SODA: Selective Optimization with Deferred BN Alignment for Efficient Dataset Distillation

Xinyue Bi[1*], **Jiacheng Cui**[1*], **Yaxin Luo**[1], **Xinyi Shang**[1,2], **Jiacheng Liu**[1], **Xiaohan Zhao**[1], **Zhiqiang Shen**[1†]

[1]VILA Lab, MBZUAI [2]University College London

[*]Equal Contribution [†]Corresponding Author

Code github.com/XinyueBi/SODA

Abstract

Recent AI research has increasingly evolved along two complementary directions: model-centric learning, which improves architectures and training algorithms, and data-centric learning, which improves the quality and compression of training data. Within data-centric learning, dataset distillation on large-scale datasets has attracted growing attention, with decoupled distillation methods like SRe^{2}L, G-VBSM, LPQLD, FADRM have emerged as a representative paradigm. These methods typically synthesize compact datasets by matching the Batch Norm statistics of a teacher network, using BN alignment as an effective training objective. However, modern networks often contain many BN layers, and enforcing alignment at every layer introduces substantial computational redundancy. More importantly, we observe that not all BN layers provide equally informative supervision or consistent performance gains. Motivated by this, we propose SODA, a novel *Selective Optimization with Deferred BN Alignment* framework for efficient dataset distillation. SODA selectively identifies and optimizes informative BN alignment objectives while deferring unnecessary or low-value alignment computations, substantially improving both distillation efficiency and synthetic data quality. We further provide a theoretical analysis explaining why selective and deferred BN alignment can simultaneously reduce optimization cost and improve generalization. Extensive experiments across multiple datasets of CIFAR-100, Tiny-ImageNet, ImageNet-1K and its subsets thereof demonstrate that SODA achieves state-of-the-art performance while offering significantly improved computational efficiency over existing BN-matching-based dataset distillation methods, surpassing FADRM+ by +1.7% on ImageNet-1K IPC=10 under ResNet-18 while delivering a $1.54\times$ speedup.

1 Introduction

The remarkable success of recent model-centric study, specifically large-scale foundation models, spanning large language models such as ChatGPT [23, 1], Gemini [31], and DeepSeek [18], as well as multimodal systems bridging vision, language and speech, has established deep models as the dominant paradigm across a broad range of artificial intelligence tasks. Central to this model-centric philosophy, performance improvements are predominantly pursued through architectural innovation and the systematic scaling of model capacity. However, this paradigm inherently couples model performance to data scale: as model capacity grows, so too does the demand for training corpora of commensurate size and quality, rendering data acquisition, storage, and repeated full-dataset optimization increasingly burdensome in both computational and economic terms. This fundamental tension has motivated a growing body of data-centric research, which reframes the optimization

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

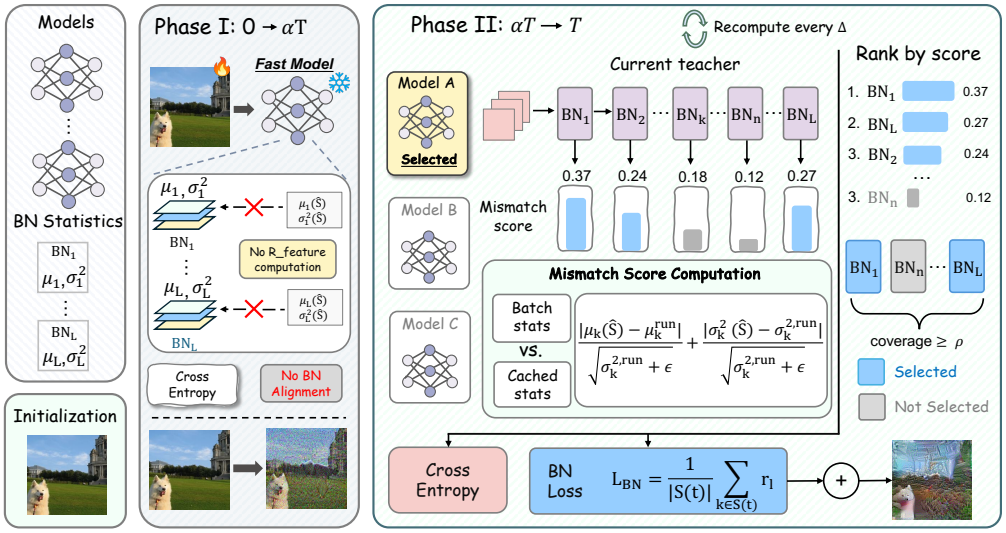

Figure 1: Overview of **SODA**. Phase I ($0 \rightarrow \alpha T$): BN alignment is deferred entirely, with optimization driven solely by cross-entropy loss under f_{fast}, allowing class-discriminative structure to form unimpeded. Phase II ($\alpha T \rightarrow T$): full ensemble supervision is activated alongside *Selective BN Alignment*, which scores each BN layer by its normalized mean-variance mismatch and selects the minimal subset satisfying coverage ratio ρ. Scores are recomputed every Δ iterations.

objective from enlarging model capacity to distilling the informational content of large-scale datasets into more compact and efficient representations, achieving competitive downstream performance at substantially reduced data volume and training overhead.

Dataset distillation [34] has emerged as a principled instantiation of this data-centric perspective, aiming to synthesize a small surrogate set $\mathcal{S}$ such that any model trained on $\mathcal{S}$ alone generalizes comparably to one trained on the full dataset $\mathcal{D}$. Recent efforts have focused on scaling dataset distillation to large-scale benchmarks such as ImageNet-1K [40], with uni-level optimization methods emerging as the dominant paradigm. By decoupling synthetic image optimization from the original dataset and instead leveraging Batch Normalization (BN) statistics as a surrogate alignment signal, uni-level methods achieve tractable optimization at scale while maintaining competitive distillation quality. Despite these advances, we identify two critical yet overlooked limitations of BN alignment, affecting both efficiency and distillation quality. *(i) Computational inefficiency.* The backward pass of BN alignment introduces low-arithmetic-intensity operations whose compute and memory traffic both scale with activation volume, consistent with a memory-bandwidth bottleneck; BN alignment accounts for 39.1–59.4% of total backward latency across six profiled backbones (Table 1). *(ii) Over-constraining.* Uniform alignment across all BN layers acts analogously to an overly large weight decay, imposing a distributional bias on synthetic data that degrades generalization.

To address both bottlenecks, we propose **SODA** (**S**elective **O**ptimization with **D**eferred BN **A**lignment), a plug-in efficiency framework for existing uni-level dataset distillation methods. **SODA** comprises two complementary components. *Deferred BN Alignment* suppresses BN alignment during the early optimization phase, when CE-gradient dominance indicates class-discriminative structure is not yet established. Deferring it to the later phase eliminates premature distributional bias and removes its backward computation in early iterations, substantially reducing per-image generation time without accuracy loss. *Selective BN Alignment* further restricts $\mathcal{L}_{\text{BN}}$ at each active iteration to the minimal subset of BN layers whose cumulative normalized mismatch scores meet a target coverage threshold. By focusing updates on the most mismatched layers and amortizing scoring cost via a reuse interval, it reduces backward computation and mitigates over-constraining, improving distillation quality with negligible overhead. Our main contributions are as follows:

- We identify and formally characterize two overlooked bottlenecks of BN alignment in existing uni-level dataset distillation: a structural computational inefficiency and over-constraining alignment across all BN layers.

- We propose **SODA**, a plug-in framework comprising *Deferred BN Alignment* and *Selective BN Alignment*, which address these bottlenecks in the temporal and spatial dimensions respectively, with negligible implementation overhead.
- **SODA** achieves state-of-the-art results across extensive benchmarks, including CIFAR-100 and ImageNet-1K, delivering significant computational speedup while consistently improving distillation quality.

2 Related Work

Dataset Distillation. Dataset Distillation aims to synthesize a compact dataset that retains the critical information of a larger original dataset, enabling efficient training while maintaining performance comparable to the full dataset. Overall, the matching criteria include *Meta-Model Matching* [34, 22, 21, 46, 9, 11], *Gradient Matching* [44, 42, 16, 12, 45], *Trajectory Matching* [2, 7, 3, 10], Distribution Matching [43, 33, 20, 15, 24, 29, 37], and *Uni-level Global Statistics Matching* [40, 39, 5, 35, 4, 17, 26, 36, 6]. Dataset distillation on large-scale datasets has recently attracted significant attention from the community. For a detailed overview, we refer readers to the survey works [25, 19] on this topic.

Efficient Dataset Distillation. Several methods improve the efficiency of dataset distillation. TESLA [7] accelerates MTT [2] via batched gradient computation, and DM [43] sidesteps bi-level optimization by directly matching feature distributions. SRe^{2}L [40] first introduces BN statistics as a surrogate alignment signal within a uni-level framework, on top of which EDC [27] adds real data initialization and FADRM [4] introduces training-level refinements. Despite these advances, the overhead of BN alignment itself remains overlooked.

Selective BN Layer. DeepInversion [38] aligns synthetic activations to BN statistics across all layers, a convention inherited by uni-level dataset distillation methods. DSG [41] relaxes this with static, hand-crafted per-layer weights in the data-free quantization setting, while PMI [32] decomposes inversion layer-by-layer for data-free continual learning but selects layers by optimization efficiency rather than distributional importance. Unlike these methods, **SODA** introduces the first data-dependent selection strategy, dynamically scoring each BN layer and selecting the minimal subset covering a target fraction of the total mismatch signal.

3 Approach

Preliminaries. Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ denote the full training set drawn from an underlying data distribution $p(x, y)$. Storing and training on $\mathcal{D}$ can be prohibitively expensive when N is large, motivating the search for a more compact alternative. The goal of dataset distillation is therefore to construct a small surrogate set $\hat{\mathcal{S}} = (\hat{x}_k, \hat{y}_k)_{k=1}^{M}$, with $M \ll N$, whose synthetic samples may not correspond to any real observation, yet any model trained on $\hat{\mathcal{S}}$ alone generalizes as well as one trained on the full dataset $\mathcal{D}$. Formally, we seek:

$$\hat{\mathcal{S}}^* = \arg\min_{\hat{\mathcal{S}}} |\hat{\mathcal{S}}| \quad \text{subject to} \quad \left| \mathbb{E}_{(x,y)\sim p}\Big[\ell\big(f_{\hat{\mathcal{S}}}(x), y\big)\Big] - \mathbb{E}_{(x,y)\sim p}\Big[\ell\big(f_{\mathcal{D}}(x), y\big)\Big] \right| \leq \varepsilon, \tag{1}$$

where $f_{\mathcal{S}}$ and $f_{\mathcal{D}}$ denote models trained on $\hat{\mathcal{S}}$ and $\mathcal{D}$, ℓ is a task-specific loss, and $\varepsilon \geq 0$ is the tolerated performance gap. A notable class of methods, referred to as *uni-level optimization*, performs synthetic data optimization without access to the original dataset by leveraging Batch Normalization (BN) statistics as a surrogate alignment signal. While this decoupling enables scalability, we identify BN alignment as a critical yet overlooked bottleneck that limits both efficiency and distillation quality, formally characterized below.

Inefficiency of BN Alignment. The BN alignment loss matches the mean and variance of feature activations against running statistics across each layer:

$$\mathcal{L}_{\text{BN}} = \sum_{l} \left(\left\| \mu_l(\hat{\mathcal{S}}) - \mu_l^{\text{run}} \right\|_2^2 + \left\| \sigma_l^2(\hat{\mathcal{S}}) - \sigma_l^{2,\text{run}} \right\|_2^2 \right) \tag{2}$$

During backpropagation, this introduces additional operations absent in CE-only optimization, specifically `VarBackward`, `MeanBackward`, and associated gradient accumulation. Let c_{var} and

c_{mean} be their per-element FLOP counts, d_l the feature map size at layer l, and ks the memory traffic per activation element. The estimated arithmetic intensity is:

$$\mathcal{I}_{\text{BN}} = \frac{\mathcal{F}_{\text{BN}}}{\mathcal{M}_{\text{BN}}} \approx \frac{(c_{\text{var}} + c_{\text{mean}})B\sum_l d_l}{ksB\sum_l d_l} = \frac{c_{\text{var}} + c_{\text{mean}}}{ks}. \tag{3}$$

Here, s is the byte width per element and k accounts for memory reads and writes. The bandwidth-bound regime corresponds to $\mathcal{I}_{\text{BN}} < \mathcal{P}/\mathcal{B}$, where $\mathcal{P}$ and $\mathcal{B}$ denote compute throughput and memory bandwidth. For float32 ($s = 4$), taking $c_{\text{var}} \approx 7$, $c_{\text{mean}} \approx 2$, and $k = 1$ gives $\mathcal{I}_{\text{BN}} \approx 2.25$ FLOPs/byte. This is approximately $36\times$ below the RTX 4090 FP32 ridge point (81.9 FLOPs/byte). BN alignment thus lies well within the bandwidth-bound region of this model. Under fixed per-element costs, activation volume cancels in Eq. (3), making arithmetic intensity invariant to batch size and network depth. Its share of backward latency, however, depends on the architecture's balance between computation and activation volume. We characterize this balance by

$$r := \frac{\text{backbone multiply–accumulate operations}}{\text{BN-input activation volume}} = \frac{\text{MACs}}{\sum_{l=1}^{L} C_l H_l W_l}, \tag{4}$$

where L is the number of BN layers and C_l, H_l, and W_l denote the input channels, height, and width of BN layer l. MACs and activation volume are computed per sample at a common input resolution. BN alignment work scales with activation volume, whereas convolutional arithmetic scales with MACs. Thus, lower r indicates more alignment work relative to backbone computation and tends to yield a larger alignment latency share. Table 1 shows that BN alignment accounts for 39.1–59.4% of backward latency across six backbones. The ratio r captures the overall trend, while exact latency also depends on kernel efficiency and memory access behavior.

