Correspondence Coverage Matters for Multi-Modal Dataset Distillation

Zhuohang Dang[1,2], Minnan Luo[1,2*], Chengyou Jia[1,2], Hangwei Qian[3,4*], Xinyu Zhang[1], Xiaojun Chang[5], Ivor Tsang[3,4]

[1]School of Computer Science and Technology, MOEKLINNS Laboratory, Xi'an Jiaotong University
[2]State Key Laboratory of Communication Content Cognition, China
[3]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore
[4]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore
[5]University of Science and Technology of China
{dangzhuohang,cp3jia,zhang1393869716}@stu.xjtu.edu.cn minnluo@xjtu.edu.cn
{qian_hangwei,ivor_tsang}@a-star.edu.sg cxj273@gmail.com

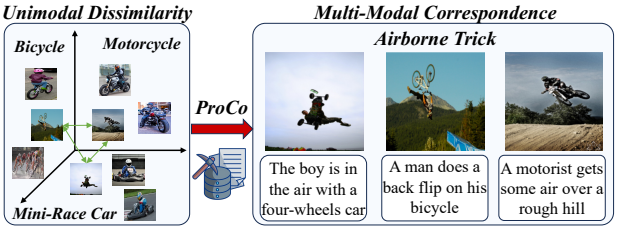
Abstract

Multi-modal dataset distillation (DD) condenses large datasets into compact ones that retain task efficacy by capturing correspondence patterns, *i.e.*, shared semantics between paired modalities. However, such patterns rely on cross-modal similarity and cannot be faithfully captured by intra-modal similarity of current unimodal strategies. As a result, current multi-modal DD methods tend to over-concentrate, redundantly encoding similar correspondence patterns and thus limiting generalizability. To this end, we propose a novel multi-modal DD framework to systematically **Pro**mote **Co**rrespondence coverage, *i.e.*, **ProCo**. Initially, we develop a correspondence consistency metric based on cross-modal retrieval distributions to cluster correspondence patterns. These clusters capture the underlying correspondence distribution, enabling ProCo to initialize distilled data with representative patterns while regularizing optimization to promote correspondence representativeness and diversity. Moreover, we employ conditional neural fields for efficient distilled data parameterization, enhancing fine-grained pattern capture while allowing more distilled data under a fixed budget to boost correspondence coverage. Extensive experiments verify that our ProCo achieves superior and elastic budget-efficacy trade-offs, surpassing prior methods by over 15% with 10× distillation budget reduction, highlighting its real-world practicality.

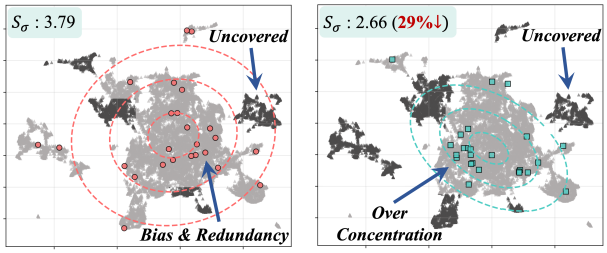
Introduction

Dataset Distillation (DD) improves training efficiency by condensing a large-scale dataset into a compact synthetic ***dataset*** that preserves model training efficacy, in contrast to knowledge distillation that produces a lightweight ***model*** for faster inference. In essence, DD comprehensively encodes informative patterns from original data into distilled data with minimal redundancy, *e.g.*, shape and texture in unimodal image classification tasks (Guo et al. 2024). Recently, DD has been extended to multi-modal scenarios to reduce the substantial computational and data storage costs of large vision-language models (Bai et al. 2025; Wu et al. 2024).

*Corresponding Author.



(a) Illustration of multi-modal correspondence patterns, where dissimilar visual patterns, *e.g.*, motorcycles and mini-race cars, capture the same correspondence pattern, *i.e.*, airborne trick.



(b) UMAP visualization of LoRS's original (gray), randomly initialized (red) and distilled (cyan) data, where uncovered correspondence patterns from original data are emphasized in dark gray. Dashed lines are Gaussian contour ellipses, with S_σ proportional to ellipse's area, indicating data distribution compactness.

Figure 1: Illustration of existing methods' drawbacks.

However, extending DD from unimodal to multi-modal domains poses unique challenges. Different from unimodal patterns based on intra-modal similarity, multi-modal correspondence patterns capture the shared semantics between paired modalities, *e.g.*, an image and its caption. As shown in Figure 1a, even samples with dissimilar intra-modal features may still convey the same correspondence pattern, making unimodal pattern mining and promotion strategies (Liu et al. 2023; Du et al. 2024) unreliable for multi-modal scenarios. This motivates a fundamental problem: ***How can we systematically ensure broad and faithful correspondence coverage in multi-modal dataset distillation?***

Recent works distill multi-modal data with trajectory matching (Wu et al. 2023), enhanced by soft correspondences proposed by LoRS (Xu et al. 2024). However, they suffer from poor correspondence coverage due to over-concentration inherent in DD, failing to ensure correspondence representativeness and diversity during distilled data's initialization and optimization (Liu et al. 2022). Figure 1b shows that random initialization in prior methods leads to severe correspondence redundancy and bias, which is further exacerbated by DD's overemphasis on pattern representativeness while neglecting diversity (Guo et al. 2024). Consequently, similar patterns are redundantly encoded across distilled data, undermining its generalization. Moreover, these methods directly optimize pixel-wise RGB values, suffering from poor scalability with increasing input size and low budget efficiency due to visual redundancy inherent in multi-modal learning (Liang et al. 2023; Yuan et al. 2024).

In light of the above, we propose ProCo, a novel multi-modal dataset distillation framework to systematically promote correspondence coverage. Our key insight is that ***samples sharing similar correspondence patterns exhibit consistent retrieval distributions***, when queried against a shared gallery. Therefore, we propose a novel correspondence consistency metric, based on retrieval distribution discrepancies, to adaptively cluster correspondence patterns. These clusters model the underlying correspondence distribution, from which we sample representative patterns to initialize distilled data to reduce redundancy and bias. We further introduce cluster-level correspondence representativeness and diversity regularization to guide distileld data's optimization, thereby mitigating DD's over-concentration. Moreover, to tackle visual redundancy, we encode each distilled image with a conditional neural field, a continuous coordinate-to-pixel mapping guided by paired captions. Crucially, this facilitates the capture and refinement of correspondence patterns while enabling more distilled samples within the same budget, significantly boosting correspondence coverage.

Extensive experiments on standard multi-modal benchmarks confirm that ProCo's distilled data improves model efficacy by over 15% with only 10% of previous competitors' distillation budget. With increased parameterization capacity, ProCo consistently yields an extra 2% gain by finer-grained correspondence capture, offering an elastic budget-efficacy trade-off. Moreover, we empirically show ProCo's strong correspondence coverage, supporting its remarkable scalability, generalization and real-world applicability.

