Efficient Multimodal Dataset Distillation via Generative Models

Zhenghao Zhao[1] **Haoxuan Wang**[1] **Junyi Wu**[1]
Yuzhang Shang[2] **Gaowen Liu**[3] **Yan Yan**[1]*
[1]University of Illinois Chicago [2]University of Central Florida [3]Cisco Research
{zzhao72, hwang339, jwu834, yyan55}@uic.edu
yuzhang.shang@ucf.edu, gaoliu@cisco.com

Abstract

Dataset distillation aims to synthesize a small dataset from a large dataset, enabling the model trained on it to perform well on the original dataset. With the blooming of large language models and multimodal large language models, the importance of multimodal datasets, particularly image-text datasets, has grown significantly. However, existing multimodal dataset distillation methods are constrained by the Matching Training Trajectories algorithm, which significantly increases the computing resource requirement, and takes days to process the distillation. In this work, we introduce EDGE, a generative distillation method for efficient multimodal dataset distillation. Specifically, we identify two key challenges of distilling multimodal datasets with generative models: 1) The lack of correlation between generated images and captions. 2) The lack of diversity among generated samples. To address the aforementioned issues, we propose a novel generative model training workflow with a bi-directional contrastive loss and a diversity loss. Furthermore, we propose a caption synthesis strategy to further improve text-to-image retrieval performance by introducing more text information. Our method is evaluated on Flickr30K, COCO, and CC3M datasets, demonstrating superior performance and efficiency compared to existing approaches. Notably, our method achieves results 18× faster than the state-of-the-art method. Our code will be made public at https://github.com/ichbill/EDGE.

1 Introduction

Dataset distillation [53] seeks to synthesize a compact dataset from a larger one, enabling a model trained on the distilled dataset to achieve strong performance on the original dataset. Typical kernel-based dataset distillation [53, 64] aims to match the performance of the synthetic dataset with the original dataset. The matching training trajectories (MTT) approaches [2, 19, 5, 9, 63, 12] have been proven effective for the distillation of small datasets. Some works [12, 63] even achieve a lossless performance from the original dataset. Recently, aiming the large-scale dataset, decoupled dataset distillation methods [45, 57], and generative distillation methods [44, 11, 49] have been proposed, which enhances scalability by improving the distillation efficiency.

With the emergence of large language models (LLM) and multimodal large language models (MLLM) [25, 23, 24, 16, 38], multimodal datasets become essential, especially image-text datasets. The first work on Vision Language Dataset Distillation (VLDD) [55] matches the expert training trajectories for both the image and text encoders and applies Low-Rank Adaptation [14] to reduce the computing requirements. The following work, LoRS [56], utilizes similarity mining during the distillation to improve the performance. However, these MTT-based methods require substantial time

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and computational resources to distill datasets. As illustrated in Figure 1, it takes about four days for MTT-VL to distill a dataset. The memory peak usage of MTT-VL reaches 200 GB for distilling 500 pairs and 320 GB for 100 pairs. For LORS [56], it takes about a week (150 hours) to distill only 500 image-text pairs. Therefore, time consumption becomes one of the main drawbacks of current VLDD methods.

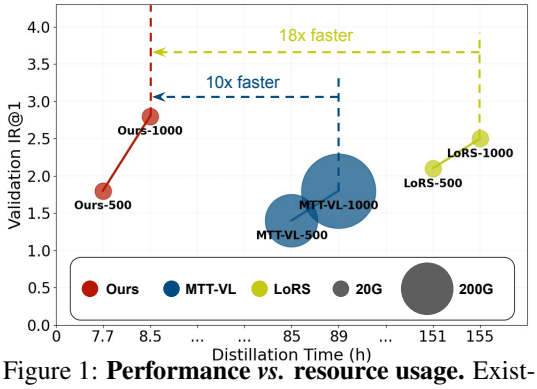

Figure 1: **Performance *vs.* resource usage.** Existing methods require substantial computation time and memory to distill datasets, whereas our method achieves up to 18× faster processing and 16× lower memory usage while delivering competitive performance. The experiments are conducted on NVIDIA RTX A5000 GPUs.

Generative dataset distillation methods [11, 44] leverage large-scale pre-trained diffusion models [34] to distill image datasets for image classification tasks. However, such single-modality generative distillation methods are unsuitable for multimodal scenarios because they are designed for a limited number of classes and cannot be generalized to image-caption datasets where the captions are not constrained by a fixed number of categories. Additionally, directly applying multimodal generative models [34] to VLDD will encounter two main issues: 1) The lack of correlation between generated images and captions. We observe that directly applying generative models to distillation yields a suboptimal performance. This is mainly because the diffusion model training concentrates on sample-wise noise prediction instead of image-text correspondence, which is the most important aspect of the image-text contrastive learning (ITC) task on vision language datasets. 2) The lack of diversity among the generated samples. In dataset distillation, the synthetic dataset only contains less than 5% of the original data, so the generalization of the dataset is essential for the distilled dataset generation.

To address the aforementioned issues, in this work, we introduce **E**fficient Multimodal Dataset **D**istillation via **GE**nerative Models (EDGE), which utilizes generative priors for distilling vision-language datasets. Our approach incorporates a novel training workflow that integrates a bi-directional contrastive loss inspired by InfoNCE [31] and diversity loss inspired by Minimax [11] to improve image-text correlation and increase the diversity of distilled datasets. The EDGE framework begins by introducing noise to the image latent representation, predicting the noise with a conditioning embedding, and obtaining the denoised image latent. Using the denoised image latent and the corresponding text embedding, we apply the proposed losses as follows: 1) Contrastive loss supervises the generative model to produce content that is highly relevant to the given conditioning, aligning image and text representations effectively. 2) Diversity loss increases the variability of the generated samples by pushing apart the sample features, encouraging the synthetic dataset to reflect the same distribution as the original dataset. Additionally, to further improve performance in text-to-image retrieval tasks, we incorporate a caption synthesis strategy. This involves generating additional captions for images in the synthetic dataset using MLLMs, providing sufficient text information for the evaluation model training and boosting retrieval performance.