Backbone	MACs (G)	Act. (M)	r	Time (%)
MobileNetV2	0.31	6.68	47	59.4
DenseNet-121	2.87	15.67	183	56.0
ResNet-50	4.11	11.11	370	51.4
ResNet-101	7.83	16.23	483	49.5
ResNet-18	1.82	2.48	732	40.0
ResNet-34	3.67	3.74	982	39.1

Table 1: BN alignment overhead. Time: percentage of total backward latency; Act.: summed BN-input activations.

Redundancy of Full BN Alignment. Beyond its computational cost, we identify a second limitation of BN alignment. We interpret $\mathcal{L}_{\text{BN}}$ as a regularizer anchoring synthetic images to the teacher's feature distribution. Applying it across all BN layers is analogous to an excessively large weight decay: it risks over-constraining the feature distribution. As shown in Fig. 2, a large fraction of BN layer pairs exhibit high pairwise cosine similarity in their image gradients, indicating that their alignment signals are largely correlated. Simultaneously constraining all of them therefore amplifies this over-constraining effect, imposing a heavier distributional bias than is necessary for effective distillation. These two bottlenecks jointly motivate a principled, data-driven approach to selective and deferred BN alignment.

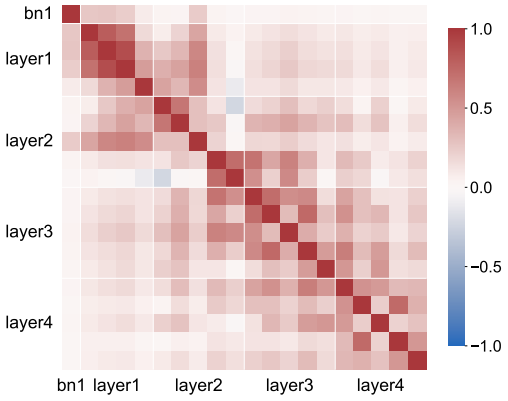

Figure 2: Pairwise cosine similarity of $\nabla_{\hat{x}}\mathcal{L}_{\text{BN}}$ across BN layers, revealing largely correlated alignment signals.

3.1 Overview of SODA

The proposed **SODA**, illustrated in Fig. 1, directly addresses both bottlenecks. *Deferred BN Alignment* suppresses $\mathcal{L}_{\text{BN}}$ during early optimization, allowing unimpeded semantic formation. *Selective BN Alignment* restricts alignment to the most informative BN layers at each iteration, simultaneously reducing computation and alleviating over-constraining. Together, these two components ensure both computational efficiency and distillation fidelity throughout optimization.

3.2 Deferred BN Alignment

Although $\mathcal{L}_{\text{BN}}$ is essential for anchoring synthetic images to the teacher's feature distribution, its effectiveness is inherently stage-dependent. Define the gradient ratio $\gamma(t) := \|\nabla_{\hat{x}}\mathcal{L}_{\text{CE}}^{(t)}\| \,/\, \|\nabla_{\hat{x}}\mathcal{L}_{\text{BN}}^{(t)}\|$.

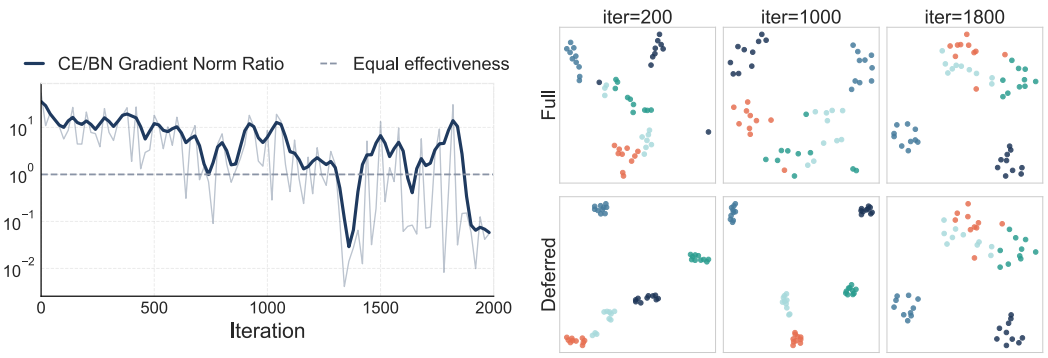

Figure 3: **Left:** CE/BN gradient norm ratio throughout optimization. Above the dashed line: CE-dominant semantic formation phase; below: BN-dominant distributional refinement phase. **Right:** t-SNE visualization of synthetic image features under Full BN alignment and Deferred BN alignment.

As shown in Fig. 3 (left), $\gamma(t) \gg 1$ during early iterations and monotonically decreases, crossing unity at some t^*. This marks a natural phase transition from CE-driven semantic formation to BN-driven distributional refinement. Introducing $\mathcal{L}_{\text{BN}}$ during the early phase ($\gamma(t) \gg 1$) is therefore premature: despite its small gradient magnitude, BN alignment imposes a persistent directional bias toward $\{\mu_l^{\text{run}}, \sigma_l^{2,\text{run}}\}$ before class-discriminative structure has been established. As shown in Fig. 3 (right), full BN alignment slows intra-class convergence, while suppressing $\mathcal{L}_{\text{BN}}$ lets $\mathcal{L}_{\text{CE}}$ cluster features unimpeded. Both trajectories converge to equivalent intra-class variance at termination, confirming deferral preserves final fidelity. Formally, given T iterations, we define a skip ratio $\alpha \in [0, 1)$ such that $\mathcal{L}_{\text{BN}}$ is suppressed for the first $\lfloor \alpha T \rfloor$. The total optimization loss at iteration t is:

$$\mathcal{L}^{(t)} = \begin{cases} \mathcal{L}_{\text{CE}}^{(t)} & t < \lfloor \alpha T \rfloor \\ \mathcal{L}_{\text{CE}}^{(t)} + \lambda \mathcal{L}_{\text{BN}}^{(t)} & t \geq \lfloor \alpha T \rfloor \end{cases} \tag{5}$$

where λ controls the BN alignment weight. This formulation decouples semantic formation from distributional constraint, ensuring $\mathcal{L}_{\text{BN}}$ intervenes only after coherent representations emerge. Since the early phase relies solely on CE supervision, we restrict it to the lightest teacher f_{fast}; full ensemble supervision is deferred to the late phase alongside $\mathcal{L}_{\text{BN}}$, with no performance degradation (Table 6d).

3.3 Selective BN Alignment

While Deferred BN Alignment eliminates redundant BN constraints in the temporal dimension, full alignment across all BN layers at each active iteration remains both computationally expensive and over-constraining. We therefore propose Selective BN Alignment, which restricts $\mathcal{L}_{\text{BN}}$ to a data-dependent subset of layers at each iteration, addressing both bottlenecks simultaneously.

Data-Dependent Layer Importance Scoring. At each iteration t, we assign an importance score to each BN layer l based on the normalized mismatch between the *current synthetic batch statistics* and the teacher's running statistics:

$$s_l^{(t)} = \frac{1}{C_l} \sum_{c=1}^{C_l} \left(\frac{|\mu_l^{(t)}(c) - \mu_l^{\text{run}}(c)|}{\sqrt{\sigma_l^{2,\text{run}}(c) + \varepsilon}} + \frac{|\sigma_l^{2,(t)}(c) - \sigma_l^{2,\text{run}}(c)|}{\sqrt{\sigma_l^{2,\text{run}}(c) + \varepsilon}} \right) \tag{6}$$

where $\mu_l^{(t)} \in \mathbb{R}^{C_l}$ and $\sigma_l^{2,(t)} \in \mathbb{R}^{C_l}$ denote the batch mean and variance of layer l at iteration t, computed over the batch and spatial dimensions of the feature map $\mathbf{F}_l^{(t)} \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$. The normalization by $\sqrt{\sigma_l^{2,\text{run}}(c) + \varepsilon}$ makes the score scale-invariant across layers with different activation magnitudes, ensuring that deeper layers with larger feature norms do not systematically dominate the selection. A higher score $s_l^{(t)}$ indicates a larger distributional gap at layer l, signaling that explicit alignment at this layer yields the greatest marginal benefit to the optimization.

Computational Cost of Scoring. Computing $\{s_l^{(t)}\}_{l=1}^{L}$ requires reading batch statistics $\{\mu_l^{(t)}, \sigma_l^{2,(t)}\}$ already accumulated during the forward pass, and comparing them against cached running statistics

$\{\mu_l^{\text{run}}, \sigma_l^{2,\text{run}}\}$. The arithmetic cost is $\mathcal{O}(\sum_l C_l)$, which is negligible relative to the forward and backward passes. However, at high iteration counts, the cumulative overhead of scoring at every step can be non-trivial. We therefore introduce a *score interval* $\Delta \in \mathbb{Z}^+$, recomputing scores only every Δ iterations and reusing the cached selection $\mathcal{S}^{(t-1)}$ otherwise:

$$\mathcal{S}^{(t)} = \begin{cases} \text{SelectLayers}\Big(\{s_l^{(t)}\}, \rho\Big) & t \equiv 0 \pmod{\Delta} \\ \mathcal{S}^{(t-1)} & \text{otherwise} \end{cases} \tag{7}$$

This amortizes the scoring cost over Δ iterations, reducing the per-iteration overhead to $\mathcal{O}(\sum_l C_l/\Delta)$.

Adaptive BN Layer Selection. Given the per-layer scores $\{s_l^{(t)}\}_{l=1}^L$, we select the minimal subset $\mathcal{S}^{(t)} \subseteq \{1, \ldots, L\}$ of layers that collectively account for a target coverage ratio ρ:

$$\mathcal{S}^{(t)} = \underset{\mathcal{S}\subseteq\{1,\ldots,L\}}{\arg\min} |\mathcal{S}| \quad \text{subject to} \quad \frac{\sum_{l\in\mathcal{S}} s_l^{(t)}}{\sum_{l=1}^L s_l^{(t)}} \geq \rho \tag{8}$$

In practice, $\mathcal{S}^{(t)}$ is obtained in $\mathcal{O}(L \log L)$ by sorting layers in descending order of $s_l^{(t)}$ and greedily accumulating until the coverage threshold ρ is satisfied. The BN alignment loss is then restricted to the selected subset, with a compensation factor $\frac{L}{|\mathcal{S}^{(t)}|}$ to preserve the effective loss scale:

$$\mathcal{L}_{\text{BN}}^{(t)} = \frac{L}{|\mathcal{S}^{(t)}|} \sum_{l\in\mathcal{S}^{(t)}} r_l^{(t)} \tag{9}$$

where $r_l^{(t)}$ denotes the per-layer BN alignment loss at iteration t. By concentrating gradient updates on the most mismatched layers and amortizing scoring cost via Δ, Selective BN Alignment simultaneously reduces backward computation and alleviates over-constraining with negligible scoring overhead. Fig. 4 validates this design: although only frequently selected layers receive explicit gradient updates, the mismatch of rarely selected layers decreases concurrently, indicating that alignment on top-mismatch layers implicitly propagates to the remaining layers. This implicit alignment effect justifies coverage-based selection; constraining every layer is redundant, since the optimization signal from high-score layers suffices to drive distributional convergence.

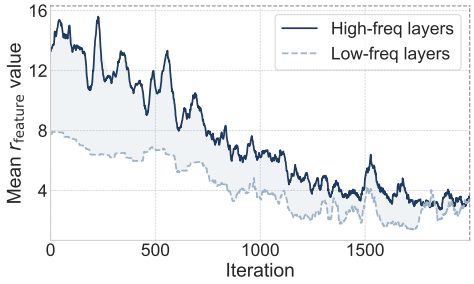

Figure 4: Mean r_{feature} of frequently-selected layers (those with selection ratios above the median) and rarely-selected layers (those with selection ratios at or below the median).

3.4 Theoretical Analysis of Selective BN Alignment

We analyze Selective BN Alignment from a generalization perspective: excluding low-mismatch layers lowers the population risk of the student trained on the distilled data, since their gradients are less shared across teachers and transfer poorly to the student (Theorem 1). Beyond this perspective, Appendix A also provides a training perspective, showing that selective alignment cannot worsen the constrained CE optimum (Proposition 1) and still aligns the skipped layers implicitly (Proposition 2).

