

A More Related Works

Dataset distillation (DD), first proposed by Wang et al. [50], aims to improve training efficiency by condensing information from large-scale datasets into a small set of synthetic samples. Building on this foundation, recent advancements have introduced a wide range of techniques for effectively and efficiently compressing representative knowledge into compact datasets. Depending on the underlying distillation objective, existing DD methods can be broadly categorized into gradient matching [61, 59, 24, 44], trajectory matching [4, 7, 13, 14], and distribution matching [48, 60, 39, 46, 9, 57, 56]. Among these, trajectory matching approaches demonstrate competitive performance without relying on additional label augmentation, making them particularly effective and efficient for practical distillation tasks.

While early efforts have predominantly focused on image data, recent works have extended DD to other domains such as text [30, 32], video [11, 51], and graph data [29, 58]. For example, DiLM [32] leverages a generative language model to produce textual synthetic data, enabling model-agnostic distillation with strong generalization. Wang et al. [51] address the underexplored challenge of temporal compression in videos by disentangling spatial and temporal information. In the graph domain, GDEM [29] aligns the eigenbasis and node features of real and synthetic graphs, achieving efficient and architecture-agnostic distillation without relying on GNN-specific supervision. These promising achievements naturally motivate exploration into multimodal scenarios. MTT-VL [53] is the first attempt in this direction, adapting trajectory matching for image-text datasets and demonstrating the feasibility of distilling multimodal information. Building upon this, LoRS [55] further investigates the unique challenge in multimodal dataset distillation (MDD), i.e., high representational variance, and proposes to construct a similarity matrix to mine associations between all matched and mismatched pairs more effectively. Despite these advances, existing methods remain focused on data structures, overlooking the fundamental impact of contrastive objectives in multimodal optimization, which can lead to modality collapse. In this paper, we propose an effective and efficient MDD framework that explicitly addresses this issue.

B Derivation of Equation 3

As defined in Equation 2,

$$\mathcal{L}_{\text{wBCE}}^{\mathcal{B}} = \sum_{i,j} w_{ij} \cdot \ell(\tilde{\mathbf{y}}_{ij}, \sigma(\hat{\mathbf{y}}_{ij}/\gamma)), \quad w_{ij} = \frac{\mathbb{I}[\tilde{\mathbf{y}}_{ij} > \beta]}{|\{(i,j) : \tilde{\mathbf{y}}_{ij} > \beta\}|} + \frac{\mathbb{I}[\tilde{\mathbf{y}}_{ij} \leq \beta]}{|\{(i,j) : \tilde{\mathbf{y}}_{ij} \leq \beta\}|},$$

where $\sigma(x)$ is the sigmoid function and $\ell(y, p) = -y \log(p) - (1-y) \log(1-p)$ is the binary cross-entropy loss. Thus, we have:

$$\begin{aligned} \ell(y, \sigma(x)) &= -y \log \frac{1}{1+e^{-x}} - (1-y) \log \frac{e^{-x}}{1+e^{-x}} \\ &= y \log(1+e^{-x}) + (1-y)x + (1-y) \log(1+e^{-x}) = \log(1+e^{-x}) + (1-y)x, \end{aligned}$$

whose derivative with respect to x is:

$$\frac{\partial \ell(y, \sigma(x))}{\partial x} = \frac{-e^{-x}}{1+e^{-x}} + (1-y) = \sigma(x) - y.$$

Given $\hat{\mathbf{y}}_{ij} = \tilde{\mathbf{x}}_i^{\top} \tilde{\boldsymbol{\tau}}_j'$, the overall gradient of wBCE is:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}_n'} = \sum_{j=1}^{|\mathcal{B}|} w_{nj} \frac{\partial}{\partial \tilde{\mathbf{x}}_n'} \ell(\tilde{\mathbf{y}}_{nj}, \sigma(\tilde{\mathbf{x}}_n^{\top} \tilde{\boldsymbol{\tau}}_j'/\gamma)) = \sum_{j=1}^{|\mathcal{B}|} \frac{w_{nj}}{\gamma} (\sigma(\hat{\mathbf{y}}_{nj}/\gamma) - \tilde{\mathbf{y}}_{nj}) \tilde{\boldsymbol{\tau}}_j'.$$