We evaluate our methods on multiple vision language datasets, including Flickr30K [33], COCO [22], and Conceptual Captions 3 Million (CC3M) [42]. Compared to existing VLDD methods, our method not only achieves more efficient distillation on Flickr30K and COCO, but also extends the capability to distill the large-scale CC3M dataset. As illustrated in Figure 1, our method significantly reduces the distillation time, achieving comparable performance to existing methods within just a few hours, compared to the days required by existing approaches. Furthermore, we successfully distill CC3M, a dataset containing approximately three million image-text pairs. Notably, most existing dataset condensation methods, including coreset selection, are infeasible for CC3M due to their extremely high computational resource requirements.

Our contributions are summarized as follows:

- We identify the efficiency issue of current MTT-based vision language dataset distillation methods and propose a generative vision language distillation method, EDGE, to efficiently distill the vision language datasets.

- We identify two key challenges in applying generative methods to VLDD, and introduce a novel training workflow incorporating contrastive loss and diversity loss. We further introduce caption synthesis to improve the text-to-image retrieval performance.
- We conduct empirical comparisons with existing methods, demonstrating competitive performance while reducing computational costs by 18 times compared to the SOTA approach.

2 Related works

2.1 Dataset Distillation

Dataset distillation (DD) [53] aims to compress a large dataset into a small synthetic dataset while preserving the performance of the models trained on it. Dataset distillation methods can be categorized into several classes: performance matching [53, 64, 27, 29, 28, 30, 3], parameter matching [61, 18, 15, 2, 5, 63, 12, 21, 50], distribution matching [60, 51, 39, 62, 36, 58, 7, 52, 6, 26, 20] and decoupled distillation methods [57, 40, 45, 41, 8, 59, 43, 48, 4]. Recently, generative distillation methods [11, 44, 49] synthesize images using generative models to efficiently distill large-scale datasets. Minimax Diffusion [11] introduces additional minimax criteria in the generative training to enhance the performance of the generated images of the diffusion model. D^4M [44] has explored the use of diffusion models to improve cross-architecture generalization and reduce computational overhead by constraining the consistency of the real and synthetic image spaces.

While generative distillation methods have proven successful for image classification tasks, existing single-modality approaches are not suitable for multimodal scenarios. These methods are designed for a fixed number of classes in both training and test datasets and cannot be generalized to image-caption datasets, where captions are not constrained to predefined categories.

2.2 Vision Language Dataset Distillation

While the majority of the dataset distillation community focuses on the image classification task in dataset distillation, recent works [55, 56] began to explore the new task of image-text contrastive learning (ITC) on vision language dataset. The first work of multimodal dataset distillation [55] matches both the image and text encoders and applies LoRA [14] to reduce the computing requirements. In particular, to overcome the challenge that the image-text contrastive learning dataset does not have discrete classes, this work first proposed jointly distilling image-text pairs in a contrastive manner. The following work, LoRS [56], utilizes similarity mining during distillation for performance improvement. Specifically, a similarity matrix for synthetic image-text pairs is calculated during data synthesis to explicitly describe the relevance of each image-text pair in the distilled dataset.

However, existing VLDD methods struggle with the distillation efficiency and the scale of the datasets, which is mainly due to the use of the Matching Training Trajectories (MTT) algorithm [2]. MTT obtains synthetic data through a bi-level optimization process, which involves an inner loop for model updates and an outer loop for synthetic data updates. Not only does the required large number of unrolled iterations during optimization cause immense computational costs, but generating expert trajectories itself is also extremely time-consuming. In this work, we propose EDGE to efficiently distill datasets via the generative method and scale up the large-scaled CC3M [42] dataset.

3 Methods

3.1 Problem Formulation

Given a large target dataset $\mathcal{T} = \{x_i, y_i\}_{i=1}^{|\mathcal{T}|}$, where x_i denotes the training data and y_i denotes the corresponding annotation, general dataset distillation method aims to synthesize a small dataset $\mathcal{S} = \{x_i, y_i\}_{i=1}^{|\mathcal{S}|}$, where $|\mathcal{S}| << \mathcal{T}$, so that the model trained on $\mathcal{S}$ has the optimal performance on the target dataset $\mathcal{T}$. Specifically for image-text contrastive learning (ITC), the target dataset consists of $|\mathcal{T}|$ image-text pairs. Note that in practice, in a dataset, an image can have multiple captions, and a caption can also have multiple corresponding images. We consider the size of the datasets as the number of image-text pairs since the number of image-text pairs determines the actual training time.

In Section 3.2, we begin by introducing the advantages of introducing diffusion model priors and highlighting the limitations of directly applying them to vision-language dataset distillation. Then we present the workflow of EDGE to overcome the limitations. Section 3.3 presents the proposed

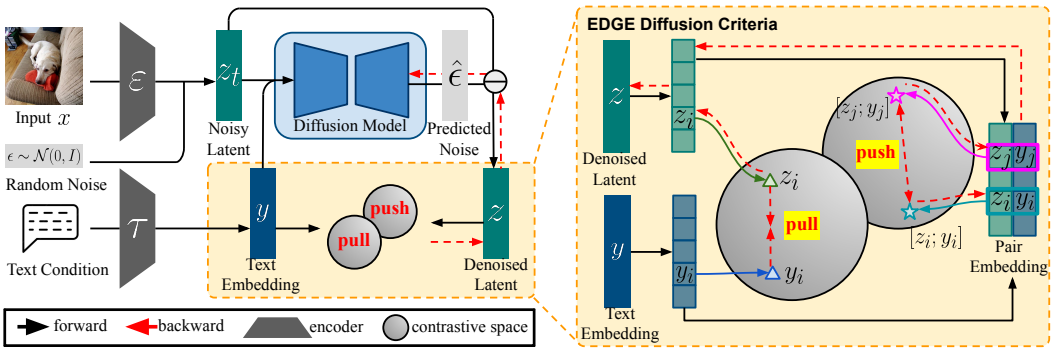

Figure 2: **The workflow of our method.** Given the input image x, text condition, we first get image latent $\mathbf{z}_0$ and text embedding $\mathbf{y}$ via the image encoder ε and text encoder τ_θ. Random noise $\epsilon \sim \mathcal{N}(0, I)$ is added to the image latent, and the noise is then predicted using a U-Net [35] conditioned on $\mathbf{y}$. The denoised latent representation $\mathbf{z}$ is obtained as $\mathbf{z} = (\mathbf{z}_t - \sqrt{\beta^t} \cdot \hat{\epsilon})/\sqrt{\alpha^t}$, where $\hat{\epsilon}$ is the predicted noise. With the denoised latent $\mathbf{z}$ and text embedding $\mathbf{y}$, we employ a contrastive loss $\mathcal{L}_C$ to align corresponding image latents and text embeddings and a diversity loss $\mathcal{L}_D$ to encourage the diversity and generalization between concatenated image-text features.