Related Works

Dataset Distillation (DD). Most methods directly treat data content as learnable parameters and optimize them via meta-learning (Loo et al. 2023) or matching-based (Cui et al. 2023; Wang et al. 2022; Sajedi et al. 2023) strategies. Recent works further incorporate generative priors (Wang et al. 2024) to enhance distillation efficiency and efficacy. However, these methods are restricted to unimodal image domain with simple classification tasks, limiting their applicability to more complex multi-modal scenarios (Zhang et al. 2025).

In multi-modal learning, Wu et al. (2023) extends trajectory matching to all modality encoders, while Xu et al. (2024) uses soft correspondences to enrich distilled data's information density. However, they overlook DD's over-concentration and visual redundancy of multi-modal data, leading to poor correspondence coverage. Conversely, our ProCo explores cluster-level correspondence distribution guidance with efficient parameterization, boosting correspondence coverage with elastic budget-efficacy trade-offs.

Neural Field. Neural field is a continuous function parameterized by a neural network that maps input coordinates to output quantities (Xie et al. 2022). Recently, they have been widely explored in various tasks, *e.g.*, representation learning (Mildenhall et al. 2021) and generative modeling (You et al. 2023). In dataset distillation, DDiF (Shin et al. 2025) adopts neural fields to improve parameterization efficiency but is restricted to unconditioned coordinate-to-pixel mappings. DDiF overlooks the collaboration between parameterization and potential pattern descriptors (*e.g.*, soft labels), thereby limiting distilled data's efficacy. Conversely, beyond mitigating over-concentration with cluster-level correspondence distribution regularization, ProCo enables distillation-parameterization collaboration via caption guidance.

Methodology

Problem Formulation

Given an image-text dataset $\mathcal{D} = (V_i, T_j, y_{ij})_{i,j=1}^{|\mathcal{D}|}$, V_i and T_j are i-th image and j-th text with a matching label $y_{ij} \in \{0, 1\}$. We aim to synthesize a distilled dataset $\mathcal{S} = (\tilde{V}_i, \tilde{T}_j, \tilde{y}_{ij})_{i,j=1}^{|\mathcal{S}|}$ with $|\mathcal{S}| \ll |\mathcal{D}|$ and learnable soft correspondences $\tilde{y}_{ij} \in [0, 1]$, such that $\mathcal{S}$ faithfully captures $\mathcal{D}$'s representative and diverse correspondence patterns. Ideally, the optimal $\mathcal{S}$ allows the model trained on it to achieve consistent performance as the one trained on $\mathcal{D}$.

Model Overview

Figure 2 and Algorithm 1 outline our ProCo framework. Specifically, we adaptively mine correspondence pattern clusters from $\mathcal{D}$ with representative sampling to initialize $\mathcal{S}$, reducing correspondence bias and redundancy. Beyond conventional trajectory matching, we explore cluster-level regularization to promote correspondence representativeness and diversity during $\mathcal{S}$'s optimization. Moreover, we introduce an efficient data parameterization strategy based on conditional neural fields, enabling compact sample representation and improved capture of correspondence patterns. We elaborate on each component in the following sections.

Input Encoding. Given multi-modal input (V_i, T_j), we encode them into a unified feature space with modality-specific encoders f and g, *i.e.*, $F_V^i = f(V_i)$ and $F_T^j = g(T_j)$. We compute the similarity between (V_i, T_j) via cosine similarity as $h(F_V^i, F_T^j)$, denoted as $h_{ij} = h(V_i, T_j)$ for brevity.

Structure-Aware Distillation

We adopt trajectory matching (Cui et al. 2023) for dataset distillation (DD) and complement its over-concentration by exploring correspondence coverage priors from $\mathcal{D}$.

Initialization. We aim to initialize $\mathcal{S}_{init}$ with balanced broad and faithful correspondence coverage to ensure stable

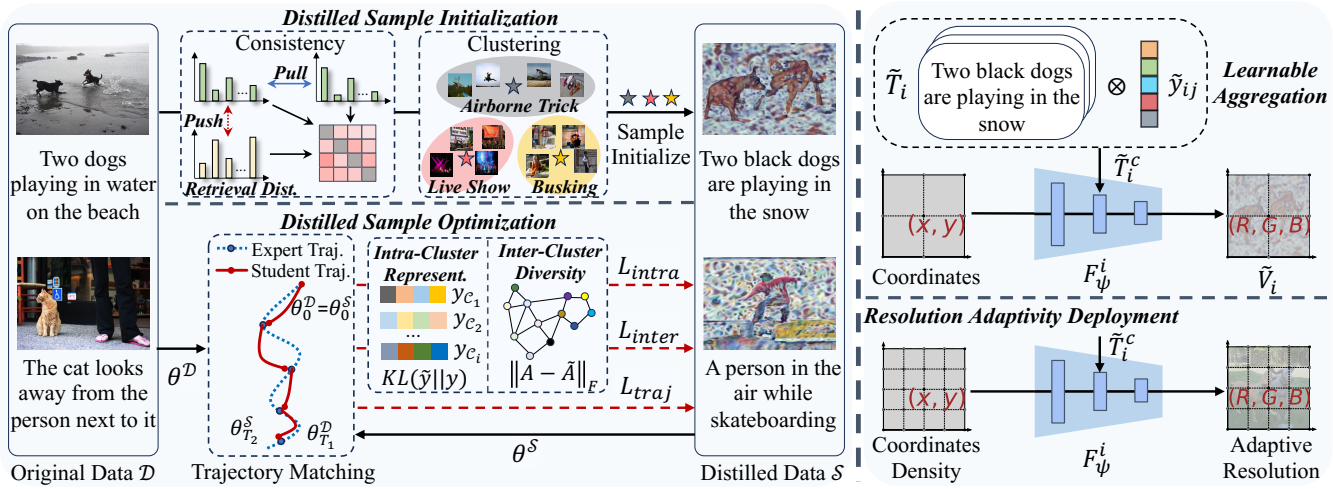


Figure 2: ProCo's overview, illustrating Structure-Aware Distillation (left) and Efficient Parameterization (right). Black/red arrows indicate forward/backward passes, while pentagrams mark samples selected from correspondence pattern clusters.

and effective optimization (Liu et al. 2023), avoiding the redundancy and bias caused by random initialization without global awareness. To this end, we model $\mathcal{D}$'s correspondence pattern distribution and adaptively select representative samples based on the distillation budget for initialization, *e.g.*, as $\mathcal{S}_{init}$'s size increases, its ideal initialization should shift from coarse representatives to fine-grained diverse ones.