Similarly,

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}_m'} = \sum_{j=1}^{|\mathcal{B}|} w_{mj} \frac{\partial}{\partial \tilde{\mathbf{x}}_m'} \ell(\tilde{\mathbf{y}}_{mj}, \sigma(\tilde{\mathbf{x}}_m^{\top} \tilde{\boldsymbol{\tau}}_j'/\gamma)) = \sum_{j=1}^{|\mathcal{B}|} \frac{w_{mj}}{\gamma} (\sigma(\hat{\mathbf{y}}_{mj}/\gamma) - \tilde{\mathbf{y}}_{mj}) \tilde{\boldsymbol{\tau}}_j'.$$

Thus,

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_n} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_m} = \sum_{i,j=1}^{|\mathcal{B}|} \frac{w_{ni}w_{mj}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mj}/\gamma) - \tilde{\mathbf{y}}_{mj}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_j,$$

which can be rewritten as:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_n} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_m} &= \sum_{i,j \neq n,m}^{|\mathcal{B}|} \frac{w_{ni}w_{mj}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mj}/\gamma) - \tilde{\mathbf{y}}_{mj}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_j \\ &+ \sum_{i \neq n, m; j=n}^{|\mathcal{B}|} \frac{w_{ni}w_{mn}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mn}/\gamma) - \tilde{\mathbf{y}}_{mn}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_n \\ &+ \sum_{i \neq n, m; j=m}^{|\mathcal{B}|} \frac{w_{ni}w_{mm}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mm}/\gamma) - \tilde{\mathbf{y}}_{mm}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_m \\ &+ \sum_{i=n; j \neq n, m}^{|\mathcal{B}|} \frac{w_{nn}w_{mj}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nn}/\gamma) - \tilde{\mathbf{y}}_{nn})(\sigma(\hat{\mathbf{y}}_{mj}/\gamma) - \tilde{\mathbf{y}}_{mj}) \tilde{\boldsymbol{\tau}}'_n{}^\top \tilde{\boldsymbol{\tau}}'_j \\ &+ \sum_{i=m; j \neq n, m}^{|\mathcal{B}|} \frac{w_{nm}w_{mj}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nm}/\gamma) - \tilde{\mathbf{y}}_{nm})(\sigma(\hat{\mathbf{y}}_{mj}/\gamma) - \tilde{\mathbf{y}}_{mj}) \tilde{\boldsymbol{\tau}}'_m{}^\top \tilde{\boldsymbol{\tau}}'_j \\ &+ \frac{w_{nn}w_{mn}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nn}/\gamma) - \tilde{\mathbf{y}}_{nn})(\sigma(\hat{\mathbf{y}}_{mn}/\gamma) - \tilde{\mathbf{y}}_{mn}) \tilde{\boldsymbol{\tau}}'_n{}^\top \tilde{\boldsymbol{\tau}}'_n \\ &+ \frac{w_{nn}w_{mm}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nn}/\gamma) - \tilde{\mathbf{y}}_{nn})(\sigma(\hat{\mathbf{y}}_{mm}/\gamma) - \tilde{\mathbf{y}}_{mm}) \tilde{\boldsymbol{\tau}}'_n{}^\top \tilde{\boldsymbol{\tau}}'_m \\ &+ \frac{w_{nm}w_{mn}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nm}/\gamma) - \tilde{\mathbf{y}}_{nm})(\sigma(\hat{\mathbf{y}}_{mn}/\gamma) - \tilde{\mathbf{y}}_{mn}) \tilde{\boldsymbol{\tau}}'_m{}^\top \tilde{\boldsymbol{\tau}}'_n \\ &+ \frac{w_{nm}w_{mm}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nm}/\gamma) - \tilde{\mathbf{y}}_{nm})(\sigma(\hat{\mathbf{y}}_{mm}/\gamma) - \tilde{\mathbf{y}}_{mm}) \tilde{\boldsymbol{\tau}}'_m{}^\top \tilde{\boldsymbol{\tau}}'_m. \end{aligned}$$

