PROFILE ONCE, DISTILL ACROSS BUDGETS: BUDGET–CHECKPOINT ALIGNMENT FOR DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoupled dataset distillation improves scalability by using pretrained checkpoints for synthetic image recovery and labeling. Its effectiveness, however, depends strongly on checkpoint maturity, with different images-per-class (IPC) budgets favoring different checkpoints. Jointly searching over recovery and labeling checkpoints at each budget requires a separate post-evaluation for every candidate pair, creating substantial alignment overhead. We introduce **P**rofile-Once **A**lignment of **C**heckpoints (**PAC**), a plug-in framework for reducing checkpoint alignment costs across IPC budgets. **PAC** replaces recovery checkpoint search with a single curriculum recovery run that uses checkpoints from all maturity levels to generate a shared image bank at the largest budget. With these images fixed, only the labeling checkpoints need to be aligned with each IPC budget, reducing the joint recovery labeling search to labeling selection. To guide this selection across IPC budgets, we identify exposure-indexed rank transfer, which links labeling checkpoint rankings at matched cumulative synthetic data exposure, defined as IPC times the number of training epochs on synthetic data. This allows us to profile labeling checkpoints only at the largest budget by tracking model performance throughout training. For smaller budgets, we sample from the recovered images and select labeling checkpoints using reference rankings at epochs matching the target budget's full-training exposure. To further reduce profiling cost, we characterize critical exposures at which checkpoint preferences change and derive sufficient conditions for permanently eliminating candidates from further evaluation. The resulting selection procedure supports previously unseen budgets within the reference range without additional recovery or profiling. Experiments across five datasets, various IPC budgets, and model architectures establish new state-of-the-art performance. At IPC 1, integrating **PAC** with FADRM+ substantially improves top-1 accuracy by **12.4%** on CIFAR-100 and by **18.1%** on ImageNet-1K.

1 INTRODUCTION

Recent advances in natural language processing and computer vision, exemplified by GPT-4 (Achiam et al., 2023), DeepSeek-V3 (Liu et al., 2024a), Gemini-2.5 (Comanici et al., 2025) and SAM (Kirillov et al., 2023), have been driven by high-capacity models trained on massive datasets. However, the data supporting these advances also impose substantial resource demands, from storing and accessing large collections to processing their contents during training. Reducing data volume can alleviate these demands, but the resulting representation must retain the information needed for effective learning. This motivates a data-centric approach that treats training examples themselves as an object of optimization, with downstream model performance serving as a measure of their utility. The goal is to generate a compact synthetic dataset whose examples are optimized to capture the knowledge of the original data and support effective learning across a range of downstream model architectures.

This idea underlies dataset distillation, which learns a compact synthetic dataset that preserves the learning value of a much larger source dataset (Wang et al., 2018). To scale this process, decoupled methods use pretrained checkpoints to recover synthetic images and generate soft labels (Yin et al., 2023). Recovery determines which visual structures are encoded in the synthetic images, whereas labeling controls the class relationships conveyed by the corresponding soft targets. The two stages

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

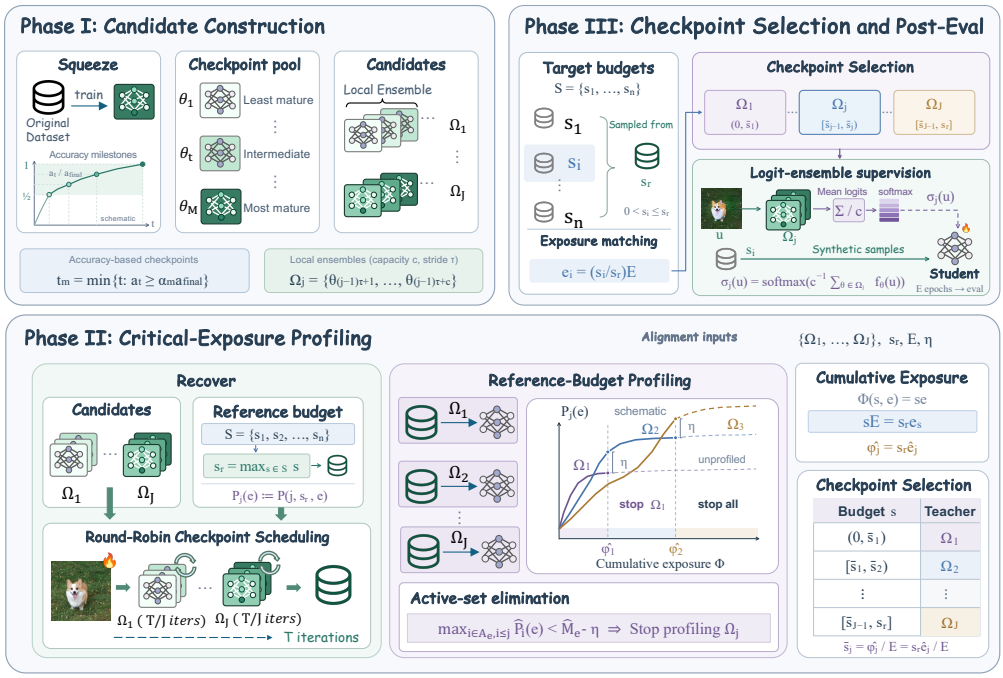

Figure 1: Overview of **PAC**. We first construct candidates from checkpoints sampled at accuracy milestones. We then recover synthetic data at the largest reference budget via full pool curriculum scheduling and profile labeling checkpoints, using critical exposures to eliminate candidates early and build a candidate selection map. Finally, exposure matching selects checkpoints for model training at target budgets. The map supports unseen budgets within the reference range without further profiling.

need not use the same checkpoints, and their preferred maturity can vary with the images-per-class (IPC) budget (Guo et al., 2024; Qin et al., 2024). Joint search must therefore account for both recovery and labeling configurations at each budget. With J candidates per role and N budgets, exhaustive search requires NJ^2 full model-training evaluations, alongside repeated image recovery. Each additional budget thus incurs another full sweep over recovery–labeling checkpoint pairs.

To address this alignment overhead, we propose **P**rofile-Once **A**lignment of **C**heckpoints (**PAC**). **PAC** replaces recovery checkpoint search with a single curriculum recovery run that uses checkpoints from all maturity levels to generate a shared image bank at the largest budget. Fixing these images reduces the joint recovery labeling search to labeling selection. We profile labeling checkpoints at this reference budget, tracking model performance with the images and training protocol fixed. Our key insight is exposure-indexed rank transfer, which links labeling checkpoint rankings across budgets at matched cumulative exposure, defined as IPC times training epochs on synthetic data. For smaller budgets, we sample from the recovered images and select labeling checkpoints using reference rankings at earlier epochs matching their full-training exposure. Critical exposures mark changes in checkpoint preferences, enabling early candidate elimination and yielding a candidate selection map. This reduces joint search to one recovery run and at most J reference training trajectories, supporting unseen budgets within the reference range without further recovery or profiling.

Our contributions are threefold:

- We identify exposure-indexed rank transfer, linking labeling checkpoint preferences across budgets through cumulative synthetic-data exposure under a shared recovery protocol.
- We propose **PAC**, which fixes full-pool recovery and reduces labeling selection costs across budgets through reference-budget profiling. We characterize critical exposures and derive sufficient conditions for permanent candidate elimination, reducing evaluation cost.
- We validate **PAC** across five datasets, multiple budgets, and model architectures. At IPC 1, it improves model accuracy over the prior state-of-the-art FADRM+ by 12.4% on CIFAR-100 with ResNet-18 and by 18.1% on ImageNet-1K with ResNet-18.

2 Related Work

Dataset Distillation. The goal of Dataset Distillation is to compress the knowledge contained in a large-scale training set into a much smaller synthetic dataset. Models trained on this distilled data are expected to achieve performance close to those trained on the complete dataset, while requiring substantially lower computational cost. Existing approaches can generally be distinguished by the type of information they aim to match, including: (1) Meta-Model Matching (Wang et al., 2018; Nguyen et al., 2021a;b; Zhou et al., 2022; Loo et al., 2022; 2023; Chen et al., 2024), (2) Trajectory Matching (Zhao et al., 2021; Zhao and Bilen, 2021; Cazenavette et al., 2022; Lee et al., 2022; Du et al., 2023a; Cui et al., 2023; Guo et al., 2024; Du et al., 2023b; Liu et al., 2024b; Zhong et al., 2025; Wang et al., 2025b), (3) Distribution Matching (Wang et al., 2022b; Zhao and Bilen, 2023; Zhao et al., 2023; Sajedi et al., 2023; Deng et al., 2024; Wang et al., 2025c; Liu et al., 2025; Li et al., 2026b), (4) Diffusion-based Generation (Gu et al., 2024; Su et al., 2024; Chen et al., 2025; Zhao et al., 2025; Santiago et al., 2025; Wang et al., 2025a; Zou et al., 2025; Moser et al., 2025; Cui et al., 2025c; Su et al., 2026; Zhou et al., 2026; Cui et al., 2026b; Xia et al., 2026; Gan et al., 2026), (5) Decoupled Distillation (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2024a; Sun et al., 2024; Shao et al., 2024b; Du et al., 2024; Shen et al., 2025; Cui et al., 2025b;a; Tran et al., 2025; Ma et al., 2025; Li et al., 2026a; Shang et al., 2026; Cui et al., 2026a). For a more detailed and comprehensive overview of this topic, we refer readers to the recent survey works by (Shang et al., 2025; Liu and Du, 2025; Yu et al., 2024), which provide extensive discussions of the relevant developments.

Checkpoint Selection for Dataset Distillation. Recent studies have shown that dataset distillation is highly sensitive to teacher maturity and that the preferred teacher can vary with the storage budget. DATM (Guo et al., 2024) observes that smaller IPC settings favor earlier, less mature expert trajectories, whereas larger IPC settings benefit from later trajectories. Subsequent studies further report that the optimal expert checkpoint and the appropriate sample difficulty vary with the storage budget (Qin et al., 2024; Lee and Chung, 2024), highlighting the importance of budget-dependent teacher maturity in dataset distillation. Related observations have also been extensively studied in knowledge distillation. RCO (Jin et al., 2019) and Pro-KD (Rezagholizadeh et al., 2022) exploit intermediate teacher states through progressive supervision, while CheckpointKD (Wang et al., 2022a) explicitly shows that intermediate checkpoints can outperform fully converged teachers and proposes a checkpoint selection criterion based on mutual information. Different from these works, which select teachers for individual student configurations or storage budgets, we focus on amortizing teacher checkpoint selection across multiple storage budgets, enabling a single profiling process to be reused for different budget settings, including previously unseen ones.

3 Approach

Preliminaries. Consider a training set $\mathcal{T} = \{(u_n, v_n)\}_{n=1}^{L}$ sampled from an unknown joint distribution $q(u, v)$, where u denotes an input image and v its corresponding class label. To reduce the storage and computational cost of training on large-scale data, dataset distillation constructs a compact synthetic set $\mathcal{C} = \{(\tilde{u}_j, \tilde{v}_j)\}_{j=1}^{K}$, where $K \ll L$, and $\tilde{u}_j$ and $\tilde{v}_j$ denote a synthetic image and its associated label, respectively. The synthetic samples need not correspond to real observations, but are optimized such that models trained on $\mathcal{C}$ achieve performance comparable to those trained on $\mathcal{T}$. For any training set $\mathcal{A}$, let $h_{\mathcal{A}}$ denote the resulting trained model, with population risk defined as

$$\mathcal{R}(\mathcal{A}) = \mathbb{E}_{(u,v)\sim q}\left[\mathcal{L}(h_{\mathcal{A}}(u), v)\right],$$

where $\mathcal{L}$ denotes the loss associated with the downstream task. The objective of dataset distillation can then be expressed as finding the smallest synthetic collection whose population risk remains sufficiently close to that of the full training set:

$$\mathcal{C}^{\star} = \operatorname*{argmin}_{\mathcal{C}} \operatorname{card}(\mathcal{C}), \qquad \text{s.t.} \quad |\mathcal{R}(\mathcal{C}) - \mathcal{R}(\mathcal{T})| \leq \delta, \tag{1}$$

where $\operatorname{card}(\mathcal{C})$ denotes the number of distilled samples and $\delta \geq 0$ specifies the maximum allowable degradation in population-level performance. Thus, the distilled set substantially reduces data size while preserving performance within tolerance δ.

