MULTIMODAL DATASET DISTILLATION MADE SIMPLE BY PROTOTYPE-GUIDED DATA SYNTHESIS

Junhyeok Choi, Sangwoo Mo & Minwoo Chae[*]
Department of Industrial and Management Engineering
Pohang University of Science and Technology
{cjunh4810, sangwoo.mo, mchae}@postech.ac.kr

ABSTRACT

Recent advances in multimodal learning have achieved remarkable success across diverse vision–language tasks. However, such progress heavily relies on large-scale image–text datasets, making training costly and inefficient. Prior efforts in dataset filtering and pruning attempt to mitigate this issue, but still require relatively large subsets to maintain performance and fail under very small subsets. Dataset distillation offers a promising alternative, yet existing multimodal dataset distillation methods require full-dataset training and joint optimization of image pixels and text features, making them architecture-dependent and limiting cross-architecture generalization. To overcome this, we propose a learning-free dataset distillation framework that eliminates the need for large-scale training and optimization while enhancing generalization across architectures. Our method uses CLIP to extract aligned image–text embeddings, obtains prototypes, and employs an unCLIP decoder to synthesize images, enabling efficient and scalable multimodal dataset distillation. Extensive experiments demonstrate that our approach consistently outperforms optimization-based dataset distillation and subset selection methods, achieving state-of-the-art cross-architecture generalization. Our code is available at https://github.com/junhyeok9712/PDS.

1 INTRODUCTION

Multimodal models, such as CLIP (Radford et al., 2021), have emerged as a central paradigm in machine learning, demonstrating remarkable generalization across diverse tasks, including zero-shot classification and image–text retrieval. At the core of these successes are large-scale image–text datasets, such as LAION-5B (Schuhmann et al., 2022), whose use entails substantial computational and memory costs during training. This challenge naturally raises a critical question: *Can we substantially reduce the number of training samples without severely compromising performance?* Addressing this question is not only crucial for resource efficiency but also offers several practical advantages. Compact distilled datasets enable faster training, which facilitates rapid benchmarking of models and training strategies, including hyperparameter tuning and neural architecture search (Zhao et al., 2021). They are also effective in continual learning settings (Zhao & Bilen, 2021) that require quick adaptation to new data or tasks. Moreover, distilling massive datasets into smaller essential forms provides new insights into the core information required for effective learning.

A variety of approaches have been proposed to reduce dataset size, including coreset selection (Farahani & Hekmatfar, 2009; Welling, 2009), dataset filtering (Gadre et al., 2023), and pruning methods (Yang et al., 2023; Mahmoud et al., 2024), which have shown competitive performance in the context of CLIP. However, since they rely on representative subsets of the original dataset, they are effective only with sufficiently large reduced datasets and fail once the subset is too small to preserve semantic diversity. In contrast, dataset distillation (Wang et al., 2018) synthesizes new samples by interpolating the semantics encoded in the original dataset, producing compact yet highly informative datasets. This makes it particularly effective with extremely reduced datasets, where it has outperformed selection-based strategies in image classification (Zhao & Bilen, 2021; Cazenavette et al., 2022), even with only a few distilled samples per class.

[*] Corresponding author

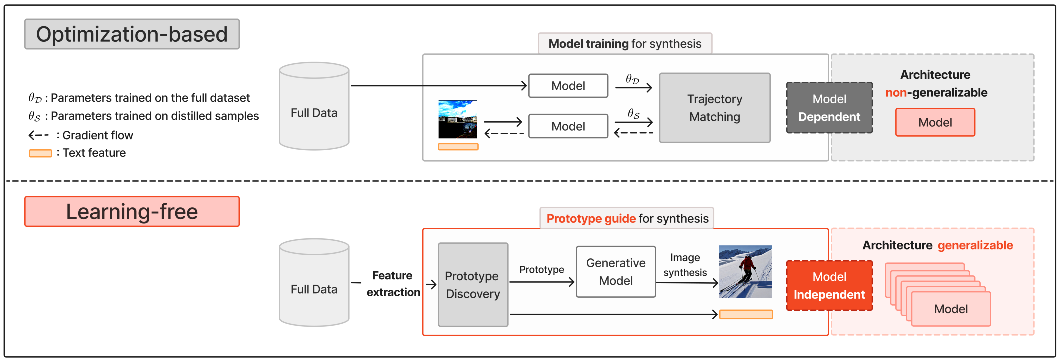


Figure 1: **Concept.** Optimization-based methods are computationally expensive and architecture-dependent. Our learning-free framework is simple, efficient, and generalizable across architectures.

However, research on multimodal dataset distillation remains underexplored, despite the increasing importance of multimodal learning. Existing approaches (Wu et al., 2024; Xu et al., 2024) are typically optimization-based, jointly optimizing image pixels and text features. Unfortunately, these methods face severe computational and scalability challenges. They require repeatedly training models on the full dataset, storing all intermediate parameters, and reusing them during sample synthesis. As datasets and model capacities grow, this process becomes prohibitively time-consuming and memory-intensive. Moreover, the number of parameters to optimize increases linearly with the number of synthetic samples, and numerous hyperparameters (e.g., learning rates for image and text) must be carefully tuned. More critically, these methods are architecture-dependent (Zhao et al., 2023b; Zhong & Liu, 2023). The resulting synthetic dataset is nearly identical to the original data, as the distillation process essentially adds architecture-dependent adversarial perturbations to the initialization images (see Figure 3). Consequently, these distilled datasets generalize poorly when applied to different model architectures, requiring the distillation process to be repeated from scratch. This requirement imposes substantial computational cost and severely limits the practical applicability of existing methods.

In this paper, we propose the prototype-guided data synthesis (PDS) framework, the first learning-free multimodal dataset distillation that is both simple and effective. Unlike optimization-based approaches, PDS requires no training or fine-tuning and avoids learning pixels and text features, making it computationally efficient, architecture-independent, and broadly generalizable (See Figure 1). Recent works have introduced learning-free methods for image classification (Su et al., 2024; Chan-Santiago et al., 2025), where images are generated for each class using explicit class labels. However, these methods cannot be extended to multimodal datasets because they do not consider cross-modal alignment. Aligned image–text pairs are essential for multimodal learning (Chen et al., 2020; Li et al., 2021). However, the VAE encoder (Kingma & Welling, 2014) commonly used in image-only distillation, produces embeddings that are not aligned with text embeddings from separate text encoders. Consequently, naively extending existing learning-free techniques produces synthetic datasets lacking semantic alignment across modalities. This motivates a new learning-free approach that enforces cross-modal alignment to synthesize semantically aligned image–text pairs.

We use CLIP (Radford et al., 2021) to capture image–text alignment in a learning-free manner. For each modality, we extract features and perform clustering to obtain semantically diverse clusters. We then formulate a linear assignment problem to align these clusters across modalities, obtaining image–text prototypes. These prototypes are then used to synthesize images. However, since standard Stable Diffusion models (Rombach et al., 2022; Podell et al., 2024) cannot generate images conditioned on CLIP image embeddings, we build on the idea of unCLIP (Ramesh et al., 2022; Patel et al., 2024), which enables such conditioning. Our ablation studies validate this design choice by confirming the importance of image prototype-based synthesis. Furthermore, experiments show that PDS generalizes well across architectures, reducing dependence on specific architectures and outperforming optimization-based multimodal dataset distillation methods. We also demonstrate the limitations of subset selection when the reduced dataset is too small, and show that image-only learning-free approaches cannot be extended to multimodal settings without cross-modal alignment.

2 Related Work

Dataset distillation. Dataset distillation (Wang et al., 2018) seeks to synthesize a compact synthetic dataset that can replace a large-scale dataset while preserving performance close to training on the full dataset. Most prior work has primarily focused on image classification, aiming to synthesize a few representative images per class, with its potential applications in continual learning (Zhao & Bilen, 2021; Zhao et al., 2021; Zhao & Bilen, 2023), privacy-preserving data sharing (Li et al., 2020; Chen et al., 2022; Dong et al., 2022), and neural architecture search (Zhao et al., 2021; Zhao & Bilen, 2023). We first review optimization-based and generative approaches in image-only settings and then address multimodal dataset distillation.

Optimization-based methods. Early work formulates dataset distillation as a bi-level optimization problem, where synthetic data are optimized to match the training dynamics induced by the full dataset. These approaches generally incur substantial computational overhead, limiting their scalability to high-resolution images. Gradient matching (Zhao & Bilen, 2021; Zhao et al., 2021; Lee et al., 2022) enforces matching at the gradient level, whereas trajectory matching (Cazenavette et al., 2022; Cui et al., 2023; Du et al., 2023) matches the parameter trajectory during training to capture long-term dynamics. TESLA (Cui et al., 2023) makes trajectory matching more practical by proposing a variant that is more efficient in memory and computation time.

To further improve scalability, recent work has proposed several extensions. Decoupled optimization methods such as SRe^{2}L (Yin et al., 2023) adopt a squeeze-recover-relabel procedure. Building on this, subsequent work explores multi-backbone generalization (Shao et al., 2024a), patch-based recovery (Sun et al., 2024), which was later extended with refinement strategies (Tran et al., 2025), and adaptive or category-aware optimization (Shao et al., 2024b). A complementary line of research parameterizes the synthetic dataset with compact factorized representations, such as shared bases with learnable recombination functions (Liu et al., 2022; Deng & Russakovsky, 2022; Wei et al., 2023) and feature-sharing mechanisms (Zheng et al., 2024; Zhang et al., 2025a). Alternative parameterizations include frequency-domain (Shin et al., 2023) and neural-field representations (Shin et al., 2025). In parallel, feature-distribution matching has also been extensively studied (Wang et al., 2022; Liu et al., 2023; Zhao & Bilen, 2023; Zhao et al., 2023a; Zhang et al., 2024).

Generative approaches. Beyond pixel-space optimization, generative models have emerged as an alternative paradigm. One line of work optimizes the latent codes of pretrained generators, including GANs (Zhao & Bilen, 2022; Zhang et al., 2023), StyleGAN variants (Cazenavette et al., 2023; Zhong et al., 2025), and latent diffusion models (Moser et al., 2024; Wang et al., 2025). Another line fine-tunes pretrained generators, such as class-conditional GANs (Zhang et al., 2023) or diffusion models (Gu et al., 2024; Su et al., 2024; Zou et al., 2025). While effective, these methods are still computationally expensive due to the large number of trainable parameters and repeated forward–backward passes. Recent work proposes learning-free approaches that leverage pretrained diffusion models without additional fine-tuning. Specifically, image embeddings are clustered within each class, and representative embeddings from each cluster are used to synthesize new images (Su et al., 2024; Chan-Santiago et al., 2025; Zhao et al., 2025). These methods achieve competitive performance while substantially reducing computational cost.

Multimodal dataset distillation. Existing multimodal dataset distillation approaches are optimization-based, while learning-free methods remain unexplored. The first work (Wu et al., 2024) adopted trajectory matching (Cazenavette et al., 2022) via bi-level optimization to jointly learn image pixels and text features. To mitigate the substantial time and memory costs, the TESLA framework (Cui et al., 2023) was later applied, and Xu et al. (2024) further extended this line of work by replacing the InfoNCE loss with a weighted binary cross-entropy loss and introducing a learnable similarity matrix between synthesized images and texts. Subsequent studies (Zhang et al., 2025b; Dang et al., 2025) refined these strategies, further advancing the optimization-based paradigm.