In high-dimensional embedding spaces, both intra-modal and inter-modal negative pairs tend to be mutually orthogonal. Specifically, for any negative pair (i, j) , where $i \neq j$,

$$\tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_j \approx 0.$$

In our case, all pairs beyond $(i, j) \in \{(m, n), (n, m), (i, i)\}$ are negatives, thus we have,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_n} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_m} &\approx \sum_{i=1}^{|\mathcal{B}|} \frac{w_{ni}w_{mi}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mi}/\gamma) - \tilde{\mathbf{y}}_{mi}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_i \\ &+ \frac{w_{nn}w_{mm}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nn}/\gamma) - \tilde{\mathbf{y}}_{nn})(\sigma(\hat{\mathbf{y}}_{mm}/\gamma) - \tilde{\mathbf{y}}_{mm}) \tilde{\boldsymbol{\tau}}'_n{}^\top \tilde{\boldsymbol{\tau}}'_m \\ &+ \frac{w_{nm}w_{mn}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nm}/\gamma) - \tilde{\mathbf{y}}_{nm})(\sigma(\hat{\mathbf{y}}_{mn}/\gamma) - \tilde{\mathbf{y}}_{mn}) \tilde{\boldsymbol{\tau}}'_m{}^\top \tilde{\boldsymbol{\tau}}'_n. \end{aligned}$$

Because (n, n) and (m, m) are strictly aligned pairs, we have $\sigma(\hat{\mathbf{y}}_{nn}/\gamma) \approx \tilde{\mathbf{y}}_{nn} \approx 1$ and $\sigma(\hat{\mathbf{y}}_{mm}/\gamma) \approx \tilde{\mathbf{y}}_{mm} \approx 1$, hence $\sigma(\hat{\mathbf{y}}_{nn}/\gamma) - \tilde{\mathbf{y}}_{nn}$ and $\sigma(\hat{\mathbf{y}}_{mm}/\gamma) - \tilde{\mathbf{y}}_{mm}$ are close to zero. Therefore, we have:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_n} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{x}}'_m} &\approx \sum_{i \neq n, m}^{|\mathcal{B}|} \frac{w_{ni}w_{mi}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{ni}/\gamma) - \tilde{\mathbf{y}}_{ni})(\sigma(\hat{\mathbf{y}}_{mi}/\gamma) - \tilde{\mathbf{y}}_{mi}) \tilde{\boldsymbol{\tau}}'_i{}^\top \tilde{\boldsymbol{\tau}}'_i \\ &+ \frac{w_{nm}w_{mn}}{\gamma^2} (\sigma(\hat{\mathbf{y}}_{nm}/\gamma) - \tilde{\mathbf{y}}_{nm})(\sigma(\hat{\mathbf{y}}_{mn}/\gamma) - \tilde{\mathbf{y}}_{mn}) \tilde{\boldsymbol{\tau}}'_m{}^\top \tilde{\boldsymbol{\tau}}'_n. \end{aligned}$$

The first term captures the aggregated influence of shared negative examples on both $\tilde{\mathbf{x}}'_n$ and $\tilde{\mathbf{x}}'_m$, which affect them similarly and thus contribute little to their relative update direction. In contrast, the second term reflects their mutual interaction and plays a dominant role in determining their representational divergence or alignment.

C Calculation of Concentration Ratio (CR)

To compute the *concentration ratio* (CR), we use the surface area of a hyperspherical cap on the unit $(d-1)$ -sphere, where d is the dimensionality of the embedding space. Given a normalized cosine similarity value $c \in [0, 1]$, we consider the set of all unit vectors that form this similarity with a fixed reference direction. These vectors define a hyperspherical cap, a region on the surface of the unit hypersphere bounded by a fixed similarity threshold. The surface area ratio of this cap is given by:

$$A = \mathcal{I}_{1-c^2} \left(\frac{d-1}{2}, \frac{1}{2} \right).$$

Here, $\mathcal{I}_x(a, b)$ denotes the regularized incomplete Beta function, defined as:

$$\mathcal{I}_x(a, b) = \frac{\int_0^x t^{a-1} (1-t)^{b-1} dt}{\int_0^1 t^{a-1} (1-t)^{b-1} dt}.$$

This function describes the cumulative distribution of the Beta distribution and is widely used in geometric probability. In our context, it measures the proportion of the unit hypersphere’s surface that lies within a given angular range, equivalently, within a given cosine similarity of a fixed direction. Specifically, when computing hyperspherical cap areas, the variable substitution $x = 1 - c^2$ arises naturally from the spherical-to-cartesian coordinate transformation.

We then define the concentration ratio as the complement of this surface ratio:

$$\text{CR} = 1 - A.$$

This value reflects the proportion of the hypersphere surface that lies outside the similarity-defined cone. A higher CR indicates that the given similarity corresponds to a narrower directional region on the hypersphere, implying stronger feature concentration in the high-dimensional embedding space.