Formally, let $\hat{g}^{(j)}(\rho) := \nabla_{\hat{x}} \mathcal{L}_{\text{BN}}^{(j)}(\hat{x}^{(t)}; \mathcal{S}^{(j)}(\rho))$ be the BN gradient of teacher j at coverage ρ ($\rho = 1$: full alignment) and $k = k(t)$ the teacher active at Phase-II iteration t, so the applied update is $\hat{g}_{\mathcal{S}}(\rho) := \hat{g}^{(k)}(\rho)$. Using the other teachers' full-coverage gradients as a ρ-independent reference, $\mathcal{U}_k := \text{span}\{\hat{g}^{(j)}(1) : j \neq k\}$, we split the update as

$$\hat{g}_{\mathcal{S}}(\rho) = \underbrace{P_{\mathcal{U}_k} \hat{g}_{\mathcal{S}}(\rho)}_{\text{shared}} + \underbrace{P_{\mathcal{U}_k^\perp} \hat{g}_{\mathcal{S}}(\rho)}_{\text{teacher-specific}}, \qquad \pi_k(\rho) := \frac{\|P_{\mathcal{U}_k^\perp} \hat{g}_{\mathcal{S}}(\rho)\|^2}{\|\hat{g}_{\mathcal{S}}(\rho)\|^2}. \tag{10}$$

Assumption 1 (Energy-proportional transfer)**.** For a student f with population risk $\mathcal{E}_f(\hat{x})$ after training on $\hat{x}$ and architectural distance τ_f to the teacher pool, the shared and teacher-specific components transfer at an architecture-independent rate $\omega_0 > 0$ and a reduced rate $\omega(\tau_f) < \omega_0$, respectively, each in proportion to its energy share and independently of ρ. With $\Gamma_\rho := \|\nabla \mathcal{E}_f\| \, \|\hat{g}_{\mathcal{S}}(\rho)\|$,

$$\big\langle \nabla \mathcal{E}_f, P_{\mathcal{U}_k} \hat{g}_{\mathcal{S}}(\rho) \big\rangle = \omega_0 \, \Gamma_\rho \big(1 - \pi_k(\rho)\big), \qquad \big\langle \nabla \mathcal{E}_f, P_{\mathcal{U}_k^\perp} \hat{g}_{\mathcal{S}}(\rho) \big\rangle = \omega(\tau_f) \, \Gamma_\rho \, \pi_k(\rho). \tag{11}$$

Dataset	IPC (Ratio)	ResNet-18					ResNet-50					ResNet-101				
		RDED	EDC	DELT	FADRM+	**SODA**	RDED	EDC	DELT	FADRM+	**SODA**	RDED	EDC	DELT	FADRM+	**SODA**
CIFAR-100	1 (0.2%)	10.2±0.3	39.7±0.1	29.1±0.6	40.4±0.2	**41.4**±0.4	9.5±1.0	36.1±0.5	22.5±0.2	37.7±1.7	**43.5**±0.3	8.2±0.3	32.3±0.3	23.9±0.2	40.8±0.7	**41.4**±0.7
	10 (2.0%)	43.1±0.2	63.7±0.3	57.2±0.5	67.7±0.4	**68.5**±0.4	40.4±0.8	62.1±0.1	57.2±0.1	66.4±0.7	**69.5**±0.1	39.6±0.9	61.7±0.1	57.7±0.4	68.0±0.7	**69.6**±0.4
	50 (10.0%)	64.2±1.8	68.6±0.2	59.3±0.2	71.4±0.4	**72.2**±0.2	64.0±0.2	69.4±0.3	60.4±0.1	71.2±0.6	**72.8**±0.1	63.7±0.4	68.5±0.1	60.5±0.1	71.0±1.1	**73.1**±0.2
	Whole Dataset			78.9					79.9					79.5		
Tiny-ImageNet	1 (0.2%)	9.7±0.4	39.2±0.4	36.4±0.6	39.3±0.8	**42.8**±0.5	7.5±0.5	35.9±0.2	33.0±0.8	37.3±1.5	**41.2**±0.2	3.8±0.1	40.6±0.3	31.4±0.4	41.9±1.3	**43.5**±0.4
	10 (2.0%)	41.9±0.2	51.2±0.5	43.0±0.1	52.3±0.2	**53.9**±0.3	41.3±2.1	50.2±0.3	49.7±0.4	53.1±0.4	**55.5**±0.3	22.9±0.3	51.6±0.2	42.9±0.9	53.6±0.3	**55.9**±0.1
	50 (10.0%)	58.2±0.1	57.2±0.2	55.7±0.1	58.6±0.3	**60.0**±0.2	45.6±0.1	58.8±0.4	55.8±0.2	59.5±0.6	**61.6**±0.4	41.2±0.4	58.6±0.1	58.5±0.3	60.7±0.5	**61.4**±0.9
	Whole Dataset			68.9					71.5					70.6		
ImageNette	1 (0.1%)	35.8±1.0	-	24.1±1.8	40.0±0.7	**41.1**±0.7	23.4±2.7	-	24.6±0.7	28.6±2.4	**34.3**±0.9	25.1±2.7	-	19.4±1.7	28.4±0.8	**29.9**±0.5
	10 (1.0%)	61.4±0.4	-	66.0±1.4	69.5±0.2	**70.9**±0.4	55.8±0.7	-	57.0±0.8	68.0±0.1	**69.8**±1.3	54.0±0.4	-	55.4±6.2	66.5±0.5	**67.0**±0.2
	50 (5.2%)	80.4±0.4	-	**88.2**±1.2	84.5±0.1	85.8±0.1	81.3±0.7	-	82.4±0.2	84.2±0.9	**86.3**±0.1	75.0±1.2	-	83.3±1.1	82.1±0.4	**86.1**±0.1
	Whole Dataset			93.8					89.8					89.3		
ImageWoof	1 (0.1%)	20.8±1.2	-	18.8±1.1	23.1±0.3	**30.6**±1.5	18.9±1.3	-	18.9±0.6	22.2±1.9	**28.0**±0.9	19.6±1.8	-	16.3±1.1	20.1±1.2	**25.8**±0.5
	10 (1.1%)	38.5±2.1	-	53.8±0.3	57.5±0.9	**57.7**±0.4	33.4±1.5	-	41.7±0.3	52.0±1.6	**53.7**±0.9	31.3±1.3	-	36.3±0.9	49.3±1.7	**50.3**±0.9
	50 (5.3%)	68.5±0.7	-	61.9±0.6	69.3±2.4	**74.1**±0.6	66.9±1.6	-	60.8±0.1	69.8±1.3	**74.9**±0.2	59.1±0.7	-	59.4±0.4	68.8±1.3	**73.6**±0.2
	Whole Dataset			88.2					77.8					82.7		
ImageNet-1K	1 (0.1%)	6.6±0.2	12.8±0.1	-	14.4±0.3	**17.6**±0.4	6.0±1.5	13.3±0.3	-	15.7±1.2	**23.4**±0.1	5.9±0.4	12.2±0.2	-	12.0±1.6	**19.7**±1.2
	10 (0.8%)	42.0±0.1	48.6±0.3	46.1±0.4	50.6±0.3	**52.3**±0.3	46.8±2.1	54.1±0.2	47.5±0.5	56.4±0.8	**58.3**±0.4	48.3±1.0	51.7±0.3	48.5±1.6	57.1±1.0	**58.7**±0.7
	50 (3.9%)	56.5±0.1	58.0±0.2	59.2±0.4	61.2±0.3	**62.5**±0.3	63.6±1.1	64.3±0.2	65.3±0.1	66.6±0.2	**67.9**±0.1	61.2±0.4	64.9±0.2	66.1±0.5	67.1±0.2	**68.5**±0.1
	Whole Dataset			72.3					78.6					79.8		

Table 2: Post-evaluation performance comparison with state-of-the-art baselines. All experiments adhere to the training protocol established in EDC [27]. Specifically, models are trained for 300 epochs on Tiny-ImageNet (IPC = 10, 50), ImageNet-1K, and its subsets, and for 1,000 epochs on CIFAR-100 and Tiny-ImageNet (IPC = 1). This evaluation protocol ensures fair comparisons while maintaining methodological consistency across all experiments. Results are reported as mean ± std over 3 random seeds. **Bold** indicates the best result and underline indicates the second best.

Theorem 1 (Proof in Appendix A.1)**.** *Under Assumption 1, let $\mathcal{E}_f \in C^2$ near $\hat{x}^{(t)}$, $\|\hat{g}_{\mathcal{S}}(\rho)\| = \|\hat{g}_{\mathcal{S}}(1)\| =: G$, and $\hat{x}^{(t+1)}(\rho) := \hat{x}^{(t)} - \eta\nabla_{\hat{x}}(\mathcal{L}_{CE} + \lambda\mathcal{L}_{BN})$ be one Phase-II step at coverage ρ. Then*

$$\mathcal{E}_f\big(\hat{x}^{(t+1)}(1)\big) - \mathcal{E}_f\big(\hat{x}^{(t+1)}(\rho)\big) = \eta\lambda G\,\|\nabla\mathcal{E}_f\|\,\underbrace{\big(\pi_k(1) - \pi_k(\rho)\big)}_{\substack{\text{extra teacher-}\\ \text{specific signal}}}\,\underbrace{\big(\omega_0 - \omega(\tau_f)\big)}_{\substack{\text{teacher–student}\\ \text{mismatch}}} + O(\eta^2). \quad (12)$$

Hence, for any student, selective alignment lowers the risk at leading order whenever $\pi_k(\rho) < \pi_k(1)$.

3.5 Theoretical Efficiency Analysis

Let $\zeta(\rho) = \mathbb{E}_t[|\mathcal{S}^{(t)}|/L]$ be the expected fraction of BN layers selected per iteration. Let N denote the number of teacher models cycling in round-robin. Define $\mathcal{C}_{\mathrm{CE}}^{\mathrm{full}} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{C}_{\mathrm{CE}}^{(k)}$ as the average per-iteration CE cost under full ensemble supervision, and $\mathcal{C}_{\mathrm{BN}}^{\mathrm{full}} = \frac{1}{N}\sum_{k=1}^{N}(\mathcal{C}_{\mathrm{BN}}^{\mathrm{fwd},(k)} + \mathcal{C}_{\mathrm{BN}}^{\mathrm{bwd},(k)})$ the corresponding average BN alignment cost, where skipping a layer eliminates both forward hooks and backward gradients. In the early phase, supervision is restricted to f_{fast} alone; let $\beta = \mathcal{C}_{\mathrm{CE}}^{(\mathrm{fast})}/\mathcal{C}_{\mathrm{CE}}^{\mathrm{full}} \le 1$ denote the ratio of fast-model CE cost to full-ensemble CE cost, so the early-phase CE cost is $\mathcal{C}_{\mathrm{CE}}^{\mathrm{fast}} = \beta\,\mathcal{C}_{\mathrm{CE}}^{\mathrm{full}}$. The total costs are:

$$\mathcal{C}^{\mathrm{base}} = T\big(\mathcal{C}_{\mathrm{CE}}^{\mathrm{full}} + \mathcal{C}_{\mathrm{BN}}^{\mathrm{full}}\big), \quad (13)$$

$$\mathcal{C}^{\mathrm{ours}} = \alpha T\,\mathcal{C}_{\mathrm{CE}}^{\mathrm{fast}} + (1-\alpha)T\,\mathcal{C}_{\mathrm{CE}}^{\mathrm{full}} + (1-\alpha)T\cdot\zeta(\rho)\cdot\mathcal{C}_{\mathrm{BN}}^{\mathrm{full}} + \tfrac{(1-\alpha)T}{\Delta}\cdot\mathcal{C}_{\mathrm{score}}, \quad (14)$$

where Δ is the score reuse interval. Defining the BN cost fraction $\phi = \mathcal{C}_{\mathrm{BN}}^{\mathrm{full}}/(\mathcal{C}^{\mathrm{base}}/T)$ and the scoring cost fraction $\psi = \mathcal{C}_{\mathrm{score}}/(\mathcal{C}^{\mathrm{base}}/T)$, the speedup ratio is:

$$\mathcal{R}(\alpha,\beta,\rho,\Delta) = \frac{1}{(1-\phi)\big[1-\alpha(1-\beta)\big] + (1-\alpha)\big[\zeta(\rho)\phi + \psi/\Delta\big]}. \quad (15)$$

The savings decompose into two complementary terms: $\alpha(1-\beta)$ from lightweight early-phase supervision, which reduces the CE cost fraction $(1-\phi)$; and $1-(1-\alpha)\zeta(\rho)$ from deferred and selective BN alignment, which reduces the BN cost fraction ϕ.

4 Experiments

4.1 Datasets and Experimental Setup

Datasets. To assess scalability across image resolutions and dataset scales, we evaluate on CIFAR-100 (32×32) [13], Tiny-ImageNet (64×64) [14], ImageNet-1K (224×224) [8], and its subsets.

Baseline Methods. We compare against representative dataset distillation methods spanning multiple paradigms: RDED [30], EDC [27], DELT [28], and FADRM+ [4]. Among these, FADRM+ is a residual-based dataset distillation method that represents the current state of the art in uni-level dataset distillation, serving as our primary baseline. As **SODA** is a plug-in framework, we adopt FADRM+ as the default backbone for instantiation and report results under this setting unless otherwise specified.

4.2 Main Results

As shown in Table 2, **SODA** achieves the best performance in the majority of settings. On ImageNet-1K, **SODA** outperforms FADRM+ by up to +7.7% at IPC=1 (ResNet-50) and 1.3–1.9% at higher IPC settings. Similar gains are observed on CIFAR-100 and Tiny-ImageNet. On the fine-grained benchmark ImageWoof, **SODA** achieves margins of up to +5.1% over FADRM+ at IPC=50, highlighting its effectiveness on fine-grained tasks with high inter-class visual similarity. Notably, these gains are achieved while aligning only a data-dependent subset of BN layers, confirming that selective alignment avoids over-constraining the synthetic distribution and yields better distillation quality.

Efficiency Comparison. As shown in Table 3, **SODA** achieves a per-image generation time of 0.71s and a total distillation cost of just 9.8 hours for the full IPC=50 set, representing a 1.54× speedup over FADRM+, 7× over EDC. Crucially, this efficiency gain does not come at the cost of performance: **SODA** outperforms all baselines, demonstrating that our data-dependent selective and deferred BN alignment eliminates redundant computation without sacrificing distillation quality.

Method	Time Cost (s)	Total Time (hrs)
G-VBSM	17.28	240.0
CV-DD	8.20	113.9
EDC	4.99	69.3
FADRM+	1.09	15.1
SODA	**0.71**	**9.8**

Table 3: Per-image generation time and total generation time on ImageNet-1K.

Theoretical and Empirical Efficiency Gain. To validate the cost model derived in Eq. (15), we instantiate $R(\alpha, \beta, \rho, \Delta)$ under the experimental configurations used for CIFAR-100 and ImageNet-1K, and compare the predicted speedup against wall-clock measurements. As reported in Table 4, the theoretical model closely tracks empirical performance, with a discrepancy of less than $0.03\times$ in both settings ($1.77\times$ vs. $1.80\times$ on CIFAR-100; $1.53\times$ vs. $1.54\times$ on ImageNet-1K), confirming that the cost model accurately captures the dominant sources of overhead.

α	β	$\zeta(\rho)$	Δ	ϕ	ψ	Theoretical Gain	Empirical Gain
0.30	0.40	0.40	5	0.50	0.1	1.77×	1.80×
0.30	0.56	0.40	5	0.38	0.05	1.53×	1.54×

Table 4: Theoretical vs. empirical speedup of SODA under two dataset configurations. Top: CIFAR-100; Bottom: ImageNet-1K.