3 Prototype-guided Data Synthesis

3.1 Problem formulation

Let a large-scale image–text dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ be given, where x_n denotes an image and y_n its paired caption. The objective of multimodal dataset distillation is to synthesize a compact yet

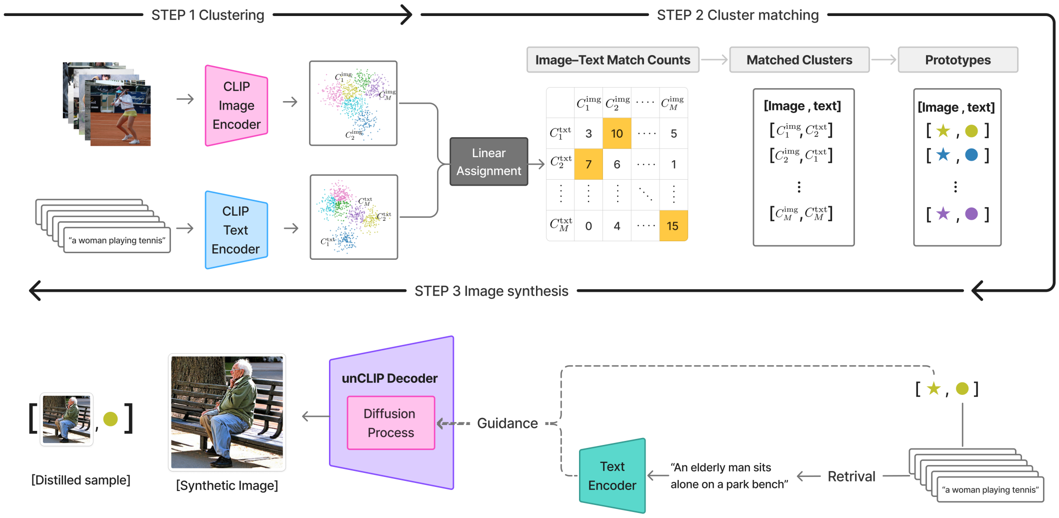


Figure 2: **Our PDS framework** distills multimodal datasets by synthesizing samples from image-text prototypes through three stages: (i) modality-specific clustering of CLIP embeddings, (ii) cross-modal cluster matching via a linear assignment to obtain prototypes, and (iii) image synthesis using an unCLIP decoder guided by image prototype and caption embeddings.

informative dataset $\mathcal{S} = \{(\tilde{x}_m, \tilde{y}_m)\}_{m=1}^{M},\ M \ll N$, such that a vision-language model trained on $\mathcal{S}$ achieves performance without significant degradation compared to training on $\mathcal{D}$. This enables efficient model training with reduced computational and storage costs. Formally, for samples $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{b}$ from the underlying data distribution $\mathcal{P}$, the objective can be formulated as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}}\ \mathbb{E}_{\mathcal{B}\sim\mathcal{P}^{\otimes b}}\Big[\big|\mathcal{L}(\mathcal{B};\boldsymbol{\theta}_{\mathcal{S}}) - \mathcal{L}(\mathcal{B};\boldsymbol{\theta}_{\mathcal{D}})\big|\Big] \quad \text{subject to} \quad \boldsymbol{\theta}_{\mathcal{X}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{X};\boldsymbol{\theta}),$$

where $\mathcal{L}$ denotes a contrastive loss such as InfoNCE. In practice, since text is inherently discrete, dataset distillation is performed in the text embedding space, while images are parameterized in the pixel space. Existing methods typically rely on optimization-based approaches, which require repeated training on the full dataset and storage of intermediate parameters, resulting in significant computational and storage overhead. Moreover, their performance is architecture-dependent, often degrading when used to train models with different backbones.

3.2 Our PDS framework

We propose a simple yet effective PDS framework for multimodal dataset distillation that uses pre-trained models without fine-tuning. The core idea of PDS is to identify image-text prototypes and synthesize new images based on these prototypes to construct a compact dataset. As shown in Figure 2, the pipeline consists of three steps: (i) clustering the CLIP embeddings of images and texts, (ii) matching the clusters to obtain cross-modal prototypes, and (iii) synthesizing images.

Modality-specific clustering. We adopt CLIP encoders (Radford et al., 2021) to obtain image and text embeddings that preserve high-level semantics and remain aligned across modalities. This alignment is particularly important for learning-free dataset distillation, where modality alignment cannot be learned. In fact, using an image encoder that is not aligned with a text encoder typically leads to poor performance, as shown by the VAE-based baselines in Table 3, thereby motivating the use of CLIP for both modalities. After extracting image and text embeddings $\{(z_n^{\text{img}}, z_n^{\text{txt}})\}_{n=1}^{N}$, we prune pairs with low similarity scores to remove noisy or weakly aligned pairs, resulting in a refined subset. We then perform clustering separately for each modality, producing clusters $\{C_m^{\text{img}}\}_{m=1}^{M}$ and $\{C_m^{\text{txt}}\}_{m=1}^{M}$ that group semantically similar embeddings and collectively capture the dataset's broad semantic diversity, where M is set to the desired number of distilled samples. To handle large datasets efficiently, we adopt the mini-batch k-means algorithm (Hartigan & Wong, 1979; Sculley, 2010) (see Appendix C.1).

Cluster matching for prototype construction. While we obtain semantically coherent clusters within each modality, semantically related clusters are not aligned across modalities. To establish cross-modal correspondences, we measure the association between C_i^{img} and C_j^{txt} by the number of shared image-text pairs whose embeddings belong to the respective clusters. Clusters with more shared pairs are regarded as more semantically related, and we therefore aim to find a one-to-one cluster matching that maximizes the total number of shared pairs.

To this end, we formulate the problem as a linear assignment problem (Burkard & Cela, 1999), a framework widely used in logistics and scheduling. Specifically, we construct a cost matrix $K \in \mathbb{R}^{M \times M}$, where each entry represents the negative of the number of shared image-text pairs between a given image cluster and a text cluster:

$$K_{ij} = -\left|\{(z_n^{\text{img}}, z_n^{\text{txt}}) \mid z_n^{\text{img}} \in C_i^{\text{img}},\, z_n^{\text{txt}} \in C_j^{\text{txt}}\}\right|.$$

Maximizing the total number of shared pairs is therefore equivalent to minimizing the total cost:

$$\min_{P \in \{0,1\}^{M \times M}} \sum_{i=1}^{M} \sum_{j=1}^{M} K_{ij} P_{ij} \quad \text{subject to} \quad \sum_{j=1}^{M} P_{ij} = 1,\ \sum_{i=1}^{M} P_{ij} = 1,\ P_{ij} \in \{0,1\},$$

where P is a permutation matrix indicating the chosen matches. We solve this problem using the Hungarian algorithm (Munkres, 1957; Jonker & Volgenant, 1987), which computes the optimal one-to-one matching between image and text clusters.

For each matched cluster pair $(C_i^{\text{img}}, C_j^{\text{txt}})$, we retain only the embeddings of the shared image–text pairs. Embeddings whose paired embedding does not belong to the matched cluster of the other modality are discarded, as they are unlikely to be semantically related to the features in the matched cluster. This filtering strategy strengthens cross-modal alignment between matched clusters (See Appendix C.2). We then obtain the prototypes $(\tilde{z}_i^{\text{img}}, \tilde{z}_j^{\text{txt}})$ for each matched cluster pair by averaging the retained features. If no shared pairs exist, we instead use the original cluster centers to preserve the intended distilled dataset size. Matched clusters without shared pairs (hereafter simply referred to as pairless clusters) rarely appear when the distilled set size is small, and keeping or discarding them yields nearly identical performance. However, as the distilled set size increases, these pairless clusters become more frequent and their centroids tend to become misaligned, weakening cross-modal alignment and degrading performance. In these larger-scale regimes, discarding pairless clusters is therefore preferable. Additional discussion is provided in Appendix C.3. The resulting text prototypes are used as the distilled text features.

Image synthesis. We aim to synthesize images that encode information captured by image prototypes, since selecting real-image subsets does not adequately preserve the dataset's semantic diversity, which leads to lower performance as shown in Table 2. To synthesize images efficiently, we use a generative model rather than pixel-space optimization, as Table 4 demonstrates that direct optimization is slow and memory-intensive. However, existing learning-free dataset distillation methods cannot leverage image prototypes, as standard Stable Diffusion models (Rombach et al., 2022; Podell et al., 2024) do not condition their denoising U-Net (Ronneberger et al., 2015; Ho et al., 2020) on CLIP image embeddings. Similarly, text-to-image generation methods based on retrieved captions rely solely on text, leaving image prototypes unused.

To address this challenge, we build on the idea of unCLIP (Ramesh et al., 2022; Razzhigaev et al., 2023; Patel et al., 2024), which adopts a two-stage pipeline: a separately trained *prior* maps text into the CLIP image embedding space, and a *decoder* then generates images from these embeddings. In our approach, we employ only the unCLIP decoder, conditioning directly on CLIP image embeddings. This design enables the model to utilize image prototypes for generation, resulting in semantically rich synthesized images, as illustrated in Figure 3. In addition, it improves retrieval performance over baselines that either do not use image prototypes or use them only to select real images, as demonstrated in Table 5. We further strengthen semantic alignment by incorporating a text prototype. Because the unCLIP decoder cannot condition on CLIP text embeddings, we instead retrieve the caption most similar to the text prototype from the training set and use it as an additional condition. During diffusion, generation is guided by both the image prototype and the retrieved caption, maintaining semantic alignment. This additional caption condition further improves performance (Appendix C.4). To the best of our knowledge, this is the first dataset distillation method that generates images directly from CLIP image embeddings using an unCLIP decoder.

Table 1: **PDS distills datasets with superior cross-architecture generalization** compared to multimodal dataset distillation baselines. We evaluate on unseen vision backbones, comparing against TESLA-VL and LoRS on Flickr30K (top) and MS-COCO (bottom). Across all distilled dataset sizes and evaluation backbones, PDS consistently outperforms both baselines, demonstrating that its distilled datasets generalize effectively to unseen backbones.

Evaluation Model	Pairs	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	100	TESLA-VL	4.1 ± 0.3	14.7 ± 0.9	22.9 ± 1.2	6.5 ± 0.4	17.8 ± 1.4	27.3 ± 1.4
		LoRS	6.3 ± 0.1	18.6 ± 0.1	28.0 ± 0.2	9.1 ± 0.2	24.3 ± 0.4	34.5 ± 0.8
		PDS	**7.9 ± 0.3**	**25.8 ± 0.4**	**37.3 ± 0.3**	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**
	300	TESLA-VL	10.3 ± 0.3	28.8 ± 0.6	40.6 ± 0.8	14.9 ± 0.9	36.2 ± 1.1	48.8 ± 1.7
		LoRS	8.6 ± 0.9	23.5 ± 1.5	33.5 ± 1.8	14.7 ± 1.2	33.5 ± 2.8	44.1 ± 3.2
		PDS	**14.4 ± 0.4**	**38.1 ± 0.2**	**51.4 ± 0.4**	**18.7 ± 0.5**	**45.0 ± 0.4**	**57.8 ± 0.6**
		Full Dataset	28.5 ± 0.2	59.6 ± 0.1	71.4 ± 0.1	46.0 ± 0.6	76.2 ± 0.3	84.4 ± 0.2
ViT	100	TESLA-VL	2.1 ± 0.3	7.8 ± 0.7	13.1 ± 1.2	2.6 ± 0.6	8.7 ± 0.9	13.7 ± 1.4
		LoRS	2.8 ± 0.1	9.9 ± 0.4	16.1 ± 0.2	5.2 ± 0.3	13.6 ± 0.5	20.5 ± 0.2
		PDS	**6.8 ± 0.3**	**19.2 ± 0.3**	**28.5 ± 0.4**	**6.6 ± 0.5**	**17.5 ± 0.5**	**26.9 ± 0.5**
	300	TESLA-VL	5.1 ± 0.7	16.2 ± 1.0	24.5 ± 1.1	6.1 ± 0.6	18.0 ± 1.0	27.3 ± 1.4
		LoRS	4.1 ± 0.5	13.1 ± 1.0	20.7 ± 1.3	6.2 ± 0.6	17.1 ± 1.2	25.7 ± 1.8
		PDS	**9.1 ± 0.1**	**27.3 ± 0.4**	**38.4 ± 0.4**	**9.6 ± 0.3**	**26.1 ± 0.5**	**37.5 ± 1.2**
		Full Dataset	22.7 ± 0.3	49.9 ± 0.3	62.3 ± 0.3	35.5 ± 0.4	62.9 ± 0.3	73.0 ± 0.4
ResNet	100	TESLA-VL	1.4 ± 0.1	5.8 ± 0.2	10.2 ± 0.4	2.1 ± 0.3	7.6 ± 0.3	12.5 ± 0.3
		LoRS	1.8 ± 0.1	6.8 ± 0.2	11.4 ± 0.4	2.3 ± 0.2	7.8 ± 0.5	12.6 ± 0.4
		PDS	**2.8 ± 0.1**	**10.0 ± 0.2**	**17.3 ± 0.3**	**4.5 ± 0.2**	**14.0 ± 0.3**	**21.4 ± 0.4**
	300	TESLA-VL	3.0 ± 0.1	10.7 ± 0.3	17.6 ± 0.4	5.7 ± 0.2	17.0 ± 0.3	25.6 ± 0.4
		LoRS	2.5 ± 0.2	8.5 ± 0.4	13.8 ± 0.6	3.2 ± 0.2	10.1 ± 0.4	15.9 ± 0.6
		PDS	**5.3 ± 0.2**	**17.2 ± 0.4**	**27.2 ± 0.6**	**7.4 ± 0.3**	**20.7 ± 0.3**	**30.2 ± 0.4**
		Full Dataset	12.6 ± 0.1	33.4 ± 0.1	46.5 ± 0.2	26.6 ± 0.4	53.1 ± 0.2	66.1 ± 0.1
ViT	100	TESLA-VL	0.5 ± 0.2	2.1 ± 0.8	3.8 ± 1.4	0.2 ± 0.1	1.0 ± 0.5	1.7 ± 0.9
		LoRS	0.8 ± 0.1	3.0 ± 0.2	5.4 ± 0.4	0.7 ± 0.2	2.4 ± 0.3	4.1 ± 0.5
		PDS	**2.3 ± 0.1**	**8.6 ± 0.1**	**14.5 ± 0.1**	**2.2 ± 0.1**	**7.7 ± 0.2**	**13.2 ± 0.3**
	300	TESLA-VL	1.5 ± 0.1	5.9 ± 0.1	10.1 ± 0.3	2.2 ± 0.2	7.8 ± 0.2	12.6 ± 0.4
		LoRS	1.0 ± 0.3	3.7 ± 1.1	6.3 ± 1.8	1.3 ± 0.4	4.6 ± 1.4	7.4 ± 2.0
		PDS	**4.1 ± 0.1**	**13.4 ± 0.1**	**21.2 ± 0.1**	**3.7 ± 0.1**	**12.6 ± 0.1**	**19.8 ± 0.2**
		Full Dataset	11.5 ± 0.1	29.8 ± 0.1	41.7 ± 0.1	19.5 ± 0.2	42.4 ± 0.4	55.3 ± 0.3

4 Experiments

This section presents extensive empirical results that validate the effectiveness of PDS. Performance comparisons highlight its strengths in cross-generalization, effectiveness with extremely reduced datasets, and suitability for multimodal datasets. In addition, the ablation studies support our design choices, including the use of generative models and prototype-based image synthesis. In all tables, the best results are highlighted in **bold**, and the second-best are underlined.