In implementation, we compute this value using the `scipy.special.betainc` function in Python.

D Implementation of Representation Blending

Algorithm 2 RepBlend:Representation Blending

Require: image and text representation $\{f^{\text{imgE}}(\tilde{\mathbf{x}}_b), \tilde{\tau}_b\}_{b=1}^B$ of one batch, Parameter α for MixUP

```

1: function REPBLEND( $\{f^{\text{imgE}}(\tilde{\mathbf{x}}_b), \tilde{\tau}_b\}_{b=1}^B, \alpha$ )
2:    $\{f^{\text{imgE}}(\tilde{\mathbf{x}}_b)^{\text{shuf}}, \tilde{\tau}_b^{\text{shuf}}\}_{b=1}^B \leftarrow \text{shuffle}(\{f^{\text{imgE}}(\tilde{\mathbf{x}}_b), \tilde{\tau}_b\}_{b=1}^B)$ 
3:   ▷ Shuffle image and text representations in one batch
4:   Sample  $\lambda$  from Beta( $\alpha, \alpha$ ) for the batch
5:   for  $b = 1$  to  $|B|$  do
6:     ▷ Linear interpolation in representation space
7:      $f^{\text{imgE}}(\tilde{\mathbf{x}}_b) \leftarrow \lambda f^{\text{imgE}}(\tilde{\mathbf{x}}_b) + (1 - \lambda) f^{\text{imgE}}(\tilde{\mathbf{x}}_b)^{\text{shuf}}$ 
8:      $\tilde{\tau}_b \leftarrow \lambda \tilde{\tau}_b + (1 - \lambda) \tilde{\tau}_b^{\text{shuf}}$ 
9:   end for return  $\{f^{\text{imgE}}(\tilde{\mathbf{x}}_b), \tilde{\tau}_b\}_{b=1}^B$ 
10: end function

```

E Comparison Methods

Coreset Selection Methods.

1) Random (Rand): Randomly selects a subset of samples from the full dataset to form a coreset. While this approach is unbiased, it may fail to capture the most informative or representative instances necessary for efficient training.

2) Herding (Herd) [52]: Selects samples based on herding dynamics to approximate the mean of the data distribution. It iteratively chooses instances that minimize the discrepancy between the coreset and the full dataset’s feature distribution.

3) K-Center (K-Cent) [16]: Selects samples that serve as representative centers in the feature space. It aims to maximize coverage by iteratively choosing points that are maximally distant from the already selected ones.

4) Forgetting (Forget) [47]: Selects samples based on how often they are forgotten during training, i.e., when correct predictions become incorrect. Samples with low forgetting counts are removed first, prioritizing the retention of harder and more informative examples.

Dataset Distillation Methods.

1) MTT-VL [53]: The first MDD approach that extends the trajectory matching framework MTT [4] to vision-language data, enabling dataset distillation in multimodal settings.

2) TESLA-VL [55]: An efficient variant of the MTT framework, TESLA [7], implemented in LoRS [55] as an ablation to evaluate the effectiveness of similarity mining in multimodal distillation.

3) LoRS [55]: A sota MDD method that distills both image-text pairs and their similarity matrix to enhance multimodal distillation, while leveraging low-rank factorization for improving efficiency.

F Hyperparameter Settings

The hyperparameter settings, summarized in Table 7 and Table 8, follow the configurations used in LoRS [55] to ensure fair and consistent comparisons.

Table 7: Hyperparameter settings for buffer.

	Flickr-30K	MS-COCO
epoch	10	10
num_experts	20	20
batch_size	128	128
lr_teacher_img	0.1	0.1
lr_teacher_txt	0.1	0.1
image_size	224×224	224×224

Table 8: Hyperparameter settings for distillation.