Cross-Architecture Generalization. The ability to generalize across diverse network architectures serves as a stringent benchmark for practical applicability. As shown in Table 5, **SODA** consistently surpasses all competing methods across nine architectures spanning a wide range of model sizes and design paradigms, demonstrating the broader applicability and robustness of our approach across both CNN-based and Transformer-based architectures.

Model	#Params	RDED	EDC	FADRM+	**SODA**
ResNet-18	11.7M	42.0	48.6	50.6	**52.3**
ResNet-50	25.6M	46.8	54.1	56.4	**58.3**
ResNet-101	44.5M	48.3	51.7	57.1	**58.7**
EfficientNet-B0	5.3M	42.8	51.1	51.9	**52.3**
MobileNetV2	3.4M	34.4	45.0	45.5	**46.6**
ShuffleNetV2-0.5x	1.4M	19.6	29.8	30.2	**32.9**
Swin-Tiny	28.0M	29.2	38.3	39.1	**41.0**
ConvNext-Tiny	29.0M	48.3	54.4	54.0	**54.8**
DeiT-Tiny	5.7M	14.0	18.4	16.5	**19.2**

Table 5: Cross-architecture generalization on ImageNet-1K IPC=10.

4.3 Ablation Study

Ablation on SODA Components. Table 6a ablates each component. *Deferred BN* (with f_{fast} in the early phase) preserves accuracy while reducing per-image generation time by 20–29%, confirming that early-phase $\mathcal{L}_{\text{BN}}$ is dispensable. Adding *Selective BN* yields further speedup and a +0.7% gain on ImageNet-1K, indicating that full alignment over-constrains the synthetic distribution. Together, the two components achieve speedup over FADRM+ with improved distillation quality.

Effect of Score Interval. Table 6b shows that $\Delta = 5$ achieves the best performance on both benchmarks. A smaller Δ causes unstable layer switching, preventing selected layers from being sufficiently aligned before the selection changes, while a larger Δ causes stale selections that fail to track the evolving synthetic distribution.

Effect of Skip Ratio. Table 6c ablates the skip ratio α with ρ fixed at 0.5. $\alpha = 0.3$ strikes the optimal balance and achieves the best accuracy on both CIFAR-100 and ImageNet-1K. With a smaller

Components		CIFAR-100		ImageNet-1K	
Deferred BN	Selective BN	Acc (%)	Time/img (s)	Acc (%)	Time/img (s)
✗	✗	63.0	0.63	51.3	1.09
✓	✗	63.0	0.45	51.6	0.87
✓	✓	**63.5**	**0.35**	**52.3**	**0.71**

(a) Effect of deferred and selective BN alignment.

Δ	2	5	10	20
CIFAR-100	63.3	**63.5**	63.0	62.8
ImageNet-1K	52.2	**52.3**	52.0	51.9

(b) Effect of Score Interval.

α	CIFAR-100	ImageNet-1K
0.2	63.3	52.2
0.3	**63.5**	**52.3**
0.4	62.8	52.0

(c) Effect of skip ratio α.

	CIFAR-100	ImageNet-1K
Ensemble	62.6	52.0
f_{fast} only	**63.5**	**52.3**

(d) Effect of Fast-only Model.

Score Type	CIFAR-100	ImageNet-1K
Mean only	62.8	51.6
Variance only	62.6	52.0
Mean + Variance	**63.5**	**52.3**

(e) Effect of score type for BN layer selection.

Selection	CIFAR-100	ImageNet-1K
Random k layers	62.7	50.9
Ours (Data-Dependent)	**63.5**	**52.3**

(f) Effect of dynamic vs. random layer selection.

Table 6: Ablation studies on **SODA** components (a) and design choices (b)–(f) under IPC=10. Performance is reported as Top-1 accuracy (%). Time is evaluated by per-image generation time (s).

value ($\alpha = 0.2$), BN alignment is introduced before the synthetic images have formed clear class semantics, so premature BN alignment disrupts semantic formation and slightly lowers accuracy. With a larger value ($\alpha = 0.4$), the alignment stage becomes too short, leaving insufficient iterations for distributional refinement and causing a more noticeable drop. These results indicate that BN alignment should be introduced neither too early nor too late. Moreover, performance varies only mildly across the tested range, suggesting that **SODA** is not overly sensitive to the exact choice of α; we therefore use $\alpha = 0.3$ for all datasets.

Effect of Coverage Ratio. Fig. 5 shows that moderate BN coverage yields the best accuracy on both benchmarks, peaking at $\rho = 0.5$. At low coverage, aligning too few layers may provide insufficient distributional guidance. Increasing coverage beyond this optimum generally reduces accuracy despite enforcing more BN constraints. This pattern is consistent with excessive alignment over-constraining the synthetic feature distribution. Selective alignment thus balances distributional guidance with the flexibility needed for generalization.

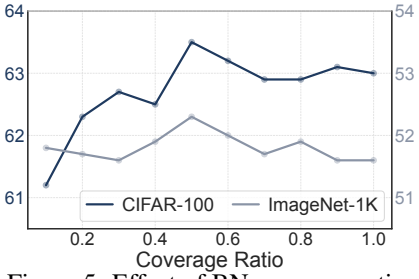

Figure 5: Effect of BN coverage ratio.

Effect of Fast-Only CE Early Training. Table 6d validates restricting early-phase supervision to f_{fast} alone. Despite real data initialization, using the full ensemble during the early phase yields no accuracy gain over fast-only training on either benchmark, confirming that ensemble diversity is redundant before class-discriminative structure has been established.

Effect of Score Type. Table 6e ablates the score formulation: mean-only or variance-only scores degrade performance, confirming both statistics are needed to compute the score. Table 6f further shows data-dependent selection consistently surpasses random selection, indicating the gain stems from principled importance scoring rather than merely reducing aligned layers.

4.4 Analysis

Verification of Theorem 1. Theorem 1 predicts a gain whenever $\pi_k(\rho) < \pi_k(1)$, i.e., when the discarded low-mismatch layers carry gradients that are less shared across teachers. Table 7 confirms this premise on ImageNet-1K: for all four teachers and at both iterations, the shared fraction $1 - \pi$ of the selected-layer gradient is 3.0–8.6× that of the

Iter.	ResNet-18	DenseNet-121	MobileNetV2	ShuffleNetV2
1k	5.3×	7.1×	5.2×	4.7×
2k	3.0×	8.3×	4.9×	8.6×

Table 7: Ratio of $1 - \pi$ between selected and discarded layers on ImageNet-1K.

discarded layers. Here, for each active teacher k, $\mathcal{U}_k$ is spanned by the full-coverage BN gradients of the other three teachers at the current iterate, and the selected layers are those chosen at the default coverage $\rho = 0.5$ (details in Appendix A.3).

Stability of Layer Selection. To examine which layers Selective BN Alignment selects, we accumulate the per-layer selection frequency over all iterations and compare the resulting layer rankings via Spearman correlation. As shown in Table 8, the ranking is highly consistent across random seeds (0.98) but varies across datasets (0.54), confirming that the selection adapts to the data rather than being fixed by the network. Across architectures, BN layers following depthwise convolutions are almost never selected, and the first BN in each block is prioritized over the second (Appendix C.1).

Comparison	Spearman
Across seeds	0.98
Across datasets	0.54

Table 8: Rank correlation of per-layer selection frequency.

Compatibility with Existing Methods. We further integrate **SODA** into SRe^{2}L to verify generality beyond residual-based uni-level frameworks. Table 9 shows gains in both accuracy and efficiency: accuracy rises from 48.5% to 48.9%, while the per-image generation time drops from 0.90 s to 0.74 s. The smaller speedup ($1.22\times$ vs. $1.54\times$) arises because SRe^{2}L's single teacher ($\beta = 1.0$) nullifies the CE saving term $\alpha(1-\beta)$ in Eq. (15); the remaining gain comes from BN backward elimination and selective alignment.

Method	SODA	Acc (%)	Time/img (s)	Speedup
SRe^{2}L	✗	48.5	0.90	$1.22\times$
	✓	**48.9**	**0.74**	

Table 9: Compatibility of **SODA**.

Extension to Gradient Matching. Selective alignment only requires the objective to decompose into per-layer residuals $\sum_l r_l$, which also holds for gradient matching [44]. We score each layer by $s_l := r_l / C_l$, where C_l is the number of output channels, and keep layer selection, compensation, and caching unchanged. On CIFAR-10 (IPC=10), **SODA** matches the baseline accuracy while reducing the matching cost by $1.22\times$ (Table 10), confirming its generality.

Method	Acc (%)	Speedup
DC	37.29$\pm$0.17	–
DC + **SODA**	37.43$\pm$0.13	$1.22\times$

Table 10: Gradient matching on CIFAR-10 (IPC=10).

4.5 Distilled Image Visualization

Fig. 6 compares distilled images from **SODA** and FADRM+ on ImageNet-1K. **SODA** produces visibly cleaner samples with smoother backgrounds and sharper foreground objects, particularly evident in the tench and goldfish classes where FADRM+ shows fragmented background textures from over-aligning to teacher BN statistics, while **SODA** keeps the fish bodies complete and clearly separated from the background. This visual gap directly echoes our quantitative findings: by relaxing redundant BN constraints, **SODA** avoids overfitting to non-discriminative distributional patterns and concentrates synthetic content on class-relevant features.

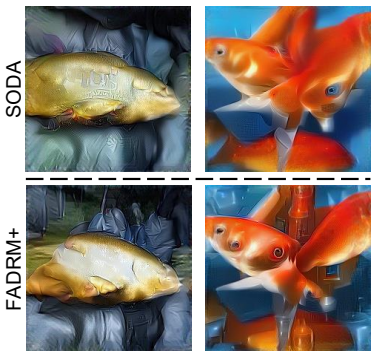

Figure 6: Visualization of distilled images from **SODA** and FADRM+.

5 Conclusion

We proposed **SODA**, a simple yet effective dataset distillation framework that addresses the computational redundancy and over-constraining issues of existing BN-matching-based decoupling methods. By selectively optimizing informative BN layers and deferring unnecessary alignment objectives, **SODA** simultaneously improves both efficiency and synthetic data quality. Our theoretical analysis explains the effectiveness of selective and deferred BN alignment, while extensive experiments across multiple datasets of CIFAR-100, Tiny-ImageNet, ImageNet-1K and its subsets thereof demonstrate that SODA achieves state-of-the-art performance with substantially reduced computational cost. In particular, excluding low-mismatch layers lowers the risk of students trained on the distilled data, as their alignment gradients are less shared across teachers and transfer poorly to the student, a premise we verify on ImageNet-1K. Future work includes adapting ρ and α during optimization and extending selective alignment to other distillation paradigms.

Acknowledgements and Disclosure

This work is supported by the MBZUAI-WIS Joint Program for Artificial Intelligence Research. The authors declare no competing interests.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10708–10717. IEEE, 2022.

[3] Mingyang Chen, Bo Huang, Junda Lu, Bing Li, Yi Wang, Minhao Cheng, and Wei Wang. Dataset distillation via adversarial prediction matching. *arXiv preprint arXiv:2312.08912*, 2023.

[4] Jiacheng Cui, Xinyue Bi, Yaxin Luo, Xiaohan Zhao, Jiacheng Liu, and Zhiqiang Shen. Fadrm: Fast and accurate data residual matching for dataset distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Jiacheng Cui, Zhaoyi Li, Xiaochen Ma, Xinyue Bi, Yaxin Luo, and Zhiqiang Shen. Dataset distillation via committee voting. *arXiv preprint arXiv:2501.07575*, 2025.

[6] Jiacheng Cui, Bingkui Tong, Xinyue Bi, Xiaohan Zhao, Jiacheng Liu, and Zhiqiang Shen. Hard labels in! Rethinking the role of hard labels in mitigating local semantic drift. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 21984–22006. PMLR, 06–11 Jul 2026.

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

[8] 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.

[9] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. *Advances in Neural Information Processing Systems*, 35: 34391–34404, 2022.

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

[11] Yang He, Lingao Xiao, Joey Tianyi Zhou, and Ivor Tsang. Multisize dataset condensation. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

[14] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[15] Hae Beom Lee, Dong Bok Lee, and Sung Ju Hwang. Dataset condensation with latent space knowledge factorization and sharing. *arXiv preprint arXiv:2208.10494*, 2022.

[16] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, pages 12352–12364. PMLR, 2022.

[17] Muquan Li, Hang Gou, Yingyi Ma, Rongzheng Wang, Ke Qin, and Tao He. Fixed anchors are not enough: Dynamic retrieval and persistent homology for dataset distillation. *arXiv preprint arXiv:2602.24144*, 2026.

[18] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[19] Ping Liu and Jiawei Du. The evolution of dataset distillation: Toward scalable and generalizable solutions. *arXiv preprint arXiv:2502.05673*, 2025.

[20] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. *Advances in Neural Information Processing Systems*, 35:1100–1113, 2022.

[21] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *Advances in Neural Information Processing Systems*, 35: 13877–13891, 2022.

[22] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34:5186–5198, 2021.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[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, October 2023.

[25] Xinyi Shang, Peng Sun, Zhiqiang Shen, Tao Lin, and Jing-Hao Xue. Dataset distillation in the era of large-scale data: Methods, analysis, and future directions. 2025.

[26] 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*, pages 16709–16718, 2024.

[27] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. *Advances in Neural Information Processing Systems*, 37:99161–99201, 2024.

[28] Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4797–4806, 2025.

[29] Donghyeok Shin, Seungjae Shin, and Il-Chul Moon. Frequency domain-based dataset distillation. *Advances in Neural Information Processing Systems*, 36:70033–70044, 2023.

[30] 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*, pages 9390–9399, 2024.