contrastive loss and diversity loss. Finally, in Section 3.4, to further improve the text-to-image retrieval performance, we illustrate the caption synthesis process.

3.2 EDGE Diffusion for VLDD

Diffusion models learn the distribution of a dataset by progressively adding Gaussian noise to images and then reversing this process to reconstruct the original data. For instance, in the latent diffusion model (LDM) [34], the training process is described as follows for a given training image x. In the forward noising process, a vision encoder ε encodes the image into the latent representation $\mathbf{z_0} = \varepsilon(x)$. Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is incrementally added to the initial latent code z_0, resulting in: $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\bar{\alpha}_t$ is a hyperparameter controlling the noise level at step t. Next, a decoder reconstructs this latent code back to the image space. With conditioning vector $\mathbf{y}$ encoding additional information, the diffusion model is trained to minimize the mean squared error between the predicted noise $\hat{\epsilon} = \epsilon_\theta(\mathbf{z}_t, \mathbf{y})$ and the added noise $\epsilon \sim \mathcal{N}(0, I)$:

$$\mathcal{L} = ||\epsilon_\theta(\mathbf{z}_t, \mathbf{y}) - \epsilon||_2^2, \tag{1}$$

where ϵ_θ is a noise prediction network parameterized by θ.

Diffusion models are proven efficient on dataset distillation methods for image classification tasks [11, 44]. However, when applying diffusion models to dataset distillation for ITC tasks, there are two main limitations to the distillation process: 1) The correspondence between image and text features is not fully explored. As mentioned above, the training of the diffusion model is sample-wise, which concentrates on noise prediction. Such a training scheme lacks the supervision of the correspondence of image and text features. 2) The lack of diversity among the synthetic samples. The number of samples is much smaller than the original datasets, so the diversity among the distilled samples is essential for dataset distillation.

To address the aforementioned issues, we propose the EDGE finetuning workflow as illustrated in Figure 2. Instead of calculating the mean squared error between the predicted noise $\epsilon_\theta(\mathbf{z}_t, \mathbf{c})$ and the ground truth ϵ, we proposed two novel criteria for the vision language dataset distillation. Given the input image x, text condition, the encoders ε and τ_θ transform them into a compact latent representation, image latent $\mathbf{z}_0$ and text embedding $\mathbf{y}$, respectively. Then we add random noise $\epsilon \sim \mathcal{N}(0, I)$ to the image latent and predict the noise with U-Net [35] with the conditioning $\mathbf{y}$. With predicted noise $\hat{\epsilon}$, we use

$$\mathbf{z} = \frac{\mathbf{z}_t - \sqrt{\beta^t} \cdot \hat{\epsilon}}{\sqrt{\alpha^t}} \tag{2}$$

to get the de-noised latent $\mathbf{z}$, where α^t and β^t are hyperparameters. The de-noised latent $\mathbf{z}$ and text embedding $\mathbf{y}$ are used to calculate the contrastive loss and the diversity loss, introduced in Section 3.3.

The contrastive loss is used to encourage the generative model to synthesize images that are highly relevant to the given text conditions, while the diversity loss is used to push away the image-text embeddings among different image-text pairs.

3.3 EDGE Diffusion Criteria

In this section, we begin by presenting the motivation behind the contrastive loss, followed by its design and impact. Next, we discuss the rationale for the diversity loss and provide its formulation. We then introduce the combined EDGE loss and conclude by explaining the process of generating the distilled dataset using our proposed method. **Contrastive Loss.** The goal of image-text contrastive learning (ITC) is to establish meaningful relationships between images and text. However, generative models, such as diffusion models, are typically trained to predict noise, without explicit supervision to capture image-text alignment. Instead of predicting the noise, we want to connect the latent space $\mathbf{z}_i$ and $\mathbf{y}_i$ for corresponding image x_i and text y_i. Inspired by InfoNCE [31], we propose the contrastive loss that encourages aligned embeddings. Specifically, with the given image embedding $\mathbf{z}$ and the text embedding $\mathbf{y}$, the similarity score between normalized features $\mathbf{z}_i$ and $\mathbf{y}_i$ can be defined as $S_{ii} = \frac{\mathbf{z}_i \cdot \mathbf{y}_i}{\tau}$, where τ is the temperature parameter, which is set to 0.5 in experiments. Then we calculate the cross-entropy loss for both directions:

$$\mathcal{L}_{I\to T} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{N}\exp(S_{ij})}, \mathcal{L}_{T\to I} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{N}\exp(S_{ji})}, \tag{3}$$

where $S_{ij} = \frac{I_i \cdot T_j}{\tau}$ represents the similarity score between $\mathbf{z}_i$ and $\mathbf{y}_j$, and $S_{ji} = \frac{I_j \cdot T_i}{\tau}$ represents the similarity score between $\mathbf{z}_j$ and $\mathbf{y}_i$. With λ controlling the weight ratio, the overall contrastive loss is defined as:

$$\mathcal{L}_{\text{C}} = \lambda\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}. \tag{4}$$

In equation 3, the loss encourages alignment between each image and its corresponding caption while also ensuring that each text matches the correct image. By incorporating this bi-directional contrastive objective, the diffusion model learns to capture the semantic relevance between images and text. Consequently, during sampling, the model generates image-text pairs with high relevance, which is beneficial for evaluation model training.