4.1 Experimental Settings

Datasets and metrics. We evaluate our approach on two widely used vision-language benchmarks: Flickr30K (Plummer et al., 2015) (29,000/1,000 images for train/test) and MS-COCO (Lin et al., 2014) (113,000/5,000 images for train/test), where each image is paired with five human-annotated captions. Model performance is evaluated using recall at k (R@k) for cross-modal retrieval, with $k \in \{1, 5, 10\}$. For each query from one modality, we compute cosine similarity with all candidates from the other modality in the test set, retrieve the top-k most similar samples, and check whether the ground-truth match is included among them. We report text-to-image retrieval as IR@k and image-to-text retrieval as TR@k.

Implementation details. Unless otherwise specified, all baselines use ViT-L/14 CLIP model (Radford et al., 2021). For PDS, images are generated with the unCLIP decoder using classifier-free guidance (Ho & Salimans, 2022) with a guidance scale of 5.0 and 100 sampling steps, and resized to 224×224 for evaluation. To assess cross-architecture generalization, we follow Xu et al. (2024) and evaluate whether the distilled datasets remain effective when transferred to alternative vision backbones such as ResNet-50 (He et al., 2016) and ViT-Ti/16 (Dosovitskiy et al., 2021). For this evaluation, we use each alternative backbone as the image encoder and attach a learnable linear projection layer after the text encoder. We then fine-tune the CLIP-style model on the distilled dataset while keeping the text encoder frozen. For all baselines, we adopt the hyperparameters reported in the original papers, and conduct all experiments on a single RTX 3090 GPU.

4.2 Performance Comparison

We evaluate PDS against three categories of baselines: multimodal dataset distillation methods, dataset subset selection methods, and learning-free distillation methods for image classification. These comparisons are designed to highlight three aspects of our contribution: (i) mitigating the cross-generalization limitations of existing multimodal dataset distillation methods, (ii) demonstrating the advantage of dataset distillation over subset selection when the reduced dataset is extremely small, and (iii) overcoming the limitations of existing learning-free distillation approaches when extending from image classification to multimodal datasets. Detailed descriptions of each baseline are provided in Appendix A.

Multimodal dataset distillation. A key limitation of architecture-dependent distilled datasets is that they require re-distillation whenever a new backbone is introduced. To examine whether the distilled dataset can generalize to unseen backbones, we compare PDS against multimodal dataset distillation baselines, TESLA-VL and LoRS (Xu et al., 2024), in a cross-architecture setting. We exclude MTT-VL (Wu et al., 2024) due to its higher computational and memory cost as well as its lower performance compared to these methods. The baselines use pretrained NFNet (Brock et al., 2021) as image encoder, following Xu et al. (2024), while we replace the original BERT text encoder (Devlin et al., 2019) with the CLIP text encoder for a fair comparison. This replacement also improves performance (see Appendix C.5).

As shown in Table 1, PDS consistently outperforms both baselines on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014) across all distilled pair sizes and evaluation backbones. For example, with 300 distilled pairs and a ResNet backbone, PDS achieves 14.4% IR@1 and 18.7% TR@1 on Flickr30K, surpassing baselines over 4.1 and 3.8 percentage points (pp), respectively. The performance gap increases at higher retrieval cutoffs (e.g., +10.8 pp in IR@10). These results reveal a key limitation of optimization-based baselines. Their distilled datasets are strongly architecture-dependent and require expensive re-distillation for each new backbone. In contrast, PDS produces distilled datasets that generalize well across architectures through a simple and efficient distillation process, enabling broad practical applicability.

Dataset subset selection. To obtain a highly reduced yet informative dataset, representative images must preserve broad semantic coverage. In this case, dataset distillation offers a more effective solution than subset selection, as it synthesizes data by interpolating diverse semantic features. To validate this, we evaluate PDS against classical coreset-selection methods, K-center (Farahani & Hekmatfar, 2009) and Herding (Welling, 2009), as well as widely used filtering schemes such as CLIP score filtering, LAION filtering, and image-based filtering (Gadre et al., 2023).

As shown in Table 2, PDS consistently achieves the highest performance, outperforming the strongest baseline (Herding) by 17.2 and 10.8 pp in IR@10 and TR@10, respectively. Although dataset filtering methods are widely used in CLIP training due to their effectiveness with moderately reduced datasets, they fail with extremely reduced datasets. These results demonstrate that dataset distillation is particularly well-suited for such scenarios, as it preserves broader semantic coverage through interpolation and achieves superior performance.

Learning-free approaches for image classification. We examine whether learning-free dataset distillation methods for image classification can be extended to the multimodal setting. To investigate this, we extend D^4M (Su et al., 2024) and MGD^3 (Chan-Santiago et al., 2025) by retaining their Stable Diffusion VAE (Rombach et al., 2022) as the image encoder and additionally incorpo-

Table 2: **The representative samples obtained by interpolating diverse semantic features are more suitable than subsets of the original dataset when the reduced dataset is extremely small.** We compare PDS with subset selection methods in the 100-pair setting on Flickr30K. Baselines restricted to real samples are limited in preserving broad semantic diversity. In contrast, PDS distills representative samples that reflect diverse semantics and consistently achieve superior results.

Evaluation Model	**Metric**	**Subset selection**					**Distillation**
		K-center	Herding	CLIP score	LAION filtering	Image-based	PDS
ResNet	IR@1	2.9 ± 0.1	3.6 ± 0.2	2.5 ± 0.1	2.4 ± 0.1	2.2 ± 0.2	**7.9 ± 0.3**
	IR@5	10.4 ± 0.3	12.6 ± 0.5	8.7 ± 0.3	8.5 ± 0.2	8.6 ± 0.3	**25.8 ± 0.4**
	IR@10	16.8 ± 0.3	20.1 ± 0.6	14.5 ± 0.3	14.5 ± 0.3	13.6 ± 0.4	**37.3 ± 0.3**
	TR@1	5.3 ± 0.3	6.7 ± 0.3	4.7 ± 0.4	4.7 ± 0.3	4.0 ± 0.3	**10.2 ± 0.3**
	TR@5	15.6 ± 0.5	19.4 ± 0.7	12.6 ± 0.7	12.4 ± 0.6	10.5 ± 0.4	**28.2 ± 0.9**
	TR@10	24.1 ± 0.4	28.2 ± 0.6	18.9 ± 0.7	19.3 ± 0.4	16.2 ± 0.6	**39.0 ± 0.3**

Table 3: **Cross-modal alignment is crucial for effective multimodal dataset distillation.** We compare PDS with learning-free methods for image classification in the 100-pair setting on Flickr30K. Both D^{4}M and MGD^3 rely on the VAE encoder, which produces image features that are not aligned with text features, leading to low retrieval accuracy. In contrast, PDS uses CLIP encoders to enforce cross-modal alignment, outperforming the baselines and highlighting the importance of alignment-aware distillation.

Evaluation Model	**Methods**	**IR@1**	**IR@5**	**IR@10**	**TR@1**	**TR@5**	**TR@10**
ResNet	D^{4}M	1.3 ± 0.1	5.8 ± 0.2	9.8 ± 0.3	1.4 ± 0.2	5.6 ± 0.5	9.5 ± 0.3
	MGD^3	2.6 ± 0.1	10.0 ± 0.2	17.2 ± 0.5	3.4 ± 0.2	11.6 ± 0.3	17.5 ± 0.4
	PDS	**7.9 ± 0.3**	**25.8 ± 0.4**	**37.3 ± 0.3**	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**

rating a CLIP text encoder. As in our method, prototypes are obtained through clustering and cluster matching, after which each baseline applies its original procedure to synthesize images.

Although these methods perform competitively on image classification, their reliance on the VAE encoder leads to poor alignment with CLIP text embeddings and consequently degrades retrieval performance (Table 3). For example, with ResNet, PDS achieves 37.3% IR@10 and 39.0% TR@10, whereas MGD^3 reaches only 17.2% and 17.5%. These findings highlight the importance of strong cross-modal alignment for effective multimodal dataset distillation and demonstrate the limitations of existing learning-free methods in multimodal settings.

4.3 Ablation Studies

Advantage of using a generative model. We analyze the benefits of using a generative model by comparing PDS with optimization-based multimodal dataset distillation methods. Such optimization-based approaches often synthesize images nearly identical to the initialization sample (Figure 3, left), essentially adding architecture-dependent adversarial perturbations rather than generating genuinely new data. In contrast, PDS employs a generative model to synthesize images from prototypes constructed by interpolating diverse semantic features.

PDS further improves efficiency by eliminating pixel-space optimization. To validate this, we compare PDS with CLIP inversion (Kazemi et al., 2024), which directly optimizes pixel values so that the CLIP embedding of a synthesized image aligns with either an image or text prototype. We evaluate two variants: (i) Image prototype alignment and (ii) Text prototype alignment. As shown in Table 4, PDS reduces peak memory usage (from 6.13 GB to 4.34 GB) and dramatically shortens the generation time (from 1,477 s to 9.7 s per image). In addition, CLIP inversion produces unrealistic images (Figure 3, middle), resulting in poor performance (Table 4).

Impact of image prototype on image generation. To assess the impact of incorporating an image prototype into image generation, we compare PDS with three baselines that obtain images through

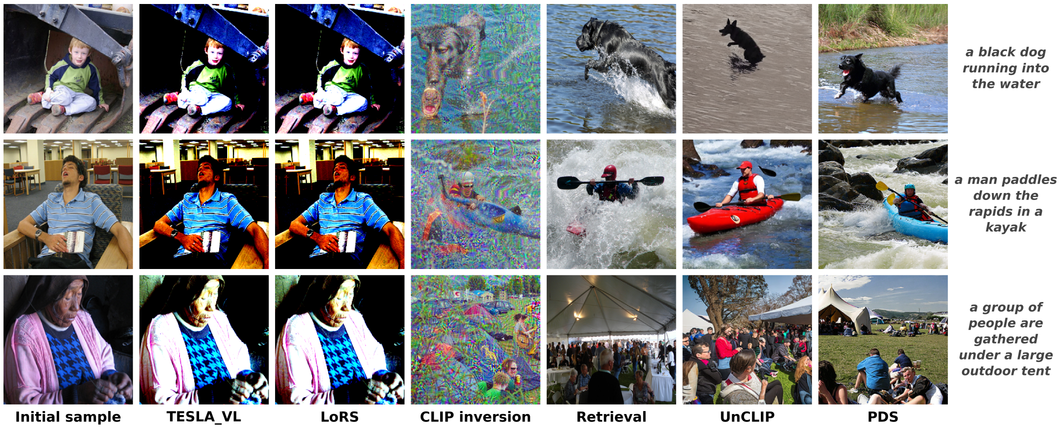

a black dog running into the water
a man paddles down the rapids in a kayak
a group of people are gathered under a large outdoor tent
Initial sample
TESLA_VL
LoRS
CLIP inversion
Retrieval
UnCLIP
PDS

Figure 3: **The synthesized images.** Left (col. 1-3): Images from multimodal dataset distillation baselines, which are nearly identical to the initialization image. Middle: Image generated via CLIP inversion from image prototypes, which is not realistic. Right (col. 5-7): Given a text prototype, we first retrieve the most similar caption and present three images: the paired real image, the UnCLIP-generated image from this caption, and the PDS-generated image. Unlike UnCLIP, which strictly follows captions, PDS produces realistic and semantically enriched images by conditioning on image prototypes.