A notable class of methods, referred to as *decoupled dataset distillation*, uses pretrained checkpoints with frozen parameters in two roles: *recovery checkpoints* guide synthetic image generation, and

labeling checkpoints provide supervision for model training. These roles need not use the same checkpoint or ensemble. Prior work has established that checkpoints at different stages of training can exhibit substantially different behaviors (Guo et al., 2024), making *training maturity* an important factor in checkpoint selection. Since the preferred maturity can vary with the data budget, both recovery and labeling configurations should be chosen to match the target budget.

3.1 Multi-Budget Checkpoint Alignment

A straightforward baseline therefore searches each budget independently: it evaluates all recovery–labeling configuration pairs and selects the pair yielding the best final post evaluation performance. To formalize this joint search and its cost, we first define a checkpoint pool spanning different maturity levels and the local ensembles that form its candidate configurations.

Checkpoint pool. For a model trained on $\mathcal{T}$, let $\theta^{(t)}$ and a_t denote epoch-t parameters and accuracy, and a_{final} its terminal accuracy. We use M equally spaced relative accuracy milestones in $[1/2, 1]$:

$$\alpha_m = \frac{1}{2}\left(1 + \frac{m-1}{M-1}\right), \qquad t_m = \min\left\{t : a_t \geq \alpha_m a_{\text{final}}\right\}, \qquad \theta_m = \theta^{(t_m)}, \tag{2}$$

for $m = 1, \ldots, M$. The ordered pool $\Theta = \{\theta_1, \ldots, \theta_M\}$ contains the first checkpoint to reach each selected accuracy milestone during training.

Candidate configurations. We form local ensembles of c consecutive checkpoints in Θ, with stride τ between adjacent windows. Each candidate spans a contiguous maturity range:

$$\Omega_j = \left\{\theta_{(j-1)\tau+1}, \ldots, \theta_{(j-1)\tau+c}\right\}, \qquad j = 1, \ldots, J, \qquad J = \left\lfloor \frac{M-c}{\tau} \right\rfloor + 1. \tag{3}$$

A larger c spans more maturity levels; smaller τ yields denser candidates. For labeling, $\sigma_j(u) = \text{softmax}\left(\frac{1}{c}\sum_{\theta\in\Omega_j} f_\theta(u)\right)$, where $f_\theta(u)$ denotes checkpoint logits.

Per-budget joint checkpoint search. Let $\mathcal{S} = \{s_1, \ldots, s_N\}$ be the budgets in images per class (IPC) and $\mathbf{R}$ the recovery procedure. Independent recovery and labeling indices i, j define

$$\mathcal{U}_s^{(i)} = \mathbf{R}(\mathcal{T}, \Omega_i, s), \qquad \mathcal{C}_s^{(i,j)} = \left\{(u, \sigma_j(u)) : u \in \mathcal{U}_s^{(i)}\right\}. \tag{4}$$

Let $Q(i, j, s, e)$ denote performance after e model-training epochs on $\mathcal{C}_s^{(i,j)}$, with architecture and training protocol fixed. At the full horizon E, joint search selects

$$(i^\star(s), j^\star(s)) = \underset{i,j\in\{1,\ldots,J\}}{\text{argmax}}\ Q(i, j, s, E). \tag{5}$$

For N target budgets, exhaustive alignment requires NJ recovery runs and NJ^2 full model-training evaluations, totaling NJ^2E post-evaluation epochs.

3.2 Single-Budget Checkpoint Alignment

The NJ^2 evaluation cost stems from pairing each of the J recovery configurations with all J labeling configurations at every budget. We find that these two choices can be decoupled through a shared image bank recovered with a curriculum spanning the full checkpoint pool. By progressively incorporating knowledge from all checkpoints in the pool, the recovered images provide a common basis for optimizing model performance under different labeling configurations. This allows **PAC** to fix recovery while selecting labeling checkpoints independently on the shared bank. We first describe how the curriculum integrates knowledge across maturity levels to recover these shared images.

Full-pool recovery. Curriculum ρ traverses the individual checkpoints $\theta_1, \ldots, \theta_M$ in maturity order over T_{rec} recovery iterations. Stage m uses only the frozen checkpoint θ_m for approximately T_{rec}/M iterations, with integer stage lengths summing to T_{rec}. Each subsequent stage starts from the images produced by the preceding stage, allowing checkpoints at later maturity levels to refine the accumulated result. The full pool thus guides a continuous recovery trajectory within the fixed

Algorithm 1 Critical-Exposure Profiling

Require: $\mathcal{U}_r^{\text{cur}}$, $\{\Omega_j\}_{j=1}^J$, s_r, E, η
1: Initialize J independent models with targets σ_j; $\mathcal{A}_0 \leftarrow \{1,\dots,J\}$, $\hat{e}_0 \leftarrow 0$, $\hat{e}_{1:J} \leftarrow E+1$
2: **for** $e = 0,\dots,E$ **do**
3: Train active models for one epoch if $e > 0$; evaluate $\widehat{P}_j(e)$ and set ℓ_e by Eq. (12)
4: $\hat{e}_j \leftarrow e$ for $j \in \mathcal{A}_e$ with $j < \ell_e$; update $\mathcal{A}_{e+1}$ by Eq. (13)
5: **if** $\mathcal{A}_{e+1} = \{J\}$ **then** set $\ell_t \leftarrow J$ $(e < t \le E)$ and **break**
6: **end for**
7: $\mathcal{B}_j \leftarrow [\hat{e}_{j-1}s_r/E, \hat{e}_j s_r/E) \cap (0, s_r]$; **return** budget intervals $\{\mathcal{B}_j\}_{j=1}^J$ for candidate selection

iteration budget. This curriculum is defined directly on Θ, independently of the labeling windows Ω_j. This exposes the shared images to all maturity levels before budget-specific labeling is applied. At $s_r = \max_{s\in\mathcal{S}} s$, we recover a shared bank:

$$\mathcal{U}_r^{\text{cur}} = \mathbf{R}(\mathcal{T}, \Theta, s_r; \rho), \qquad \mathcal{U}_s^{\text{cur}} = \Pi_s(\mathcal{U}_r^{\text{cur}}), \tag{6}$$

where Π_s selects s images per class from the shared bank. Recovery is performed once at the largest budget, and smaller budgets reuse subsets of these images without additional image optimization. At each budget, all labeling candidates are compared on the same selected images. This removes the recovery index i from subsequent alignment while retaining budget-dependent selection of labeling configurations. With recovery fixed, we examine whether labeling alignment can be reused across budgets. To this end, we introduce a measure combining data budget and model training duration.

Definition 1 (Cumulative exposure)**.** For e model epochs at budget s, define

$$\Phi(s, e) = se, \tag{7}$$

which counts synthetic sample presentations per class, including repeated visits across epochs.

Exposure-indexed rank transfer. Let $P(j, s, e)$ denote model performance after e epochs on $\mathcal{U}_s^{\text{cur}}$ with supervision σ_j, holding the model architecture and training protocol fixed. Since checkpoint selection depends on relative rather than absolute performance, we consider the decomposition

$$P(j, s, e) \approx G_j(\Phi(s, e)) + \Delta(s, e), \tag{8}$$

where G_j captures exposure-dependent configuration effects and $\Delta(s, e)$ is shared across configurations. At matched exposure, the shared offsets cancel, leaving the same exposure-dependent contrasts in pairwise comparisons. Configuration-dependent approximation errors can still alter rankings. When these errors are small relative to pairwise gaps, we obtain

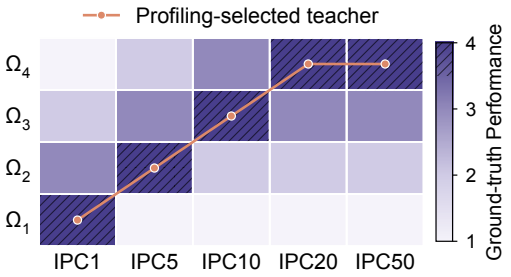

Figure 2: Labeling preferences under fixed full-pool recovery. The heatmap shows per-budget performance; the orange line marks configurations selected by reference profiling.

$$\Phi(s_1, e_1) = \Phi(s_2, e_2) \implies \text{Rank}_j\, P(\cdot, s_1, e_1) \approx \text{Rank}_j\, P(\cdot, s_2, e_2). \tag{9}$$

We call this relation *exposure-indexed rank transfer*, illustrated in Figure 2.

Single-budget readout. At s_r, we profile one model per labeling configuration on the shared distilled image bank. Each trajectory tracks the evolution of model performance under a fixed labeling configuration and provides intermediate evaluations throughout training. For a target run at budget $s \le s_r$ and horizon E, exposure matching gives

$$\Phi(s, E) = \Phi(s_r, e_s), \qquad e_s = \frac{sE}{s_r} \le E. \tag{10}$$

Applying the rank-transfer relation at these matched exposures gives

$$\text{Rank}_j\, P(\cdot, s, E) \approx \text{Rank}_j\, P(\cdot, s_r, e_s). \tag{11}$$

We use the ranking at e_s to select a labeling configuration for budget s. Different budgets query different epochs of the same reference trajectories, so these trajectories can be reused without additional reference training. The NJ^2 joint sweep is thus replaced by one shared recovery run and J reference trajectories. Fully profiling these trajectories still costs JE model-epochs.

3.3 Efficient Profiling via Critical Exposures

When the preferred configuration shifts monotonically toward more mature checkpoints as exposure increases (Figure 3), superseded candidates can be retired. We handle evaluation uncertainty with a margin-based frontier whose advances mark the *critical exposures* recorded by Algorithm 1.

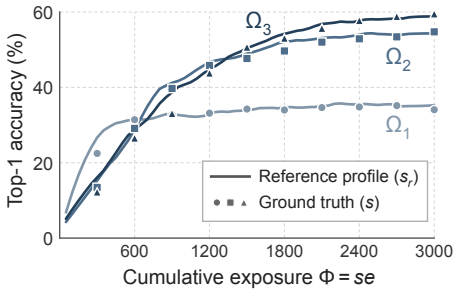

Figure 3: Profiling curves (solid) and matched-exposure budget evaluations (markers) under full-pool recovery; $s = \Phi/300$.

Candidate selection and elimination. Let $P_j(e) \coloneqq P(j, s_r, e)$ and $\widehat{P}_j(e)$ denote reference performance and its estimate on $\mathcal{E} = \{0, \dots, E\}$. Independent model trajectories are unaffected by stopping other candidates. Initialize $\mathcal{A}_0 = \{1, \dots, J\}$ and let $\widehat{M}_e = \max_{j \in \mathcal{A}_e} \widehat{P}_j(e)$. For $\eta \geq 0$, we select the smallest active candidate index ℓ_e with estimated performance at least $\widehat{M}_e - \eta$:

$$\ell_e = \min \left\{ j \in \mathcal{A}_e : \widehat{P}_j(e) \geq \widehat{M}_e - \eta \right\}. \tag{12}$$

We record ℓ_e as the selection frontier and retire active candidates with smaller indices:

$$\mathcal{A}_{e+1} = \{ j \in \mathcal{A}_e : j \geq \ell_e \}. \tag{13}$$

Theorem 1 (Margin-Certified Prefix Elimination)**.** *Assume that $m(e) = \min \operatorname{argmax}_j P_j(e)$ is non-decreasing on $\mathcal{E}$, and that $|\widehat{P}_j(e) - P_j(e)| \leq \beta$ holds jointly over all evaluated candidates and epochs. If $\eta \geq 2\beta$, the updates in Eqs. (12)–(13) permanently remove only candidates that are strictly suboptimal throughout the remaining horizon. Moreover, ℓ_e is non-decreasing and*

$$\max_j P_j(e) - P_{\ell_e}(e) \leq \eta + 2\beta. \tag{14}$$

For $\eta = \beta = 0$, the readout is exact: $\ell_e = m(e)$.