Diversity Loss. Compared to the original dataset, the distilled dataset distilled by generative methods [11] has limited diversity. This problem is even more evident in image-text contrastive (ITC) tasks compared to image classification, since the captions are not constrained to predefined categories. Diffusion model training is in a sample-wise manner and neglects this requirement for dataset distillation. To address this issue and ensure the distilled dataset accurately represents the original distribution, we focus on increasing the diversity of generated samples and enhancing the generalization of models trained on the synthetic dataset. To achieve this, we introduce a diversity loss that leverages the de-noised embedding $\mathbf{z}$ and the text embedding $\mathbf{y}$:

$$\mathcal{L}_{\text{D}} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\frac{[z_i; y_i]}{\|[z_i; y_i]\|} \times \frac{[z_j; y_j]}{\|[z_j; y_j]\|}\right), \tag{5}$$

where $[z_i; y_i]$ represents the concatenation of the image and text embeddings z_i and y_i. The term $\left(\frac{[z_i;y_i]}{\|[z_i;y_i]\|} \times \frac{[z_j;y_j]}{\|[z_j;y_j]\|}\right)$ indicates the similarity matrix of normalized concatenated embeddings. The diversity loss is designed to maximize the distance between these concatenated embeddings, effectively pushing them away in the embedding space. This encourages diversity in the representations of image-text pairs, promoting improved generalization and robustness.

Finally, by incorporating the contrastive loss $\mathcal{L}_{\text{C}}$ and the diversity loss $\mathcal{L}_{\text{D}}$ together, we formulate the EDGE loss as:

$$\mathcal{L}_{\text{EDGE}} = \mathcal{L}_{\text{C}} + \lambda\mathcal{L}_{\text{D}}. \tag{6}$$

The EDGE loss is used to train the diffusion model on the target dataset, to get the corresponding generative model. To synthesize the distilled dataset, we directly use diffusion model sampling for the required number of images. We first randomly select captions from the original dataset, and then use each caption as a condition to sample the image.

3.4 Caption Synthesis

We observed that when training a model on the synthetic dataset, it is often more challenging for the model to retrieve the correct image with a given text than to retrieve the correct text for a given image. As discussed in Section 1, applying the generative-based dataset distillation method to ITC tasks will counter the issue that the captions are not sufficient for the model training. To address this issue, we propose caption synthesis to generate more captions for each image. We developed a scalable approach to create as many captions as we want. Our method involves crafting specific prompt engineering templates that guide the Multimodal Large Language Models (MLLM) [25, 23, 24] to produce the captions for given images.

We start by gathering the images from the synthetic dataset. For each image, we consider a straightforward prompt template to generate captions effectively. As illustrated in Figure 3, we can use MLLM such as LLaVA [25] or GPT [32] model to generate any number of captions for a given image. The prompts we pass to different MLLMs are a little bit different. For LLaVA, we directly pass the prompt "*Describe the image in one sentence*" and we can get the required caption. GPT models tend to generate long sequences of text, and it often contains the phase that is used for interactions, such as "*the image contains*". Thus we change the prompt to "*Describe the image briefly in one sentence. Do not start with 'the image.*'"

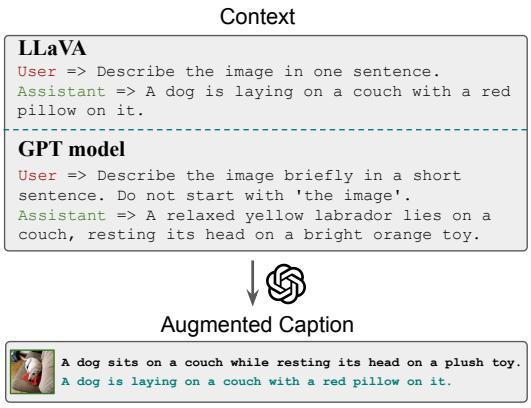

Figure 3: **Illustration of the caption synthesis.** An image and a prompt are passed to a Multimodal Large Language Model (MLLM) to generate captions. Then the dataset contains $|\mathcal{S}|/\eta$ images with $|\mathcal{S}|$ captions, where η is the caption per image (CPI) and each image is captioned by η captions. Caption synthesis aims to provide sufficient text information for the model training.

During the model training stage, even though our method produces fewer images than captions, we consider the same image with different caption pairs as distinct image-text pairs. This is due to the similar computational resources required from the model training perspective.

4 Experiments

4.1 Datasets and Metrics

Following previous methods [55, 56], we evaluate our method on Flickr30k [33] and COCO [22] datasets for a fair comparison with existing methods. Flickr30K and COCO are image-text datasets with 31K and 123K images, where each image is paired with five captions. We also evaluated methods on a larger image-caption dataset, Conceptual Captions 3 Million (CC3M) [42], to validate the effectiveness of our method on large-scale datasets. CC3M contains about 3.3 million image-text pairs. The evaluation involves the metrics of top-K retrieval from both the image and the text perspectives. We denote the text-to-image retrieval as IR@K and image-to-text retrieval as TR@K.

4.2 Implementation Details

Following previous works [55, 56] on Vision Language Dataset Distillation, for the evaluation model, we adopt NFNet [1] as image encoder and BERT-base [17] as text encoder. The pre-trained weights are used, and the text encoder is frozen during evaluation. Our entire distillation process can be done on a single NVIDIA RTX A5000 GPU. For evaluation on the CC3M dataset, since the dataset is constructed with URLs and by the time we conduct experiments, some URLs are no longer available. Eventually, we collected 2.3M of 3.3M images for the CC3M training data. In addition, since the validation set is large, we first subsample a 1000 image-text pair validation subset and then evaluate the validation subset. In the distillation stage, following previous works [56], the images are resized to 224×224 resolution, and text embeddings are of 768 dimension.

4.3 Main Results

We compare our method with coreset selection methods [54, 10, 46], existing VLDD methods [55, 56] and a pre-trained generative model, Stable-diffusion v1.5 [34] on Flickr30K and COCO datasets.