Table 4: **Using a generative model for image synthesis improves both efficiency and performance.** We evaluate two variants of CLIP inversion: (i) Image prototype alignment and (ii) Text prototype alignment, both of which directly optimize pixel values, in the 100-pair setting on Flickr30K. By employing a generative model, PDS accelerates image generation, reduces memory usage, and produces realistic images, achieving superior performance over CLIP inversion.

Evaluation Model	Methods	IR@1	TR@1	Memory (GB)	Time (s)
ResNet	Text alignment	1.4 ± 0.1	2.0 ± 0.5	6.13	1477.71
	Image alignment	4.4 ± 0.2	4.2 ± 0.3		
	PDS	**7.9 ± 0.3**	**10.2 ± 0.3**	**4.34**	**9.71**

Table 5: **Incorporating an image prototype through the unCLIP decoder improves performance.** We compare PDS with two retrieval baselines and an unCLIP-based text-to-image generation method in the 100-pair setting on Flickr30K. While the retrieval baselines rely on real images and unCLIP generates images from text alone, PDS incorporates image prototypes and textual information through the unCLIP decoder, leading to superior performance.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	Text-prototype	5.2±0.1	17.5±0.2	27.1±0.1	6.4±0.2	19.4±0.4	28.2±0.6
	Image-prototype	5.5±0.3	19.0±0.1	28.7±0.4	8.0±0.4	20.4±0.6	30.2±0.6
	unCLIP	5.2±0.2	17.1±0.4	26.7±0.3	6.4±0.5	19.5±0.8	28.9±0.6
	PDS	**7.9±0.3**	**25.8±0.4**	**37.3±0.3**	**10.2±0.3**	**28.2±0.9**	**39.0±0.3**

different mechanisms. First, we consider two baselines that do not use image prototypes. The first is text-prototype-based retrieval, which retrieves the caption in the training set whose embedding has the highest cosine similarity to the given text prototype and then selects its paired real image. The second is an unCLIP-based text-to-image generation method (Ramesh et al., 2022), which applies the full unCLIP pipeline (prior and decoder) to generate an image from the retrieved caption. In addition, we consider image-prototype-based retrieval, which uses the image prototype only to retrieve the most similar real image from the training set without using it for generation.

Figure 3 (right) illustrates the benefits of incorporating an image prototype into image generation. PDS generates images with richer visual details, producing backgrounds that not only align with the caption but also reflect the interpolated semantic information encoded in the image prototype, resulting in scenes contextually consistent with both. Quantitatively, PDS consistently outperforms baselines that do not use image prototypes (Table 5). For example, with ResNet, IR@10 improves from 26.7% (unCLIP) and 27.1% (Text-prototype) to 37.3%, with similar gains observed for TR@10. The performance remains substantially below that of PDS even when using the image prototype to retrieve real images (Image-prototype). These results demonstrate that using an image prototype for image generation is crucial for producing compact yet highly informative synthetic samples and achieving strong retrieval performance beyond what can be obtained from captions alone or from retrieving real images.

Additional ablation studies and extended analyses are provided in Appendix C, including comprehensive analyses of PDS-related hyperparameters, distilled dataset sizes, and alternative decoder conditioning strategies. This appendix also examines robustness on rare samples, evaluates generalization to a different CLIP backbone, and reports results on an additional dataset. Appendix D presents two applications of our method, demonstrating how the distilled dataset can be used within ASIF (Norelli et al., 2023) and how PDS can be incorporated into an optimization-based distillation framework.

5 Conclusion

Summary. In this work, we introduce PDS, a learning-free approach to multimodal dataset distillation that considers image–text alignment and synthesizes images using image–text prototypes. PDS is scalable and efficient, achieving strong cross-generalization and maintaining significant advantages even with an extremely reduced dataset. This work offers a new perspective on multimodal dataset distillation and demonstrates a more effective and generalizable strategy.

Limitations. PDS leverages a pre-trained generative model capable of synthesizing images conditioned on embeddings. Although stronger models such as SigLIP (Zhai et al., 2023) provide better image–text alignment, the absence of generative models that can be conditioned on these embeddings prevents their direct use for learning-free dataset distillation, suggesting a promising direction for future work. Additionally, the prototypes may be more influenced by dominant classes corresponding to frequently occurring concepts, while rare or long-tail classes may be underrepresented. However, this limitation is shared by subset selection and other distillation baselines, and PDS shows stronger robustness on rare samples than these methods (see Appendix C.9). A separate challenge arises when distilling datasets from domains that differ significantly from those used to train CLIP and the unCLIP decoder. Since these models are trained primarily on natural images, they perform poorly in specialized areas such as medical imaging, and image generation conditioned on such embeddings may fail. In such cases, both CLIP and the unCLIP decoder need to be fine-tuned.

Acknowledgments

Funding. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00240861), a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea government (MOTIE) (RS-2024-00409092, 2024 HRD Program for Industrial Innovation), Institute of Information & Communications Technology Planning & Evaluation (IITP)-Global Data-X Leader HRD program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00441244), and Korea Planning & Evaluation Institute of Industrial Technology (KEIT) funded by the Ministry of Trade, Industry and Resources (No. RS-2025-25458052, Development of Core Technologies for Manufacturing Foundation Models)

References

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *Proc. International Conference on Machine Learning*, pp. 1059–1071. PMLR, 2021.

Rainer E Burkard and Eranda Cela. Linear assignment problems and extensions. In *Handbook of Combinatorial Optimization: Supplement volume A*, pp. 75–149. Springer, 1999.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 4750–4759, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 3739–3748, 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 *Proc. International Conference on Machine Learning*, 2025.

Dingfan Chen, Raouf Kerkouche, and Mario Fritz. Private set generation with discriminative information. In *Proc. Neural Information Processing Systems*, pp. 14678–14690, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-TExt representation learning. In *Proc. European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

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

Zhuohang Dang, Minnan Luo, Chengyou Jia, Hangwei Qian, Xiaojun Chang, and Ivor W Tsang. Multi-modal dataset distillation in the wild. *arXiv preprint arXiv:2506.01586*, 2025.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 39(1):1–22, 1977.

Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *Proc. Neural Information Processing Systems*, pp. 34391–34404, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *Proc. International Conference on Machine Learning*, pp. 5378–5396. PMLR, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations*, 2021.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 3749–3758, 2023.

Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility Location: Concepts, Models, Algorithms and Case Studies*. Physica Heidelberg, 2009.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Proc. Neural Information Processing Systems*, pp. 27092–27112, 2023.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 15793–15803, 2024.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C. Appl.Statist.*, 28(1):100–108, 1979.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* , 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Neural Information Processing Systems*, pp. 6840–6851, 2020.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547, 2019.

Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

Hamid Kazemi, Atoosa Chegini, Jonas Geiping, Soheil Feizi, and Tom Goldstein. What do we learn from inverting clip models? *arXiv preprint arXiv:2403.02580*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. International Conference on Learning Representations*, 2014.

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

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Soft-label anonymous gastric x-ray image distillation. In *Proc. IEEE International Conference on Image Processing*, pp. 305–309. IEEE, 2020.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. Neural Information Processing Systems*, pp. 9694–9705, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *Proc. Neural Information Processing Systems*, pp. 1100–1113, 2022.

Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. DREAM: Efficient dataset distillation by representative matching. In *Proc. International Conference on Computer Vision*, pp. 17314–17324, 2023.

Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S Morcos. Sieve: Multimodal dataset pruning using image captioning models. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 22423–22432, 2024.

Brian B Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models. *arXiv preprint arXiv:2403.03881*, 2024.

James Munkres. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.*, 5(1):32–38, 1957.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. In *Proc. Neural Information Processing Systems*, 2023.

Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 9069–9078, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. International Conference on Computer Vision*, pp. 2641–2649, 2015.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proc. International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proc. Neural Information Processing Systems*, pp. 25278–25294, 2022.

David Sculley. Web-scale k-means clustering. In *Proc. International World Wide Web Conference*, pp. 1177–1178, 2010.

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

Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *Proc. Neural Information Processing Systems*, pp. 99161–99201, 2024b.

Donghyeok Shin, Seungjae Shin, and Il-Chul Moon. Frequency domain-based dataset distillation. In *Proc. Neural Information Processing Systems*, pp. 70033–70044, 2023.

Donghyeok Shin, HeeSun Bae, Gyuwon Sim, Wanmo Kang, and Il-Chul Moon. Distilling dataset into neural field. In *Proc. International Conference on Learning Representations*, 2025.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D^{4}M: Dataset distillation via disentangled diffusion model. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 5809–5818, 2024.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 9390–9399, 2024.

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

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. CaO_2: Rectifying inconsistencies in diffusion-based dataset distillation. In *Proc. International Conference on Computer Vision*, 2025.

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 *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 12196–12205, 2022.

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

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, 58 (301):236–244, 1963.

Xing Wei, Anjia Cao, Funing Yang, and Zhiheng Ma. Sparse parameterization for epitomic dataset distillation. In *Proc. Neural Information Processing Systems*, pp. 50570–50596, 2023.

Max Welling. Herding dynamical weights to learn. In *Proc. International Conference on Machine Learning*, pp. 1121–1128, 2009.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transact. mach. learn. res.*, 2024.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *Proc. International Conference on Machine Learning*, 2024.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. In *Proc. International Conference on Learning Representations*, 2023.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. International Conference on Computer Vision*, pp. 11975–11986, 2023.

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

Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *Proc. AAAI Conference on Artificial Intelligence*, pp. 9314–9322, 2024.

Xin Zhang, Jiawei Du, Ping Liu, and Joey Tianyi Zhou. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. In *Proc. International Conference on Learning Representations*, 2025a.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representations blending for multimodal dataset distillation. *arXiv preprint arXiv:2505.14705*, 2025b.

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

Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022.

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

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

Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 7856–7865, 2023a.

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 *Proc. International Conference on Machine Learning*, 2025.

Lirui Zhao, Yuxin Zhang, Fei Chao, and Rongrong Ji. Boosting the cross-architecture generalization of dataset distillation through an empirical study. *arXiv preprint arXiv:2312.05598*, 2023b.

Haizhong Zheng, Jiachen Sun, Shutong Wu, Bhavya Kailkhura, Z Morley Mao, Chaowei Xiao, and Atul Prakash. Leveraging hierarchical feature sharing for efficient dataset condensation. In *Proc. European Conference on Computer Vision*, pp. 166–182. Springer, 2024.

Xinhao Zhong, Hao Fang, Bin Chen, Xulin Gu, Meikang Qiu, Shuhan Qi, and Shu-Tao Xia. Hierarchical features matter: A deep exploration of progressive parameterization method for dataset distillation. In *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 30462–30471, 2025.

Xuyang Zhong and Chen Liu. Towards mitigating architecture overfitting in dataset distillation. *arXiv preprint arXiv:2309.04195*, 2023.

Yawen Zou, Guang Li, Duo Su, Zi Wang, Jun Yu, and Chao Zhang. Dataset distillation via vision-language category prototype. In *Proc. International Conference on Computer Vision*, 2025.

LLM Usage Statement

We employed large language models solely for polishing the writing and checking grammar.

A Baselines

Dataset subset selection

- **Herding** (Welling, 2009). This method selects samples so that the subset mean approximates the full-dataset mean in the feature space. Starting from an empty set, it iteratively adds the image-text pair whose inclusion minimizes the distance between the updated subset mean and the full-dataset mean.
- **K-center** (Farahani & Hekmatfar, 2009). This method selects samples to maximize dataset coverage by choosing samples that are maximally separated. At each step, it selects the sample whose minimum distance to the selected set is maximized.
- **CLIP score filtering**. This method selects image-text pairs whose CLIP-based cosine similarity exceeds a predefined threshold. In our experiments, we retain the top-k pairs from the candidate set.
- **LAION filtering**. This method follows the LAION-2B filtering scheme in two steps. It first applies cld3 English filtering, retaining pairs whose text is classified as English with probability above a predefined threshold. It then applies CLIP score filtering to select pairs with high CLIP score.
- **Image-based filtering**. This method constructs a subset aligned with the ImageNet classes. It first applies fastText-based English filtering and caption-length filtering. CLIP image embeddings are extracted from the remaining dataset and clustered using FAISS (Johnson et al., 2019). CLIP embeddings of ImageNet images are then assigned to these clusters, and only clusters containing at least one ImageNet embedding are retained. Samples in these retained clusters form the subset, which is further refined using CLIP score filtering.

Multimodal dataset distillation

- **TESLA-VL** (Xu et al., 2024). This method applies the TESLA framework (Cui et al., 2023) to reduce the computational and memory costs of trajectory matching in the original work (Wu et al., 2024), and further replaces the InfoNCE loss with a weighted binary cross-entropy loss.
- **LoRS** (Xu et al., 2024). This method extends TESLA-VL by learning an additional similarity matrix between synthesized images and texts.