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[32] Ruilin Tong, Haodong Lu, Yuhang Liu, and Dong Gong. Model inversion with layer-specific modeling and alignment for data-free continual learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

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

[35] Lingao Xiao and Yang He. Are large-scale soft labels necessary for large-scale dataset distillation? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[36] Lingao Xiao and Yang He. Soft label pruning and quantization for large-scale dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[37] Eric Xue, Yijiang Li, Haoyang Liu, Peiran Wang, Yifan Shen, and Haohan Wang. Towards adversarially robust dataset distillation by curvature regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9041–9049, 2025.

[38] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8715–8724, 2020.

[39] Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024.

[40] 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.

[41] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15658–15667, 2021.

[42] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021.

[43] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6514–6523, 2023.

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

[45] Binglin Zhou, Linhao Zhong, and Wentao Chen. Improve cross-architecture generalization on dataset distillation. *arXiv preprint arXiv:2402.13007*, 2024.

[46] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022.

Appendix of SODA

Contents

A Proofs and Empirical Verification **15**

A.1 Proof of Theorem 1 15

A.2 Analysis under the Training Objective 16

A.2.1 Feasibility Relaxation 16

A.2.2 Implicit Alignment of Unselected Layers 16

A.3 Empirical Verification of Theoretical Premises 18

B Experimental Details **18**

B.1 Hyper-Parameters 19

B.2 Computer Resources 19

C Additional Experiments **20**

C.1 Stability of BN Layer Selection 20

C.2 Breakdown of the Speedup 20

D Limitations **21**

E Distilled Image Visualization **21**

A Proofs and Empirical Verification

Appendix A.1 proves the generalization result of Sec. 3.4 (Theorem 1). Appendix A.2 gives the training-objective perspective referenced there: selective alignment cannot worsen the constrained CE optimum (Proposition 1) and still aligns the skipped layers implicitly (Proposition 2). Appendix A.3 verifies the empirical premises of Theorem 1 and Proposition 2.

A.1 Proof of Theorem 1

Setup. We use the notation of Sec. 3.4. Fix a Phase-II iteration $t \geq \lfloor \alpha T \rfloor$ with active teacher k and a coverage $\rho \in (0, 1)$. $P_{\mathcal{U}_k}$ denotes the orthogonal projection onto $\mathcal{U}_k$ and $P_{\mathcal{U}_k^\perp} = I - P_{\mathcal{U}_k}$. Since $\mathcal{U}_k$ is spanned by the other teachers' full-coverage gradients at the current iterate, it does not depend on ρ, so both candidate updates are decomposed in the same reference frame. We treat $\mathcal{E}_f$ as a smooth function of $\hat{x}$ (e.g., the population risk averaged over the randomness of student training), and assume $\nabla\mathcal{E}_f(\hat{x}^{(t)}) \neq 0$ and $\hat{g}_\mathcal{S}(\rho), \hat{g}_\mathcal{S}(1) \neq 0$, so that π_k is well defined. Assumption 1 posits $\omega(\tau_f) < \omega_0$ even when f shares an architecture with a teacher, since the student is trained from scratch and does not inherit the statistics specific to an individual teacher.

Step 1: First-order expansion. Let $d(\rho) := \nabla_{\hat{x}}\mathcal{L}_{\text{CE}}(\hat{x}^{(t)}) + \lambda\, \hat{g}_\mathcal{S}(\rho)$, so that $\hat{x}^{(t+1)}(\rho) = \hat{x}^{(t)} - \eta\, d(\rho)$. Since $\mathcal{E}_f \in C^2$ near $\hat{x}^{(t)}$, Taylor's theorem gives

$$\mathcal{E}_f\big(\hat{x}^{(t+1)}(\rho)\big) = \mathcal{E}_f(\hat{x}^{(t)}) - \eta\, \big\langle \nabla\mathcal{E}_f, d(\rho)\big\rangle + O(\eta^2), \tag{16}$$

where the remainder is bounded by $\frac{\eta^2}{2} \sup_\xi \|\nabla^2\mathcal{E}_f(\xi)\|_{\text{op}}\|d(\rho)\|^2$, with ξ ranging over the segment between $\hat{x}^{(t)}$ and $\hat{x}^{(t+1)}(\rho)$. The CE gradient is evaluated at the same iterate and with the same teacher for both coverages, so it cancels in the difference:

$$\mathcal{E}_f\big(\hat{x}^{(t+1)}(1)\big) - \mathcal{E}_f\big(\hat{x}^{(t+1)}(\rho)\big) = \eta\lambda\Big(\big\langle \nabla\mathcal{E}_f, \hat{g}_\mathcal{S}(\rho)\big\rangle - \big\langle \nabla\mathcal{E}_f, \hat{g}_\mathcal{S}(1)\big\rangle\Big) + O(\eta^2). \tag{17}$$

Step 2: Shared/teacher-specific decomposition. For $\rho' \in \{\rho, 1\}$, linearity of the inner product and Assumption 1 give

$$\begin{aligned}\big\langle \nabla\mathcal{E}_f, \hat{g}_\mathcal{S}(\rho')\big\rangle &= \big\langle \nabla\mathcal{E}_f, P_{\mathcal{U}_k}\hat{g}_\mathcal{S}(\rho')\big\rangle + \big\langle \nabla\mathcal{E}_f, P_{\mathcal{U}_k^\perp}\hat{g}_\mathcal{S}(\rho')\big\rangle \\ &= \Gamma_{\rho'}\Big(\omega_0 - \big(\omega_0 - \omega(\tau_f)\big)\,\pi_k(\rho')\Big).\end{aligned} \tag{18}$$

Step 3: Conclusion. Under $\|\hat{g}_\mathcal{S}(\rho)\| = \|\hat{g}_\mathcal{S}(1)\| = G$, we have $\Gamma_\rho = \Gamma_1 = G\,\|\nabla\mathcal{E}_f\|$. Substituting (18) into (17), the ω_0 terms cancel and

$$\mathcal{E}_f\big(\hat{x}^{(t+1)}(1)\big) - \mathcal{E}_f\big(\hat{x}^{(t+1)}(\rho)\big) = \eta\lambda G\,\|\nabla\mathcal{E}_f\|\,\big(\pi_k(1) - \pi_k(\rho)\big)\big(\omega_0 - \omega(\tau_f)\big) + O(\eta^2), \tag{19}$$

which is (12). Since $\omega_0 - \omega(\tau_f) > 0$, the leading term is positive whenever $\pi_k(\rho) < \pi_k(1)$. □

Remark (unequal norms). The equal-norm condition can be relaxed. If $\|\hat{g}_\mathcal{S}(\rho)\| \geq \|\hat{g}_\mathcal{S}(1)\|$ and $\omega(\tau_f) \geq 0$, the bracket in (18) is a convex combination of ω_0 and $\omega(\tau_f)$ and hence non-negative, so Γ_ρ may be replaced by the smaller Γ_1 in the first inner product of (17). The same argument then yields (12) with "$=$" replaced by "$\geq$" and $G := \|\hat{g}_\mathcal{S}(1)\|$.

Remark (role of the energy-proportional form). For $\pi_k(\rho) \in (0, 1)$, define the effective transfer rates $\omega_{\text{sh}}(\rho) := \langle \nabla\mathcal{E}_f, P_{\mathcal{U}_k}\hat{g}_\mathcal{S}(\rho)\rangle / \big(\Gamma_\rho(1 - \pi_k(\rho))\big)$ and $\omega_{\text{sp}}(\rho) := \langle \nabla\mathcal{E}_f, P_{\mathcal{U}_k^\perp}\hat{g}_\mathcal{S}(\rho)\rangle / \big(\Gamma_\rho\,\pi_k(\rho)\big)$. Then $\langle \nabla\mathcal{E}_f, \hat{g}_\mathcal{S}(\rho)\rangle = \Gamma_\rho\big(\omega_{\text{sh}}(\rho)(1 - \pi_k(\rho)) + \omega_{\text{sp}}(\rho)\,\pi_k(\rho)\big)$ holds by definition, and Assumption 1 states precisely that these rates do not depend on ρ and satisfy $\omega_{\text{sp}} < \omega_{\text{sh}}$. The energy-proportional form is thus the substantive modeling choice. If the two contributions were instead proportional to the component norms, i.e., $\omega_0\Gamma_\rho\sqrt{1 - \pi_k(\rho)}$ and $\omega(\tau_f)\Gamma_\rho\sqrt{\pi_k(\rho)}$, the leading term would become $\eta\lambda G\|\nabla\mathcal{E}_f\|\big(h(\pi_k(\rho)) - h(\pi_k(1))\big)$ with $h(\pi) := \omega_0\sqrt{1 - \pi} + \omega(\tau_f)\sqrt{\pi}$. Since h is strictly decreasing on $(\bar{\pi}, 1]$ with $\bar{\pi} := \omega(\tau_f)^2/(\omega_0^2 + \omega(\tau_f)^2)$, the conclusion of Theorem 1 continues to hold whenever $\pi_k(\rho) > \bar{\pi}$.

Remark (scope). Like Proposition 2, Theorem 1 compares single updates from a common iterate at first order. Its cumulative effect over Phase II follows only heuristically, provided the premise $\pi_k(\rho) < \pi_k(1)$ persists along the trajectory, which we verify at two iterations (Appendix A.3); it is not a guarantee on the final distilled set. Both results analyze a plain gradient step, whereas the Adam optimizer used in practice rescales coordinates adaptively. Finally, the size of the gap scales with $\|\nabla \mathcal{E}_f\|$, G, and $\pi_k(1) - \pi_k(\rho)$, all of which vary across students; Theorem 1 therefore predicts the sign of the gain for every student rather than an ordering of gains across architectures.

A.2 Analysis under the Training Objective

While Theorem 1 concerns the population risk of the student, this section analyzes Selective BN Alignment with respect to the distillation objective itself: selective alignment cannot worsen the constrained CE optimum (Proposition 1), and the skipped layers are still aligned implicitly (Proposition 2).

A.2.1 Feasibility Relaxation

Define the per-layer feasibility set as $\mathcal{F}_l(\epsilon) := \{\hat{x} : r_l(\hat{x}) \le \epsilon\}$, the full-alignment feasibility as $\mathcal{F}_{\text{full}}(\epsilon) := \bigcap_{l=1}^{L} \mathcal{F}_l(\epsilon)$, and the selective-alignment feasibility as $\mathcal{F}_{\mathcal{S}}(\epsilon) := \bigcap_{l\in\mathcal{S}} \mathcal{F}_l(\epsilon)$.

Proposition 1 (Feasibility relaxation)**.** Assume $\mathcal{X}$ is compact, $\mathcal{L}_{\text{CE}}$ and $\{r_l\}_{l=1}^{L}$ are continuous on $\mathcal{X}$, and $\mathcal{F}_{\text{full}}(\epsilon) \neq \emptyset$. For any $\mathcal{S} \subseteq \{1, \dots, L\}$ and $\epsilon \ge 0$, $\mathcal{F}_{\text{full}}(\epsilon) \subseteq \mathcal{F}_{\mathcal{S}}(\epsilon)$, and

$$\min_{\hat{x}\in\mathcal{F}_{\mathcal{S}}(\epsilon)} \mathcal{L}_{\text{CE}}(\hat{x}) \;\le\; \min_{\hat{x}\in\mathcal{F}_{\text{full}}(\epsilon)} \mathcal{L}_{\text{CE}}(\hat{x}). \tag{20}$$

We prove two claims in order: (i) set inclusion, and (ii) the CE inequality.

(i) Set Inclusion: $\mathcal{F}_{\text{full}}(\epsilon) \subseteq \mathcal{F}_{\mathcal{S}}(\epsilon)$. Fix $\hat{x} \in \mathcal{F}_{\text{full}}(\epsilon)$. By definition of intersection,

$$\hat{x} \in \mathcal{F}_{\text{full}}(\epsilon) = \bigcap_{l=1}^{L} \mathcal{F}_l(\epsilon) \iff r_l(\hat{x}) \le \epsilon \;\forall l \in \{1, \dots, L\}. \tag{21}$$

Since $\mathcal{S} \subseteq \{1, \dots, L\}$, the implication $r_l(\hat{x}) \le \epsilon$ holds in particular for every $l \in \mathcal{S}$, hence

$$\hat{x} \in \bigcap_{l\in\mathcal{S}} \mathcal{F}_l(\epsilon) = \mathcal{F}_{\mathcal{S}}(\epsilon). \tag{22}$$

As $\hat{x}$ was arbitrary, $\mathcal{F}_{\text{full}}(\epsilon) \subseteq \mathcal{F}_{\mathcal{S}}(\epsilon)$. □

(ii) CE Inequality. By compactness of $\mathcal{X}$ and continuity of each r_l, the sublevel set $\mathcal{F}_l(\epsilon) = r_l^{-1}((-\infty, \epsilon])$ is closed in $\mathcal{X}$, and so is the finite intersection $\mathcal{F}_{\text{full}}(\epsilon)$. Hence $\mathcal{F}_{\text{full}}(\epsilon)$ is compact and non-empty. Continuity of $\mathcal{L}_{\text{CE}}$ together with the Weierstrass extreme value theorem ensures $\arg\min_{\mathcal{F}_{\text{full}}(\epsilon)} \mathcal{L}_{\text{CE}}$ is non-empty. Pick any $\hat{x}^*_{\text{full}} \in \arg\min_{\mathcal{F}_{\text{full}}(\epsilon)} \mathcal{L}_{\text{CE}}$. By (i), $\hat{x}^*_{\text{full}} \in \mathcal{F}_{\mathcal{S}}(\epsilon)$, so

$$\min_{\hat{x}\in\mathcal{F}_{\mathcal{S}}(\epsilon)} \mathcal{L}_{\text{CE}}(\hat{x}) \;\le\; \mathcal{L}_{\text{CE}}(\hat{x}^*_{\text{full}}) \;=\; \min_{\hat{x}\in\mathcal{F}_{\text{full}}(\epsilon)} \mathcal{L}_{\text{CE}}(\hat{x}), \tag{23}$$

where the existence of the minimum on the left follows from the same compactness-and-continuity argument applied to $\mathcal{F}_{\mathcal{S}}(\epsilon)$. □

A.2.2 Implicit Alignment of Unselected Layers

Proposition 2 (Implicit alignment of unselected layers)**.** Let $g_l := \nabla_{\hat{x}} r_l(\hat{x}^{(t)})$ and $g_{\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{l'\in\mathcal{S}} g_{l'}$, and assume $r_l \in C^2$ in a neighborhood of $\hat{x}^{(t)}$. After a gradient step $\hat{x} \leftarrow \hat{x} - \eta\nabla_{\hat{x}}\mathcal{L}_{BN}^{(t)}$, the change in mismatch at any unselected layer $l \notin \mathcal{S}$ satisfies, to first order,

$$\Delta r_l \;=\; -\eta L \,\langle g_l, g_{\mathcal{S}}\rangle + O(\eta^2) \;=\; -\eta L \,\|g_l\|\,\|g_{\mathcal{S}}\|\,\cos(g_l, g_{\mathcal{S}}) + O(\eta^2). \tag{24}$$

In particular, $\Delta r_l < 0$ at leading order whenever $\langle g_l, g_{\mathcal{S}}\rangle > 0$; a sufficient condition expressible in terms of pairwise cosines is given in Corollary 1.