To begin with, we propose a novel correspondence consistency metric tailored for multi-modal scenarios, instead of conventional intra-modal similarity. In particular, we adopt an indirect, distributional perspective with the key insight: ***cross-modal retrieval distributions implicitly encode high-level correspondence patterns***. For instance, dissimilar visual patterns (*e.g.*, bike and mini-race car) may retrieve a similar set of captions describing "airborne trick", which serves as a semantic anchor to unify them under the same correspondence pattern. Thus, if two queries yield similar cross-modal retrieval distributions over a shared gallery, they are likely to share the same correspondence pattern. Formally, we compute image-to-text (i2t) and text-to-image (t2i) similarities over the original data $\mathcal{D}$:

$$\hat{H}_{ij}^{i2t} = h(V_i, T_j), \hat{H}_{ij}^{t2i} = h(T_j, V_i) \ \ \forall i, j \in [1, |\mathcal{D}|], \quad (1)$$

where $\hat{H}_i^{i2t}$ are normalized to obtain V_i's retrieval distribution H_i^{i2t} over all captions and vice versa for H_i^{t2i}. Accordingly, we define correspondence consistency between (V_i, T_i) and (V_j, T_j) as the Jensen-Shannon divergence between retrieval distributions across modalities, capturing the bidirectional nature of multi-modal alignment, *i.e.*,

$$C_{ij} = \text{JSD}(H_i^{i2t}, H_j^{i2t}) + \text{JSD}(H_i^{t2i}, H_j^{t2i}), \quad (2)$$

Then, we use Agglomerative Clustering over C to obtain correspondence pattern clusters, with cluster size matching $|\mathcal{S}_{init}|$ for adaptive control over pattern granularity, *i.e.*,

$$\mathcal{C}_1, \cdots, \mathcal{C}_{|\mathcal{S}_{init}|} = \text{AgglomerativeClustering}(C). \quad (3)$$

For each cluster $\mathcal{C}_i$, we select sample $(\tilde{V}_i, \tilde{T}_i)$ that is most consistent with its centroid $(V_{\mathcal{C}_i}, T_{\mathcal{C}_i})$ to initialize $\mathcal{S}_{init}$, ensuring broad correspondence coverage without redundancy.

Optimization. Initially, we refine $\mathcal{S}_{init}$ with trajectory matching to obtain the final distilled data $\mathcal{S}$. We randomly initialize *identical* model parameters for expert $\theta^{\mathcal{D}}$ and student $\theta^{\mathcal{S}}$, *i.e.*, $\theta_0^{\mathcal{D}}$=$\theta_0^{\mathcal{S}}$. The $\theta^{\mathcal{D}}$ is trained on $\mathcal{D}$ with standard InfoNCE loss (Radford et al. 2021), while $\theta^{\mathcal{S}}$ is trained on $\mathcal{S}$ with binary cross-entropy (BCE) loss (Xu et al. 2024), *i.e.*,

$$L_{\mathcal{S}} = \sum_{i,j=1}^{B} w_{ij} \cdot l_{\text{BCE}}(\tilde{y}_{ij}, h(\tilde{V}_i, \tilde{T}_j)/\tau),$$
$$w_{ij} = \frac{\mathbb{I}[\tilde{y}_{ij} > \beta]}{|(i,j) : \tilde{y}_{ij} > \beta|} + \frac{\mathbb{I}[\tilde{y}_{ij} \leq \beta]}{|(i,j) : \tilde{y}_{ij} \leq \beta|}, \quad (4)$$

where τ and β are the temperature and positive-negative threshold. After training $\theta_{\mathcal{D}}$ and $\theta_{\mathcal{S}}$ for T_1 and T_2 steps, we use updated parameters $\theta_{T_1}^{\mathcal{D}}$ and $\theta_{T_2}^{\mathcal{S}}$ to optimize $\mathcal{S}$ by minimizing the accumulated parameter trajectory discrepancy:

$$L_{traj} = \|\theta_{T_2}^{\mathcal{S}} - \theta_{T_1}^{\mathcal{D}}\|^2 / \|\theta_{T_1}^{\mathcal{D}} - \theta_0^{\mathcal{D}}\|^2. \quad (5)$$

However, L_{traj} fails to balance correspondence representativeness and diversity (Du et al. 2024), due to DD's over-concentration. To this end, we exploit intra-cluster structure to maintain local correspondence representativeness over $\mathcal{C}_i$:

$$y_{\mathcal{C}_i}^{i2t} = \frac{exp(h(V_{\mathcal{C}_i}, T_{\mathcal{C}_i})/\tau)}{\sum_{T_j \in \mathcal{C}_i} exp(h(V_{\mathcal{C}_i}, T_j)/\tau)}, \tilde{y}_{\mathcal{C}_i}^{i2t} = \frac{exp(h(\tilde{V}_i, T_{\mathcal{C}_i})/\tau)}{\sum_{T_j \in \mathcal{C}_i} exp(h(\tilde{V}_i, T_j)/\tau)},$$

with $y_{\mathcal{C}_i}^{t2i}$ and $\tilde{y}_{\mathcal{C}_i}^{t2i}$ denfined analogously. The intra-cluster structural guidance is defined with KL divergence, *i.e.*,

$$L_{intra} = \sum_{i=1}^{|\mathcal{S}|} \text{KL}(y_{\mathcal{C}_i}^{i2t}, \tilde{y}_{\mathcal{C}_i}^{i2t}) + \text{KL}(y_{\mathcal{C}_i}^{t2i}, \tilde{y}_{\mathcal{C}_i}^{t2i}). \quad (6)$$

Moreover, we further exploit inter-cluster topological relationships to maintain global correspondence diversity, *i.e.*,

$$A_{ij}^{i2t} = \frac{exp(h(V_{\mathcal{C}_i}, T_{\mathcal{C}_j})/\tau)}{\sum_{k=1}^{|\mathcal{S}|} exp(h(V_{\mathcal{C}_i}, T_{\mathcal{C}_k})/\tau)}, \tilde{A}_{ij}^{i2t} = \frac{exp(h(\tilde{V}_i, T_{\mathcal{C}_j})/\tau)}{\sum_{k=1}^{|\mathcal{S}|} exp(h(\tilde{V}_i, T_{\mathcal{C}_k})/\tau)},$$

with A_{ij}^{t2i} and $\tilde{A}_{ij}^{t2i}$ denfined similarly. The inter-cluster structural guidance is defined with matrix Frobenius norm:

$$L_{inter} = \|A^{i2t} - \tilde{A}^{i2t}\|_F + \|A^{t2i} - \tilde{A}^{t2i}\|_F, \quad (7)$$

Algorithm 1: ProCo's distillation process.

Input: Original data $\mathcal{D}$, interval T_1 and T_2, learning rate α.
Output: Compact distilled data $\mathcal{S}$.