Learning-free dataset distillation for image classification

- **D^{4}M** (Su et al., 2024). This method leverages a latent diffusion model (LDM) and introduces an initialization strategy based on cluster prototypes. Features are extracted from the dataset using a VAE encoder, and k-means clustering is performed for each class to obtain class-specific prototypes. These prototypes generate noise through the forward diffusion process, and the resulting noise is used to initialize the reverse diffusion process.
- **MGD^3** (Chan-Santiago et al., 2025). This approach follows a strategy similar to D^{4}M. Features are extracted with a VAE encoder, and k-means clustering is applied within each class to obtain prototypes. These prototypes then serve as explicit guidance during the reverse diffusion process.

B Algorithm

Algorithm 1 PDS framework

```
Require: Original real dataset $\mathcal{D}$; CLIP image ans text encoder; unCLIP decoder.
▷ Modality-Specific Clustering
Calculate the CLIP image and text embeddings $(z^{\text{img}}, z^{\text{txt}})$.
Perform clustering separately on the image and text embeddings to obtain $C^{\text{img}}$ and $C^{\text{txt}}$.
▷ Cluster Matching for Prototype Construction
Solve the linear assignment problem:
    $$\min_{P \in \{0,1\}^{M \times M}} \sum_{i=1}^{M} \sum_{j=1}^{M} K_{ij} P_{ij} \text{ subject to } \sum_{j=1}^{M} P_{ij} = 1, \ \sum_{i=1}^{M} P_{ij} = 1, \ P_{ij} \in \{0, 1\},$$
    where $K_{ij}$ denotes the negative of the number of shared pairs between cluster $i$ and $j$.
For each matched cluster pair, obtain the prototypes $(\bar{z}^{\text{img}}, \bar{z}^{\text{txt}})$.
▷ Image Synthesis
Synthesize images from the prototypes using the unCLIP decoder.
Ensure: Synthetic dataset $\mathcal{S}$
```

C Additional ablation studies

C.1 Effect of the clustering algorithm

We investigate how the choice of clustering algorithm affects overall performance by applying Gaussian mixture models (GMM; (Dempster et al., 1977)), agglomerative clustering (Ward Jr, 1963), and mini-batch k-means (Sculley, 2010). As shown in Table 6, PDS achieves robust performance across different clustering methods, indicating low sensitivity to the choice of clustering algorithm. Although GMM and agglomerative clustering achieve slightly better performance, their standard implementations require full-batch processing without mini-batch support, making them computationally prohibitive for large-scale data (see Table 7, measured on single-modality clustering). In contrast, mini-batch k-means scales efficiently and remains practical for large datasets, making it a suitable choice for real-world scenarios.

Table 6: **Comparison of clustering methods.** We evaluate PDS in the 100-pair setting on Flickr30K using different clustering algorithms. PDS achieves robust performance, indicating low sensitivity to the choice of clustering algorithm.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	GMM	8.3±0.1	26.2±0.3	37.4±0.2	10.9±0.4	30.1±0.4	41.1±0.6
	Agglomerative	**9.3±0.2**	**28.9±0.3**	**41.4±0.1**	**11.6±0.6**	**30.2±0.3**	**41.6±0.6**
	Mini-batch k-means	7.9±0.3	25.8±0.4	37.3±0.3	10.2±0.3	28.2±0.9	39.0±0.3
ViT	GMM	6.0±0.4	19.2±0.3	28.3±0.1	6.1±0.4	18.2±0.9	26.9±0.7
	Agglomerative	6.3±0.2	**19.9±0.2**	**29.5±0.4**	6.4±0.5	**18.5±0.3**	**27.4±0.4**
	Mini-batch k-means	**6.8±0.3**	19.2±0.3	28.5±0.4	**6.6±0.5**	17.5±0.5	26.9±0.5

Table 7: **Clustering time comparison.** Clustering time (in seconds) across different methods.

	GMM	Agglomerative	Mini-batch k-means
Clustering Time (s)	1,469.7	4,387.9	**9.0**

C.2 Effect of the Filtering strategy

When obtaining image-text prototypes from the matched cluster pairs, we use only the embeddings of the shared image-text pairs. In this section, we compare this filtering strategy against a baseline that uses all embeddings without exclusion. Figure 4 presents histograms of cosine similarity between paired image and text prototypes under both settings. Filtering results in higher similarity, indicating stronger cross-modal alignment and ultimately leading to performance gains, as shown in Table 8. Applying the filtering strategy significantly improves performance over the non-filtering baseline. For example, with ResNet, IR@10 improves from 28.1% to 37.3% and TR@10 from 30.6% to 39.0%.

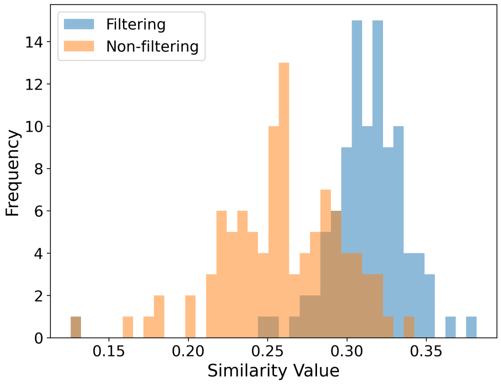


Figure 4: **Histograms of cosine similarity** between paired image and text prototypes in the 100-pair setting on Flickr30K, shown with and without filtering. Filtering leads to higher similarity.

Table 8: **Filtering enhances cross-modal alignment.** We compare our filtering strategy against a non-filtering baseline in the 100-pair setting on Flickr30K. Prototypes obtained from embeddings of the shared image–text pairs consistently improve performance.

Evaluation Model	**Methods**	**IR@1**	**IR@5**	**IR@10**	**TR@1**	**TR@5**	**TR@10**
ResNet	Non-filtering	5.3 ± 0.1	18.4 ± 0.1	28.1 ± 0.3	7.4 ± 0.5	21.2 ± 0.7	30.6 ± 0.5
	Filtering	**7.9 ± 0.3**	**25.8 ± 0.4**	**37.3 ± 0.3**	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**
ViT	Non-filtering	3.9 ± 0.0	14.0 ± 0.0	21.3 ± 0.0	4.5 ± 0.0	12.4 ± 0.0	19.6 ± 0.0
	Filtering	**6.8 ± 0.3**	**19.2 ± 0.3**	**28.5 ± 0.4**	**6.6 ± 0.5**	**17.5 ± 0.5**	**26.9 ± 0.5**

C.3 Effects of pairless clusters under large-scale distilled dataset

To more comprehensively evaluate the effect of matched clusters without shared pairs (pairless clusters), the two strategies (Keeping and Discarding) are applied across a wide range of distilled dataset sizes. In the small-scale settings, both approaches show nearly identical performance because only a few pairless clusters arise. However, as the distilled size increases, the number of such clusters grows substantially, and the pairless clusters tend to have misaligned centroids (see Table 9 and Figure 5). This misalignment weakens cross-modal alignment and leads to the performance degradation reported in Table 10. At larger distilled sizes, discarding pairless clusters is therefore preferable.

Table 9: **Larger distilled sizes yield more pairless clusters.** We distill Flickr30K into various numbers of pairs. As the distilled size increases, the number of pairless clusters grows substantially.

	Pairs				
	100	300	500	1000	1500
Number of pairless clusters	1	8	24	87	166

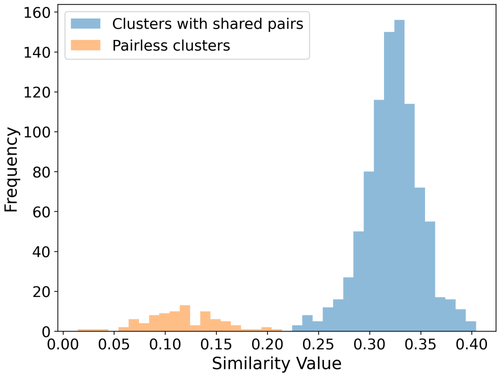


Figure 5: **Histograms of cosine similarity** between paired image and text prototypes in the 1000-pair setting on Flickr30K, shown for pairless clusters and clusters with shared pairs. Pairless clusters exhibit lower similarity.

Table 10: **Discarding pairless clusters becomes increasingly beneficial at larger distilled sizes.** We distill Flickr30K to a wide range of distilled dataset sizes and compare two strategies, keeping all clusters or discarding the pairless clusters. While both strategies perform similarly in small-scale settings, discarding pairless clusters becomes preferable when the distilled size becomes large.

Evaluation Model	**Pairs**	**Methods**	**IR@1**	**IR@5**	**IR@10**	**TR@1**	**TR@5**	**TR@10**
ResNet	100	Including	**7.9 ± 0.3**	25.8 ± 0.4	37.3 ± 0.3	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**
		Discarding	7.8 ± 0.2	**25.9 ± 0.4**	**37.5 ± 0.2**	10.0 ± 0.3	27.5 ± 1.2	38.8 ± 0.9
	300	Including	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	18.7 ± 0.5	**45.0 ± 0.4**	**57.8 ± 0.6**
		Discarding	**14.9 ± 0.2**	**38.6 ± 0.5**	**51.9 ± 0.3**	**19.3 ± 0.4**	**45.0 ± 0.5**	57.2 ± 0.6
	500	Including	16.3 ± 0.2	41.3 ± 0.2	54.9 ± 0.4	22.5 ± 0.3	49.3 ± 0.8	61.4 ± 0.4
		Discarding	**17.1 ± 0.2**	**43.0 ± 0.2**	**56.4 ± 0.1**	**23.9 ± 0.9**	**50.5 ± 0.6**	**62.0 ± 0.3**
	1000	Including	18.2 ± 0.3	43.5 ± 0.2	56.4 ± 0.5	24.3 ± 0.4	52.7 ± 0.6	64.7 ± 0.6
		Discarding	**19.5 ± 0.3**	**46.4 ± 0.4**	**58.9 ± 0.2**	**27.4 ± 0.4**	**55.3 ± 0.6**	**67.5 ± 0.6**
	1500	Including	18.9 ± 0.2	45.6 ± 0.2	59.4 ± 0.3	26.5 ± 0.3	55.5 ± 0.8	67.8 ± 1.2
		Discarding	**20.5 ± 0.2**	**48.7 ± 0.3**	**62.5 ± 0.2**	**28.8 ± 0.5**	**59.2 ± 0.8**	**71.2 ± 0.4**

C.4 Effect of incorporating captions retrieved from the text prototype

We generate images conditioned not only on the image prototype but also on a caption retrieved from the text prototype, to enhance semantic alignment between the generated image and the text prototype. To evaluate the benefit of this design, we compare our method with a baseline that generates images conditioned solely on the image prototype. As shown in Table 11, incorporating the caption improves performance.

Table 11: **Joint conditioning improves performance.** We compare our method with a baseline that generates images conditioned solely on the image prototype in the 100-pair setting on Flickr30K. Conditioning on both the image prototype and a caption retrieved from the text prototype improves performance.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	Image prototype	6.0 ± 0.3	19.2 ± 0.3	28.3 ± 0.1	6.1 ± 0.4	18.2 ± 0.9	26.9 ± 0.7
	Both prototypes	**7.9 ± 0.3**	**25.8 ± 0.4**	**37.3 ± 0.3**	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**
ViT	Image prototype	5.4 ± 0.1	17.5 ± 0.2	26.5 ± 0.5	5.4 ± 0.3	**17.8** ± 0.5	25.7 ± 1.1
	Both prototypes	**6.8 ± 0.3**	**19.2 ± 0.3**	**28.5 ± 0.4**	**6.6 ± 0.5**	17.5 ± 0.5	**26.9 ± 0.5**

C.5 Performance comparison between BERT and CLIP text encoder

To ensure a fair comparison of PDS with optimization-based multimodal dataset distillation baselines, we replace the BERT encoder used in prior work with the CLIP text encoder. As shown in Table 12, this replacement not only ensures a fair comparison but also consistently improves performance.

Table 12: **Comparison of BERT and CLIP as text encoders in the 100-pair setting on Flickr30K.** Within each method, we compare BERT and CLIP as text encoders under the same vision backbone (ResNet or ViT). In all cases, replacing BERT with CLIP consistently improves performance.

