobo Zhang, and Wei Wang. Influence-guided diffusion for dataset distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3, 4, 5, 6

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

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

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[9] Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier Henaff. Data curation via joint example selection further accelerates multimodal learning. *Advances in Neural Information Processing Systems*, 37:141240–141260, 2024. 2, 4

[10] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15793–15803, 2024. 1, 2, 3, 5, 6

[11] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[12] Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7713–7722, 2024. 2, 7

[13] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *Transactions on Machine Learning Research*, 2024. 2

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

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

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

[17] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11102–11118, 2022. 5

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[19] Longzhen Li, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Generative Dataset Distillation: Balancing global structure and local details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Workshop*, pages 7664–7671, 2024. 6

[20] Zhiheng Ma, Anjia Cao, Funing Yang, Yihong Gong, and Xing Wei. Curriculum dataset distillation. *IEEE Transactions on Image Processing*, 2025. 2

[21] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR, 2022. 4

[22] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 2

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

[24] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z. Liu, Yuri A. Lawryshyn, and Konstantinos N. Plataniotis. DataDAM: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17097–17107, 2023. 2

[25] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16709–16718, 2024. 2

[26] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[27] Seungjae Shin, Heesun Bae, Donghyeok Shin, Weonyoung Joo, and Il-Chul Moon. Loss-curvature matching for dataset selection and condensation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. 1, 2

[28] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D4M: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5809–5818, 2024. 1, 2, 3

[29] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9390–9399, 2024. 2, 5, 6

[30] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, 2020. Association for Computational Linguistics. 7

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

[32] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 6

[33] David Junhao Zhang, Heng Wang, Chuhui Xue, Rui Yan, Wenqing Zhang, Song Bai, and Mike Zheng Shou. Dataset condensation via generative model. *arXiv preprint arXiv:2309.07698*, 2023. 1, 2, 3

[34] Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, and Joey Tianyi Zhou. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26223–26232, 2024. 2

[35] Bo Zhao and Hakan Bilen. Dataset condensation with gradient matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[36] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Workshop*, 2022. 2, 3

[37] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6514–6523, 2023. 2, 5

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

[39] Qing Zhou, Junyu Gao, and Qi Wang. Scale efficient training for large datasets. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20458–20467, 2025. 2