//Correspondence pattern mining via clustering.
Compute i2t and t2i retrieval distribution H (Eq.1)
Compute correspondence consistency C (Eq. 2)
$\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{S}|} = \text{AgglomerativeClustering}(C)$
//Initialize distilled data with efficient parameterization.
for $i = 1 : |\mathcal{S}|$ **do**
  $\tilde{V}_i, \tilde{T}_i \leftarrow \text{Sampling}(\mathcal{C}_i)$, $F^i_\psi \leftarrow \tilde{V}_i$, $\tilde{\mathcal{Y}} \leftarrow \mathbf{I}$
for $j = 1 : num\_steps$ **do**
  //Expert-Student trajectory matching.
  Initialize $\theta^{\mathcal{D}}$ and $\theta^{\mathcal{S}}$ with identical value, $\theta^{\mathcal{D}}_0 = \theta^{\mathcal{S}}_0$.
  Train $\theta^{\mathcal{D}}$ on $\mathcal{D}$ for T_1 steps with InfoNCE loss.
  Train $\theta^{\mathcal{S}}$ on $\mathcal{S}$ for T_2 steps with Eq. 4.
  $L_{traj} = \|\theta^{\mathcal{S}}_{T_2} - \theta^{\mathcal{D}}_{T_1}\|^2 / \|\theta^{\mathcal{D}}_{T_1} - \theta^{\mathcal{D}}_0\|^2$
  //Intra- and Inter-Cluster Regularization.
  Compute L_{intra} and L_{inter} via Eq. 6 and 7.
  $L = L_{traj} + L_{intra} + L_{inter}$
  //Distilled data optimization.
  $\mathcal{S} \leftarrow \mathcal{S} - \alpha \cdot \nabla_{\mathcal{S}} \mathcal{L}$

While these regularizations can be pre-computed for efficiency, they are static and fail to adapt to evolving $\mathcal{S}$ during optimization. Therefore, we introduce a momentum correction to adaptively update such regularizations at step t, *i.e.*,

$$\begin{aligned} y^{i2t}_{\mathcal{C}_i}(t) &= \alpha \cdot y^{i2t}_{\mathcal{C}_i}(t-1) + (1-\alpha) \cdot \tilde{y}^{i2t}_{\mathcal{C}_i}(t), \\ A^{i2t}(t) &= \alpha \cdot A^{i2t}(t-1) + (1-\alpha) \cdot \tilde{A}^{i2t}(t), \end{aligned} \tag{8}$$

where $\alpha \in (0, 1)$ is the momentum coefficient. Overall, the $\mathcal{S}$ is jointly supervised by trajectory matching and our structural guidance, *i.e.*, $L = L_{traj} + L_{intra} + L_{inter}$.

Efficient Distilled Sample Parameterization

Evidently, L directly optimizes distilled image $\tilde{V}_i$'s pixel-wise RGB values, suffering from poor scalability and budget efficiency (Liang et al. 2023). In this sense, we seek to efficiently parameterize $\tilde{V}_i$ to facilitate distillation by alleviating visual redundancy, allowing more samples within the same budget to further boost correspondence coverage.

We encode $\tilde{V}_i$ from raw RGB-space into an efficient conditional neural field F^i_ψ, where captions serve as natural correspondence pattern conditions, enabling more effective parameterization than vanilla neural field (Shin et al. 2025). Given $\tilde{V}_i \in \mathbb{R}^{H \times W \times C}$ with resolution (H, W) and $C = 3$ for RGB, F^i_ψ is a continuous function that maps each pixel coordinate (x, y) to its RGB value (R_{xy}, G_{xy}, B_{xy}):

$$(R_{xy}, G_{xy}, B_{xy}) = F^i_\psi(x, y, \tilde{T}^c_i), \tag{9}$$

where $\tilde{T}^c_i$ is a text-based conditional signal. Notably, we use learnable soft correspondence $\tilde{y}_{ij}$ to greatly enrich $\tilde{T}^c_i$ by aggregating the relevant distilled captions, *i.e.*,

$$\tilde{T}^c_i = \text{Norm}(\sum\nolimits_{j=1}^{|\mathcal{S}|} \tilde{y}_{ij} \cdot \tilde{T}_j), \tag{10}$$

which effectively facilitate the capture and refinement of fine-grained correspondence patterns in $\tilde{V}_i$, enriching distilled data's information density. Moreover, benefiting from neural fields's continuous nature, Figure 2 shows that our ProCo supports flexible resolution adaptation during deployment by simply adjusting the coordinate sampling density, effectively alleviating previous methods' resolution rigidity.

Efficiency Discussion

We highlight ProCo's efficiency and scalability in promoting correspondence coverage, pivotal for large-scale real-world applications. The analyses are focused on ProCo's core components, with empirical validations in ablation studies.

Initialization. Previous methods initialize samples by difficulty estimation (Lee and Chung 2024; Chen et al. 2025), gradually including harder samples as the budget increases. However, this requires training multiple models on data subsets (Jiang et al. 2021), which is impractical at scale. Moreover, difficulty scores often correlate with correspondence patterns, leading to severe redundancy. Conversely, ProCo initializes samples with comprehensive yet compact coverage by clustering correspondence patterns based on retrieval consistency, which can be efficiently pre-computed via $\theta_{\mathcal{D}}$.

Optimization. Beyond trajectory matching, ProCo proposes lightweight regularizers: L_{inter} models inter-cluster topology to maintain global correspondence diversity and L_{intra} explores intra-cluster consistency for correspondence representativeness, accelerated by sampling k representative samples per cluster. Overall, ProCo incurs a cost of $\mathcal{O}(|\mathcal{S}|^2 + k|\mathcal{S}|)$, less than 4% of the total distillation overhead.

Parameterization. Previous methods explore latent codes with pretrained generators to synthesize distilled data (Chan-Santiago et al. 2025; Wang et al. 2024), struggling with considerable budget and computational overhead, *e.g.*, a single Flux (Batifol et al. 2024) model consumes over 23GB budget (1000× than ours) with high image synthesis costs. Conversely, our ProCo employs lightweight MLP-based conditional neural fields with negligible computational overhead, offering a more favorable budget-efficacy trade-off.

Experiments

Experimental Setup

Datasets. We adopt two multi-modal datasets: Flickr30K (Young et al. 2014) and COCO (Lin et al. 2014) and data splits are the same as Xu et al. (2024); Dang et al. (2025).

Evaluation Metrics. We train multiple randomly initialized models using the distilled data and evaluate them on the standard test set. Performance is reported as the mean and standard deviation of Recall@K (R@K, K = 1/5/10).

Comparison with State-of-the-Art (SOTA)

To verify our ProCo's efficacy, we compare it against SOTA multi-modal dataset distillation methods, including two categories: (1) Coreset selection methods, which directly select representative subsets from the original data for training, such as K-Center (Sener and Savarese 2018) and Forgetting (Toneva et al. 2018); (2) Dataset distillation (DD) methods, including LoRS (Xu et al. 2024), TESLA (Cui et al. 2023)

