Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation

Xin Zhang[1,2] **Ziruo Zhang**[3] **Jiawei Du**[1,2] **Zuozhu Liu**[4] **Joey Tianyi Zhou**[1,2] ✉
[1]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore
[2]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[3]National University of Singapore, Singapore [4]Zhejiang University, China
{zhangx7, dujw, Joey_Zhou}@a-star.edu.sg
ziruo.z@u.nus.edu zuozhuliu@intl.zju.edu.cn

Abstract

Multimodal Dataset Distillation (MDD) seeks to condense large-scale image-text datasets into compact surrogates while retaining their effectiveness for cross-modal learning. Despite recent progress, existing MDD approaches often suffer from ***Modality Collapse***, characterized by over-concentrated intra-modal representations and enlarged distributional gap across modalities. In this paper, for the first time, we identify this issue as stemming from a fundamental conflict between the over-compression behavior inherent in dataset distillation and the cross-modal supervision imposed by contrastive objectives. To alleviate modality collapse, we introduce **RepBlend**, a novel MDD framework that weakens overdominant cross-modal supervision via representation blending, thereby significantly enhancing intra-modal diversity. Additionally, we observe that current MDD methods impose asymmetric supervision across modalities, resulting in biased optimization. To address this, we propose symmetric projection trajectory matching, which synchronizes the optimization dynamics using modality-specific projection heads, thereby promoting balanced supervision and enhancing cross-modal alignment. Experiments on Flickr-30K and MS-COCO show that RepBlend consistently outperforms prior state-of-the-art MDD methods, achieving significant gains in retrieval performance (e.g., +9.4 IR@10, +6.3 TR@10 under the 100-pair setting) and offering up to 6.7× distillation speedup. Our code is publicly available at https://github.com/zhangxin-xd/RepBlend.

1 Introduction

The unprecedented expansion of large-scale datasets has catalyzed recent breakthroughs in deep learning [6, 2, 1], but has also introduced considerable storage and computational overhead [20, 22]. Thus, reducing dataset size to streamline the development process has emerged as an important research focus. Among various solutions, Dataset Distillation (DD) [50] has emerged as a compelling strategy, achieving high compression ratios by synthesizing a compact surrogate dataset that approximates the training efficacy of the original dataset. The effectiveness of DD has been demonstrated across various modalities, including images [4, 57], text [30, 32], videos [11, 51], and graphs [29, 58]. These unimodal successes motivate its extension to increasingly prominent multimodal scenarios [37, 28, 35, 5].

The pioneering effort in multimodal dataset distillation (MDD) is MTT-VL [53], which first validates the feasibility of extending existing vanilla DD techniques to the image-text setting. Building on this baseline, LoRS [55] further proposes to mine cross-modal similarity to calibrate the supervision

✉ Corresponding author.

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

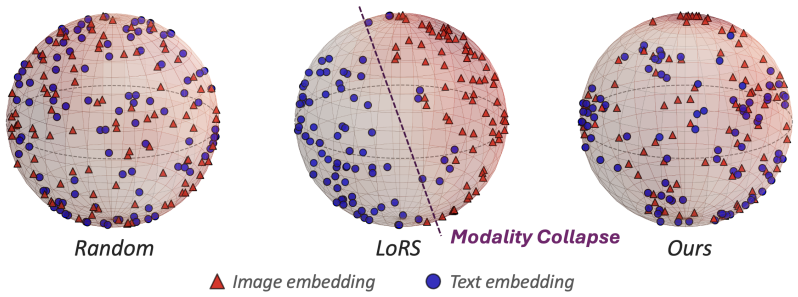

Figure 1: Multimodal embedding distributions across various distillation methods. We extract image and text embeddings from a finetuned CLIP [37] and project them into a shared representation space using DOSNES [31]. Red triangles and blue circles denote image and text embeddings, respectively. **Left**: Embeddings from randomly sampled data in the original dataset exhibit a well-spread and modality-aligned distribution. **Middle**: The distilled dataset generated by a SOTA MDD method (LoRS [55]) suffers from *Modality Collapse*, where image and text embeddings are poorly aligned and concentrated in separate regions. **Right**: Our method effectively mitigates modality collapse, resulting in a distribution with improved cross-modal alignment and higher representational diversity.