Only the optimal index must be monotone; non-optimal curves may cross freely. To transfer this guarantee across budgets, we quantify the approximation in Eq. (8) at $q_s = \lfloor sE/s_r \rfloor$ by

$$\max_j \left| P(j, s, E) - P_j(q_s) - c_s \right| \leq \delta_s, \tag{15}$$

where c_s is shared across candidates and $\delta_s \geq 0$ bounds transfer and grid-readout error.

Corollary 1 (Critical-Exposure Profiling)**.** *Under Theorem 1, let $\hat{e}_j = \min\{e \in \mathcal{E} : \ell_e > j\}$ for $1 \leq j < J$, with $\hat{e}_0 = 0$ and $\hat{e}_J = \min \varnothing = E + 1$. Then $\hat{e}_1 \leq \cdots \leq \hat{e}_{J-1}$ and $\ell_e = j \iff \hat{e}_{j-1} \leq e < \hat{e}_j$, $e \in \mathcal{E}$. At most J nonempty segments suffice. Under Eq. (15), the readout $\hat{\jmath}(s) = \ell_{q_s}$ satisfies*

$$\max_j P(j, s, E) - P(\hat{\jmath}(s), s, E) \leq 2\delta_s + \eta + 2\beta. \tag{16}$$

These guarantees concern labeling under fixed full-pool recovery; proofs appear in Appendix A.

Profiling cost. Algorithm 1 profiles active models in parallel, recording each first frontier passage $\hat{e}_j$ and its exposure $\hat{\phi}_j = s_r \hat{e}_j$. Unreached boundaries retain $E + 1$. If only Ω_J remains, Theorem 1 permits stopping and assigning J to all later readouts; otherwise profiling continues to E. For $J \geq 2$, let $T_{\text{stop}} = \min(\hat{e}_{J-1}, E)$. The reference profiling cost is

$$C_{\text{prof}} = \sum_{j=1}^{J-2} \min(\hat{e}_j, E) + 2T_{\text{stop}} \leq JE. \tag{17}$$

Candidate selection. The recorded thresholds define budget intervals $\mathcal{B}_j = [\hat{e}_{j-1} s_r/E, \hat{e}_j s_r/E) \cap (0, s_r]$, each associated with labeling checkpoints Ω_j. Skipped candidates yield empty intervals and are excluded from selection. For a target budget $0 < s \leq s_r$ and training horizon E, we select the checkpoints associated with the interval containing s. Equivalently, we read the recorded frontier at the corresponding reference epoch $q_s = \lfloor sE/s_r \rfloor$:

$$\hat{\jmath}(s) = \ell_{\lfloor sE/s_r \rfloor} = j \quad \iff \quad s \in \mathcal{B}_j. \tag{18}$$

We train the model for E epochs on $\Pi_s(\mathcal{U}_r^{\text{cur}})$ with soft supervision $\sigma_{\hat{\jmath}(s)}$ from $\Omega_{\hat{\jmath}(s)}$. With the architecture and training protocol fixed, unseen budgets in $(0, s_r]$ require no further recovery or reference profiling. Corollary 1 bounds the selection regret under its stated conditions.

Table 1: Post-evaluation performance comparison with state-of-the-art baselines. Within each evaluation setting, all models are trained for 300 epochs using identical training configurations. For FADRM+, DELT, and RDED, the training set at each target IPC is sampled from the corresponding distilled dataset. **Bold** and underline indicate the best and second-best results, respectively.

Dataset	IPC (Ratio)	ResNet-18				ResNet-101				MobileNet-V2		
		RDED	DELT	FADRM+	**PAC**	RDED	DELT	FADRM+	**PAC**	DELT	FADRM+	**PAC**
CIFAR-10	1 (0.02%)	22.9±0.4	24.0±0.8	22.4±1.6	**27.5**±0.3	18.7±0.1	20.4±1.0	20.7±1.7	**23.7**±2.3	20.2±0.4	22.7±0.8	**23.2**±1.1
	5 (0.1 %)	28.4±0.3	31.2±2.3	31.6±1.2	**38.7**±2.0	27.0±0.9	26.9±1.6	30.7±0.6	**33.7**±0.5	24.4±0.8	34.3±2.5	**35.7**±1.1
	10 (0.2%)	37.1±0.3	43.0±0.9	44.0±0.1	**59.7**±0.6	33.7±0.3	37.4±1.2	39.5±0.8	**44.2**±1.2	29.3±0.3	38.9±1.1	**49.7**±1.1
	20 (0.4 %)	44.0±2.2	50.5±1.9	54.4±0.5	**68.5**±1.2	40.6±0.5	41.7±0.4	49.4±1.1	**62.9**±2.0	39.9±0.6	48.2±1.6	**59.8**±1.6
	50 (1.0%)	62.1±0.1	64.9±0.9	71.0±1.5	**79.5**±0.5	51.6±0.4	54.1±0.8	65.7±0.7	**79.1**±0.4	42.9±2.2	60.9±1.0	**76.1**±0.4
	Whole Dataset	95.0				95.6				94.4		
CIFAR-100	1 (0.2%)	11.0±0.3	14.4±0.8	19.0±0.2	**31.4**±0.4	10.8±0.1	18.6±0.5	14.7±0.5	**22.6**±0.5	9.7±0.3	12.5±0.7	**17.9**±1.2
	5 (1.0 %)	29.2±0.7	32.0±0.3	53.4±0.5	**57.7**±0.3	23.6±1.6	31.3±1.2	53.2±0.9	**55.1**±0.1	15.5±0.8	36.8±0.3	**44.0**±0.6
	10 (2.0%)	42.6±0.2	48.0±0.1	60.7±0.3	**64.7**±0.2	41.1±0.2	43.5±0.3	62.2±0.1	**64.8**±0.4	24.7±0.2	52.3±0.2	**54.5**±0.1
	20 (4.0 %)	54.8±0.7	59.4±0.1	66.4±0.1	**68.5**±0.1	53.4±1.0	53.7±0.5	67.3±0.1	**68.4**±0.4	31.4±0.4	62.1±0.2	**63.9**±0.3
	50 (10.0%)	62.6±0.1	65.7±0.2	68.4±0.1	**71.5**±0.1	63.4±0.3	64.2±0.2	69.2±0.2	**71.0**±0.4	44.0±0.5	66.4±0.1	**68.5**±0.1
	Whole Dataset	78.9				79.5				76.4		
Tiny-ImageNet	1 (0.2%)	9.7±0.4	9.3±0.5	17.7±1.0	**30.0**±0.6	3.8±0.1	5.6±1.0	17.9±1.5	**26.8**±1.0	3.5±0.5	13.6±0.4	**22.1**±0.3
	5 (1.0%)	31.9±1.0	29.8±0.8	46.3±0.3	**48.8**±0.4	14.6±1.1	26.2±0.5	46.6±0.5	**50.2**±0.4	13.6±0.5	42.0±0.5	**44.0**±0.4
	10 (2.0%)	41.9±0.2	43.0±0.9	52.3±0.2	**53.8**±0.4	22.9±3.3	42.8±0.9	53.6±0.3	**56.9**±0.3	26.5±0.5	51.0±0.1	**52.1**±0.4
	20 (4.0%)	51.0±0.3	50.6±0.4	55.5±0.1	**57.0**±0.1	30.8±0.7	52.7±0.4	57.5±0.5	**59.9**±0.3	40.8±0.8	55.7±0.3	**57.1**±0.2
	50 (10.0%)	58.2±0.1	55.7±0.5	58.6±0.3	**60.4**±0.1	41.2±0.4	58.5±0.3	60.7±0.5	**62.0**±0.7	51.3±0.5	58.7±0.2	**60.3**±0.1
	Whole Dataset	68.9				70.6				65.8		
ImageNette	1 (0.1%)	35.8±1.0	24.1±1.8	40.0±0.7	**51.7**±0.6	25.1±2.7	19.4±1.7	28.4±0.8	**34.1**±1.0	19.1±1.0	30.2±1.8	**38.3**±0.6
	5 (0.5%)	45.9±2.4	55.7±2.5	63.5±0.3	**70.2**±0.2	35.1±1.3	41.4±2.4	55.0±1.3	**63.4**±1.8	50.1±1.6	52.7±0.8	**63.4**±0.7
	10 (1.0%)	61.4±0.4	66.0±1.4	69.5±0.2	**77.0**±0.1	54.0±0.4	55.4±6.2	67.0±0.5	**72.6**±0.3	64.7±1.4	63.4±0.8	**72.5**±0.5
	20 (2.0%)	72.9±1.3	69.1±0.5	78.8±0.6	**81.1**±0.4	63.0±2.0	58.6±2.4	76.1±0.7	**79.5**±0.1	71.9±0.8	75.1±0.3	**78.7**±0.7
	50 (5.3%)	80.4±0.4	**88.2**±1.2	84.5±0.1	85.8±0.2	75.0±1.2	83.3±1.1	82.1±0.4	**85.6**±0.3	**85.7**±0.4	82.5±0.4	84.7±0.1
	Whole Dataset	93.8				89.3				90.0		
ImageNet-1K	1 (0.1%)	6.6±0.2	3.9±0.4	13.4±0.6	**31.5**±0.4	5.9±0.4	2.6±0.5	12.0±1.6	**30.2**±1.0	1.0±0.2	7.3±0.2	**18.2**±0.6
	5 (0.4%)	23.2±0.4	30.3±0.8	41.0±0.5	**45.8**±0.4	25.2±1.2	30.2±0.7	45.9±0.9	**51.7**±1.6	18.4±1.0	34.0±1.1	**41.3**±0.4
	10 (0.8%)	42.0±0.1	46.1±0.4	50.9±0.3	**52.5**±0.2	48.3±1.0	48.5±1.6	57.1±1.0	**58.7**±1.0	35.5±0.7	45.5±0.5	**50.1**±0.2
	20 (1.6%)	50.4±0.1	53.2±0.3	54.9±0.2	**56.3**±0.1	54.7±0.7	58.6±1.2	60.5±1.3	**63.1**±0.4	46.1±0.4	49.7±1.1	**54.6**±0.1
	50 (3.9%)	56.5±0.1	59.2±0.4	61.0±0.1	**61.8**±0.1	61.2±0.4	66.1±0.5	67.1±0.2	**68.0**±0.2	56.2±0.3	58.3±0.1	**59.4**±0.2
	Whole Dataset	69.8				77.4				71.9		

4 EXPERIMENTS

4.1 EXPERIMENTAL DATASETS AND CONFIGURATION

Datasets. To assess the scalability of our method across different image resolutions and dataset sizes, we conduct experiments on CIFAR-10 and CIFAR-100 (32×32) (Krizhevsky and Hinton, 2009), Tiny-ImageNet (64×64) (Le and Yang, 2015), ImageNet-1K (224×224) (Deng et al., 2009), and ImageNette, a subset of ImageNet-1K.

Baseline Methods. We adopt RDED (Sun et al., 2024), DELT (Shen et al., 2025), and FADRM+ (Cui et al., 2025a) as our primary baselines, as they represent state-of-the-art performance. Since **PAC** is a plug-in component, we integrate it into the strongest baseline, FADRM+. The consistent performance gains across different settings directly demonstrate the effectiveness of our model selection technique.