Pairs	Methods	Flickr30K						MS-COCO 1K					
		Image⟶Text			Text⟶Image			Image⟶Text			Text⟶Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
100	K-center	0.6	5.0	7.6	0.7	3.1	6.1	1.4	3.7	5.5	0.4	1.4	2.5
	Forget	1.2	4.2	9.7	0.7	2.4	5.6	0.7	2.6	4.8	0.3	1.5	2.5
	vl-distill	$9.9_{\pm 0.3}$	$28.3_{\pm 0.5}$	$39.1_{\pm 0.7}$	$4.7_{\pm 0.2}$	$15.7_{\pm 0.5}$	$24.6_{\pm 1.0}$	$2.5_{\pm 0.3}$	$10.0_{\pm 0.5}$	$15.7_{\pm 0.4}$	$1.3_{\pm 0.1}$	$5.4_{\pm 0.3}$	$9.5_{\pm 0.5}$
	LoRS	$11.8_{\pm 0.2}$	$35.8_{\pm 0.6}$	$49.2_{\pm 0.5}$	$8.3_{\pm 0.2}$	$24.1_{\pm 0.2}$	$35.1_{\pm 0.3}$	$3.3_{\pm 0.2}$	$12.2_{\pm 0.3}$	$19.6_{\pm 0.3}$	$1.8_{\pm 0.1}$	$7.1_{\pm 0.2}$	$12.2_{\pm 0.2}$
	Ours-S	$14.9_{\pm 0.4}$	$41.2_{\pm 0.4}$	$55.7_{\pm 0.6}$	$9.7_{\pm 0.3}$	$28.9_{\pm 0.1}$	$41.0_{\pm 0.4}$	$6.0_{\pm 0.3}$	$18.9_{\pm 0.1}$	$28.6_{\pm 0.3}$	$3.1_{\pm 0.1}$	$11.3_{\pm 0.1}$	$18.9_{\pm 0.2}$
	Ours-L	$\mathbf{15.5}_{\pm \mathbf{1.1}}$	$\mathbf{42.0}_{\pm \mathbf{0.6}}$	$\mathbf{56.9}_{\pm \mathbf{0.5}}$	$\mathbf{10.2}_{\pm \mathbf{0.1}}$	$\mathbf{29.2}_{\pm \mathbf{0.2}}$	$\mathbf{41.6}_{\pm \mathbf{0.3}}$	$\mathbf{6.6}_{\pm \mathbf{0.2}}$	$\mathbf{19.9}_{\pm \mathbf{0.4}}$	$\mathbf{30.1}_{\pm \mathbf{0.2}}$	$\mathbf{3.3}_{\pm \mathbf{0.1}}$	$\mathbf{12.0}_{\pm \mathbf{0.1}}$	$\mathbf{19.6}_{\pm \mathbf{0.2}}$
200	K-center	2.2	8.2	13.5	1.5	5.4	9.9	1.2	3.8	7.5	0.7	2.1	5.8
	Forget	1.5	8.4	10.2	1.2	3.1	8.4	1.1	3.5	7.0	0.6	2.8	4.9
	vl-distill	$10.2_{\pm 0.8}$	$28.7_{\pm 1.0}$	$41.9_{\pm 1.9}$	$4.6_{\pm 0.9}$	$16.0_{\pm 1.6}$	$25.5_{\pm 2.6}$	$3.3_{\pm 0.2}$	$11.9_{\pm 0.6}$	$19.4_{\pm 1.2}$	$1.7_{\pm 0.1}$	$6.5_{\pm 0.4}$	$12.3_{\pm 0.8}$
	LoRS	$14.5_{\pm 0.5}$	$38.7_{\pm 0.5}$	$53.4_{\pm 0.5}$	$8.6_{\pm 0.3}$	$25.3_{\pm 0.2}$	$36.6_{\pm 0.3}$	$4.3_{\pm 0.1}$	$14.2_{\pm 0.3}$	$22.6_{\pm 0.2}$	$2.4_{\pm 0.1}$	$9.3_{\pm 0.2}$	$15.5_{\pm 0.2}$
	Ours-S	$17.4_{\pm 0.5}$	$44.5_{\pm 0.4}$	$59.3_{\pm 0.5}$	$10.7_{\pm 0.2}$	$30.8_{\pm 0.4}$	$42.9_{\pm 0.4}$	$7.4_{\pm 0.2}$	$22.8_{\pm 0.1}$	$33.5_{\pm 0.2}$	$4.2_{\pm 0.1}$	$14.1_{\pm 0.2}$	$22.5_{\pm 0.3}$
	Ours-L	$\mathbf{18.1}_{\pm \mathbf{0.3}}$	$\mathbf{45.1}_{\pm \mathbf{0.9}}$	$\mathbf{60.7}_{\pm \mathbf{1.0}}$	$\mathbf{11.4}_{\pm \mathbf{0.2}}$	$\mathbf{31.8}_{\pm \mathbf{0.6}}$	$\mathbf{44.5}_{\pm \mathbf{0.7}}$	$\mathbf{7.9}_{\pm \mathbf{0.1}}$	$\mathbf{24.4}_{\pm \mathbf{0.2}}$	$\mathbf{35.5}_{\pm \mathbf{0.3}}$	$\mathbf{4.5}_{\pm 0.1}$	$\mathbf{14.9}_{\pm \mathbf{0.2}}$	$\mathbf{23.5}_{\pm \mathbf{0.3}}$
500	K-center	4.9	16.4	23.3	3.5	10.4	17.3	2.5	8.7	14.3	1.1	6.3	10.5
	Forget	3.6	12.3	19.3	1.8	9.0	15.9	2.1	8.2	13.0	0.8	5.8	8.2
	vl-distill	$13.3_{\pm 0.6}$	$32.8_{\pm 1.8}$	$46.8_{\pm 0.8}$	$6.6_{\pm 0.3}$	$20.2_{\pm 1.2}$	$30.0_{\pm 2.1}$	$5.0_{\pm 0.4}$	$17.2_{\pm 1.3}$	$26.0_{\pm 1.9}$	$2.5_{\pm 0.5}$	$8.9_{\pm 0.7}$	$15.8_{\pm 1.5}$
	LoRS	$15.5_{\pm 0.7}$	$39.8_{\pm 0.4}$	$53.7_{\pm 0.3}$	$10.0_{\pm 0.2}$	$28.9_{\pm 0.7}$	$41.6_{\pm 0.6}$	$5.3_{\pm 0.5}$	$18.3_{\pm 1.5}$	$27.9_{\pm 1.4}$	$2.8_{\pm 0.2}$	$9.9_{\pm 0.5}$	$16.5_{\pm 0.7}$
	Ours-S	$19.6_{\pm 0.5}$	$48.4_{\pm 0.4}$	$62.8_{\pm 0.6}$	$13.5_{\pm 0.2}$	$36.0_{\pm 0.5}$	$48.6_{\pm 0.5}$	$7.3_{\pm 0.2}$	$21.7_{\pm 0.4}$	$32.5_{\pm 0.3}$	$6.2_{\pm 0.1}$	$19.4_{\pm 0.5}$	$29.8_{\pm 0.6}$
	Ours-L	$\mathbf{21.5}_{\pm \mathbf{0.7}}$	$\mathbf{51.1}_{\pm \mathbf{0.7}}$	$\mathbf{65.3}_{\pm \mathbf{0.9}}$	$\mathbf{13.9}_{\pm \mathbf{0.2}}$	$\mathbf{36.8}_{\pm \mathbf{0.6}}$	$\mathbf{49.5}_{\pm \mathbf{0.5}}$	$\mathbf{7.6}_{\pm \mathbf{0.2}}$	$\mathbf{22.9}_{\pm \mathbf{0.4}}$	$\mathbf{34.3}_{\pm \mathbf{0.5}}$	$\mathbf{6.9}_{\pm \mathbf{0.2}}$	$\mathbf{20.9}_{\pm \mathbf{0.6}}$	$\mathbf{31.6}_{\pm \mathbf{0.7}}$

Table 1: Distillation on Flickr30K and COCO. Model on original dataset training achieves R@1/5/10 of 33.9/65.1/75.2 (Image) and 27.3/57.1/69.7 (Text) on Flickr30K; while 19.6/45.6/59.5 (Image) and 16.9/41.9/55.9 (Text) on COCO.