Table 1: **Results of 500 pairs on Flickr30K and COCO dataset.** For Flickr30K, the dataset condensation ratio is 1.7%. For COCO, the condensation ratio is 4.4‰. **Bold** values indicate the best results, while underlined values denote the second-best results.

Dataset	Ratio	Metric	Coreset Selection				Dataset Distillation				
			Rand	HERD	K-Cent	Forget	MTT-VL	TESLA	LoRS	SD	**EDGE**
Flickr30K	1.7%	IR@1	2.4	3.0	3.5	1.8	6.6	1.1	**10.0**	3.1	6.7
		IR@5	10.5	10.0	10.4	9.0	20.2	7.3	**28.9**	11.5	21.0
		IR@10	17.4	17.0	17.3	15.9	30.0	12.6	**41.6**	18.5	30.5
		TR@1	5.2	5.1	4.9	3.6	13.3	5.1	**15.5**	4.6	13.3
		TR@5	18.3	16.4	16.4	12.3	32.8	15.3	**39.8**	15.1	35.6
		TR@10	25.7	24.3	23.3	19.3	46.8	23.8	**53.7**	22.2	47.5
COCO	4.4‰	IR@1	1.1	1.7	1.1	0.8	1.4	0.8	**2.1**	1.2	1.8
		IR@5	5.0	5.3	6.3	5.8	5.1	3.6	**8.0**	5.1	6.5
		IR@10	8.7	9.9	10.5	8.2	8.9	6.9	**13.7**	9.0	11.2
		TR@1	1.9	1.9	2.5	2.1	2.3	1.7	2.8	2.2	**2.9**
		TR@5	7.5	7.8	8.7	8.2	8.4	5.9	**9.9**	8.7	9.5
		TR@10	12.5	13.7	14.3	13.0	13.8	10.2	**16.2**	14.9	15.7

Table 2: **Results of 1000 pairs on Flickr30K and COCO dataset. Bold** values indicate the best results, while underlined values denote the second-best results.

Dataset	Ratio	Metric	Rand	LoRS	SD	**EDGE**
Flickr30K	3.4%	IR@1	3.9	**11.2**	3.3	9.9
		IR@5	13.0	**31.3**	11.9	28.2
		IR@10	20.7	**42.9**	19.1	40.5
		TR@1	5.3	14.4	5.0	**14.5**
		TR@5	17.1	37.5	15.9	**38.3**
		TR@10	27.7	50.5	24.8	**51.7**
COCO	8.8‰	IR@1	1.6	2.5	1.8	**2.8**
		IR@5	6.5	9.4	6.9	**9.8**
		IR@10	11.5	15.4	11.9	**16.2**
		TR@1	2.4	3.6	2.8	**3.9**
		TR@5	9.9	11.8	10.2	**13.0**
		TR@10	16.7	19.1	16.8	**21.0**

Table 3: **Results on CC3M.** We compare our method with random selection and a pre-trained Stable Diffusion v1.5 model. The synthetic dataset contains 1000 image-text-pairs, which is only 0.3‰ of the original dataset. Other existing methods are not applicable due to long computation time or high memory requirements.

Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Rand	0.1	0.4	0.9	0.0	0.3	0.9
SD	0.1	0.4	0.9	0.0	0.3	0.9
EDGE	**0.2**	**0.5**	**1.0**	**0.1**	**0.7**	**1.1**

Table 4: **CLIP score.**

Methods	Flickr-30K	COCO	CC3M
MTT-VL	0.2966	0.2264	-
LoRS	0.0077	0.0105	-
Ours	**0.3227**	**0.3178**	**0.3086**

Table 5: **FID score.**

Methods	Flickr-30K	COCO
MTT-VL	210.0	276.3
Ours	**88.1**	**83.1**

Due to extremely long computing time, coreset selection methods [54, 10, 46] are hard to be applied on a large number of pairs. As for MTT-based DD methods [55, 5], distilling to a large number of pairs will lead to extremely high GPU memory usage.

Flickr30K results. We evaluate our method on the Flickr30K dataset and compare it with existing approaches. The distilled dataset contains 500 and 1000 image-text pairs, representing 1.7% and 3.4% of the original dataset, respectively. As shown in Table 1 and Table 2, our method achieves performance comparable to or even outperforms that of time-intensive SOTA methods across several metrics. Notably, compared to the baseline—a pre-trained Stable Diffusion v1.5 model—our approach demonstrates significant improvements, highlighting the effectiveness of the proposed techniques.

COCO results. We also evaluate our method on the COCO dataset in comparison with existing methods in Table 1 and Table 2. The distilled dataset contains 500 and 1000 image-text pairs, representing 4.4‰ and 8.8‰ of the original dataset, respectively. With 500 pairs, it achieves competitive performance relative to SOTA methods, which require a significantly longer distillation time. When distilling 1000 image-text pairs, our method shows an obvious advantage from both performance and efficiency perspectives. Similar to the Flickr30K results, our method substantially improves over the baseline pre-trained Stable Diffusion model, further validating its effectiveness.

CC3M results. For larger datasets like Conceptual Captions 3M (CC3M) [42], existing dataset distillation methods are not applicable due to the high resource demands of trajectory matching algorithms. Even coreset selection methods are difficult to apply, as their time consumption and memory usage scale rapidly. In contrast, our method is capable of distilling the CC3M dataset efficiently using only a single GPU. The distilled dataset contains 500 image-text pairs, representing 0.3‰ of the original dataset. We compare our approach with the random selection method and the pre-trained Stable Diffusion v1.5 model in Table 3. Our method outperforms the baseline methods, demonstrating the effectiveness of our method on CC3M dataset.

More aspects of evaluation. Furthermore, we provide the CLIP scores on Flickr30K, COCO, and CC3M, comparing with existing dataset distillation methods in Table 4. Table 5 assesses image

Table 6: **Comparison of required computing resources.** We provide the GPU hours required for different method to distill different datasets. "OOM" in the table indicates the method is not applicable due to extremely high memory usage.