4.2 MAIN RESULTS

Results Analysis. As shown in Table 1, **PAC** consistently improves over FADRM+ and outperforms RDED and DELT in most settings. With ResNet-18 at IPC 1, **PAC** outperforms FADRM+ by impressive margins of 18.1% and 12.4% on ImageNet-1K and CIFAR-100, respectively. At IPC 50, the corresponding gains remain 0.8% and 3.1%. We attribute the low-IPC gains mainly to successfully selecting checkpoint candidates whose maturity matches the limited data budget. At larger IPC budgets, ensembling checkpoints from nearby transitional maturity stages combines complementary knowledge and provides richer supervision than fully trained checkpoints alone.

Table 2: Cross-architecture generalization.

Model	RDED	DELT	FADRM+	**PAC**
ResNet-18	23.2	30.3	41.0	**45.8**
ResNet-50	27.1	38.3	50.5	**54.2**
ResNet-101	25.2	30.2	45.9	**51.7**
EfficientNet-B0	32.0	31.4	45.1	**49.2**
MobileNetV2	20.3	18.4	34.0	**41.3**
ShuffleNetV2-0.5x	9.6	8.6	24.8	**30.0**
Swin-Tiny	17.1	11.1	25.1	**30.3**
ConvNext-Tiny	19.7	14.6	37.4	**39.8**
DeiT-Tiny	7.0	2.9	7.9	**10.5**

Cross-Architecture Generalization. Table 2 compares top-1 accuracy under **PAC**, RDED, DELT, and FADRM+ on ImageNet-1K at IPC 5 across nine architectures spanning CNNs and transformers. **PAC** outperforms all baselines on all of the architectures. For example, accuracy rises from 45.1% to

49.2% on EfficientNet-B0 and from 25.1% to 30.3% on Swin-Tiny. These gains across CNN and transformer models of varying capacities support budget-aware checkpoint selection in the low-IPC regime, where teacher maturity is most mismatched to the data budget.

Comparison with Joint Search. To test whether recovery and labeling require joint checkpoint alignment, we compare **PAC** with *Joint Search*, which exhaustively selects the best recovery–labeling pair within its candidate set for each budget (equation 5) as shown in Table 3. **PAC** replaces window-specific recovery with a shared bank from full-pool curriculum recovery, enabling independent labeling alignment. This design outperforms Joint Search by 2% at IPC 5, with a 19.4× alignment speedup by reusing the shared bank across labeling candidates.

Table 3: Comparison with Joint Search on CIFAR-100.

Top-1 (%) ↑	Joint Search	**PAC**
IPC=1	30.6	**31.4**
IPC=5	55.7	**57.7**
IPC=50	71.4	**71.5**
Alignment (h)↓	37.93	**1.96**

4.3 Ablation Study

Impact of Checkpoint Selection. To evaluate the effectiveness of our checkpoint selection strategy, we compare **PAC** with *Ensemble all*, which uses the full checkpoint pool as a labeling ensemble. We fix recovery and vary only the labeling checkpoints. As shown in Table 4, **PAC** consistently outperforms this baseline across all evaluated IPCs, with gains of 4.7% at IPC 1. These results suggest that **PAC** benefits from selecting checkpoints whose maturity matches the target data budget, as indiscriminate aggregation may introduce mismatched supervision.

Table 4: Impact of Checkpoint Selection.

IPC	1	2	3	4	5	6	10	20	50
Ensemble all	26.7	40.6	48.2	52.6	54.7	55.3	59.5	63.7	65.0
PAC	**31.4**	**43.9**	**50.0**	**54.3**	**57.7**	**58.9**	**64.7**	**68.5**	**71.5**

Impact of Checkpoint Pool Size (M). Increasing M samples teacher maturity more finely and, for fixed c and τ, introduces more candidate windows to profile. Table 5 shows that accuracy peaks at $M = 6$ across different IPCs, and then declines at large M. Denser sampling narrows the maturity span of each candidate, so its checkpoints may provide redundant supervision and less complementary knowledge. Larger pools can increase profiling cost without improving candidate quality or downstream model performance.

Table 5: Impact of Pool Size.

M	4	6	8	10
IPC=1	28.9	**31.4**	30.3	29.8
IPC=5	54.9	**57.7**	56.2	55.9
IPC=50	70.8	**71.5**	71.3	71.2

Impact of Ensemble Capacity (c). Table 6 shows that accuracy peaks at $c = 2$ and declines for larger windows, with $c = 5$ underperforming $c = 1$. Moderate windows can combine complementary knowledge from nearby maturity stages, whereas larger windows approach *Ensemble all* and may introduce supervision poorly matched to the target budget, reducing the benefit of combining multiple teachers.

Table 6: Impact of Capacity.

c	1	2	3	4	5
IPC=1	29.8	**31.4**	30.6	29.5	30.1
IPC=5	54.2	**57.7**	55.7	55.6	55.4
IPC=50	70.9	**71.5**	71.0	68.5	66.8

Impact of Stride (τ). Stride τ controls the spacing between window starting positions. With M and c fixed, changing τ preserves the checkpoint spacing and maturity span within each candidate. Table 7 shows that $\tau = 1$ achieves the highest accuracy at all three IPC budgets. Increasing τ reduces the number of candidates to profile but skips some valid windows, potentially excluding ensembles suited to a target budget. Setting $\tau = 1$ retains every valid window in the fixed pool, giving checkpoint selection more options for matching supervision to each budget.

Table 7: Impact of Stride.

τ	1	2	3
IPC=5	**57.7**	56.3	56.7
IPC=20	**68.5**	68.3	67.9
IPC=50	**71.5**	71.2	70.6

Impact of Selection Tolerance η. Table 8 reports top-1 agreement and regret (mean accuracy gap) against full 300-epoch optima across CIFAR-100 IPCs. Agreement peaks at 100% for $\eta = 1.0$: smaller tolerances risk noisy elimination, while larger ones delay transitions. Regret is at most 0.56% across the tested η values, indicating that model accuracy loss remains small even when the selection tolerance is not optimally chosen.

Table 8: Selection quality.

η	0	0.5	1	1.5	2
Agree. (%)	98%	98%	**100%**	92%	86%
Regret (pp)	0.34	0.34	**0**	0.25	0.56

Impact of Recovery Strategy. With labeling fixed per IPC, we compare **PAC**'s full-pool round-robin recovery with *Matched*, which uses only the selected labeling checkpoints for image recovery. **PAC** achieves higher accuracy at every IPC (Table 9), including an increase from 55.0% to 57.7% at IPC 5. These results support full-pool recovery, which uses checkpoints from all maturity levels to progressively refine the same synthetic images within a fixed overall number of recovery iterations.

Table 9: Impact of Recovery Strategy.

	IPC=1	IPC=5	IPC=10	IPC=20	IPC=50
Matched	31.3	55.0	63.7	67.7	69.9
PAC	**31.4**	**57.7**	**64.7**	**68.5**	**71.5**

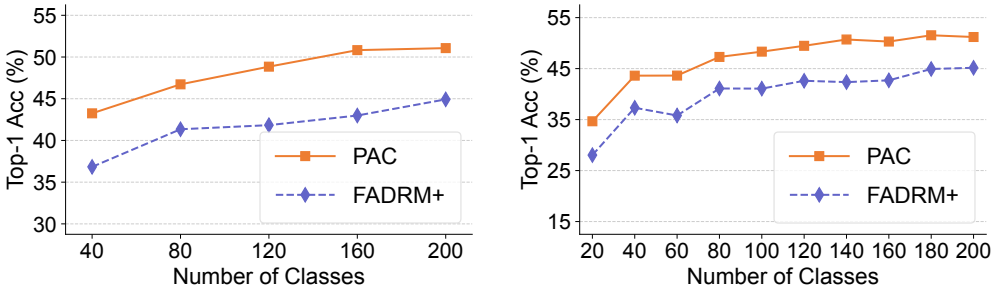

Figure 4: Five-step and ten-step class-incremental learning on Tiny-ImageNet with IPC=5.

4.4 Analysis

Effect of Prefix Elimination. Table 10 shows a reduction in profiling time from 8.52 h to 1.96 h (4.35× speedup), with 100% selection agreement across reference epochs. As the selection frontier advances, superseded candidates are retired, avoiding further training and evaluation of their models at later epochs. The unchanged selections suggest that prefix elimination avoids redundant candidate evaluation while preserving all the budget-dependent choices of full profiling across all evaluated reference epochs.

Table 10: Prefix elimination.

	Time (h)	Agreement (%)
w/o	8.52	Reference
w/	1.96	100%

Computational Efficiency of PAC. Table 11 compares alignment time across IPC 1, 5, 10, 20, and 50, excluding final model training. **PAC** reduces the total from 74.38 to 1.96 hours (37.95×). Labeling selection drops from 69.75 to 0.93 hours, below the 1.03 hours required for recovery. This reduction addresses *Joint Search*'s main bottleneck: repeated model training to evaluate recovery–labeling pairs. Reusing the bank and reference profile reduces this cost while retaining budget-specific checkpoint choices.

Table 11: Alignment time (h).

	Recovery	Labeller
Joint Search	4.63	69.75
PAC	1.03	0.93

Compatibility with Other Frameworks. We integrate **PAC** into SRe^{2}L (Yin et al., 2023) and evaluate it on CIFAR-100. Table 12 shows that **PAC** improves top-1 accuracy over SRe^{2}L alone across all four IPC budgets. At IPC 5, accuracy increases by 21.8 percentage points, from 31.1% to 52.9%. The gain is also substantial at IPC 1 (14.9 points), compared with 3.8 points at IPC 20, suggesting greater benefits under tighter synthetic data budgets and supporting the transfer of **PAC** beyond FADRM+.

Table 12: Compatibility of **PAC**.

IPC	1	5	10	20
SRe^{2}L	11.2	31.1	47.5	60.5
+ **PAC**	**26.1**	**52.9**	**59.2**	**64.3**

4.5 Application: Continual Learning

Following DM (Zhao and Bilen, 2023), we evaluate five- and ten-step class-incremental learning on Tiny-ImageNet at IPC 5. The two schedules introduce 40 and 20 new classes per stage, respectively. **PAC** outperforms FADRM+ under both schedules (Fig. 4). The gains persist as new classes are introduced, including at the final stage of each schedule. These results support applying **PAC** to continual learning under a fixed IPC budget as the set of observed classes expands with each stage.

5 Conclusion

We presented **P**rofile-Once **A**lignment of **C**heckpoints (**PAC**), a plug-in framework for reducing checkpoint alignment costs across IPC budgets in decoupled dataset distillation. By combining a shared image bank from full-pool curriculum recovery with exposure-indexed rank transfer, **PAC** uses a single profile at the largest budget to select labeling checkpoints suited to each target budget. Critical-exposure profiling further reduces evaluation cost by retiring superseded candidates, with our analysis establishing sufficient conditions for permanent elimination and bounds on selection regret. The resulting procedure supports checkpoint selection at previously unseen budgets within the reference range without additional recovery or profiling. Experiments demonstrate strong performance across datasets, IPC budgets, and model architectures, supporting budget-aware checkpoint selection.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval and discovery, including identifying potentially relevant prior work. Additionally, we used these tools to assist with drafting and polishing portions of the manuscript, including improvements to grammar, wording, clarity, and readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates dataset distillation to improve the efficiency of data storage and model training. Our experiments use publicly available benchmarks under their respective usage terms and do not involve additional data collection from human participants. To limit direct exposure of the original training data during downstream training, we use distilled images and obtain their supervision from frozen pretrained teacher checkpoints, without directly accessing the original training samples in these training and labeling steps. Before releasing synthetic samples, we conduct nearest-neighbor searches and perceptual similarity checks against the source data, removing samples that closely reproduce original images to mitigate memorization risks. The accompanying release terms permit lawful research use only and expressly prohibit surveillance and identification applications. Downstream users should additionally assess potential biases, comply with source dataset licenses, and avoid applications that could harm individuals or communities. We commit to conducting and reporting this research in accordance with the principles outlined in the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Our experiments use the publicly available benchmarks described in Section 4. Appendix C provides the checkpoint pool construction procedure, model configurations, training schedules, and hyperparameters for pretraining, synthetic data generation, and downstream evaluation. The assumptions underlying our theoretical results are stated in the main text, with proofs provided in Appendix A. An anonymized implementation is included in the supplementary material, together with configuration files and scripts for teacher pretraining, synthetic data generation, checkpoint profiling and selection, and downstream evaluation. The accompanying README specifies the software dependencies and commands for reproducing the main experiments. We will also release these materials publicly.