We prove the result in two parts. Lemma 1 gives an exact first-order expression for Δr_l in terms of $\langle g_l, g_{\mathcal{S}} \rangle$, which directly implies Proposition 2. Corollary 1 translates the inner-product condition into a pairwise-cosine condition, recovering the $\bar{\kappa}_{l,\mathcal{S}}$ form under uniform-magnitude assumptions and providing two-sided bounds in the non-uniform case.

Lemma 1 (Exact First-Order Implicit Alignment)**.** *Let $\hat{x}^{(t+1)} = \hat{x}^{(t)} - \eta \nabla_{\hat{x}} \mathcal{L}_{BN}^{(t)}$, where $\mathcal{L}_{BN}^{(t)} = \frac{L}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} r_{l'}$. Assume each $r_l \in C^2$ in a neighborhood of $\hat{x}^{(t)}$. Then for any layer $l \notin \mathcal{S}$,*

$$r_l(\hat{x}^{(t+1)}) - r_l(\hat{x}^{(t)}) = -\eta L \langle g_l, g_{\mathcal{S}} \rangle + O(\eta^2), \tag{25}$$

where $g_l := \nabla r_l(\hat{x}^{(t)})$ and $g_{\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} g_{l'}$. Equivalently,

$$\Delta r_l = -\eta L \, \|g_l\| \, \|g_{\mathcal{S}}\| \, \cos(g_l, g_{\mathcal{S}}) + O(\eta^2). \tag{26}$$

Proof. By linearity of the gradient operator,

$$\nabla_{\hat{x}} \mathcal{L}_{BN}^{(t)} = \frac{L}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} g_{l'} = L \cdot g_{\mathcal{S}}. \tag{27}$$

Hence the displacement is $\Delta \hat{x} = \hat{x}^{(t+1)} - \hat{x}^{(t)} = -\eta L \, g_{\mathcal{S}}$, with $\|\Delta \hat{x}\| = \eta L \|g_{\mathcal{S}}\|$.

Since $r_l \in C^2$, Taylor's theorem with remainder yields

$$r_l(\hat{x}^{(t)} + \Delta \hat{x}) = r_l(\hat{x}^{(t)}) + \langle g_l, \Delta \hat{x} \rangle + \tfrac{1}{2} \Delta \hat{x}^\top H_l(\xi) \Delta \hat{x}, \tag{28}$$

for some ξ on the segment between $\hat{x}^{(t)}$ and $\hat{x}^{(t+1)}$, where $H_l = \nabla^2 r_l$. The remainder is bounded by $\frac{1}{2} \|H_l(\xi)\|_{\mathrm{op}} \|\Delta \hat{x}\|^2 = O(\eta^2)$. Substituting:

$$\Delta r_l = \langle g_l, -\eta L \, g_{\mathcal{S}} \rangle + O(\eta^2) = -\eta L \langle g_l, g_{\mathcal{S}} \rangle + O(\eta^2), \tag{29}$$

which is (25). The cosine form (26) follows from $\langle u, v \rangle = \|u\| \, \|v\| \cos(u, v)$. □

Corollary 1 ($\bar{\kappa}$-Form, with Two-Sided Bounds)**.** *Let $m_l := \|g_l\|$, $m_{\min} := \min_{l' \in \mathcal{S}} m_{l'}$, $m_{\max} := \max_{l' \in \mathcal{S}} m_{l'}$. Define the positive and negative cosine averages*

$$\bar{\kappa}_{l,\mathcal{S}}^{+} := \frac{1}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} \max\big(\cos(g_l, g_{l'}), 0\big) \geq 0, \tag{30}$$

$$\bar{\kappa}_{l,\mathcal{S}}^{-} := \frac{1}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} \min\big(\cos(g_l, g_{l'}), 0\big) \leq 0, \tag{31}$$

so that the mean pairwise cosine $\bar{\kappa}_{l,\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} \cos(g_l, g_{l'})$ satisfies $\bar{\kappa}_{l,\mathcal{S}} = \bar{\kappa}_{l,\mathcal{S}}^{+} + \bar{\kappa}_{l,\mathcal{S}}^{-}$. Under the conditions of Lemma 1, the first-order change satisfies the two-sided bound

$$-\eta L \, m_l \big(m_{\max} \bar{\kappa}_{l,\mathcal{S}}^{+} + m_{\min} \bar{\kappa}_{l,\mathcal{S}}^{-}\big) \leq \Delta r_l + O(\eta^2) \leq -\eta L \, m_l \big(m_{\min} \bar{\kappa}_{l,\mathcal{S}}^{+} + m_{\max} \bar{\kappa}_{l,\mathcal{S}}^{-}\big). \tag{32}$$

In particular:

1. Uniform-magnitude case. *If $m_{l'} = m_{\mathcal{S}}$ for all $l' \in \mathcal{S}$, then exactly:*

$$\Delta r_l = -\eta L \, m_l \, m_{\mathcal{S}} \, \bar{\kappa}_{l,\mathcal{S}} + O(\eta^2). \tag{33}$$

2. Sign of leading-order term. *The leading-order term in Δr_l is strictly negative whenever*

$$m_{\min} \bar{\kappa}_{l,\mathcal{S}}^{+} > m_{\max} |\bar{\kappa}_{l,\mathcal{S}}^{-}|. \tag{34}$$

A sufficient condition is $\bar{\kappa}_{l,\mathcal{S}}^{-} = 0$ (all pairwise cosines non-negative) together with $\bar{\kappa}_{l,\mathcal{S}}^{+} > 0$.

Proof. By Lemma 1, $\Delta r_l = -\eta L \langle g_l, g_{\mathcal{S}} \rangle + O(\eta^2)$. Expanding the inner product:

$$\langle g_l, g_{\mathcal{S}} \rangle = \frac{1}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} \langle g_l, g_{l'} \rangle = \frac{m_l}{|\mathcal{S}|} \sum_{l' \in \mathcal{S}} m_{l'} \cos(g_l, g_{l'}). \tag{35}$$

Let $c_{l'} := \cos(g_l, g_{l'})$, $c_{l'}^+ := \max(c_{l'}, 0)$, $c_{l'}^- := \min(c_{l'}, 0)$, so $c_{l'} = c_{l'}^+ + c_{l'}^-$. Splitting the sum:

$$\sum_{l' \in \mathcal{S}} m_{l'} c_{l'} = \sum_{l' \in \mathcal{S}} m_{l'} c_{l'}^+ + \sum_{l' \in \mathcal{S}} m_{l'} c_{l'}^- . \tag{36}$$

Since $c_{l'}^+ \geq 0$ and $c_{l'}^- \leq 0$, applying $m_{\min} \leq m_{l'} \leq m_{\max}$ on each part separately gives

$$m_{\min} \sum_{l'} c_{l'}^+ + m_{\max} \sum_{l'} c_{l'}^- \;\leq\; \sum_{l'} m_{l'} c_{l'} \;\leq\; m_{\max} \sum_{l'} c_{l'}^+ + m_{\min} \sum_{l'} c_{l'}^- . \tag{37}$$

Dividing by $|\mathcal{S}|$ and using the definitions of $\bar{\kappa}^{\pm}$:

$$m_{\min} \bar{\kappa}_{l,\mathcal{S}}^+ + m_{\max} \bar{\kappa}_{l,\mathcal{S}}^- \;\leq\; \frac{1}{|\mathcal{S}|} \sum_{l'} m_{l'} c_{l'} \;\leq\; m_{\max} \bar{\kappa}_{l,\mathcal{S}}^+ + m_{\min} \bar{\kappa}_{l,\mathcal{S}}^- . \tag{38}$$

Combining with (35) and $\Delta r_l = -\eta L \langle g_l, g_{\mathcal{S}} \rangle + O(\eta^2)$ (where the negation flips the inequality directions) yields (32).

For (1): when all $m_{l'} = m_{\mathcal{S}}$, (35) simplifies to $m_l \cdot m_{\mathcal{S}} \cdot \bar{\kappa}_{l,\mathcal{S}}$, giving (33).

For (2): from the lower bound of $\langle g_l, g_{\mathcal{S}} \rangle$ implicit in (32),

$$\langle g_l, g_{\mathcal{S}} \rangle \;\geq\; m_l \big(m_{\min} \bar{\kappa}_{l,\mathcal{S}}^+ + m_{\max} \bar{\kappa}_{l,\mathcal{S}}^- \big) = m_l \big(m_{\min} \bar{\kappa}_{l,\mathcal{S}}^+ - m_{\max} |\bar{\kappa}_{l,\mathcal{S}}^-| \big). \tag{39}$$

Under (34), the right-hand side is strictly positive, so $\langle g_l, g_{\mathcal{S}} \rangle > 0$. By Lemma 1, $\Delta r_l = -\eta L \langle g_l, g_{\mathcal{S}} \rangle + O(\eta^2)$, so $\Delta r_l < 0$ at leading order. □ □

Remark. The premise $\langle g_l, g_{\mathcal{S}} \rangle > 0$ of Proposition 2 is verified directly in Appendix A.3, where $\cos(g_l, g_{\mathcal{S}}) > 0$ holds in 98.9% of measurements. Corollary 1 relates this premise to the pairwise cosines in Fig. 2, which are predominantly non-negative; this is expected since BN layers within the same network process semantically related features and thus tend to produce correlated alignment gradients. Specifically: under uniform magnitudes, $\cos(g_l, g_{\mathcal{S}}) > 0$ whenever $\bar{\kappa}_{l,\mathcal{S}} > 0$ via (33); under non-uniform magnitudes, the sufficient condition (34) holds whenever $\bar{\kappa}_{l,\mathcal{S}}^- \approx 0$. The two-sided bound (32) thus guarantees implicit alignment ($\Delta r_l < 0$ at leading order) under empirically verified conditions, while quantifying how gradient-magnitude heterogeneity across layers affects the strength of implicit alignment.

A.3 Empirical Verification of Theoretical Premises

Premise of Theorem 1. On ImageNet-1K, for each teacher k at iterations 1k and 2k (both in Phase II), we construct $\mathcal{U}_k$ from the full-coverage BN gradients of the other three teachers at the current iterate, and compute the shared fraction $1 - \pi$ separately for the BN gradient of the selected layers $\mathcal{S}^{(k)}(\rho)$ with $\rho = 0.5$ and for that of the discarded layers. Table 7 in Sec. 4 reports the ratio of the former to the latter: selected-layer gradients are 3.0–8.6× more concentrated in the shared subspace, for every teacher and at both iterations. Since π is scale-invariant, $\pi_k(\rho)$ equals the teacher-specific fraction of the selected-layer gradient, whereas the full-coverage update is the sum of the selected- and discarded-layer gradients. When the cross terms between the two groups are negligible, $\pi_k(1)$ reduces to the energy-weighted average of the two groups' teacher-specific fractions, so a ratio above 1 implies $\pi_k(\rho) < \pi_k(1)$.

Premise of Proposition 2. Since only $\mathcal{S}$ is aligned, the BN update moves the synthetic data along $-\eta L\, g_{\mathcal{S}}$, and Proposition 2 requires $\langle g_l, g_{\mathcal{S}} \rangle > 0$ for $l \notin \mathcal{S}$. The pairwise cosines $\cos(g_l, g_{l'})$ in Fig. 2 are only a proxy for this quantity; Fig. 2 remains evidence for the layer-wise redundancy that motivates selective alignment. We therefore record $\cos(g_l, g_{\mathcal{S}})$ on CIFAR-100 for every unselected BN layer throughout Phase II. As shown in Table 11, the premise holds in 98.9% of measurements, confirming that alignment applied to $\mathcal{S}$ propagates to the skipped layers rather than leaving them unimproved.

B Experimental Details

We strictly follow the training configuration established in EDC to ensure fair and consistent comparisons across all evaluated methods. For DELT, we re-run the method under this unified configuration and report the higher of the two results obtained between their original setup and the EDC configuration, ensuring that neither method is disadvantaged by the choice of training protocol.

Quantity	Fraction > 0
$\cos(g_l, g_{\mathcal{S}}), l \notin \mathcal{S}$ (premise of Proposition 2)	98.9%

Table 11: Direct verification of the premise of Proposition 2 on CIFAR-100, over all unselected BN layers throughout Phase II.

B.1 Hyper-Parameters

Pretraining Phase. We follow the standard pretraining procedure established in prior uni-level dataset distillation methods, where a teacher network is trained on the full dataset to obtain reliable batch normalization (BN) running statistics, which subsequently serve as the alignment target during synthetic data generation. The hyperparameters shared across all settings are listed in Table 12, while the training epochs, which vary by dataset and IPC setting, are reported in Table 13.