Dataset	Flickr30K			COCO			CC3M		
# Pairs	500	1000	1500	500	1000	1500	500	1000	1500
MTT-VL [55]	77.8	109.1	141.9	162.3	194.1	OOM	OOM	OOM	OOM
LoRS [56]	54.2	57.6	59.0	151.6	155.6	172.5	OOM	OOM	OOM
EDGE (ours)	**13.6**	**14.3**	**14.9**	**7.7**	**8.5**	**9.2**	**31.7**	**32.5**	**33.2**

Table 7: **Cross architecture evaluation.** For baseline method, we use NFNet+BERT to distill, and evaluate with various architectures. Compared to matching training trajectory based approaches, the dataset generated by our method has a better generalization.

# Pairs	Ratio	Method	Evaluation model	COCO					
				IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
1000	8.8‰	LoRS	NfNet+Bert	2.5	9.4	15.4	3.6	11.8	19.1
			ResNet+Bert	0.1	0.6	1.1	0.4	1.6	2.8
			RegNet+Bert	0.1	0.5	0.9	0.1	0.5	0.9
1000	8.8‰	EDGE	NfNet+Bert	2.8	9.8	16.2	3.9	13.0	21.0
			ResNet+Bert	2.5	9.5	16.0	3.7	12.7	20.1
			RegNet+Bert	2.3	8.4	14.1	2.6	10.7	17.7

quality using FID. Results show our method improves both text-image alignment and image quality, where the former is more crucial for retrieval tasks.

Computational efficiency. We provide the computing efficiency comparison of our method and the SOTA methods [55, 56]. As depicted in Table 6, for Flickr30K, our method is approximately 9 times faster than MTT-VL, and 4 times faster than LoRS. For the COCO dataset, our method is 22 times faster than MTT-VL and 18 times faster than LoRS. Not even mentioning that both existing methods [55, 56] are not applicable for large dataset CC3M [42]. For CC3M, our method is the only applicable dataset distillation method, and our method is able to distill the dataset in an efficient manner. The GPU hour usage is even less than the baseline methods on the Flickr30K dataset.

Cross-architecture experiments. In this work, we also evaluate the cross-architecture generalization of our method. As shown in Table 7, we evaluate performance by changing the image encoder, and our method significantly outperforms the baseline. The results highlight that the cross-architecture performance of the baseline method [56] is highly dependent on the model used during distillation. For example, when synthetic datasets are distilled with NFNet and BERT, evaluation using the exact same model architectures yields strong performance. However, substituting NFNet with ResNet [13] or RegNet [37] causes a significant performance drop in both image and text retrieval tasks. In contrast, our method, which is not related to a specific model architecture during training, maintains consistent performance across various model architectures.

4.4 Ablation Study

Ablation on different components. Table 8 states the effectiveness of each component of our method. The experiments are performed on the COCO [22] dataset, distilling into 500 image-text pairs. The baseline "SD v1.5" in the table indicates the baseline model, a pretrained stable-diffusion model v1.5 [34] used in our method. "+ $\mathcal{L}_{\text{C}}$" and "+ $\mathcal{L}_{\text{D}}$" indicates the generative model fine-tuned with our proposed workflow in Figure 2 and corresponding losses.

The pre-trained stable diffusion model shows low performance on the evaluation dataset, indicating that the vanilla generative models, which are not designed for dataset distillation tasks, do not guarantee the correspondence for image-text pairs. By adding our proposed fine-tuning workflow and the corresponding losses, both IR and TR have been improved. By further adding the post-training caption synthesis, the performance of image-text retrieval is further improved, indicating the effectiveness of proposed methods.

Effect of different diffusion models. The diffusion model used to distill the dataset is an important aspect of our method. We evaluate two widely used diffusion models: Stable Diffusion v1.5 (denoted as "SD v1.5" in the table) and Stable Diffusion 3 (denoted as "SD 3") in this work. We compare two models from both perspectives of performance and efficiency. The performance of the methods is depicted in Table 9, where the performance of two diffusion models is similar. However, the sampling

Table 8: **Effect of different components.** The experiments are conducted on COCO dataset. The synthetic dataset contains 500 image-text pairs. The evaluation model is NFNet+BERT. “CS" denotes the Caption Synthesis strategy.

Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
SD v1.5	1.2	5.1	9.0	2.2	8.7	14.9
+ $\mathcal{L}_{\text{C}}$	1.3	5.1	9.2	2.3	8.3	13.7
+ $\mathcal{L}_{\text{C}}$ + $\mathcal{L}_{\text{D}}$	1.4	5.6	10.1	2.3	8.9	14.4
+ $\mathcal{L}_{\text{EDGE}}$ + CS	1.8	6.5	11.2	2.9	9.5	15.7

Table 9: **Comparison of different generative models.** SD represents Stable Diffusion. Stable Diffusion v1.5 takes about 0.4 hour to distill 500 images, while it takes approximately 1.6 hours to distill 500 images for Stable Diffusion 3.

Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
SD v1.5	1.2	5.1	9.0	2.2	8.7	14.9
SD 3	1.2	5.2	9.4	2.4	8.6	14.7

Table 10: **Ablation study on caption per image (CPI).** On COCO dataset, when CPI=2, the model achieves the best performance.

Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
CPI = 1	1.9	6.9	11.8	2.7	9.7	16.0
CPI = 2	**2.8**	**9.8**	**16.2**	**3.9**	**13.0**	**21.0**
CPI = 5	1.2	4.7	8.2	2.0	7.0	12.3

Figure 4: **Examples of images and captions of the distilled dataset.** For each example image, two corresponding captions are provided. The **green** captions are generated by caption synthesis.

Table 11: **Comparison of different MLLMs.** Post-training strategies perform better than pre-processing strategies. With captions by LLaVA, the model achieves the best performance.

Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
LLaVA	**2.8**	**9.8**	**16.2**	**3.9**	**13.0**	**21.0**
GPT-4o-mini	2.5	8.8	14.9	3.7	12.5	20.0
Llama	1.2	5.0	8.8	2.4	8.5	13.6

time of stable diffusion 3 is four times longer than that of Stable Diffusion v1.5. Consequently, we selected Stable Diffusion v1.5 as the primary model for our results.