REFERENCES

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.

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 10718–10727. IEEE, 2022.

Mingyang Chen, Jiawei Du, Bo Huang, Yi Wang, Xiaobo Zhang, and Wei Wang. Influence-guided diffusion for dataset distillation. In *International Conference on Learning Representations*, volume 2025, pages 85143–85164, 2025.

Yilan Chen, Wei Huang, and Tsui-Wei Weng. Provable and efficient dataset distillation for kernel ridge regression. *Advances in Neural Information Processing Systems*, 37:88739–88771, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jiacheng Cui, Xinyue Bi, Yaxin Luo, Xiaohan Zhao, Jiacheng Liu, and Zhiqiang Shen. Fadrm: Fast and accurate data residual matching for dataset distillation. *Advances in Neural Information Processing Systems*, 38:62610–62655, 2025a.

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

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 International Conference on Machine Learning (ICML)*, 2026a.

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.

Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optimizing distributional geometry alignment with optimal transport for generative dataset distillation. *Advances in Neural Information Processing Systems*, 38:121368–121408, 2025c.

Xiao Cui, Yulei Qin, Mo Zhu, Wengang Zhou, Hongsheng Li, and Houqiang Li. Geometry-aware dataset condensation for diffusion model training. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=P7RGcAOZZ3.

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.

Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. Exploiting inter-sample and inter-feature relations in dataset distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17057–17066. IEEE, 2024.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3758. IEEE, 2023a.

Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. *Advances in Neural Information Processing Systems*, 36:67487–67504, 2023b.

Jiawei Du, Xin Zhang, Juncheng Hu, Wenxing Huang, and Joey T Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. *Advances in neural information processing systems*, 37:119443–119465, 2024.

Ziang Gan, Qi Zhu, and Libao Zhang. Set-coupled guidance: Set-level coordination in diffusion-based dataset distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=onosLtwoiK.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15793–15803. IEEE, 2024.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *International Conference on Learning Representations*, volume 2024, pages 56999–57020, 2024.

Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1345–1354. IEEE, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pages 4015–4026. IEEE, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

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

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.

Yongmin Lee and Hye Won Chung. SelMatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26546–26567. PMLR, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 33963–33972, 2026a.

Xuhui Li, Zhengquan Luo, Zihui Cui, Kai Zhao, and Zhiqiang Xu. GeoDM: Geometry-aware distribution matching for dataset distillation. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=ANVPCRmZQQ.

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*, 2024a.

Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. Dataset distillation by automatic training trajectories. In *European Conference on Computer Vision*, pages 334–351. Springer, 2024b.

Haoyang Liu, Yijiang Li, Tiancheng Xing, Peiran Wang, Vibhu Dalal, Luwei Li, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1205–1215. IEEE, 2025.

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

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.

Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. In *International Conference on Machine Learning*, pages 22649–22674. PMLR, 2023.

Zhiheng Ma, Anjia Cao, Funing Yang, Yihong Gong, and Xing Wei. Curriculum dataset distillation. *IEEE Transactions on Image Processing*, 34:4176–4187, 2025.

Brian Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Unlocking dataset distillation with diffusion models. *Advances in Neural Information Processing Systems*, 38: 72581–72614, 2025.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=l-PrrQrK0QR.

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, 2021b.

Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. *Advances in Neural Information Processing Systems*, 37:131946–131971, 2024.

Mehdi Rezagholizadeh, Aref Jafari, Puneeth SM Saladi, Pranav Sharma, Ali Saheb Pasand, and Ali Ghodsi. Pro-kd: Progressive distillation by following the footsteps of the teacher. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4714–4727, 2022.

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

Jeffrey A Chan Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. MGD^3 : Mode-guided dataset distillation using diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52861–52876. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/santiago25a.html.

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.

Xinyi Shang, Peng Sun, Bei Shi, Zixuan Wang, and Tao Lin. Condensing large-scale datasets directly with minimal information loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

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

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, 2024b.

Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4797–4806. IEEE, 2025.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D 4 m: Dataset distillation via disentangled diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5809–5818. IEEE, 2024.

Duo Su, Huyu Wu, Huanran Chen, Yiming Shi, Yuzhu Wang, Xi Ye, and Jun Zhu. Diffusion models as dataset distillation priors. In *International Conference on Learning Representations*, volume 2026, pages 26532–26552, 2026.

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

Minh-Tuan Tran, Trung Le, Xuan-May Le, Thanh-Toan Do, and Dinh Phung. Enhancing dataset distillation via non-critical region refinement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10015–10024. IEEE, 2025.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619, 2022a.

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Cao 2: Rectifying inconsistencies in diffusion-based dataset distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4731. IEEE, 2025a.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12196–12205. IEEE, 2022b.

Kai Wang, Zekai Li, Zhi-Qi Cheng, Samir Khaki, Ahmad Sajedi, Ramakrishna Vedantam, Konstantinos N Plataniotis, Alexander Hauptmann, and Yang You. Emphasizing discriminative features for dataset distillation in complex scenarios. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30451–30461. IEEE, 2025b.

Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25570–25580. IEEE, 2025c.

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

Qianxin Xia, Zhiyong Shu, Wenbo Jiang, Jiawei Du, Jielei Wang, and Guoming Lu. DIVER: Diving deeper into distilled data via expressive semantic recovery. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=PL8qP67p6E.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PlaZD2nGCl.

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.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE transactions on pattern analysis and machine intelligence*, 46(1):150–170, 2024.

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

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

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mSAKhLYLSsl.

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

Lin Zhao, Yushu Wu, Xinru Jiang, Jianyang Gu, Yanzhi Wang, Xiaolin Xu, Pu Zhao, and Xue Lin. Taming diffusion for dataset distillation with high representativeness. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=P1RMiAn4Tr.

Wenliang Zhong, Haoyu Tang, Qinghai Zheng, Mingzhu Xu, Yupeng Hu, and Weili Guan. Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25581–25589. IEEE, 2025.

Letian Zhou, Songhua Liu, and Xinchao Wang. Coda: From text-to-image diffusion models to training-free dataset distillation. In *International Conference on Learning Representations*, volume 2026, pages 81947–81980, 2026.

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

Yawen Zou, Guang Li, Duo Su, Zi Wang, Jun Yu, and Chao Zhang. Dataset distillation via vision-language category prototype. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2941–2950. IEEE, 2025.

Appendix of PAC

CONTENTS

A Proofs **16**

A.1 Proof of Theorem 1 16

A.2 Proof of Corollary 1 17

B Additional Experiments **18**

B.1 Full Profiling without Prefix Elimination 18

B.2 Impact of Profiling Hyperparameters. 18

C Experimental Details **19**

C.1 Hyper-Parameters 19

D Distilled Image Visualization **23**

A PROOFS

A.1 PROOF OF THEOREM 1

Proof of Theorem 1. Fix the potential full reference trajectories $P_j(e) = P(j, s_r, e)$ for $j = 1, \dots, J$ and $e \in \mathcal{E} = \{0, \dots, E\}$. These are the trajectories that would result if each student were trained through the full horizon; terminating one student does not change the trajectories of the others. Write

$$M(e) = \max_{1 \le j \le J} P_j(e), \qquad m(e) = \min \operatorname*{argmax}_{1 \le j \le J} P_j(e).$$

The assumed monotonicity concerns $m(e)$, not the ordering of every pair of candidate curves. In particular, candidates that are currently suboptimal may overtake one another repeatedly.

Work on the event that $|\widehat{P}_j(e) - P_j(e)| \le \beta$ for every evaluation actually performed. All arguments below are deterministic on this event. If a statistical procedure supplies this simultaneous event with a specified probability, the conclusions hold with at least that probability; no independence assumption on evaluation errors is required. The simultaneous quantifier matters because the active set depends on previous estimates: a bound for one fixed candidate and epoch alone would not establish the event used after adaptive selection. No evaluation-error bound is needed for a retired candidate at later epochs; its future exclusion follows from the structural assumption and the invariant below.

We prove the following invariant: at the start of epoch e, $\mathcal{A}_e$ is a nonempty suffix and contains every global maximizer at every epoch $t \ge e$. Initially $\mathcal{A}_0 = \{1, \dots, J\}$, so the invariant holds. Suppose it holds at epoch e. In particular, $m(e) \in \mathcal{A}_e$, and hence

$$\widehat{M}_e = \max_{j \in \mathcal{A}_e} \widehat{P}_j(e) \le M(e) + \beta, \qquad \widehat{P}_{m(e)}(e) \ge M(e) - \beta.$$

Consequently,

$$\widehat{P}_{m(e)}(e) \ge \widehat{M}_e - 2\beta \ge \widehat{M}_e - \eta.$$

Thus the true leftmost maximizer belongs to the near-best set in Eq. (12). This set is nonempty, and its minimum satisfies $\ell_e \le m(e)$.

Every candidate removed by Eq. (13) has index $i < \ell_e \le m(e)$. For every subsequent epoch $t \ge e$, the monotonicity assumption gives $i < m(e) \le m(t)$. By the definition of the *leftmost* maximizer,

$$P_i(t) < M(t), \qquad t = e, \dots, E.$$

This strict inequality does not require a unique maximizer: an index strictly below the smallest maximizing index cannot itself maximize. Previously removed candidates already satisfy the same property by the induction hypothesis. The updated set is therefore the suffix $\mathcal{A}_{e+1} = \{\ell_e, \dots, J\}$ and retains every future maximizer, completing the induction. In particular, $\ell_{e+1} \ge \ell_e$ whenever the next epoch is evaluated.

For the performance bound, the retained true maximizer also gives $\widehat{M}_e \ge M(e) - \beta$. The definition of ℓ_e then yields

$$P_{\ell_e}(e) \ge \widehat{P}_{\ell_e}(e) - \beta \ge \widehat{M}_e - \eta - \beta \ge M(e) - \eta - 2\beta.$$

Rearranging proves Eq. (14). The closed inequality in the near-best set is deliberate: candidates at exactly the tolerance boundary are retained, and choosing its smallest index preserves the inequality $\ell_e \le m(e)$ even in the presence of ties. When $\eta = \beta = 0$, this set is precisely the active maximizer set, so $\ell_e = m(e)$.

The additive bound is sharp under the stated assumptions. For two constant trajectories, take

$$(P_1, P_2) = (0, \eta + 2\beta), \qquad (\widehat{P}_1, \widehat{P}_2) = (\beta, \eta + \beta).$$

Both errors have magnitude β, and candidate 1 lies exactly on the near-best boundary. Thus $\ell_e = 1$ and its regret equals $\eta + 2\beta$. For scores restricted to $[0, 1]$, this example applies whenever the displayed values lie in that range.