	Flickr-30K			MS-COCO		
	100 pairs	200 pairs	500 pairs	100 pairs	200 pairs	500 pairs
syn_steps	8	8	8	8	8	8
expert_epochs	1	1	1	1	1	1
max_start_epoch	2	2	3	2	2	2
iteration	2000	2000	2000	2000	2000	2000
lr_img	100	1000	1000	1000	1000	5000
lr_txt	100	1000	1000	1000	1000	5000
lr_lr	1e-2	1e-2	1e-2	1e-2	1e-2	1e-2
lr_teacher_img	0.1	0.1	0.1	0.1	0.1	0.1
lr_teacher_txt	0.1	0.1	0.1	0.1	0.1	0.1
lr_sim	10.0	10.0	100.0	5.0	50.0	500.0
sim_type	lowrank	lowrank	lowrank	lowrank	lowrank	lowrank
sim_rank	10	5	20	10	20	40
sim_alpha	3.0	1.0	0.01	1.0	1.0	1.0
num_queries	99	199	499	99	199	499
mini_batch_size	20	20	40	20	20	30
loss_type	WBCE	WBCE	WBCE	WBCE	WBCE	WBCE
beta_distribution	$\alpha = 1.0$	$\alpha = 1.0$	$\alpha = 1.0$	$\alpha = 1.0$	$\alpha = 1.0$	$\alpha = 1.0$

G Scaling to Larger Distillation Budget Setting

We extended the comparison to larger pair settings and provide a detailed table since figures are not allowed in the rebuttal. Both LoRS and our method show only marginal gains as pairs increase, reflecting the well-known saturation issue in dataset distillation where added pairs bring limited diversity. Nevertheless, our method maintains consistent superiority over LoRS under the same conditions. In the future, we will further explore ways to overcome this saturation toward lossless or nearly lossless MDD.

Table 9: Performance comparison on Flickr-30k using NFNet + BERT. The model trained on the full dataset performs: IR@1=23.16, IR@5=53.98, IR@10=66.62; TR@1=33.8, TR@5=65.7, TR@10=76.9.

Pairs	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	LoRS	8.3	24.1	35.1	11.8	35.8	49.2
	Ours	11.5	32.0	44.5	16.2	41.7	55.5
200	LoRS	8.6	25.3	36.6	14.5	38.7	53.4
	Ours	12.7	34.7	47.6	18.6	46.0	60.0
300	LoRS	9.6	27.7	39.3	13.8	39.3	53.5
	Ours	14.3	36.1	50.6	20.6	49.2	62.3
500	LoRS	10.0	28.9	41.6	15.5	39.8	53.7
	Ours	17.0	42.5	55.9	22.5	53.2	66.7
700	LoRS	11.6	31.5	44.1	16.3	41.9	57.0
	Ours	17.4	43.5	56.8	22.8	53.4	67.1
800	LoRS	11.5	31.5	44.1	15.4	40.6	56.8
	Ours	17.6	44.0	56.9	22.2	53.7	67.0
900	LoRS	11.3	31.0	43.9	15.6	40.9	56.9
	Ours	17.1	44.2	57.1	22.0	54.0	67.0

Table 10: Performance comparison on MS-COCO using NFNet + BERT. The model trained on the full dataset performs: IR@1=14.6, IR@5=38.9, IR@10=53.2; TR@1=20.6, TR@5=46.8, TR@10=61.3.

Pairs	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	LoRS	1.8	7.1	12.2	3.3	12.2	19.6
	Ours	4.1	13.9	22.3	5.2	17.9	28.0
200	LoRS	2.4	9.3	15.5	4.3	14.2	22.6
	Ours	6.1	19.3	29.8	6.9	21.8	32.3
300	LoRS	2.5	9.5	15.9	4.4	15.8	25.1
	Ours	6.6	20.5	29.5	6.9	21.7	32.2
500	LoRS	2.8	9.9	16.5	5.3	18.3	27.9
	Ours	6.2	19.9	30.6	7.0	22.0	32.9
700	LoRS	3.0	10.9	17.7	6.0	19.0	28.1
	Ours	6.5	20.1	31.0	7.5	22.2	33.4
800	LoRS	3.1	10.7	17.4	5.9	19.1	28.5
	Ours	6.7	20.8	31.7	7.7	22.5	33.7
900	LoRS	3.0	10.4	17.6	6.0	18.9	28.0
	Ours	6.9	21.4	32.2	7.2	22.3	33.6

H Generalization to Audio-Text Datasets

To explore the generalizability of our multimodal dataset distillation approach beyond image-text data, we extend our experiments to the audio-text domain using the AudioCaps [23] dataset. AudioCaps is

a widely used dataset for audio-text contrastive learning, derived from AudioSet [17]. It comprises approximately 44,000 audio clips paired with human-annotated captions that vividly describe the auditory content. The distillation process follows a similar protocol to that used in the image-text experiments. We employ BERT as the text encoder and EfficientAT (mn20_as) [41] as the audio encoder. EfficientAT is a state-of-the-art audio classification model based on MobileNet [21], designed to achieve high representational quality with low computational overhead.