from matched and mismatched pairs, thereby achieving better adaptation to high-variance image-text data. Despite achieving promising results, existing studies remain confined to the data structure level, without probing the underlying conflict between DD and contrastive learning. Specifically, to prevent significant performance deterioration, vanilla DD prioritizes capturing representative features under limited distillation budgets, often sacrificing diversity and distributional coverage [14, 18, 15]. While this compromise is tolerable in unimodal classification tasks, naively applying such strategies to multimodal contrastive learning, which places great importance on instance-level discriminability, inevitably leads to ***Modality Collapse***. As illustrated in Figure 1 (middle), the distilled dataset exhibits pronounced intra-modality aggregation and inter-modality separation.

This modality collapse leads to two critical issues. First, *it induces excessive intra-modal similarity*, where embeddings within each modality become increasingly concentrated as distillation progresses. This over-concentration gradually suppresses representational diversity, making semantically distinct instances harder to separate, and eroding the fine-grained discrimination ability within each modality. Second, *it widens the inter-modal gap*, resulting in a large divergence between the feature distributions of different modalities. Insufficient cross-modal interaction fragments the embedding spaces and weakens semantic alignment, compromising the correct matching of positive pairs and the separation of negative pairs across modalities.

Recognizing these limitations, we propose **RepBlend**, a novel framework for MDD aimed at alleviating modality collapse. First, we theoretically identify that the collapse results from the joint effect of the over-compressive nature of DD, where optimization converges toward a small set of dominant features, and the cross-modal contrastive supervision, which further reinforces this convergence, leading to intra-modal collapse. To address this issue, RepBlend introduces Representation Blending within each modality to weaken the overly strong cross-modal supervision, thereby promoting intra-modal diversity. Furthermore, we observe that existing MDD approaches exhibit asymmetric supervision between modalities, with the image branch receiving significantly weaker update signals than the text branch. To address this, we propose Symmetric Projection Trajectory Matching, a mechanism that aligns the optimization trajectories of both projection heads, thereby enhancing cross-modal alignment and improving overall distillation efficiency. Extensive evaluations on Flickr-30K and MS-COCO demonstrate that RepBlend consistently surpasses existing MDD methods. Notably, under the 100-pair setting on Flickr-30K, it achieves improvements of +9.4 in IR@10 and +6.3 in TR@10, along with a 6.7$\times$ distillation speedup over the SOTA baseline. Beyond these benchmarks, RepBlend also exhibits strong generalization to other multimodal scenarios, such as audio-text.

Our contributions are summarized as follows:

- For the first time, we identify the modality collapse issue in current MDD solutions, where the distilled dataset exhibits high intra-modal similarity and a large inter-modal gap. Through theoretical analysis, we attribute this to a mutually reinforcing effect between the over-compression behavior of dataset distillation and the cross-modal supervision enforced by contrastive objectives.

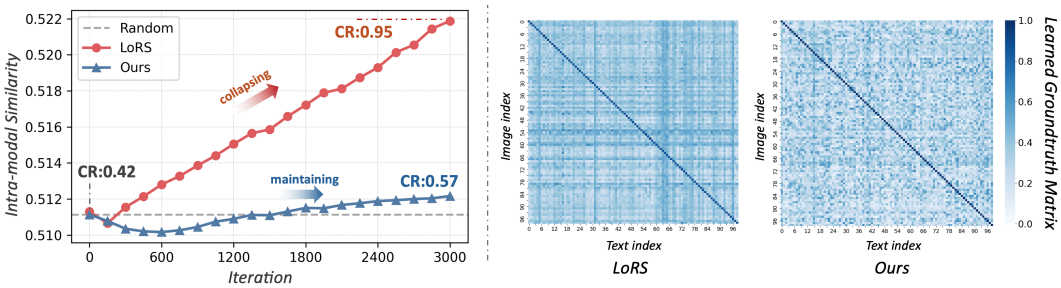

Figure 2: **Left**: Increasing intra-modal similarity as distillation progresses. We run optimization for 3000 iterations and track the intra-modal cosine similarity, which increases from 0.512 to 0.522 (red curve). Though small in magnitude, this rise leads to a more than twofold increase in concentration ratio (CR)[2] due to the high dimensionality of the embedding space. **Right**: Modality collapse undermines the effectiveness of learned soft cross-modal correspondence. The non-matching image-text pairs exhibit nearly uniform similarity scores, forming horizontal and vertical stripes.