Finally, if $\ell_e = J$, then $\ell_e \le m(e) \le J$ implies $m(e) = J$. Monotonicity forces $m(t) = J$ for every $t \ge e$, and all other candidates are strictly suboptimal. Profiling can stop and the recorded frontier can be extended by $\ell_t = J$ through E. Its subsequent reference regret is zero, so the stated guarantee remains valid without further evaluations. □

Why the condition concerns prefixes. For each boundary $j = 1, \dots, J-1$, define the prefix and suffix envelopes

$$L_j(e) = \max_{i \le j} P_i(e), \qquad R_j(e) = \max_{i > j} P_i(e), \qquad H_j(e) = R_j(e) - L_j(e).$$

Then $H_j(e) > 0$ if and only if $m(e) > j$: strict suffix dominance excludes every prefix index from the maximizer set, while a maximizer in the prefix would make $L_j(e) \ge R_j(e)$. Hence non-decreasing $m(e)$ is equivalent to persistence of strict positivity for each envelope difference H_j. For the converse, if $m(t) < m(e)$ for some $t > e$, the boundary $j = m(t)$ has $H_j(e) > 0$ but $H_j(t) \le 0$, violating persistence. This equivalence also explains why comparing arbitrary adjacent candidates is insufficient: the inequality $P_{j+1}(e) > P_j(e)$ need not exclude an earlier prefix candidate from being optimal. The update in Eq. (13) uses the performance of all active candidates to certify a prefix, while retaining every configuration to its right. No ordering of crossings between individually suboptimal curves is imposed or used in the proof.

A.2 Proof of Corollary 1

Proof of Corollary 1. Use the non-decreasing recorded frontier ℓ_e from Theorem 1, including its constant continuation after any early termination. For $j = 1, \dots, J-1$, define

$$B_j = \{e \in \mathcal{E} : \ell_e > j\}.$$

The inclusion $B_{j+1} \subseteq B_j$ is automatic. Taking first elements, with the common convention $\min \varnothing = E + 1$, therefore gives $\widehat{e}_1 \le \cdots \le \widehat{e}_{J-1}$ for the first-passage thresholds defined in Corollary 1. Moreover, monotonicity of ℓ_e makes each nonempty B_j a terminal interval of the discrete horizon. Hence

$$\ell_e > j \quad \Longleftrightarrow \quad e \ge \widehat{e}_j, \qquad e \in \mathcal{E}.$$

This equivalence also holds at the boundary indices with $\widehat{e}_0 = 0$ and $\widehat{e}_J = E + 1$. Applying it to $j - 1$ and j gives

$$\ell_e = j \quad \Longleftrightarrow \quad \widehat{e}_{j-1} \le e < \widehat{e}_j,$$

which proves the reference-epoch interval characterization in Corollary 1.

These half-open intervals cover every epoch of the reference grid, including 0 and E. Equal thresholds produce empty intervals and represent skipped candidates. A threshold of $E + 1$ denotes a boundary never crossed within the horizon; it imposes no requirement that the final frontier equal J. The intervals describe the recorded selection rule exactly. For positive tolerance or observation error, their performance guarantee is Eq. (14); they are not claimed to be exact oracle intervals. They become oracle intervals when $\eta = \beta = 0$.

For a target budget $0 < s \le s_r$, let $q_s = \lfloor sE/s_r \rfloor$ as in Eq. (18), and set $\widehat{\jmath}(s) = \ell_{q_s}$. Because the thresholds are integers,

$$\widehat{e}_{j-1} \le q_s < \widehat{e}_j \quad \Longleftrightarrow \quad \frac{s_r \widehat{e}_{j-1}}{E} \le s < \frac{s_r \widehat{e}_j}{E}.$$

Intersecting these budget intervals with $(0, s_r]$ proves Eq. (18), including the correct assignment of $s = s_r$. Only thresholds $\widehat{e}_j \le E$ correspond to observed critical exposures $\widehat{\phi}_j = s_r \widehat{e}_j$. The value $E + 1$ is a sentinel for constructing the finite-horizon partition, not a measured exposure or an additional supported budget.

To establish the target-budget bound, abbreviate $T_j = P(j, s, E)$ and $q = q_s$. The exposure bridge in Eq. (15) states

$$P_j(q) + c_s - \delta_s \le T_j \le P_j(q) + c_s + \delta_s.$$

Therefore, with $a = \widehat{\jmath}(s) = \ell_q$,

$$\begin{aligned} \max_j T_j - T_a &\le \max_j \{P_j(q) + c_s + \delta_s\} - \{P_a(q) + c_s - \delta_s\} \\ &= \max_j P_j(q) - P_a(q) + 2\delta_s \\ &\le \eta + 2\beta + 2\delta_s. \end{aligned}$$

The configuration-independent shift c_s cancels exactly, proving Eq. (16). □

Relating the exposure bridge to the performance decomposition. The approximation in Eq. (8) can be made explicit by defining its residual:

$$P(j, s, e) = G_j(se) + \Delta(s, e) + r_j(s, e).$$

For $q_s = \lfloor sE/s_r \rfloor$, choose $c_s = \Delta(s, E) - \Delta(s_r, q_s)$. Then

$$P(j, s, E) - P_j(q_s) - c_s = G_j(sE) - G_j(s_r q_s) + r_j(s, E) - r_j(s_r, q_s).$$

Define $\epsilon_{\text{target},s} = \max_j |r_j(s, E)|$ and $\epsilon_{\text{ref},s} = \max_j |r_j(s_r, q_s)|$. An admissible bridge bound is

$$\delta_s = \epsilon_{\text{target},s} + \epsilon_{\text{ref},s} + \max_j \left| G_j(sE) - G_j(s_r q_s) \right|.$$

When the matched epoch lies on the evaluation grid, the final term is zero. Otherwise it records the discrepancy introduced by rounding down to an evaluated reference epoch. The decomposition alone does not bound this term by the size of the exposure gap; such a rate would require additional regularity of the functions G_j, which is not assumed.

Finite-horizon termination and profiling cost. For $J \geq 2$, the first epoch with $\ell_e = J$ is $\widehat{e}_{J-1}$ if this threshold is observed. The theorem justifies stopping there and assigning J to all subsequent frontier readouts; otherwise profiling continues through E. Thus the stopping epoch is $T_{\text{stop}} = \min(\widehat{e}_{J-1}, E)$. Candidate $j < J$ is trained through the epoch at which the frontier first exceeds j, or through E if that never occurs. Its training cost is $\min(\widehat{e}_j, E)$ epochs, while candidate J is trained through T_{stop}. The ordered thresholds ensure that no other candidate requires training beyond T_{stop}. Consequently,

$$\begin{aligned} C_{\text{prof}} &= \sum_{j=1}^{J-1} \min(\widehat{e}_j, E) + T_{\text{stop}} \\ &= \sum_{j=1}^{J-2} \min(\widehat{e}_j, E) + 2\min(\widehat{e}_{J-1}, E) \leq JE, \end{aligned}$$

as stated in Eq. (17). Epoch 0 requires no training, so thresholds at zero are consistent with this count; evaluation overhead is not counted in student-epochs. Simultaneous deletions and unobserved thresholds are both covered. For $J = 1$, the selection is trivial and no reference profiling is needed.

B Additional Experiments

B.1 Full Profiling without Prefix Elimination

Table 13 reports complete profiling trajectories on CIFAR-100 at IPC 50, with all labeling candidates trained for 300 epochs,where epoch 6 corresponds to IPC=1 and epoch 12 corresponds to IPC=2 and so forth. Bold entries indicate the labeling candidate Ω selected by profiling at each checkpoint, while asterisks (*) mark the ground-truth selections. Gray entries correspond to the computations that would be skipped under Prefix Elimination. Crucially, full profiling identifies identical critical IPC boundaries to those obtained with Prefix Elimination (specifically, IPCs 2, 3, 4, and 7), achieving 100% agreement. This confirms that Prefix Elimination significantly reduces profiling overhead without compromising the accuracy of boundary detection or candidate selection.

B.2 Impact of Profiling Hyperparameters.

To examine the sensitivity of the hyperparameters used for reference profiling at IPC 50. The profiling learning rate follows the cosine schedule

$$\alpha(t) = \frac{\alpha_0}{2}\left(1 + \cos\left(\frac{\pi t}{\lambda E}\right)\right), \qquad 0 \leq t \leq E, \tag{19}$$

where α_0 is the initial learning rate, E is the total number of profiling epochs, and t denotes training progress in epochs. The parameter λ controls the decay rate: a larger λ slows the decay and maintains a higher learning rate for longer. We evaluate four combinations of $\alpha_0 \in 0.001, 0.0005$ and $\lambda \in \{1, 2\}$, while keeping the reference profiling evaluation protocol fixed. As shown in Table 14, $\alpha_0 = 0.001$ with $\lambda = 1$ achieves **0** regret. The other configurations yield mean regrets of 0.82% to 0.84%, suggesting that selection quality is relatively stable across the tested settings.

Table 13: Detailed profiling trajectories on CIFAR-100 at IPC 50 without Prefix Elimination. All labeling candidates are trained for 300 epochs. **Bold** entries and asterisks (*) indicate profiling selections and ground-truth selections, respectively. Black and gray entries indicate computations required and skipped under Prefix Elimination, respectively.

Epoch	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5
6	**33.88***	32.86	22.76	22.44	21.50
12	41.50	**43.14***	42.42	40.77	35.47
18	43.82	46.82	**50.40***	50.91	46.47
24	45.33	50.58	53.38	**55.06**	52.85
30	44.13	49.90	54.61	**57.85***	56.28
42	44.12	50.69	56.42	59.02	**60.11***
60	45.60	52.46	58.17	63.42	**63.46***
120	47.84	54.72	61.14	67.25	**68.44***
180	46.95	54.72	61.90	68.22	**70.31***
300	47.62	54.97	62.39	69.03	**70.93***

Table 14: Sensitivity to profiling hyperparameters at IPC 50.

Setting	Learning Rate	λ	Mean Regret (pp) $\downarrow$
S1	0.001	1	**0.00**
S2	0.001	2	0.82
S3	0.0005	1	0.84
S4	0.0005	2	0.84

C Experimental Details

We adopt a unified training and evaluation protocol to ensure fair and consistent comparisons across all evaluated methods. Within each evaluation setting, all student models are trained for 300 epochs using identical training configurations. We integrate **PAC** into FADRM+ and maintain the same student-training protocol to assess the effectiveness of our checkpoint alignment strategy. To facilitate reproducibility, we provide the detailed hyperparameters and implementation settings for model pretraining, synthetic data generation, and post-evaluation below.

C.1 Hyper-Parameters

Pretraining Phase. To establish a full-dataset performance reference across different backbone architectures, we train teacher models on each original dataset using the hyperparameters specified in Table 15. Intermediate checkpoints are retained during training to capture different stages of teacher maturity. After training, we construct a checkpoint pool of size $M = 6$ for each teacher by selecting the first retained checkpoint to reach each of six uniformly spaced accuracy milestones, ranging from 50% to 100% of the teacher's final accuracy.

Synthetic Data Generation. We adopt a consistent hyperparameter configuration across datasets for synthetic data generation. Image recovery follows a curriculum that incorporates all checkpoints from the checkpoint pools of different teacher architectures, progressing from less mature to more mature training stages. For each dataset, we directly synthesize a shared image bank at IPC 50. Distilled datasets at smaller IPC budgets are subsequently constructed by sampling the required number of images per class from this shared bank, without additional image optimization. This allows a single recovery process to support multiple target budgets.

Critical-Exposure Profiling. We conduct critical-exposure profiling using the hyperparameters specified in Table 17. The search is performed at IPC 50 using the checkpoint pools of size $M = 6$ constructed during pretraining. For each profiling run, we set the maximum training budget to 300 epochs and evaluate performance every 6 epochs to track the training progress. The ensemble capacity, stride, and selection tolerance are set to $c = 2$, $\tau = 1$, and $\eta = 1.0$, respectively. For each

student architecture, all profiling runs employ the post-evaluation settings assigned to IPC 50; the remaining training settings follow the post-evaluation configurations.