The results presented in Table 11, compare our method against LoRS [55] on the AudioCaps dataset for 100, 200, and 500 synthetic pairs. Our approach consistently outperforms LoRS across all metrics and data scales. In 500 pairs settings, our method achieves AR@10 of 46.8 and TR@10 of 54.1, compared to LoRS’s 36.7 and 41.3, respectively. Notably, our method achieves around 65% of the full dataset’s performance using only 1.13% of the data. Superior results demonstrate that our proposed approach successfully generalizes to audio-text datasets, extending beyond the image-text domain. By achieving significant performance gains over existing baseline, our method establishes a more robust framework for multimodal dataset distillation across diverse modality pairs.

Table 11: Results on AudioCaps [23]. Both distillation and validation are performed using pre-trained EfficientAT+BERT. The model trained on full dataset performs: AR@1=21.3, AR@5=53.2, AR@10=68.5; TR@1=25.2, TR@5=58.8, TR@10=71.6.

Method	Pairs	Ratio	Audio to Text			Text to Audio		
			AR@1	AR@5	AR@10	TR@1	TR@5	TR@10
LoRS [55]	100	0.23%	2.7±0.3	8.6±0.3	14.7±0.4	5.9±0.3	13.0±0.4	21.8±0.5
	200	0.45%	3.8±0.2	14.8±0.2	21.8±0.2	8.0±0.2	21.2±0.2	33.1±0.2
	500	1.13%	7.1±0.1	24.7±0.2	36.7±0.2	9.2±0.2	27.4±0.3	41.3±0.3
Ours	100	0.23%	4.1±0.2	14.2±0.3	23.7±0.4	8.9±0.1	24.3±0.2	34.7±0.3
	200	0.45%	6.8±0.2	20.6±0.2	31.4±0.3	9.7±0.2	29.1±0.4	41.2±0.4
	500	1.13%	9.7±0.1	32.2±0.3	46.8±0.2	13.8±0.3	38.6±0.3	54.1±0.4

I More Experiments on Various Architectures

We further implement our method using different combinations of image encoders (e.g., ResNet-50 [19], ViT [12], RegNet [38], NFNet [3]) and text encoders (e.g., BERT [10], DistilBERT [40]) to assess the robustness and generality of our framework. The corresponding results are presented in Figure 6 and Figure 7. Across all architecture combinations, our method consistently outperforms the baseline LoRS [55], demonstrating its adaptability to various vision and language backbones.