Method	Evaluation Model	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
TESLA-VL	ResNet+BERT	0.3 ± 0.1	1.9 ± 0.2	3.4 ± 0.2	2.0 ± 0.4	7.9 ± 0.5	12.2 ± 0.5
	ResNet+CLIP	**4.1 ± 0.3**	**14.7 ± 0.9**	**22.9 ± 1.2**	**6.5 ± 0.4**	**17.8 ± 1.4**	**27.3 ± 1.4**
	ViT+BERT	0.2 ± 0.1	1.0 ± 0.2	1.9 ± 0.5	0.5 ± 0.2	2.0 ± 0.6	3.9 ± 1.4
	ViT+CLIP	**2.1 ± 0.3**	**7.8 ± 0.7**	**13.1 ± 1.2**	**2.6 ± 0.6**	**8.7 ± 0.9**	**13.7 ± 1.4**
LoRS	ResNet+BERT	2.0 ± 0.1	7.3 ± 0.3	11.8 ± 0.2	3.7 ± 0.4	13.4 ± 0.5	19.8 ± 0.8
	ResNet+CLIP	**6.3 ± 0.1**	**18.6 ± 0.1**	**28.0 ± 0.2**	**9.1 ± 0.2**	**24.3 ± 0.4**	**34.5 ± 0.8**
	ViT+BERT	0.4 ± 0.1	2.7 ± 0.1	5.1 ± 0.2	1.7 ± 0.2	6.2 ± 0.5	10.3 ± 0.5
	ViT+CLIP	**2.8 ± 0.1**	**9.9 ± 0.4**	**16.1 ± 0.2**	**5.2 ± 0.3**	**13.6 ± 0.5**	**20.5 ± 0.2**

C.6 Sensitivity Analysis of the hyperparameters

We conduct extensive ablation studies on decoder hyperparameters such as the guidance scale and the number of sampling steps, as well as on the PDS-specific hyperparameters, including the clustering seeds and the similarity-pruning threshold ratio. We first examine the sensitivity to the decoder hyperparameters. Varying the guidance scale $\in \{3, 5, 7, 10, 15\}$ and the number of sampling steps $\in \{25, 50, 100, 150, 200\}$ yields consistent retrieval accuracy, with only slight degradation at smaller values where the effect of prototype conditioning weakens. These trends are shown in Figure 6, which reports the aggregated retrieval accuracy, computed as the average of IR@1, IR@5, IR@10, TR@1, TR@5, and TR@10, using distilled sets of 100 and 300 pairs.

We next analyze the sensitivity to the PDS-specific hyperparameters. Performance remains consistent across different clustering seeds, as illustrated by the tight box plot shown in Figure 6, using 20 different seeds. The same figure also presents the effect of the similarity-pruning threshold ratio $\in \{0.0, 0.1, 0.2, 0.3, 0.5\}$, showing that moderate pruning reliably improves performance by removing weakly aligned samples.

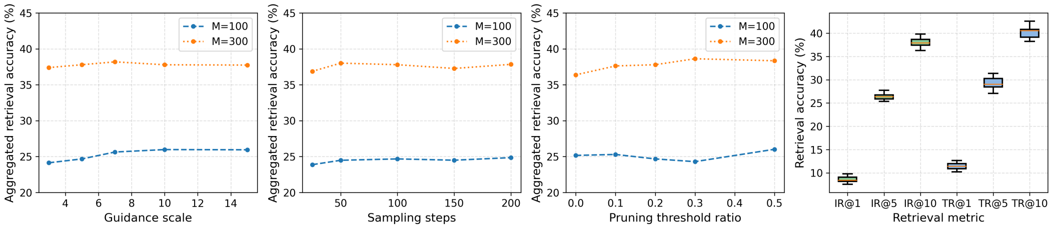


Figure 6: **Ablation study on hyperparameters.** The first three panels show the results obtained by varying the guidance scale, the number of sampling steps, and the pruning threshold ratio, respectively. Each panel reports the aggregated retrieval accuracy, computed as the average of IR@1, IR@5, IR@10, TR@1, TR@5, and TR@10, using distilled sets of 100 and 300 pairs on Flickr30K. The final panel presents a boxplot of retrieval accuracy across 20 different clustering seeds. All results in the figure were evaluated using a ResNet backbone.

C.7 Effect of the distilled dataset size on performance

Table 13 presents a monotonic improvement in performance as the distilled dataset size increases, with the results gradually approaching those obtained from the full dataset. In practice, the distilled dataset size may be chosen heuristically based on the structure of the data embeddings. When the embeddings form well-separated and compact clusters, a relatively small number of pairs is often sufficient because each cluster generally corresponds to a coherent semantic group. In contrast, when the embedding space is more diffuse or spread out, using a larger number of pairs is usually beneficial, as it allows the distilled dataset to capture more local structure.

Table 13: **Increasing the distilled dataset size consistently improves performance.** We distill Flickr30K using a wide range of distilled dataset sizes and evaluate the resulting distilled datasets. In all cases, larger distilled sets lead to better performance.

Evaluation Model	Pairs	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	100	7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3
	300	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	18.7 ± 0.5	45.0 ± 0.4	57.8 ± 0.6
	500	17.1 ± 0.2	43.0 ± 0.2	56.4 ± 0.1	23.9 ± 0.9	50.5 ± 0.6	62.0 ± 0.3
	1000	19.5 ± 0.3	46.4 ± 0.4	58.9 ± 0.2	27.4 ± 0.4	55.3 ± 0.6	67.5 ± 0.6
	1500	20.5 ± 0.2	48.7 ± 0.3	62.5 ± 0.2	28.8 ± 0.5	59.2 ± 0.8	71.2 ± 0.4
	Full Dataset	28.5 ± 0.2	59.6 ± 0.1	71.4 ± 0.1	46.0 ± 0.6	76.2 ± 0.3	84.4 ± 0.2

C.8 Alternative CLIP backbone

For our main experiments, we adopt the ViT-L/14 CLIP model because it is more computationally efficient for extracting image and text features. We additionally evaluate PDS using the ViT-H/14 CLIP to examine whether the improvement generalizes to a different CLIP backbone. As shown in Table 14, using the ViT-H/14 CLIP results in improved performance, likely due to its stronger alignment between image and text features. Importantly, our PDS framework continues to outperform all other baselines even with this CLIP variant, indicating that the observed gains are not specific to ViT-L/14.

Table 14: **Results with an alternative CLIP backbone in the 300-pair setting on Flickr30K.** We evaluate PDS using the ViT-H/14 CLIP model to examine whether the improvements generalize across different backbones. The ViT-H/14 CLIP achieves higher performance, likely due to its stronger image-text alignment.

Evaluation Model	CLIP backbone	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	ViT-L/14	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	18.7 ± 0.5	45.0 ± 0.4	57.8 ± 0.6
	ViT-H/14	**17.2 ± 0.5**	**42.6 ± 0.3**	**55.6 ± 0.4**	**23.7 ± 0.9**	**47.3 ± 0.8**	**60.0 ± 0.8**
ViT	ViT-L/14	9.1 ± 0.1	27.3 ± 0.4	38.4 ± 0.4	9.6 ± 0.3	26.1 ± 0.5	37.5 ± 1.2
	ViT-H/14	**12.0 ± 0.1**	**32.8 ± 0.3**	**43.2 ± 0.6**	**12.7 ± 0.6**	**32.4 ± 0.2**	**44.3 ± 0.2**

C.9 Robustness of the PDS to rare samples

The prototypes from PDS may be more influenced by features from dominant classes, which could lead to underrepresentation of rare or long-tail classes. However, we note that this limitation is shared by subset selection and other distillation baselines, and PDS empirically demonstrates stronger robustness in such cases. For this evaluation, we define rare samples as the 200 Flickr30K test samples farthest from the mean embedding of the training dataset. Using the 100 samples selected or distilled from Flickr30K by each method, we measure the retrieval accuracy on this rare subset. As shown in Table 15, PDS consistently outperforms both subset selection and other distillation baselines. These results indicate that, despite the inherent limitation, PDS provides a better coverage of rare examples compared to all other baselines.

Table 15: **PDS is more robust to rare samples than existing methods.** We evaluate each method on 200 rare test samples from Flickr30K, defined as those farthest from the mean embedding of the training dataset. In the 100-pair setting on Flickr30K, PDS consistently outperforms both subset selection and other distillation baselines, demonstrating stronger robustness to rare samples.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	K-center	8.5 ± 1.4	26.7 ± 0.9	40.0 ± 0.8	15.4 ± 0.9	40.3 ± 0.5	54.0 ± 1.6
	Herding	9.4 ± 0.5	28.7 ± 0.6	43.7 ± 1.7	15.6 ± 1.2	38.8 ± 2.0	51.3 ± 1.4
	CLIP score	7.2 ± 0.4	22.3 ± 0.4	33.5 ± 0.5	12.8 ± 1.6	30.6 ± 1.2	42.9 ± 2.1
	LAION filtering	6.6 ± 0.3	20.2 ± 0.5	31.8 ± 0.3	8.8 ± 0.5	24.3 ± 1.0	36.3 ± 1.8
	Image-based	6.7 ± 0.5	22.4 ± 0.3	33.3 ± 0.3	12.3 ± 1.8	30.2 ± 1.4	42.1 ± 1.5
	TESLA-VL	11.0 ± 1.7	35.0 ± 1.5	50.9 ± 2.1	15.0 ± 2.9	39.6 ± 3.9	52.5 ± 3.0
	LoRS	14.1 ± 0.5	37.2 ± 1.0	54.0 ± 0.5	22.6 ± 1.9	47.4 ± 1.4	62.2 ± 2.3
	PDS	**23.2 ± 0.8**	**58.4 ± 0.6**	**74.5 ± 0.8**	**25.2 ± 1.0**	**58.8 ± 1.2**	**73.6 ± 2.7**

C.10 Result on the Flickr8K Dataset

We further evaluate PDS on the Flickr8K (Hodosh et al., 2013) dataset and compare it against subset selection methods. As shown in Table 16, PDS consistently outperforms the subset selection baselines.

Table 16: **Comparison of PDS with subset selection methods on Flickr8K.** We evaluate PDS and subset selection baselines in the 100-pair setting on Flickr8K, and PDS consistently outperforms all subset selection methods.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	K-center	0.8 ± 0.8	3.4 ± 3.1	6.1 ± 5.2	1.2 ± 1.3	3.9 ± 3.5	6.6 ± 5.1
	Herding	0.8 ± 0.9	3.3 ± 3.4	5.6 ± 5.5	1.0 ± 1.4	3.8 ± 4.4	6.0 ± 6.5
	CLIP score	3.5 ± 0.2	12.5 ± 0.2	19.4 ± 0.3	5.5 ± 0.4	14.4 ± 0.6	22.1 ± 0.5
	LAION filtering	3.5 ± 0.1	13.0 ± 0.3	20.4 ± 0.3	6.4 ± 0.5	15.2 ± 0.9	22.6 ± 0.7
	Image-based	3.4 ± 0.2	12.3 ± 0.2	18.8 ± 0.2	5.6 ± 0.3	14.1 ± 0.6	21.9 ± 0.4
	PDS	**8.5 ± 0.2**	**27.0 ± 0.3**	**39.8 ± 0.3**	**10.3 ± 0.2**	**27.1 ± 0.7**	**38.3 ± 0.4**

C.11 Comparison of Text, Image, and Average Prototype Conditioning in the UnCLIP Decoder

We additionally generate images conditioned either on the text prototype or on the average of the text and image prototypes. As shown in Table 17, conditioning on the text prototype results in degraded performance, while averaging the text and image prototypes produces results comparable to conditioning on the image prototype, with no noticeable improvements.

Table 17: **Comparison of conditioning on text, image, and averaged prototypes.** We evaluate different conditioning strategies in the 100-pair setting on Flickr30K. Conditioning on the text prototype degrades performance, while using the average of the text and image prototypes achieves results comparable to conditioning on the image prototype, providing no additional gains.

Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	Text prototype	**8.1 ± 0.2**	24.9 ± 0.2	36.6 ± 0.4	10.8 ± 0.4	**28.2 ± 0.7**	38.3 ± 0.7
	Average prototype	7.9 ± 0.2	**26.2 ± 0.5**	**37.9 ± 0.5**	**11.1 ± 0.3**	**28.2 ± 0.5**	**39.4 ± 0.3**
	Image prototype	7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	**28.2 ± 0.9**	39.0 ± 0.3
ViT	Text prototype	5.6 ± 0.2	17.3 ± 0.2	26.4 ± 0.1	4.6 ± 0.4	16.3 ± 0.7	23.9 ± 0.6
	Average prototype	5.9 ± 0.1	18.5 ± 0.1	27.7 ± 0.3	6.5 ± 0.6	**18.4 ± 0.3**	**27.3 ± 0.3**
	Image prototype	**6.8 ± 0.3**	**19.2 ± 0.3**	**28.5 ± 0.4**	**6.6 ± 0.5**	17.5 ± 0.5	26.9 ± 0.5

C.12 Comparison of Joint and Separate Clustering Approaches

It is certainly possible to perform clustering directly in a joint embedding space, but the two modalities can differ in how their embeddings are distributed, which can sometimes cause the clustering to be influenced more by one modality than the other. For this reason, we cluster each modality separately and then match the resulting clusters across modalities. As shown in Table 18, the two approaches perform comparably, and the separate clustering approach is slightly better in a few cases.