Effect of caption per image. For caption synthesis, the number of captions per image (CPI) is essential. To investigate its impact on the model's performance, we vary the CPI during caption synthesis while keeping the total number of image-text pairs constant. The evaluation is done on COCO dataset, distilling the dataset into 500 image-text pairs. For CPI=1, CPI=2, CPI=5, we sample 500, 250, 100 images.

There are two main observations from Table 10: 1) Compared to the “vanilla" synthetic dataset, where each image is paired with only one caption, generating additional captions assists with the evaluation model training. This observation is consistent with the motivation of our caption synthesis, harm the performance of the evaluation model. This occurs because, with a fixed number of image-text pairs, increasing the CPI leads to the reduction of the total number of unique images. Consequently, the reduced number of images becomes insufficient for effective model training.

Effect of different MLLMs. The MLLM used for caption synthesis is also essential in our method. In this work, we compare three MLLM/LLM models with two different strategies. The first strategy, pre-processing, involves rephrasing captions before image generation. The second strategy, post-training caption synthesis, involves rephrasing captions after image generation, as illustrated in Section 3.4. The models evaluated in our study include LLava [25], GPT-4o-mini [32], and Llama [47]. For LLava, we adopt the post-training strategy, while for Llama, we use the pre-processing strategy. For the GPT model, both pre-processing and post-training strategies are applicable.

As illustrated in Table 11, the pre-processing strategy, regardless of the model used for rephrasing captions, proves to be less effective compared to post-training strategies. For the post-training methods, captions generated by the LLaVA model are more beneficial for training the evaluation model than those generated by the GPT model. Therefore, we adopt the post-training caption synthesis strategy with the LLaVA model as our primary approach.

4.5 Dataset Visualization

We visualized the images with corresponding captions generated by our method in Figure 4. Compared to previous VLDD methods [55, 56], the images generated by our method exhibit a realistic, high-quality appearance. The **green** captions are generated by our caption synthesis. The captions exhibited

in the figure are generated by LLaVA [25]. We can observe that compared to the image caption from the original dataset, the synthesized caption tends to capture more features from the image.

5 Conclusion

In this work, we identify key limitations in current vision language dataset distillation methods, including poor efficiency and limited scalability. To address these issues, we propose EDGE, a generative vision language dataset distillation approach designed to efficiently distill large-scale multimodal datasets. Specifically, we introduce a novel training workflow for generative models tailored to the image-text dataset distillation task. Additionally, we implement a post-training caption synthesis step to further enhance performance. Our method not only outperforms existing approaches on small datasets but also enables effective distillation of large datasets.

Acknowledgments: This research is supported by NSF IIS-2525840, CNS-2432534, ECCS-2514574, NIH 1RF1MH133764-01 and Cisco Research unrestricted gift. This article solely reflects opinions and conclusions of authors and not funding agencies.

References

[1] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *ICML*. PMLR, 2021.

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, 2022.

[3] Yilan Chen, Wei Huang, and Lily Weng. Provable and efficient dataset distillation for kernel ridge regression. In *NeurIPS*, 2024.

[4] Jiacheng Cui, Xinyue Bi, Yaxin Luo, Xiaohan Zhao, Jiacheng Liu, and Zhiqiang Shen. Fadrm: Fast and accurate data residual matching for dataset distillation. *arXiv preprint arXiv:2506.24125*, 2025.

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *ICML*, 2023.

[6] Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In *CVPR*, 2025.

[7] Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. Exploiting inter-sample and inter-feature relations in dataset distillation. In *CVPR*, 2024.

[8] Jiawei Du, Juncheng Hu, Wenxin Huang, Joey Tianyi Zhou, et al. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *NeurIPS*, 2024.

[9] Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *CVPR*, 2023.

[10] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media, 2009.

[11] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *CVPR*, 2024.

[12] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[15] Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z Pan. Delving into effective gradient matching for dataset condensation. In *COINS*, 2023.

[16] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. *arXiv preprint arXiv:2410.00255*, 2024.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*. Minneapolis, Minnesota, 2019.

[18] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *ICML*, 2022.

[19] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation using parameter pruning. *IEICE Transactions*, 2023.

[20] Wenyuan Li, Guang Li, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Hyperbolic dataset distillation. *arXiv preprint arXiv:2505.24623*, 2025.

[21] Zekai Li, Ziyao Guo, Wangbo Zhao, Tianle Zhang, Zhi-Qi Cheng, Samir Khaki, Kaipeng Zhang, Ahmad Sajedi, Konstantinos N Plataniotis, Kai Wang, et al. Prioritize alignment in dataset distillation. *arXiv preprint arXiv:2408.03360*, 2024.

[22] 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 *ECCV*. Springer, 2014.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[26] Haoyang Liu, Yijiang Li, Tiancheng Xing, Vibhu Dalal, Luwei Li, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. *arXiv preprint arXiv:2311.18531*, 2023.

[27] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *NeurIPS*, 2022.

[28] Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. In *ICML*. PMLR, 2023.

[29] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *arXiv preprint arXiv:2011.00050*, 2020.

[30] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *NeurIPS*, 2021.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] OpenAI. Chatgpt. `https://chat.openai.com`, 2024. [Online; accessed 2024-07-18].

[33] 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 *ICCV*, 2015.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[36] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *ICCV*, 2023.

[37] Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. Regnet: Multimodal sensor registration using deep neural networks. In *IV*. IEEE, 2017.

[38] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[39] Yuzhang Shang, Zhihang Yuan, and Yan Yan. Mim4dd: Mutual information maximization for dataset distillation. In *NeurIPS*, 2024.

[40] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *CVPR*, 2024.

[41] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *NeurIPS*, 2024.

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[43] Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. In *CVPR*, 2025.

[44] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *CVPR*, 2024.

[45] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *CVPR*, 2024.

[46] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[48] Minh-Tuan Tran, Trung Le, Xuan-May Le, Thanh-Toan Do, and Dinh Phung. Enhancing dataset distillation via non-critical region refinement. In *CVPR*, 2025.

[49] Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Cao _2: Rectifying inconsistencies in diffusion-based dataset distillation. *arXiv preprint arXiv:2506.22637*, 2025.