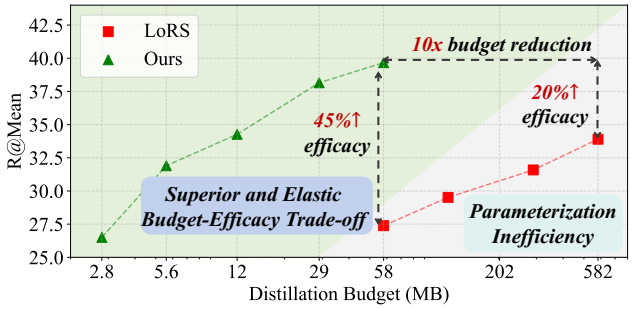


Figure 3: Budget-efficacy trade-off on Flickr30K.

and VL-Distill (Wu et al. 2023). Moreover, we implement two variants of ProCo, denoted as Ours-S and Ours-L, based on conditional neural field's parameter size. Notably, their distillation budgets are only 10% and 30% of prior SOTA methods, given the same distilled data size.

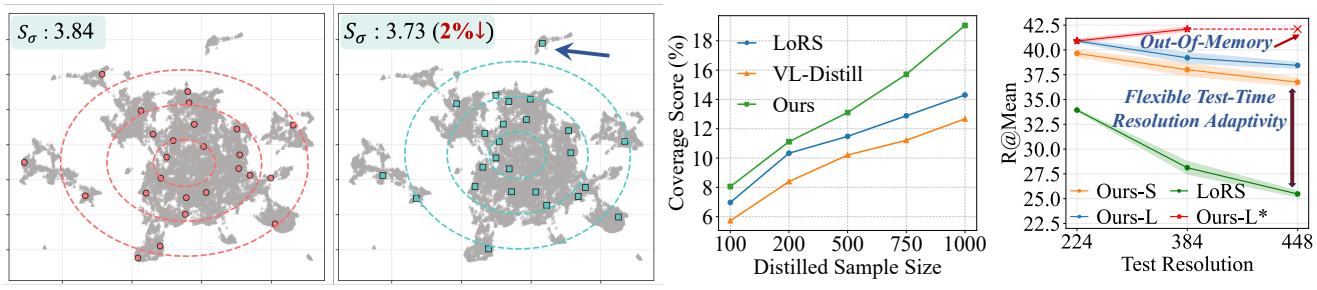
Table 1 compares distillation efficacy on Flickr30K and COCO datasets. Specifically, coreset selection methods suffer from correspondence patterns loss, as they merely select data subsets. Although DD methods attempt to enhance distilled data, their over-concentration limits performance gains when budget scales up, as redundant encoding of similar patterns severely reduces distilled data's generalizability. Moreover, their excessive input redundancy misleads optimization to instability, further hindering distillation efficacy. Conversely, our ProCo explores cluster-level correspondence distribution priors with efficient parameterization for representative and diverse correspondence coverage, surpassing prior SOTA by over 15% with 10× distillation budget reduction. Moreover, with larger parameterization capacity, ProCo yields an extra 2% gain by finer correspondence pattern capture, offering an elastic budget-efficacy trade-off.

Practicality and Scalability. Figure 3 evaluates ProCo's scalability by increasing distillation budget. Notably, LoRS incurs a high distillation budget with only marginal gains, due to persistent over-concentration. In contrast, ProCo exhibits consistent scalability, ***achieving 75% of the original data performance with only 1K distilled samples (500× storage reduction than original data)***, highlighting its practicality in budget-sensitive scenarios, *e.g.*, edge deployment.

Correspondence Coverage Analyses

We conduct qualitative and quantitative experiments to comprehensively evaluate ProCo's correspondence coverage efficacy. Compared to Figure 1b, Figure 4a shows that ProCo explores correspondence pattern cluster priors to alleviate DD's over-concentration, validated by stable S_σ. Furthermore, the blue arrow highlights that ProCo effectively adapts distilled samples during optimization, thereby promoting representative and diverse correspondence coverage.

To quantitatively assess ProCo's correspondence coverage, we extend unimodal coverage metric (Lee and Chung 2024) to multi-modal scenario. For fairness, all methods are evaluated under the same distilled data size, while ProCo uses only 10% of distillation budget. Figure 4b shows that ProCo consistently achieves superior correspondence coverage across various distilled data sizes, supporting its performance and scalability highlighted in Table 1. Moreover, as coverage score increases with distilled data size, ProCo's efficient parameterization can enable more samples within the same budget to further boost correspondence coverage.



(a) (b) (c)

Figure 4: (a) illustrates our ProCo's UMAP visualization, where distilled data exhibits representative and diverse correspondence coverage without over-concentration. (b) quantitatively analyzes correspondence coverage. (c) illustrates cross-resolution generalization efficacy. Ours-L* denotes results distilled at the corresponding evaluation resolution, whereas others are distilled at the standard CLIP pre-training resolution (224×224) for higher-resolution evaluation.

Methods				Image⟶Text	Text⟶Image
Param	Init	L_{intra}	L_{inter}	R@1/5/10	R@1/5/10
				11.8/35.8/49.2	8.3/24.1/35.1
✓				12.3/36.9/50.7	8.4/24.6/36.0
✓	✓			12.8/38.2/51.9	8.7/26.6/38.9
✓	✓	✓		14.1/40.2/54.4	9.4/28.0/40.5
✓	✓		✓	13.7/39.5/53.8	9.2/27.4/40.3
✓	✓	✓	✓	**14.9/41.2/55.7**	**9.7/28.9/41.0**

Table 2: Component analyses on Flickr30K (100 samples).

Methods	Model	Image⟶Text R@1/5/10	Text⟶Image R@1/5/10
LoRS	NFNet+BERT	15.0/40.5/53.6	10.4/29.5/42.2
	ResNet+BERT	5.9/15.3/23.6	3.6/12.3/18.9
	ViT+BERT	7.3/19.5/28.5	4.7/14.2/21.1
Ours	NFNet+BERT	19.3/49.2/63.6	14.0/35.5/47.3
	ResNet+BERT	10.8/28.2/39.2	6.7/19.2/28.7
	ViT+BERT	11.8/30.2/41.9	6.4/19.4/29.2

Table 3: Cross-architecture generalization on Flickr30K. 1K samples are distilled with NFNet+BERT for evaluation.