Table 18: **Comparison of joint-embedding clustering (Joint) and separate clustering (Separate).** We evaluate different clustering strategies by distilling the Flickr30K and MS-COCO into 100 and 300 pairs, and evaluating the resulting distilled sets using a ResNet backbone. Overall, both clustering strategies perform comparably, with the separate clustering approach showing slight improvements in a few cases.

Dataset	Pairs	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Flickr30K	100	Joint	**8.3 ± 0.1**	25.3 ± 0.2	36.8 ± 0.4	**11.0 ± 0.4**	**29.5 ± 0.4**	**40.7 ± 0.4**
		Separate	7.9 ± 0.3	**25.8 ± 0.4**	**37.3 ± 0.3**	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3
	300	Joint	**14.6 ± 0.5**	**38.8 ± 0.3**	**52.3 ± 0.4**	17.5 ± 0.4	43.5 ± 0.4	56.4 ± 0.8
		Separate	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	**18.7 ± 0.5**	**45.0 ± 0.4**	**57.8 ± 0.6**
MS-COCO	100	Joint	2.5 ± 0.1	**10.2 ± 0.2**	**17.5 ± 0.3**	4.0 ± 0.1	12.9 ± 0.2	20.6 ± 0.2
		Separate	**2.8 ± 0.1**	10.0 ± 0.2	17.3 ± 0.3	**4.5 ± 0.2**	**14.0 ± 0.3**	**21.4 ± 0.4**
	300	Joint	4.9 ± 0.1	16.6 ± 0.2	25.9 ± 0.3	6.6 ± 0.2	19.3 ± 0.3	29.2 ± 0.4
		Separate	**5.3 ± 0.2**	**17.2 ± 0.4**	**27.2 ± 0.6**	**7.4 ± 0.3**	**20.7 ± 0.3**	**30.2 ± 0.4**

D Application

D.1 Application of the distilled dataset to ASIF

We apply our distilled dataset to ASIF (Norelli et al., 2023). Although ASIF and PDS are developed for different purposes, they are complementary. ASIF aligns pre-trained unimodal encoders without additional training by computing relative representations whose elements are cosine similarities between a sample and each anchor in the anchor set. Consequently, the quality of the anchor set is crucial, and the computational cost of ASIF increases with the number of anchors. The original ASIF paper reports that more than one million anchors are required to reach CLIP-level performance, highlighting the importance of identifying compact yet informative anchors.

PDS can address this point by producing a small set of highly informative samples that can serve as stronger anchors than those obtained by selecting subsets of the original dataset. In Table 2 of the manuscript, we already show that PDS produces more representative samples than subset selection baselines. Based on this result, we evaluate the performance of ASIF with 300 anchors obtained from Flickr30K using (i) subset selection methods, (ii) other distillation methods, and (iii) PDS. For this comparison, we use a ResNet image encoder and a CLIP text encoder to compute the relative representations, and performance is evaluated through retrieval accuracy on Flickr30K. As shown in Table 19, the anchors produced by PDS lead to the strongest performance among all compared methods, implying that PDS allows ASIF to operate effectively with fewer anchors.

Table 19: **Application of the distilled dataset as anchors in ASIF.** To evaluate how well a distilled dataset can serve as compact anchors for ASIF, we construct 300 anchors from Flickr30K for each method, including subset selection, other distillation approaches, and PDS. Using a ResNet image encoder and a CLIP text encoder to compute relative representations, we measure retrieval accuracy on Flickr30K. Anchors produced by PDS consistently outperform those obtained from all other baselines, demonstrating that PDS yields more informative and effective anchors for ASIF.

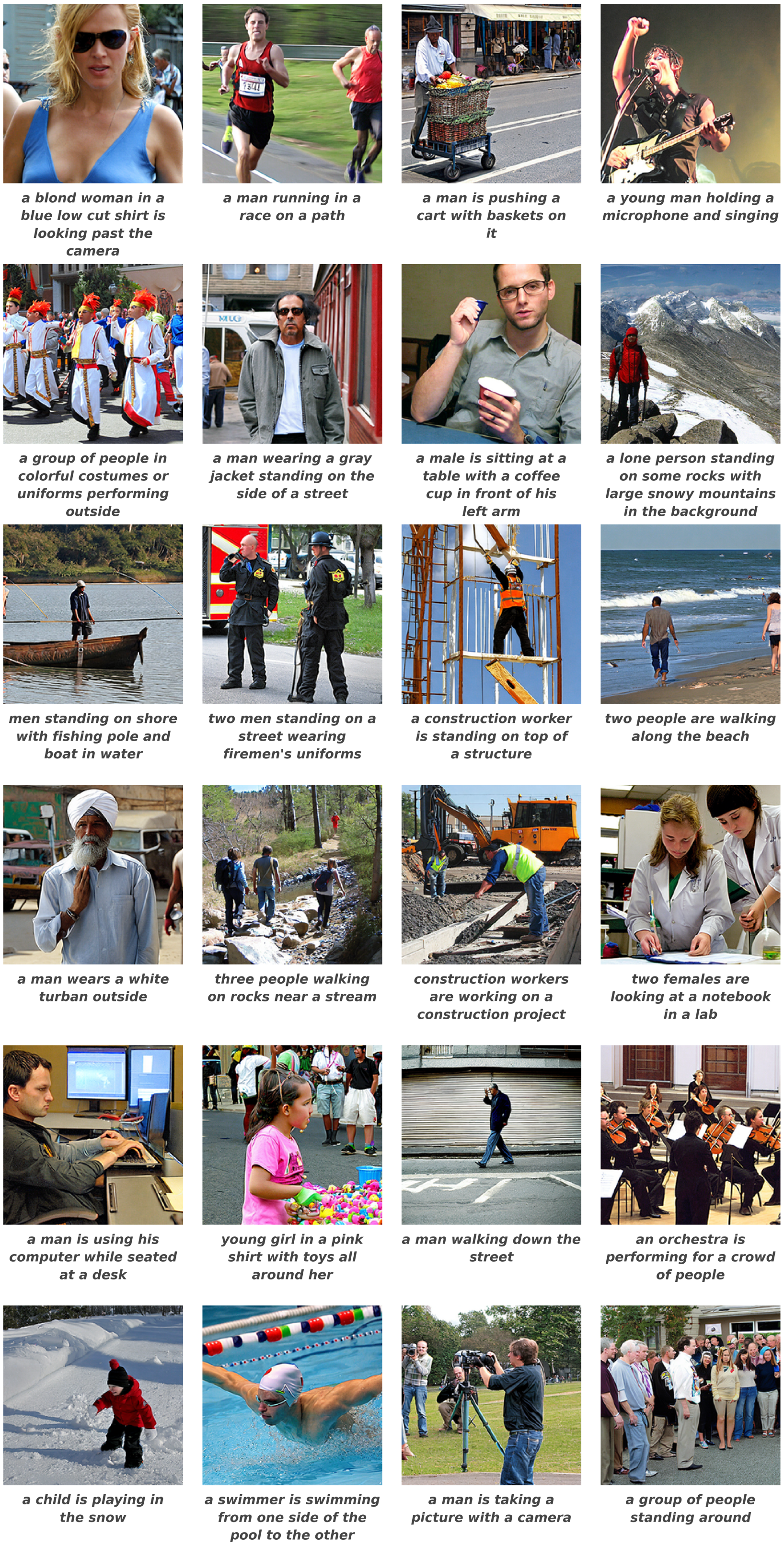
Evaluation Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	K-center	3.8	14.4	24.2	4.0	14.8	22.7
	Herding	6.0	21.5	32.5	6.9	19.1	28.1
	CLIP score	2.1	9.5	16.9	3.5	11.5	18.7
	LAION filtering	3.1	11.2	17.9	3.4	12.5	17.4
	Image-based	2.3	9.6	17.1	4.1	12.6	18.3
	TESLA-VL	5.3	19.4	29.8	5.0	16.8	25.9
	LoRS	4.9	18.6	28.2	5.7	16.6	25.9
	PDS	**9.9**	**29.8**	**41.8**	**9.6**	**27.7**	**38.6**

D.2 Integration of PDS into Optimization-based multimodal distillation

We integrate PDS into existing optimization-based multimodal distillation methods by using the PDS-distilled dataset as an initialization, which is known to be crucial for optimization-based distillation. Although replacing real samples used for initialization with PDS-distilled dataset provides an even more informative starting point, the subsequent optimization process leads to an architecture-dependent distilled dataset, which undermines the cross-architecture generalization achieved by PDS. As shown in Table 20, initializing the optimization-based method with the PDS-distilled dataset tends to improve its performance, but it still underperforms compared to using PDS alone.

Table 20: **Integrating PDS into optimization-based distillation baselines provides little advantage.** We evaluate optimization-based multimodal distillation methods initialized with the PDS-distilled dataset in the 100-pair setting on Flickr30K. Although this initialization provides a more informative starting point, the subsequent optimization makes the distilled dataset architecture-dependent, limiting the cross-architecture generalization achieved by PDS. Consequently, while PDS initialization improves these baselines, this integrated approach still underperforms compared to using PDS alone.

Evaluation Model	**Method**	**IR@1**	**IR@5**	**IR@10**	**TR@1**	**TR@5**	**TR@10**
ResNet	TESLA-VL	4.1 ± 0.3	14.7 ± 0.9	22.9 ± 1.2	6.5 ± 0.4	17.8 ± 1.4	27.3 ± 1.4
	TESLA-VL +PDS	5.2 ± 0.4	17.7 ± 0.6	27.5 ± 0.4	7.5 ± 1.0	22.4 ± 1.3	33.1 ± 1.3
	LoRS	6.3 ± 0.1	18.6 ± 0.1	28.0 ± 0.2	9.1 ± 0.2	24.3 ± 0.4	34.5 ± 0.8
	LoRS+PDS	5.7 ± 0.2	19.5 ± 0.6	29.2 ± 0.9	8.3 ± 1.2	24.1 ± 1.3	34.3 ± 1.4
	PDS	**7.9 ± 0.3**	**25.8 ± 0.4**	**37.3 ± 0.3**	**10.2 ± 0.3**	**28.2 ± 0.9**	**39.0 ± 0.3**



a blond woman in a blue low cut shirt is looking past the camera

a man running in a race on a path

a man is pushing a cart with baskets on it

a young man holding a microphone and singing

a group of people in colorful costumes or uniforms performing outside

a man wearing a gray jacket standing on the side of a street

a male is sitting at a table with a coffee cup in front of his left arm

a lone person standing on some rocks with large snowy mountains in the background