Shared Hyperparameters for Model Pre-training	
Optimizer	SGD
Learning Rate	0.1
Weight Decay	1e-4
Momentum	0.9
Batch Size	128
Scheduler	Cosine Annealing
Augmentation	RandomCrop, Horizontal Flip
Loss Function	Cross-Entropy

Table 12: Shared hyperparameters for pre-trained models across all IPC settings.

Dataset	**IPC = 1**	**IPC = 10**	**IPC = 50**
CIFAR-100	50	50	50
Tiny-ImageNet	20	50	50
ImageWoof	30	30	300
ImageNette	50	50	300

Table 13: Training epochs for pre-trained models across different datasets and IPC settings.

Synthetic Data Generation. For synthetic data generation, we set the skip ratio $\alpha = 0.3$ and coverage ratio $\rho = 0.5$ uniformly across all datasets and experimental settings. All other shared hyperparameters are listed in Table 14.

Post-Evaluation. For post-evaluation, we follow the EDC protocol [27], which is also adopted by FADRM [4]. Specifically, models are trained from scratch on the distilled dataset under a fixed training protocol, and the resulting top-1 accuracy on the original validation set is reported as the evaluation metric. This ensures that all methods are assessed under identical post-evaluation conditions, guaranteeing a fair and consistent comparison.

B.2 Computer Resources

- **Hardware.** Our experimental platform consisted of a Linux-based workstation powered by an AMD Ryzen Threadripper PRO 5995WX processor, 256GB of system memory, and six NVIDIA RTX 4090 graphics cards (24GB VRAM each), running Ubuntu 22.04.5 LTS. This configuration provided sufficient computational capacity to support all training and evaluation workflows reliably.
- **Software.** The proposed framework was developed and tested under Python 3.10, leveraging PyTorch 2.0.1 as the deep learning backbone and CUDA 11.8 for GPU-accelerated computation. Preprocessing and augmentation routines were built upon the `torchvision` library,

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.25
Beta	(0.5, 0.9)
Epsilon	1×10^{-8}
Batch Size	100 or 10 (if $C < 100$)
Iterations (N_{iter})	2,000
Merge Ratio	0.5
Number of *ARC* (k)	3
SODA Recover Model	ResNet18, DenseNet121, ShuffleNetV2, MobileNetV2
Scheduler	Cosine Annealing
Augmentation	RandomResizedCrop, Horizontal Flip

Table 14: Hyperparameters for generating the distilled datasets.

while experimental metrics and training dynamics were monitored through the `wandb` logging platform. A complete list of package dependencies is included in the accompanying repository to facilitate reproducibility.

C Additional Experiments

C.1 Stability of BN Layer Selection

We accumulate the selection frequency of each BN layer over all iterations of a run and compare the resulting layer rankings via Spearman correlation. As shown in Table 8 of the main text, the ranking is highly stable across random seeds but varies across datasets, indicating that the selection responds to the data rather than to the network alone. Table 15 further groups the BN layers of a single teacher in one run by selection frequency; since the teacher is frozen, only the synthetic data changes during optimization. A stable core of layers is selected repeatedly, while adaptation concentrates in the intermittent group. Although layer indices are not comparable across architectures, the same rule holds for every teacher: BN layers following depthwise convolutions are almost never selected, and the first BN in each block is consistently prioritized over the second.

Group	Layers
Always (freq ≥ 0.9)	L5, L3, L1
Intermittent ($0.1 <$ freq < 0.9)	L10, L7, L12, L8, L15, L17, L18, L4, L2, L13, L11
Rarely (freq ≤ 0.1)	L0, L19, L9, L6, L16, L14

Table 15: BN layers of a single teacher grouped by selection frequency over one run.

C.2 Breakdown of the Speedup

Since **SODA** also restricts the early phase to the lightest teacher f_{fast}, part of its speedup could stem from cheaper CE supervision rather than from the BN-side changes. To separate the two, we add the components one at a time on ImageNet-1K (IPC=10, the setting of Table 6a) and record the per-image generation time after each step. As shown in Table 16, the two BN-side mechanisms, skipping $\mathcal{L}_{\text{BN}}$ in the early phase and selective BN alignment, account for 79% of the total time reduction, whereas the fast-teacher substitution contributes only 21%. Even without the fast teacher, **SODA** still achieves a $1.38\times$ speedup over FADRM+. This is consistent with the SRe^{2}L experiment in Table 9: with a single teacher, SRe^{2}L cannot benefit from fast-teacher acceleration, yet **SODA** still improves its efficiency by $1.22\times$.

Configuration	Time/img (s)	Saving (s)	Contribution
FADRM+ (baseline)	1.09	–	–
+ Early $\mathcal{L}_{\text{BN}}$ skip	0.95	0.14	37%
+ f_{fast} in the early phase	0.87	0.08	21%
+ Selective BN (**SODA**)	**0.71**	0.16	42%

Table 16: Breakdown of the per-image generation time reduction of **SODA** on ImageNet-1K (IPC=10).

D Limitations

While SODA achieves significant improvements in both efficiency and distillation quality, several limitations warrant discussion. First, SODA is designed as a plug-in framework for uni-level dataset distillation methods. Although its selective mechanism also extends to gradient matching (Table 10), this is verified only on CIFAR-10, and its applicability to trajectory matching and to teachers without BN layers (e.g., LayerNorm- or GroupNorm-based architectures such as ViTs) remains unexplored. Second, SODA does not explicitly audit for biases present in the original dataset, and synthetic data generated from biased sources may inherit or amplify such biases, raising fairness concerns in downstream applications. Third, our generalization analysis (Theorem 1) is a first-order, single-step comparison that rests on a modeling assumption about how shared and teacher-specific gradient components transfer to the student (Assumption 1); it explains the direction in which coverage affects student risk rather than guaranteeing final test accuracy.

From a broader societal perspective, the positive impact of SODA lies in its ability to substantially reduce the computational and storage costs of training data synthesis, democratizing access to large-scale machine learning for researchers and practitioners with limited resources, and contributing positively to energy efficiency and sustainability. On the negative side, the increased accessibility enabled by SODA could lower barriers to deploying powerful models in sensitive or high-risk scenarios such as surveillance or automated decision-making without adequate ethical oversight. Responsible deployment, continued research into bias mitigation, and domain generalization remain essential to ensure the safe and equitable application of SODA.

E Distilled Image Visualization

Additional visualizations of the distilled data generated by **SODA** are provided in Fig. 7 (CIFAR-100), Fig. 8 (Tiny-ImageNet), Fig. 9 (ImageNette), Fig. 10 (ImageWoof), and Fig. 11 (ImageNet-1K). Across all benchmarks, **SODA** consistently produces synthetic images with clean backgrounds, well-defined foreground objects, and rich class-discriminative features.

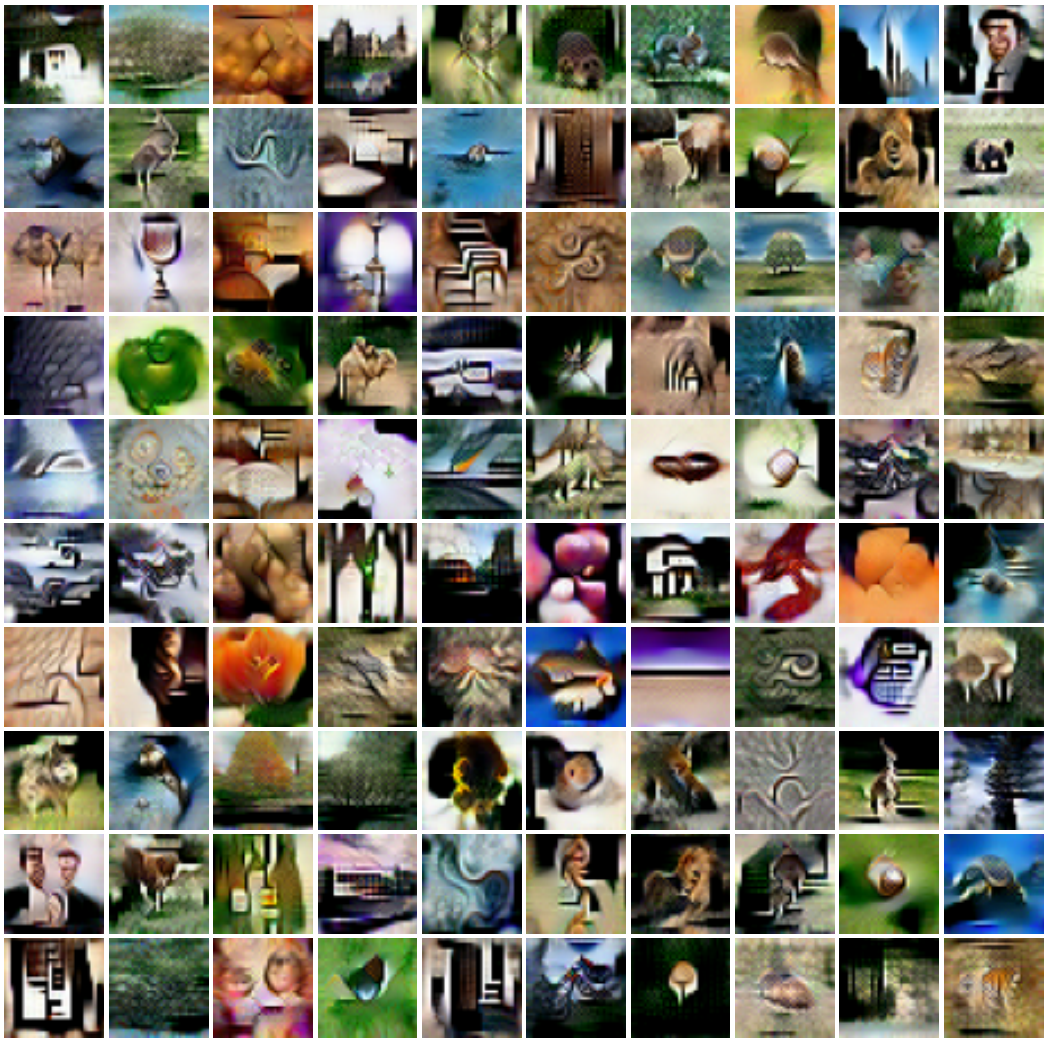

Figure 7: Visualization of synthetic data on CIFAR-100 generated by **SODA**.

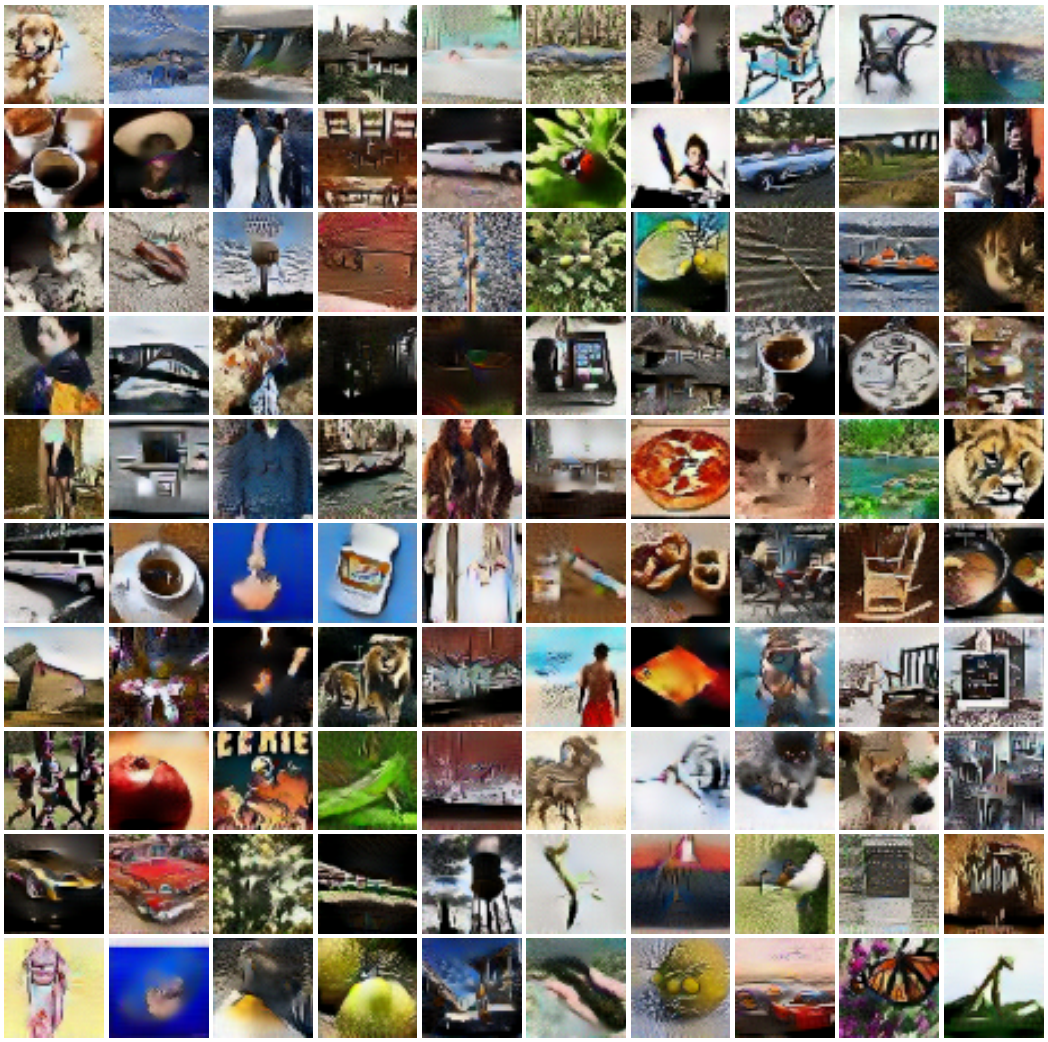

Figure 8: Visualization of synthetic data on Tiny-ImageNet generated by **SODA**.