Post-Evaluation. For each dataset and student architecture, the critical IPC boundaries partition the target budgets into regions, each associated with a labeling ensemble Ω_j. We relabel the sampled synthetic images using the ensemble assigned to the corresponding budget region, following the BSSL soft-label generation protocol. We evaluate the resulting distilled datasets using three student architectures: ResNet-18, ResNet-101, and MobileNetV2. Each student is trained for 300 epochs using AdamW and KL-divergence loss, with RandomResizedCrop, horizontal flipping, and CutMix augmentations. Table 18 lists the initial learning rates and values of λ used in the cosine learning-rate schedule defined in Eq. equation 19, with $E = 300$. Larger λ slows the decay over the fixed training horizon. The assignments of these hyperparameter configurations across architectures and IPC budgets, together with batch sizes and critical IPC boundaries, are provided in Tables 19, 20, 21, 22, and 23 for CIFAR-100, CIFAR-10, Tiny-ImageNet, ImageNette, and ImageNet-1K, respectively.

Table 15: Shared hyperparameters for pre-trained models across all IPC settings.

Hyperparameter	**Value**
Optimizer	SGD
Learning Rate	0.1
Weight Decay	1e-4
Momentum	0.9
Epochs	300
Batch Size	128
Saved Checkpoints	16 (ImageNet-1K) / 40 (Other Datasets)
Scheduler	Cosine Annealing
Augmentation	RandomCrop, Horizontal Flip
Loss Function	Cross-Entropy

Table 16: Hyperparameters shared across all datasets for synthetic data generation.

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
Number of ARC	4
Scheduler	Cosine annealing
PAC Model	ResNet18, DenseNet-121, Mobilenet-V2, ShuffleNet-V2
Augmentation	RandomResizedCrop, Horizontal flip

Table 17: Hyperparameters for Critical-Exposure Profiling.

Hyperparameter	**Value**
Search IPC	50
Checkpoint Pool Size M	6
Max epochs	300
Eval-step	6 epoch
Ensemble Capacity c	2
Stride τ	1
Selection Tolerance η	1.0

Table 18: Hyperparameter settings with learning rate and λ.

Setting	Learning Rate	λ
S1	0.001	1
S2	0.001	2
S3	0.0005	1
S4	0.0005	2

Table 19: Hyperparameters for post-evaluation task for CIFAR-100.

Hyperparameter	Value
Optimizer	AdamW
S1	IPC50 (R18); IPC1, IPC5, IPC10, IPC20, IPC50 (MBV2)
S2	IPC10, IPC20 (R18)
S3	IPC1, IPC5, IPC10,IPC20, IPC50 (R101)
S4	IPC1 IPC5 (R18)
Soft Label Generation	BSSL
Loss Function	KL-Divergence
Batch Size	16
Epochs	300
Critical Points (IPC)	2,3,4,7 (R18); 2,4,4,11 (R101); 2,4,8,14 (MBV2)
Augmentation	RandomResizedCrop, Horizontal Flip, CutMix

Table 20: Hyperparameters for post-evaluation task for CIFAR-10.

Hyperparameter	Value
Optimizer	AdamW
S1	IPC50 (R18); IPC1, IPC5, IPC10, IPC20, IPC50 (MBV2)
S2	IPC10, IPC20 (R18)
S3	IPC1, IPC5, IPC10, IPC20, IPC50 (R101), IPC5 (R18)
S4	IPC1 (R18)
Soft Label Generation	BSSL
Loss Function	KL-Divergence
Batch Size	10
Epochs	300
Critical Points (IPC)	1,13,28,51 (R18); 7,18,25,51 (R101); 4,22,26,51 (MBV2)
Augmentation	RandomResizedCrop, Horizontal Flip, CutMix

Table 21: Hyperparameters for post-evaluation task on for Tiny-ImageNet.

Hyperparameter	Value
Optimizer	AdamW
S1	IPC50 (R18); IPC1, IPC5, IPC10, IPC20, IPC50 (MBV2)
S2	IPC1, IPC5, IPC10, IPC20, (R18)
S3	IPC1, IPC5, IPC10, IPC20, IPC50 (R101)
Soft Label Generation	BSSL
Loss Function	KL-Divergence
Batch Size	16
Epochs	300
Critical Points (IPC)	1,3,3,9 (R18); 2,2,3,7 (R101); 2,5,5,18 (MBV2)
Augmentation	RandomResizedCrop, Horizontal Flip, CutMix

Table 22: Hyperparameters for post-evaluation task on for ImageNette.

Hyperparameter	Value
Optimizer	AdamW
S1	IPC1, IPC5, IPC10, IPC20, IPC50 (MBV2)
S3	IPC5, IPC10, IPC20 (R18); IPC1, IPC5, IPC10, IPC20, IPC50 (R101)
S4	IPC1, IPC50 (R18)
Soft Label Generation	BSSL
Loss Function	KL-Divergence
Batch Size	10
Epochs	300
Critical Points (IPC)	1,8,9,25 (R18); 2,7,11,21 (R101); 4,4,10,31 (MBV2)
Augmentation	RandomResizedCrop, Horizontal Flip, CutMix

Table 23: Hyperparameters for post-evaluation task for ImageNet-1K.

Hyperparameter	Value
Optimizer	AdamW
S1	IPC50 (R18); IPC1, IPC5, IPC10, IPC20, IPC50 (MBV2)
S2	IPC1, IPC5, IPC10, IPC20 (R18)
S3	IPC1, IPC5, IPC10, IPC20, IPC50 (R101)
Soft Label Generation	BSSL
Loss Function	KL-Divergence
Batch Size	16
Epochs	300
Critical Points (IPC)	2,6,6,7 (R18); 2,2,2,3 (R101); 4,5,5,5 (MBV2)
Augmentation	RandomResizedCrop, Horizontal Flip, CutMix

D DISTILLED IMAGE VISUALIZATION

Additional visualizations of the data distilled by **PAC** are presented for CIFAR-100 (Fig. 5),CIFAR-10 (Fig. 6), Tiny-ImageNet (Fig. 7), ImageNette (Fig. 8), and ImageNet-1K (Fig. 9).

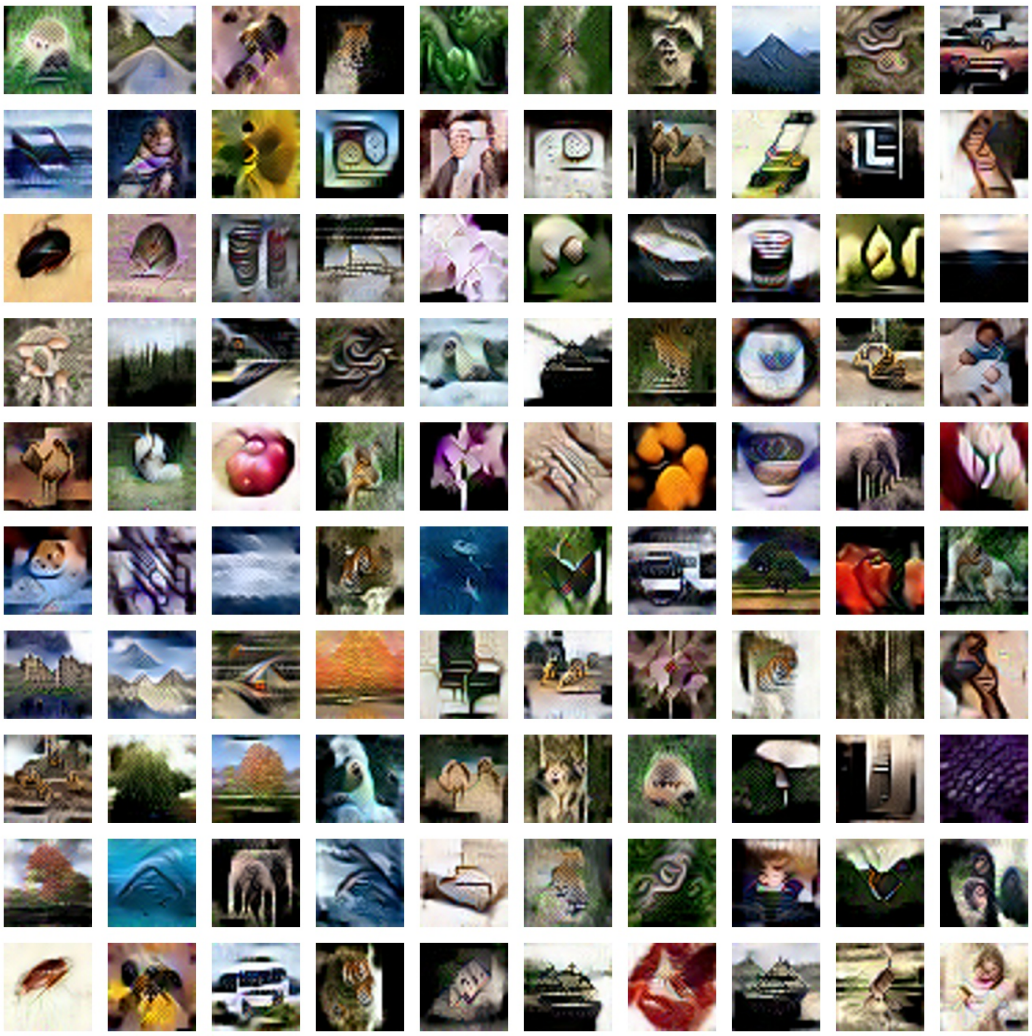

Figure 5: Visualization of synthetic data on CIFAR-100 generated by **PAC**.

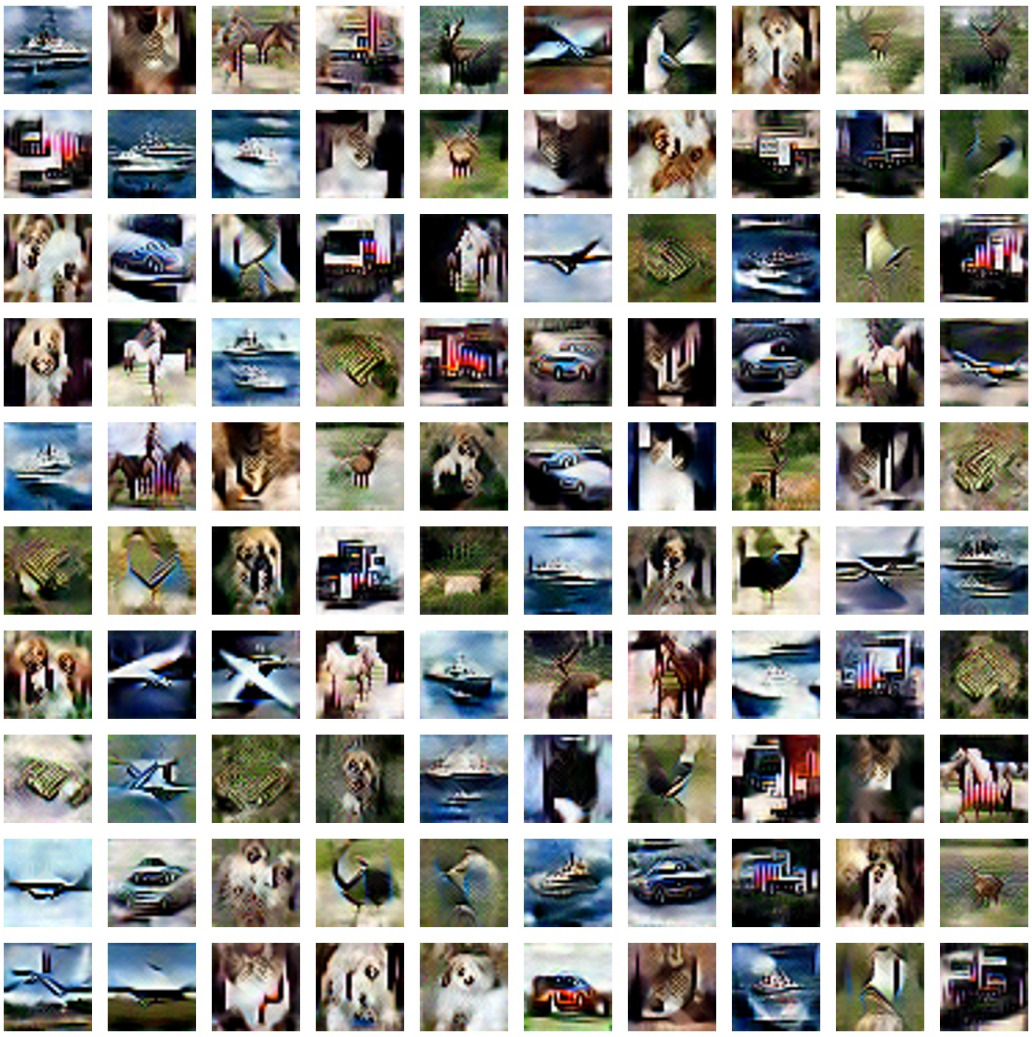

Figure 6: Visualization of synthetic data on CIFAR-10 generated by **PAC**.