men standing on shore with fishing pole and boat in water

two men standing on a street wearing firemen's uniforms

a construction worker is standing on top of a structure

two people are walking along the beach

a man wears a white turban outside

three people walking on rocks near a stream

construction workers are working on a construction project

two females are looking at a notebook in a lab

a man is using his computer while seated at a desk

young girl in a pink shirt with toys all around her

a man walking down the street

an orchestra is performing for a crowd of people

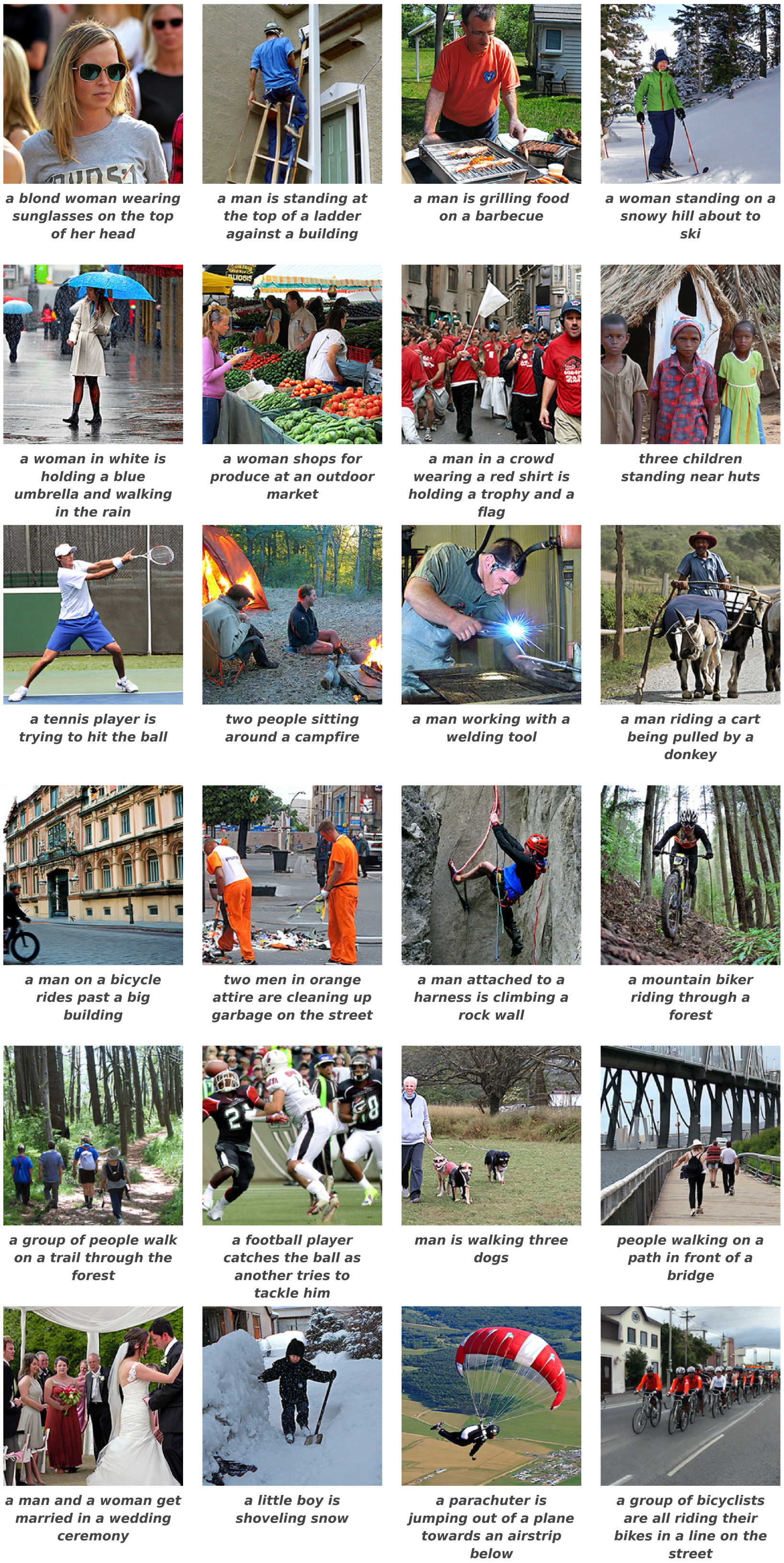
a child is playing in the snow

a swimmer is swimming from one side of the pool to the other

a man is taking a picture with a camera

a group of people standing around

Figure 7: Examples of synthetic images and retrieved captions in the 100-pair setting on Flickr30K.



a blond woman wearing sunglasses on the top of her head

a man is standing at the top of a ladder against a building

a man is grilling food on a barbecue

a woman standing on a snowy hill about to ski

a woman in white is holding a blue umbrella and walking in the rain

a woman shops for produce at an outdoor market

a man in a crowd wearing a red shirt is holding a trophy and a flag

three children standing near huts

a tennis player is trying to hit the ball

two people sitting around a campfire

a man working with a welding tool

a man riding a cart being pulled by a donkey

a man on a bicycle rides past a big building

two men in orange attire are cleaning up garbage on the street

a man attached to a harness is climbing a rock wall

a mountain biker riding through a forest

a group of people walk on a trail through the forest

a football player catches the ball as another tries to tackle him

man is walking three dogs

people walking on a path in front of a bridge

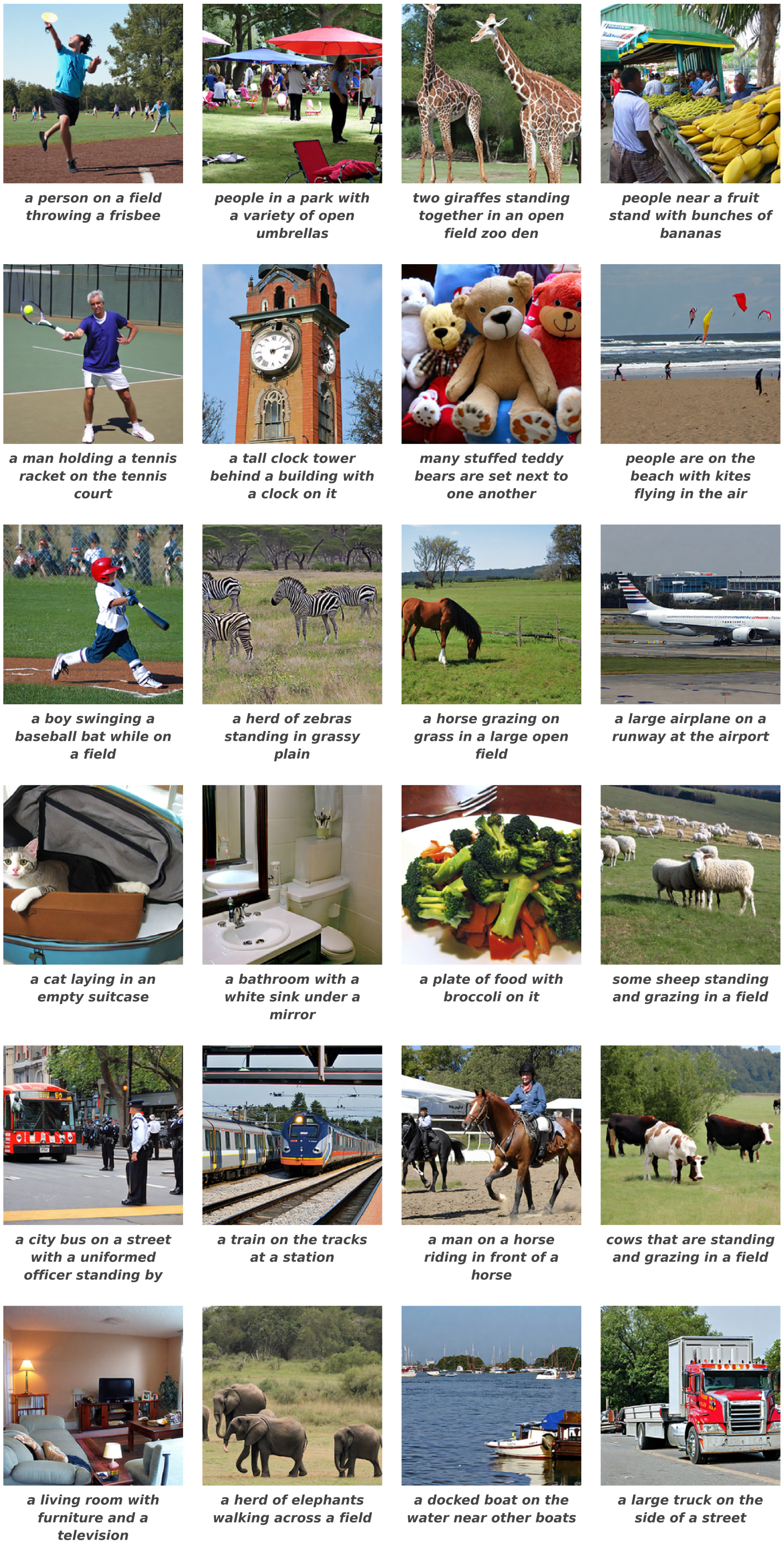
a man and a woman get married in a wedding ceremony

a little boy is shoveling snow

a parachuter is jumping out of a plane towards an airstrip below

a group of bicyclists are all riding their bikes in a line on the street

Figure 8: Examples of synthetic images and retrieved captions in the 300-pair setting on Flickr30K.



a person on a field throwing a frisbee

people in a park with a variety of open umbrellas

two giraffes standing together in an open field zoo den

people near a fruit stand with bunches of bananas

a man holding a tennis racket on the tennis court

a tall clock tower behind a building with a clock on it

many stuffed teddy bears are set next to one another

people are on the beach with kites flying in the air

a boy swinging a baseball bat while on a field

a herd of zebras standing in grassy plain

a horse grazing on grass in a large open field

a large airplane on a runway at the airport

a cat laying in an empty suitcase

a bathroom with a white sink under a mirror

a plate of food with broccoli on it

some sheep standing and grazing in a field

a city bus on a street with a uniformed officer standing by

a train on the tracks at a station

a man on a horse riding in front of a horse

cows that are standing and grazing in a field

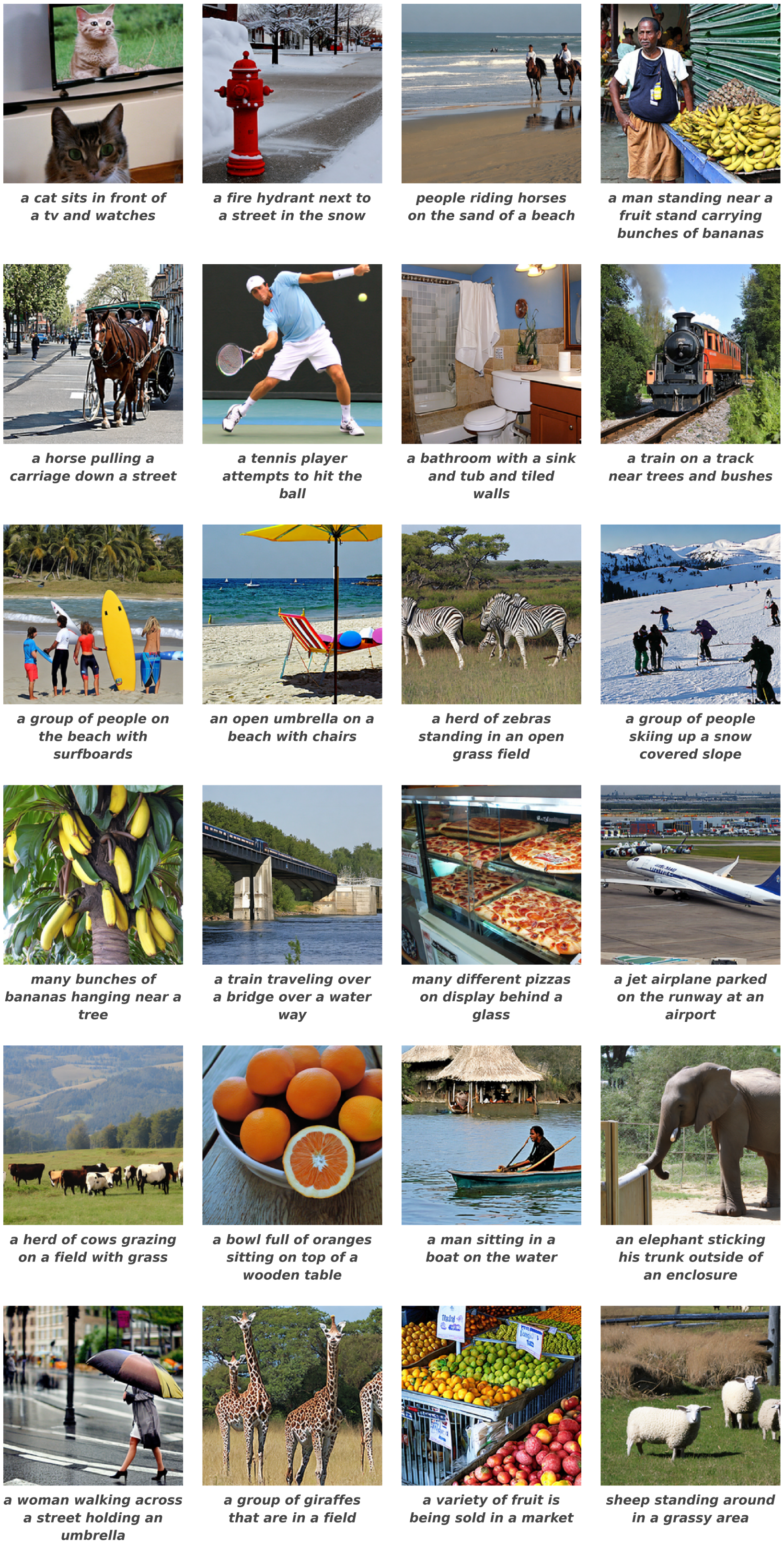
a living room with furniture and a television

a herd of elephants walking across a field

a docked boat on the water near other boats

a large truck on the side of a street

Figure 9: Examples of synthetic images and retrieved captions in the 100-pair setting on MS-COCO.



a cat sits in front of a tv and watches

a fire hydrant next to a street in the snow

people riding horses on the sand of a beach

a man standing near a fruit stand carrying bunches of bananas

a horse pulling a carriage down a street

a tennis player attempts to hit the ball

a bathroom with a sink and tub and tiled walls

a train on a track near trees and bushes

a group of people on the beach with surfboards

an open umbrella on a beach with chairs

a herd of zebras standing in an open grass field

a group of people skiing up a snow covered slope

many bunches of bananas hanging near a tree

a train traveling over a bridge over a water way

many different pizzas on display behind a glass

a jet airplane parked on the runway at an airport

a herd of cows grazing on a field with grass

a bowl full of oranges sitting on top of a wooden table

a man sitting in a boat on the water

an elephant sticking his trunk outside of an enclosure

a woman walking across a street holding an umbrella

a group of giraffes that are in a field

a variety of fruit is being sold in a market

sheep standing around in a grassy area

Figure 10: Examples of synthetic images and retrieved captions in the 300-pair setting on MS-COCO.