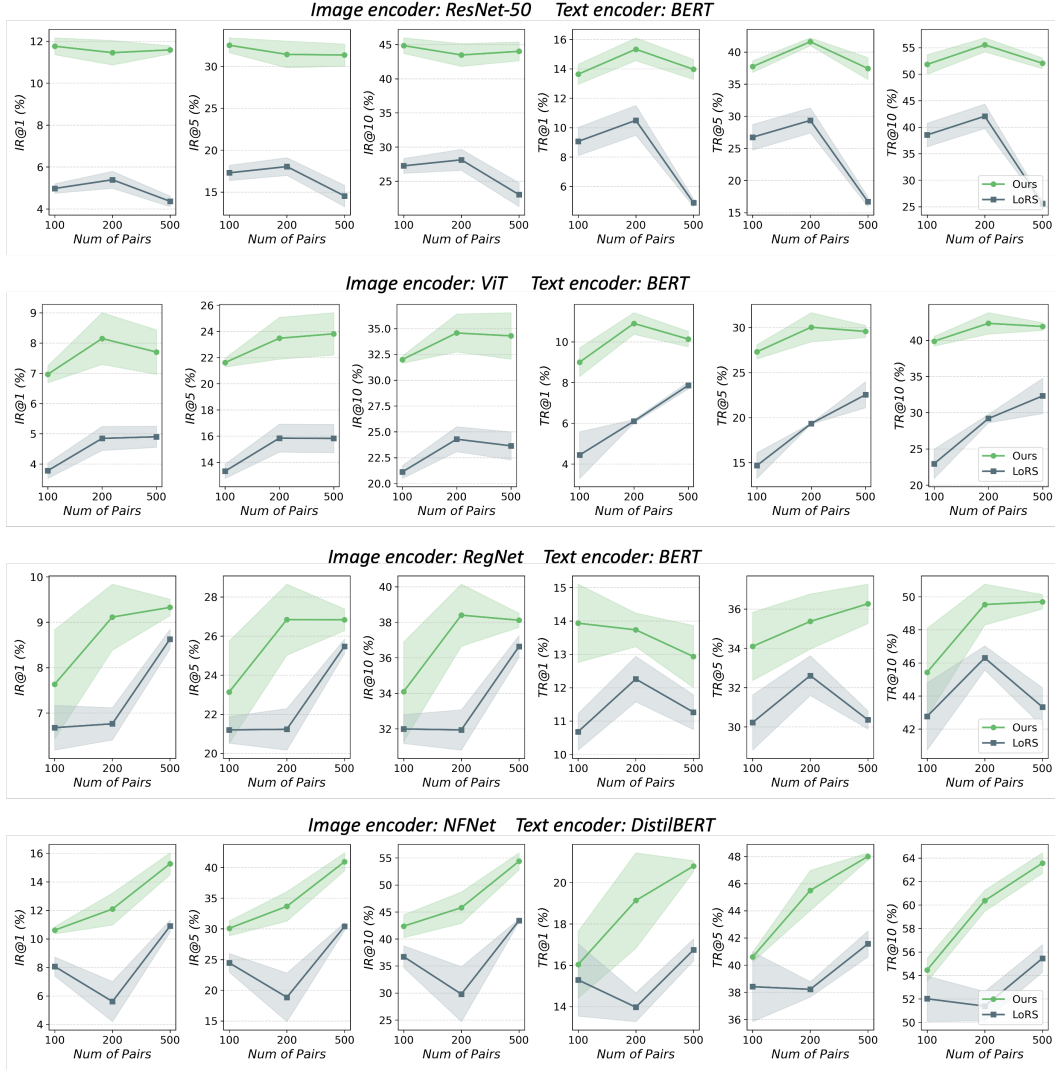


Figure 6: Performance on Flickr-30K with different combinations of image and text encoders.

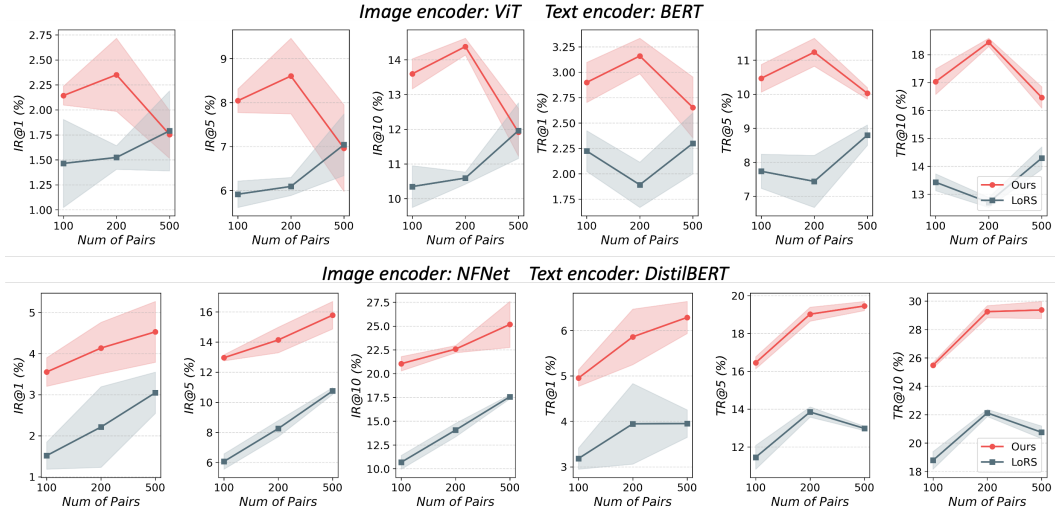


Figure 7: Performance on MS-COCO with different combinations of image and text encoders.



Figure 8: Flickr-30K before and after distillation. (Left) The original image-text pairs before the distillation. (Right) The image-text pairs after distillation.

J Visualization of Distilled Data

Here we provide visualizations of distilled image-text pairs. Figure 8 and Figure 9 present the original and distilled data on Flickr-30K and MS-COCO. The displayed texts are the closest matching sentences from the training set to the distilled text embeddings, following [53].



Figure 9: MS-COCO before and after distillation. (Left) The original image-text pairs before the distillation. (Right) The image-text pairs after distillation.