- We propose Representation Blending to mitigate modality collapse by weakening the overly strong cross-modal supervision and enhancing intra-modal representational diversity. Furthermore, we introduce Symmetric Projection Trajectory Matching to enable more balanced multimodal distillation, which not only strengthens cross-modal alignment but also improves overall distillation efficiency.

2 Preliminaries and Related Works

Dataset Distillation (DD) [50] aims to synthesize a compact surrogate dataset that emulates the key properties of the original large-scale dataset. These properties include distributional characteristics, such as feature-level statistics [60, 48, 49] and batch normalization parameters [57, 42, 15], and training dynamics, including gradients [61, 59] and optimization trajectories [4, 7, 14, 18, 25]. While DD achieves promising results on unimodal benchmarks, extending it to multimodal scenarios remains challenging due to unique data structures and learning strategies [53, 55]. We begin by formalizing the problem of Multimodal Dataset Distillation (MDD).

Problem Formulation. Given a large-scale image-text dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{\tau}_i), \boldsymbol{y}_i\}_{i=1}^{|\mathcal{D}|}$, where $\boldsymbol{x}_i \in \mathbb{R}^{d_{\text{img}}}$ and $\boldsymbol{\tau}_i \in \mathbb{R}^{d_{\text{text}}}$ denote the i-th image and its paired caption representation[1], and each pair is independently sampled from a natural data distribution $\mathcal{P}$. Each $\boldsymbol{y}_i \in \{0,1\}^{|\mathcal{D}|}$ is a one-hot vector indicating the correspondence between $\boldsymbol{x}_i$ and the caption set $\{\boldsymbol{\tau}_j\}_{j=1}^{|\mathcal{D}|}$, with the i-th entry activated. Similar to DD, MDD also aims to minimize the loss on original dataset using the model trained on its distilled synthetic counterpart $\mathcal{S} = \{(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\tau}}_i), \tilde{\boldsymbol{y}}_i\}_{i=1}^{|\mathcal{S}|}$:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{\tau})\sim\mathcal{P}} [\mathcal{L}(f_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{x}, \boldsymbol{\tau}), \boldsymbol{y})] \quad \text{s.t.} \quad \boldsymbol{\theta}_\mathcal{S} = \arg\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{\tau}})\sim\mathcal{S}} [\mathcal{L}(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\tau}}), \tilde{\boldsymbol{y}})], \quad (1)$$

where $|\mathcal{S}| \ll |\mathcal{D}|$, and $\mathcal{L}$ denotes the contrastive learning loss. The model $f_{\boldsymbol{\theta}}(\cdot)$ represents a CLIP-style network parameterized by $\boldsymbol{\theta}$. Each distilled sample consists of a synthetic image-text pair $(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\tau}}_i)$, where $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{d_{\text{img}}}$ and $\tilde{\boldsymbol{\tau}}_i \in \mathbb{R}^{d_{\text{text}}}$, accompanied by a learned soft label $\tilde{\boldsymbol{y}}_i$.

MDD vs. Vanilla DD. According to the Equation 1, the generalization from vanilla DD to MDD involves two key modifications: 1) introducing soft ground-truth vectors $\tilde{\boldsymbol{y}}_i$, and 2) optimizing under a contrastive learning loss $\mathcal{L}$ for image-text alignment. While learning soft labels is common in vanilla DD [7], optimizing $\tilde{\boldsymbol{y}}_i$ in MDD is more challenging, as both image and text representations are updated simultaneously. Besides, in practice, the contrastive loss $\mathcal{L}$ is typically instantiated as InfoNCE [33], extended InfoNCE (eNCE), or weighted BCE (wBCE) [55], all aiming to strengthen positive alignments while penalizing mismatched pairs. However, these extensions only make the multimodal adaptation feasible, overlooking the essence of dataset distillation: effective information

[1] Given the discrete nature of text, all subsequent analysis is conducted in the representation space, while images remain processed in the pixel space. Here, $d_{\text{img}} = W \times H \times 3$ and $d_{\text{text}} = 768$ (for BERT [10]).