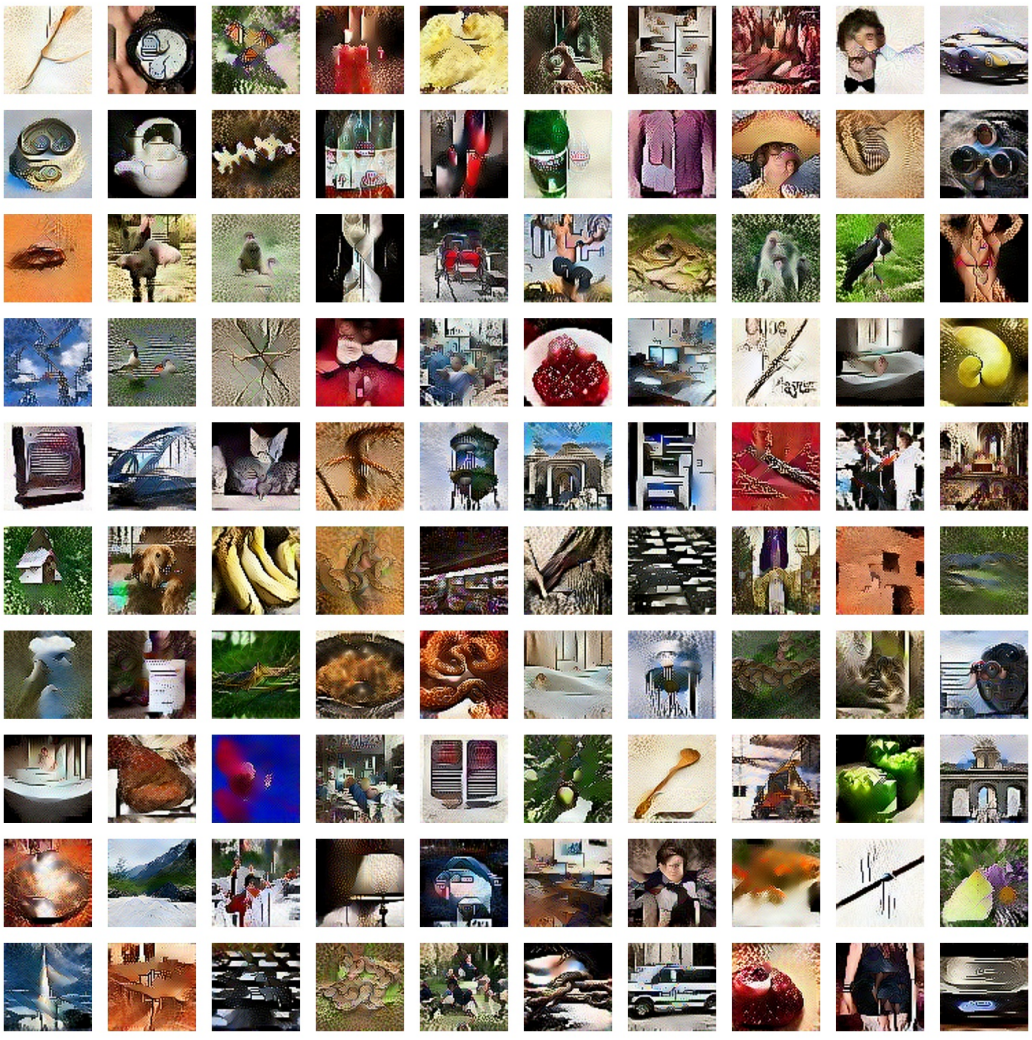

Figure 7: Visualization of synthetic data on Tiny-ImageNet generated by **PAC**.

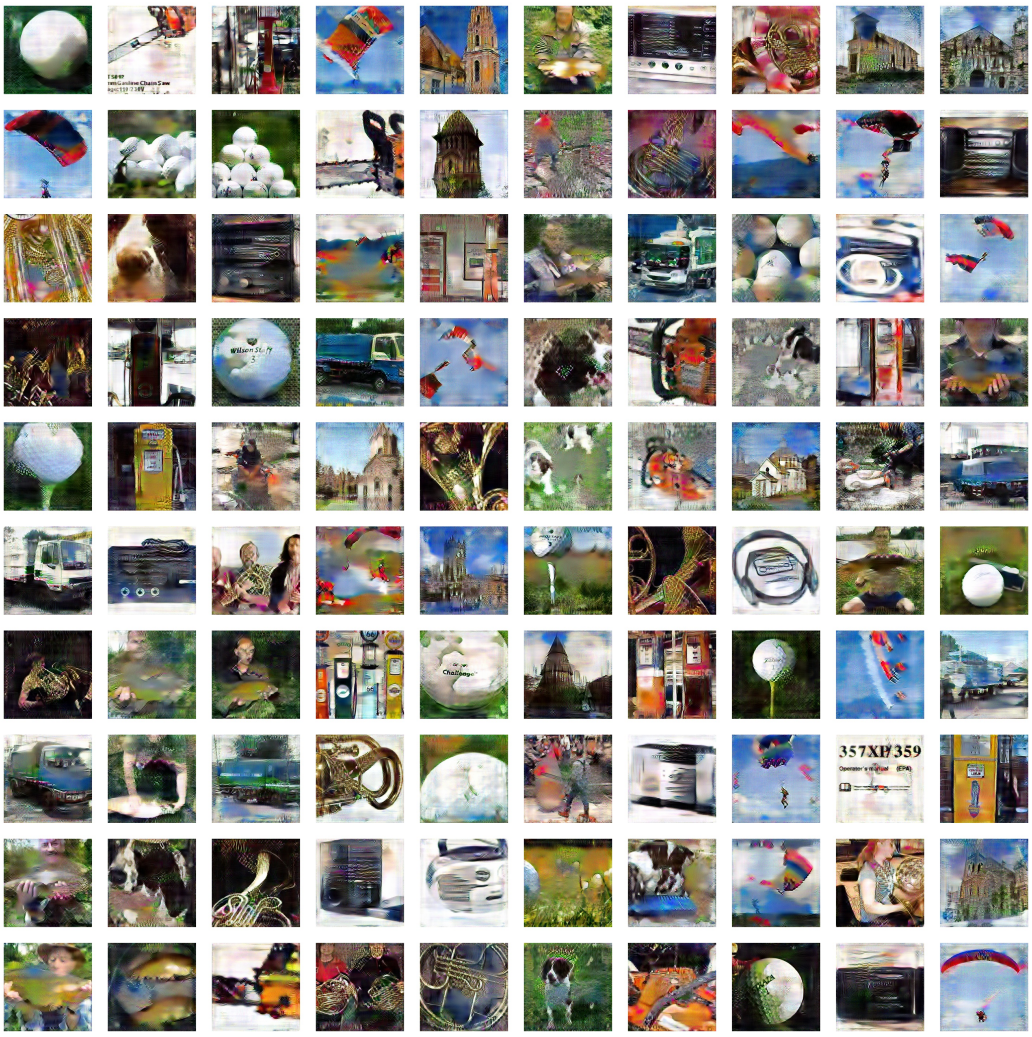

Figure 8: Visualization of synthetic data on ImageNette generated by **PAC**.

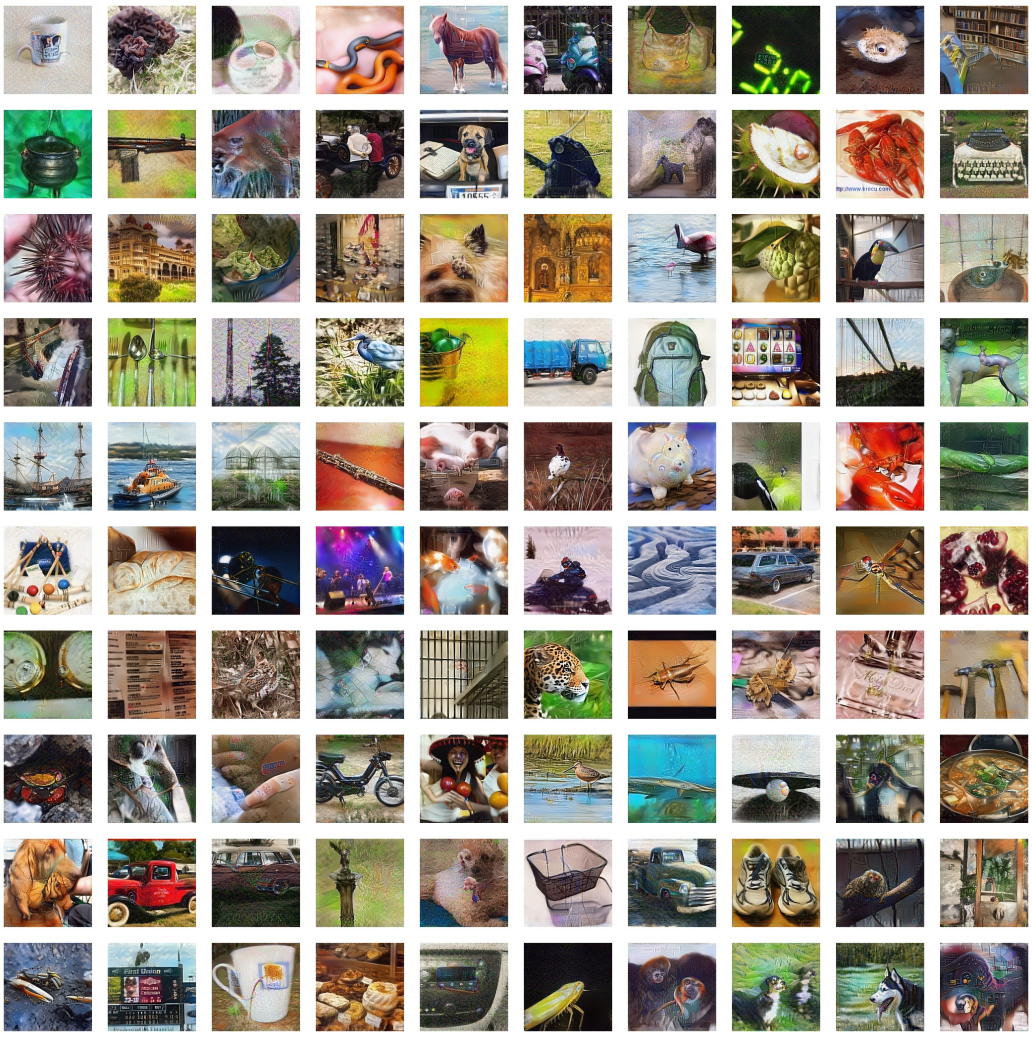

Figure 9: Visualization of synthetic data on ImageNet-1K generated by **PAC**.