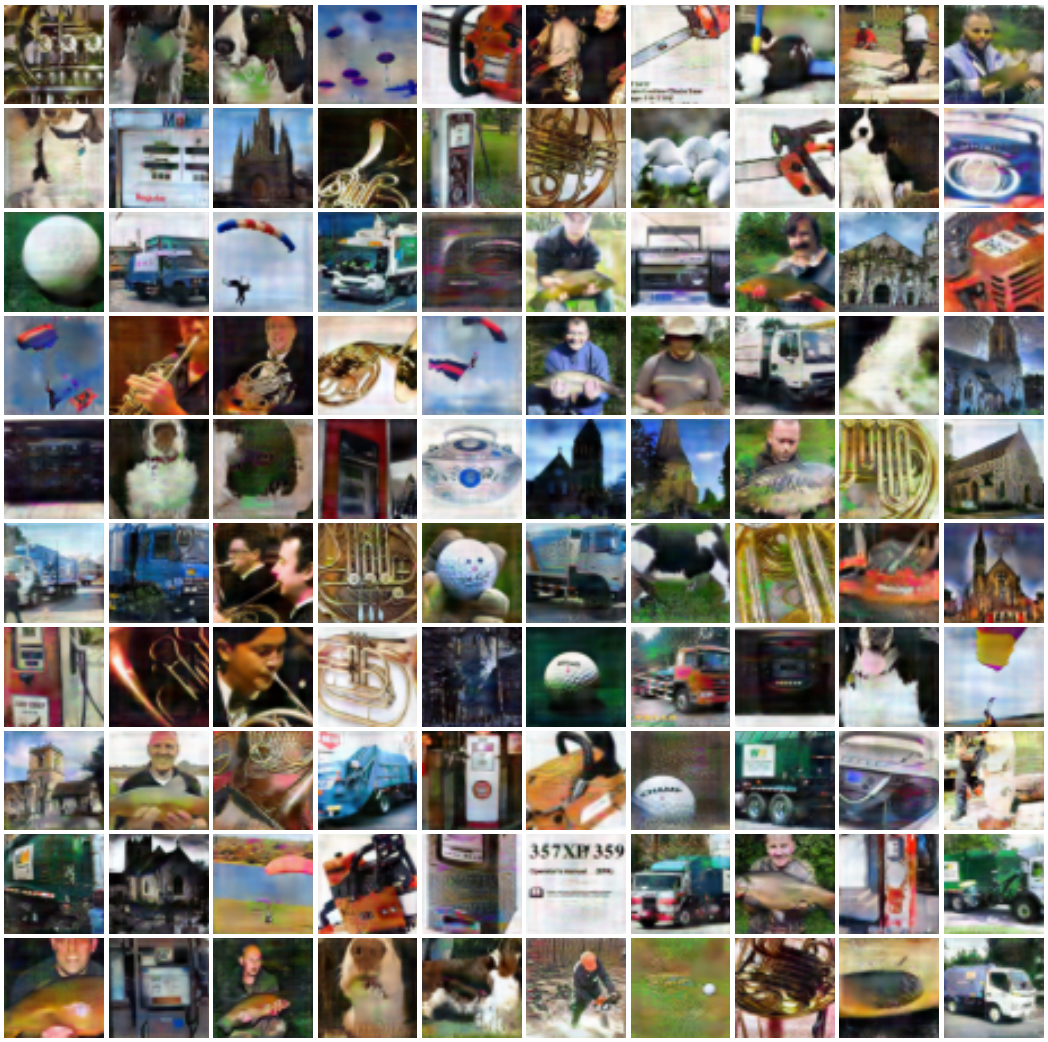

Figure 9: Visualization of synthetic data on ImageNette generated by **SODA**.

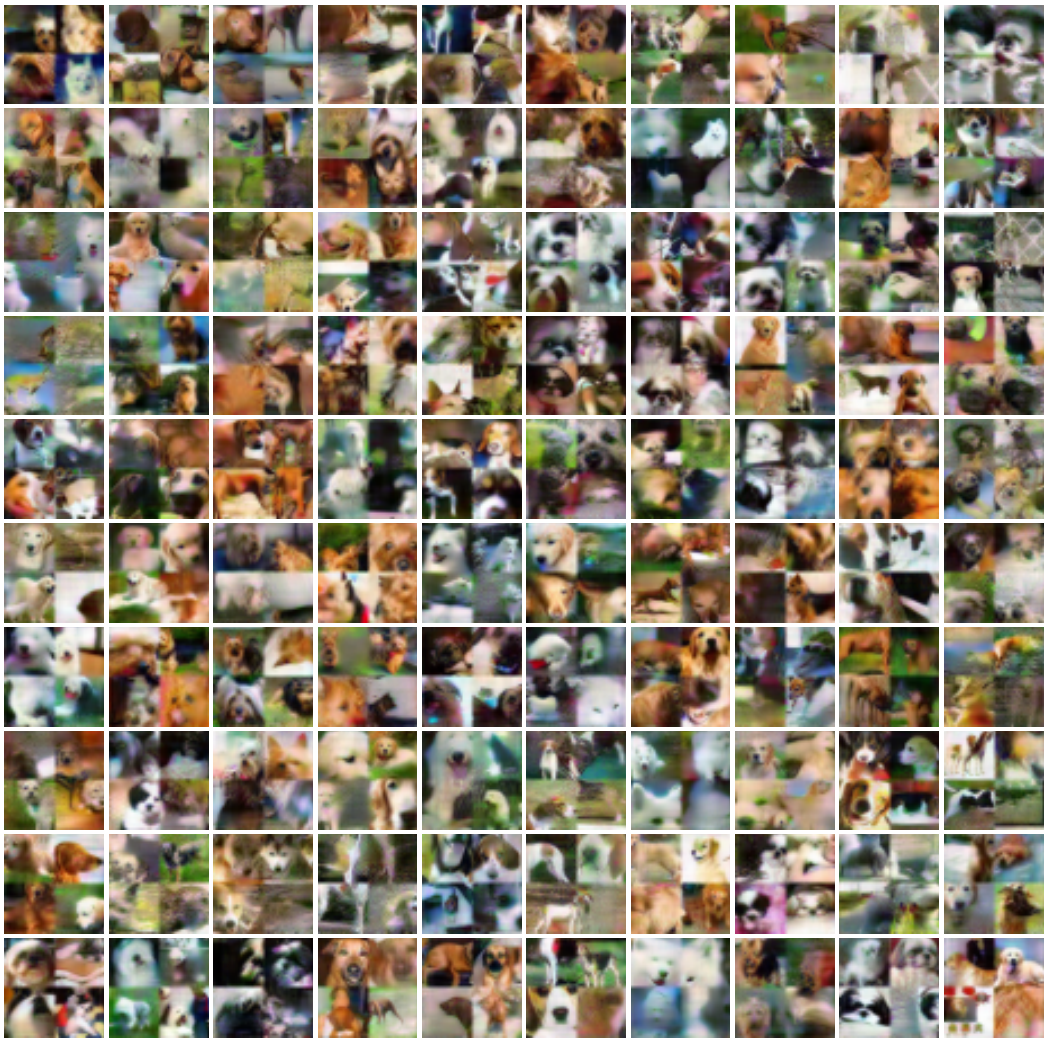

Figure 10: Visualization of synthetic data on ImageWoof generated by **SODA**.

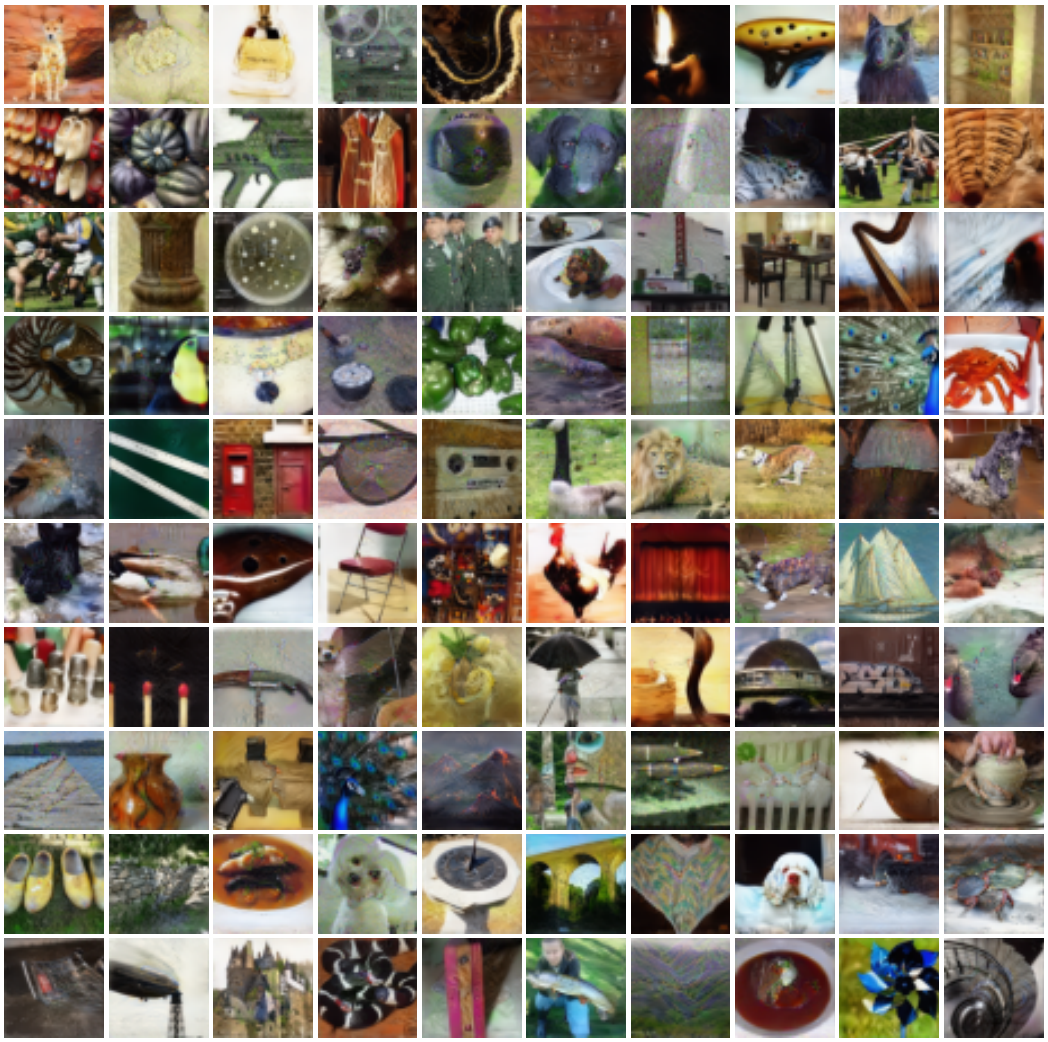

Figure 11: Visualization of synthetic data on ImageNet-1K generated by **SODA**.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The statements in the abstract and introduction are consistent with what the paper actually delivers. Specifically, the abstract outlines the proposed method, the findings, and the performance gains over existing baselines. The contributions are clearly detailed in Sec. 1, and empirical evidence corroborating each of them is provided in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. A thorough discussion of the limitations is provided in Sec. D; please check there for further details.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes. Sec. 3.4 states Theorem 1 formally together with its modeling assumption (Assumption 1), and summarizes Propositions 1 and 2, whose formal statements are given in Appendix A. Complete proofs of all three results are provided in Appendix A, all assumptions are stated within the theorem/proposition statements, and the empirical premises of Theorem 1 and Proposition 2 are verified in Appendix A.3. Supporting Lemma 1 and Corollary 1 are properly numbered and cross-referenced throughout.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes. The paper provides sufficient information to reproduce the main results. The experimental setup, datasets (CIFAR-100, Tiny-ImageNet, ImageNet-1K and its subsets), baseline methods, and evaluation protocol are described in Sec. 4.1. The full algorithm, including the mismatch scoring formulation, layer selection criterion, coverage threshold, score reuse interval, and the deferred alignment schedule, is detailed in Sec. 3.3, with all key hyperparameters reported in Appendix B.1. The paper also builds directly on publicly available baselines (FADRM+, SRe^{2}L, EDC), all of which use open datasets, providing a clear path for replication.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes. The complete codebase and data preparation scripts will be made publicly available upon the conclusion of the double-blind review process. Detailed instructions for reproducing the main experimental results are provided in Sec. 4.1 and Appendix B.1. In addition, the distilled images, pretrained generative models, and full codebase are included in the Supplementary Material to facilitate immediate reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. The core experimental settings are described in Sec. 4, including datasets, baseline methods, and evaluation protocols. Full training and implementation details, including hyperparameters are provided in Appendix B.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. The main results in Table 2 are reported as mean $\pm$ standard deviation (1-sigma) over 3 independent runs with different random seeds. Ablation studies are reported as single runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. The compute resources used in our experiments, including GPU type, memory are reported in Appendix B.2. Per-image generation time and total distillation cost are also documented in Table 3, providing a clear picture of the computational requirements for reproducing the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: Yes. The research conducted in this paper conforms with the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes. We explicitly discuss both the potential positive and negative societal impacts of our work in Appendix D.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work focuses on dataset distillation methodology and does not release any pre-trained models, image generators, or scraped datasets that would pose a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. All datasets, baseline methods, and code packages used in this paper are properly credited via citations to their original papers. The datasets used, CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet-1K, which are publicly available benchmarks widely adopted in the research community, and all baseline methods are referenced accordingly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes. This paper introduces new assets, specifically distilled datasets and pretrained models. We have included them in supplemental material. Comprehensive documentation will be provided together with the assets upon their release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The research presented in this paper involves no crowdsourcing nor any form of human subject participation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work involves neither crowdsourcing nor research with human subjects, and accordingly no IRB approval or equivalent ethical review is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology of this work does not incorporate LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.