Methods	Para	Init	L_{intra}+L_{inter}	L_{traj}
LoRS	0s/582 MB	-	-	∼25s/iter
Ours	5s/58 MB	67s	∼1s/iter	∼25s/iter

Table 4: Efficiency analyses on Flickr30K with 1K distilled data. Para and Init are parameterization and initialization.

Ablation Studies

Component Analyses. Table 2 ablates ProCo's core components: efficient parameterization (Param) and structure-aware distillation, including initialization (Init) and optimization (L_{intra}, L_{inter}). Without all components, ProCo degrades to vanilla distillation, suffering from poor correspondence coverage due to over-concentration and visual redundancy. The Param improves $10\times$ budget efficiency and 3% performance, highlighting the importance of reducing redundancy during distillation. Although promoting correspondence coverage during initialization improves performance by 3%, Init fails to fully resolve over-concentration as DD overemphasizes pattern representativeness during optimization. Moreover, L_{intra} and L_{inter} offer limited gains in isolation, as neither adequately balances correspondence representativeness and diversity. The full ProCo yields the best results, confirming each component's essential role.

Cross-Resolution Generalization. Resolution mismatches between distillation and deployment require costly full re-distillation, or even infeasible due to the quadratic growth of computation and memory with image resolution. We show that ProCo achieves strong cross-resolution generalization, allowing distilled data synthesized at low resolution to retain high efficacy when evaluated at higher resolutions. Specifically, we distill at the standard CLIP pre-training resolution (224^2) and evaluate at resolutions used in CLIP fine-tuning (384^2) and large backbone (448^2). Figure 4c shows that prior methods struggle due to their fixed input-sized parameterization, requiring naive interpolation for resizing and thus suffering from information distortion. In contrast, our ProCo enables resolution-adaptive generation via neural fields' continuous nature, highlighting its scalability and practicality by maximal task efficacy at higher resolutions.

Cross-Architecture Generalization. We further assess the generalization of our ProCo's distilled data across model architectures. We focus on the visual encoder following (Xu et al. 2024), as the text encoder remains frozen during training and distillation. Table 3 shows that LoRS's performance drops severely when applied to different model architectures for training, due to poor correspondence coverage caused by over-concentration and visual redundancy. In contrast, ProCo explores cluster-level correspondence distribution with efficient parameterization to boost correspondence coverage, thereby enabling stronger cross-architecture generalization, *e.g.*, it generalizes well across architectures with entirely different inductive biases such as ResNet and ViT.

Efficiency Analyses. Efficiency is critical for large-scale dataset distillation. Table 4 reports the training overhead using an NVIDIA Tesla L40, where ProCo's cost mainly stems from: (1) ***Parameterization***: ProCo pre-computes a

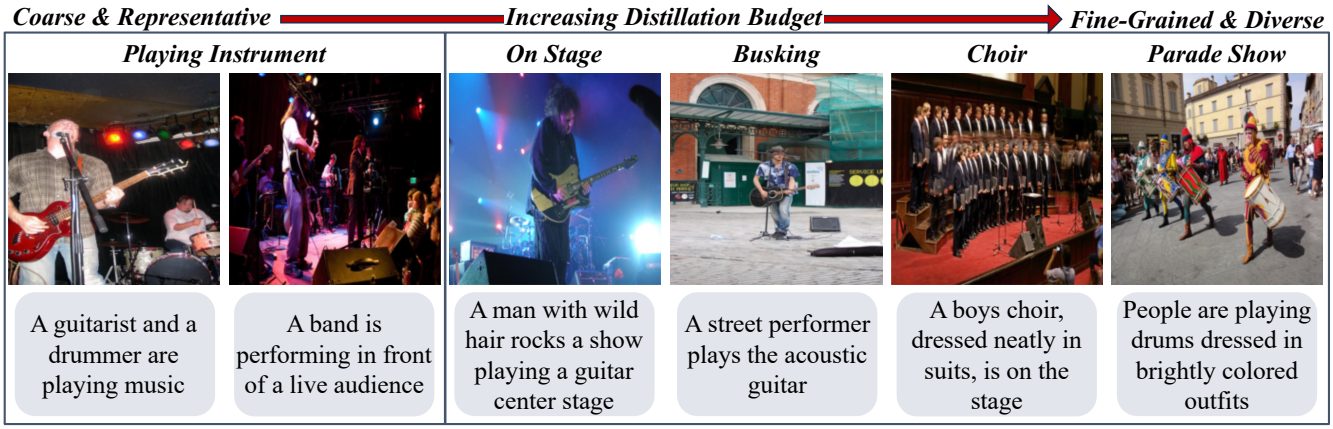


Figure 5: Illustration of evolving correspondence patterns mined by our ProCo from 100 (left) to 1000 (right) distilled data.

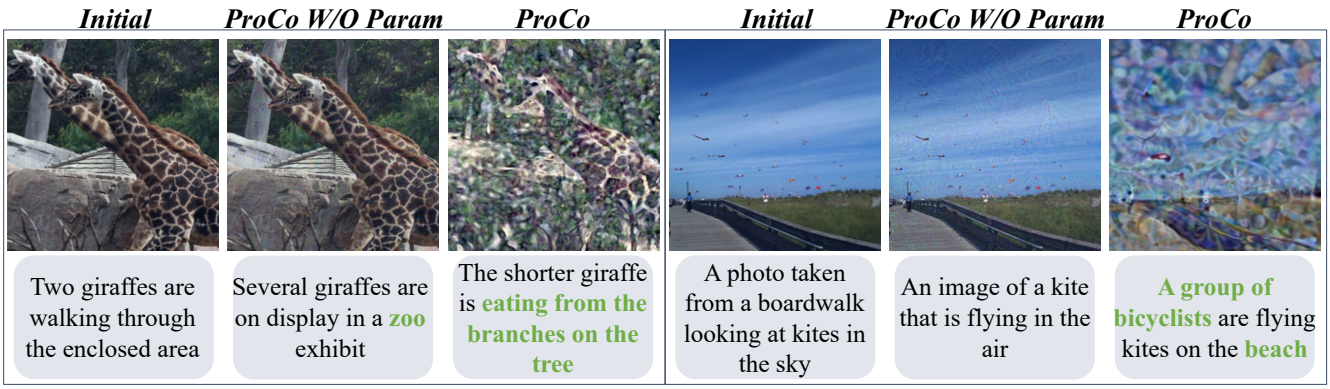


Figure 6: Visualization of synthetic distilled pairs on COCO. "ProCo w/o param" shows distillation without efficient parameterization, *i.e.*, directly optimizing pixel-wise RGB values. The newly captured correspondence patterns are emphasized in green.

conditional neural field for each sample, incurring 5s overhead each. This enables a 10× parameter size reduction and greatly improves optimization stability. (2) ***Optimization***: ProCo pre-computes retrieval distributions and performs efficient clustering with *sklearn* during initialization, requiring ∼70s. During optimization, L_{intra} and L_{inter} incur less than 4% of the conventional trajectory matching cost. Overall, our ProCo achieves substantial performance gains with minimal computation overhead, validating its practicality for scalable multi-modal dataset distillation.