[50] Kai Wang, Zekai Li, Zhi-Qi Cheng, Samir Khaki, Ahmad Sajedi, Ramakrishna Vedantam, Konstantinos N Plataniotis, Alexander Hauptmann, and Yang You. Emphasizing discriminative features for dataset distillation in complex scenarios. In *CVPR*, 2025.

[51] 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 *CVPR*, 2022.

[52] Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *CVPR*, 2025.

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

[54] Max Welling. Herding dynamical weights to learn. In *ICML*, 2009.

[55] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *arXiv preprint arXiv:2308.07545*, 2023.

[56] Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. *arXiv preprint arXiv:2406.03793*, 2024.

[57] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *NeurIPS*, 2024.

[58] Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *AAAI*, 2024.

[59] Xin Zhang, Jiawei Du, Ping Liu, and Joey Tianyi Zhou. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. *arXiv preprint arXiv:2408.06927*, 2024.

[60] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, 2023.

[61] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

[62] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *CVPR*, 2023.

[63] Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. *arXiv preprint arXiv:2408.14506*, 2024.

[64] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *NeurIPS*, 2022.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the contributions of the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The are fully discussed.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The code will be released publicly.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: They are discussed in details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive for baselines.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: It is properly discussed.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See Appendix C.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All external assets such as datasets and code libraries used in the research are properly credited, and their licenses are respected and clearly mentioned.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: The usage of LLMs is detailed in Section 4.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Training details

We provide the detailed hyperparameter settings used during diffusion model training. Following the approach in [11], we fine-tune only a small subset of the diffusion model's parameters. The weight λ in Equation 6 is set to 1. For Flickr30K dataset, the learning rate is set to 1×10^{-4}, with a batch size of 8 and a total of 16 training epochs. For COCO dataset, the learning rate is set to 1×10^{-4}, with a batch size of 8 and a total of 8 training epochs.

B Qualitative results

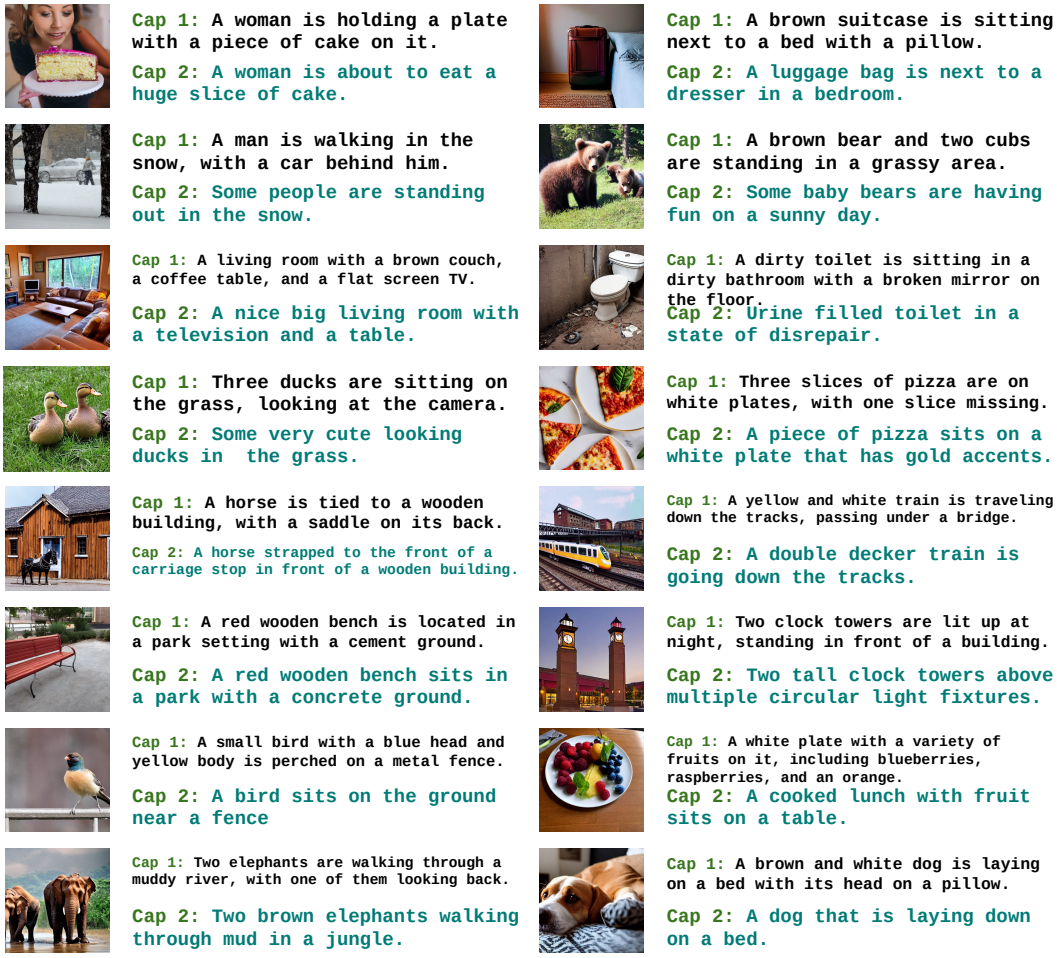

Figure 5: More qualitative results of the distilled dataset.

C Broader Impacts and Limitations

The societal impacts of our work are multifaceted. Positively, it contributes to greater resource efficiency, accelerates advancements in artificial intelligence, improves accessibility, and can enhance data privacy. However, it also introduces potential drawbacks, such as the possibility of information loss, the introduction of security vulnerabilities, and the complication of intellectual property rights. A careful balance between these advantages and challenges is paramount to ensure the responsible and ethical application of this technology.

We recognize that the current approach may exhibit limitations when applied to highly specialized domains such as medical imaging, where the data distribution, structural characteristics, and annotation

protocols differ substantially from those of general-purpose datasets used in our experiments. These domain-specific complexities may not be fully captured by the existing framework. Future research could address this limitation by incorporating domain-informed priors, adapting the optimization objectives, or designing customized representations that better align with the underlying structure and semantics of such specialized data.