Correspondence Pattern Visualization. Figure 5 shows the correspondence patterns mined by ProCo, where we visualize representative samples from each cluster. As the distillation budget increases, the mined patterns evolve from coarse-grained representatives (*i.e.*, playing instruments) to more fine-grained diverse ones (*i.e.*, busking and choir). Notably, ProCo discriminates correspondence patterns despite high intra-modal similarity, *e.g.*, guitar of on stage and busking patterns, thereby effectively modeling correspondence distribution to alleviate over-concentration.

Distilled Data Visualization. Figure 6 shows that distilled data are substantially refined to enrich correspondence patterns, *e.g.*, distilled images incorporate high-frequency details to boost generalization (Liang et al. 2023), exhibiting a DeepDream-Style (Zeiler 2014) typical in dataset distillation. This further verifies our efficient parameterization's efficacy, which reduces visual redundancy while facilitating correspondence pattern's capture and refinement. Notably, compared to input-sized pixel-wise optimization, ***ProCo better preserves critical semantics (e.g., giraffes and kites), while transforming redundant information to informative patterns (e.g., tree branches and beach) guided by learnable aggregation of relevant distilled captions***.

Conclusion

We propose ProCo, a novel multi-modal dataset distillation framework to enhance correspondence coverage. Specifically, ProCo clusters correspondence patterns with retrieval distributions, then explores cluster-level correspondence distribution to guide distilled data's initialization and optimization, promoting representative and diverse correspondence coverage. Moreover, ProCo reduces visual redundancy with conditional neural fields, boosting correspondence coverage by fine-grained pattern capture and enhanced budget efficiency. Extensive experiments validate ProCo's efficacy, scalability and generalization. We envision ProCo as a step toward advancing green, secure and responsible AI.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (No. 2024JC-JCQN-62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024), the National Natural Science Foundation of China (No. U25A20530), the Project of China Knowledge Center for Engineering Science and Technology, the K. C. Wong Education Foundation and the Program of China Scholarship Council. This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is supported by A*STAR Career Development Fund ⟨Project No. C243512010⟩.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2024. FLUX. https://github.com/black-forest-labs/flux.

Chan-Santiago, J. A.; Tirupattur, P.; Nayak, G. K.; Liu, G.; and Shah, M. 2025. MGD^3: Mode-Guided Dataset Distillation using Diffusion Models. *arXiv preprint arXiv:2505.18963*.

Chen, Y.; Chen, G.; Zhang, M.; Guan, W.; and Nie, L. 2025. Curriculum Coarse-to-Fine Selection for High-IPC Dataset Distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20437–20446.

Cui, J.; Wang, R.; Si, S.; and Hsieh, C.-J. 2023. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, 6565–6590. PMLR.

Dang, Z.; Luo, M.; Jia, C.; Qian, H.; Chang, X.; and Tsang, I. W. 2025. Multi-Modal Dataset Distillation in the Wild. *arXiv preprint arXiv:2506.01586*.

Du, J.; Hu, J.; Huang, W.; Zhou, J. T.; et al. 2024. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. *Advances in neural information processing systems*, 37: 119443–119465.

Guo, Z.; Wang, K.; Cazenavette, G.; LI, H.; Zhang, K.; and You, Y. 2024. Towards Lossless Dataset Distillation via Difficulty-Aligned Trajectory Matching. In *The Twelfth International Conference on Learning Representations*.

Jiang, Z.; Zhang, C.; Talwar, K.; and Mozer, M. C. 2021. Characterizing Structural Regularities of Labeled Data in Overparameterized Models. In *International Conference on Machine Learning*, 5034–5044. PMLR.

Lee, Y.; and Chung, H. W. 2024. SelMatch: Effectively Scaling Up Dataset Distillation via Selection-Based Initialization and Partial Updates by Trajectory Matching. In *International Conference on Machine Learning*, 26546–26567. PMLR.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.-P.; and Salakhutdinov, R. 2023. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36: 32971–32998.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Wang, K.; Yang, X.; Ye, J.; and Wang, X. 2022. Dataset distillation via factorization. *Advances in neural information processing systems*, 35: 1100–1113.

Liu, Y.; Gu, J.; Wang, K.; Zhu, Z.; Jiang, W.; and You, Y. 2023. Dream: Efficient dataset distillation by representative matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17314–17324.

Loo, N.; Hasani, R.; Lechner, M.; and Rus, D. 2023. Dataset distillation with convexified implicit gradients. In *International Conference on Machine Learning*, 22649–22674. PMLR.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sajedi, A.; Khaki, S.; Amjadian, E.; Liu, L. Z.; Lawryshyn, Y. A.; and Plataniotis, K. N. 2023. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17097–17107.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Shin, D.; Bae, H.; Sim, G.; Kang, W.; and Moon, I.-c. 2025. Distilling Dataset into Neural Field. In *The Thirteenth International Conference on Learning Representations*.

Toneva, M.; Sordoni, A.; Combes, R. T. d.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Wang, K.; Gu, J.; Gu, J.; Zhang, H.; Zhou, D.; Zhu, Z.; Jiang, W.; and You, Y. 2024. Dim: Distilling dataset into generative model. In *European Conference on Computer Vision*, 42–59. Springer.

Wang, K.; Zhao, B.; Peng, X.; Zhu, Z.; Yang, S.; Wang, S.; Huang, G.; Bilen, H.; Wang, X.; and You, Y. 2022. Cafe:

Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12196–12205.

Wu, X.; Deng, Z.; Russakovsky, O.; and Russakovsky, O. 2023. Multimodal dataset distillation for image-text retrieval. *arXiv preprint arXiv:2308.07545*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xie, Y.; Takikawa, T.; Saito, S.; Litany, O.; Yan, S.; Khan, N.; Tombari, F.; Tompkin, J.; Sitzmann, V.; and Sridhar, S. 2022. Neural fields in visual computing and beyond. In *Computer graphics forum*, volume 41, 641–676. Wiley Online Library.

Xu, Y.; Lin, Z.; Qiu, Y.; Lu, C.; and Li, Y.-L. 2024. Low-Rank Similarity Mining for Multimodal Dataset Distillation. In *Forty-first International Conference on Machine Learning*.

You, T.; Kim, M.; Kim, J.; and Han, B. 2023. Generative neural fields by mixtures of neural implicit functions. *Advances in Neural Information Processing Systems*, 36: 20352–20370.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yuan, B.; Wang, Z.; Baktashmotlagh, M.; Luo, Y.; and Huang, Z. 2024. Color-oriented redundancy reduction in dataset distillation. *Advances in Neural Information Processing Systems*, 37: 53237–53260.

Zeiler, M. 2014. Visualizing and Understanding Convolutional Networks. In *European conference on computer vision/arXiv*, volume 1311.

Zhang, J.; Wang, Z.; Zhu, H.; Liu, J.; Lin, Q.; and Cambria, E. 2025. MARS: A Multi-Agent Framework Incorporating Socratic Guidance for Automated Prompt Optimization. *arXiv preprint arXiv:2503.16874*.