[2] CR measures how tightly the features are clustered, based on how much of the hypersphere is covered at the given cosine similarity. (Refer to Appendix C for more calculation details).

condensation. More specifically, they prioritize cross-modal alignment, while failing to preserve intra-modal diversity and discriminability under severe data compression.

3 Methodology

In this section, we introduce **RepBlend**, a novel approach for MDD. We begin by identifying the phenomenon of ***Modality Collapse***, which emerges when vanilla DD methods are naively applied to multimodal settings. Through theoretical and empirical analysis, we uncover its underlying causes. To address this issue, we propose Representation Blending to enhance intra-modal diversity. In addition, we introduce Symmetric Projection Trajectory Matching, which balances the distillation process across modalities and further strengthens cross-modal alignment. The overall pipeline of RepBlend is outlined in Algorithm 1.

3.1 Modality Collapse

LoRS [55] is a representative MDD method built upon Equation 1, where $\mathcal{L}$ is defined as:

$$\mathcal{L}_{\text{wBCE}}^{\mathcal{B}} = \sum_{i,j}^{|\mathcal{B}|} w_{ij} \cdot \ell\left(\tilde{\boldsymbol{y}}_{ij}, \sigma\left(\hat{\boldsymbol{y}}_{ij}/\gamma\right)\right), \quad w_{ij} = \frac{\mathbb{I}[\tilde{\boldsymbol{y}}_{ij} > \beta]}{|\{(i,j) : \tilde{\boldsymbol{y}}_{ij} > \beta\}|} + \frac{\mathbb{I}[\tilde{\boldsymbol{y}}_{ij} \leq \beta]}{|\{(i,j) : \tilde{\boldsymbol{y}}_{ij} \leq \beta\}|}. \tag{2}$$

Here, $\mathcal{B} \subset \mathcal{S}$ denotes a sampled batch. $\hat{\boldsymbol{y}}_{ij}$ represents the cosine similarity between the normalized image and text embeddings, where $\tilde{\boldsymbol{x}}'_i = \text{Normalize}(f^{\text{imgE}}(\tilde{\boldsymbol{x}}_i))$[3] and $\tilde{\boldsymbol{\tau}}'_j = \text{Normalize}(f^{\text{textP}}(\tilde{\boldsymbol{\tau}}_j))$, with $f^{\text{imgE}}(\cdot)$ and $f^{\text{textP}}(\cdot)$ denoting the image encoder and text projection head, respectively. The threshold β is used to determine positive and negative pairs, $\sigma(\cdot)$ denotes the sigmoid function, and γ is the temperature. $\ell(\cdot,\cdot)$ refers to the binary cross-entropy loss. While this supervision primarily aims to mine cross-modal relationships, it inadvertently reinforces intra-modal similarities, ultimately leading to ***Modality Collapse***, as shown in Figure 1, where instances within each modality excessively concentrate. Without loss of generality, the following analysis focuses on the image modality.

Proposition: Cross-modal supervision reinforces intra-modal similarity. During dataset distillation, if $\{\tilde{\boldsymbol{x}}_n, \tilde{\boldsymbol{\tau}}_n\}$ and $\{\tilde{\boldsymbol{x}}_m, \tilde{\boldsymbol{\tau}}_m\}$ exhibit some non-negligible similarity, i.e., $\tilde{\boldsymbol{y}}_{nm} \approx \tilde{\boldsymbol{y}}_{mn} > \beta$, then the direction of their subsequent updates $\frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{x}}'_n} \frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{x}}'_m}$ is determined by

$$\frac{w_{nm} w_{mn}}{\gamma^2} [\sigma(\hat{\boldsymbol{y}}_{nm})/t - \tilde{\boldsymbol{y}}_{nm}][\sigma(\hat{\boldsymbol{y}}_{mn})/t - \tilde{\boldsymbol{y}}_{mn}] \tilde{\boldsymbol{\tau}}'^{\top}_m \tilde{\boldsymbol{\tau}}'_n, \tag{3}$$

which indicates that the optimization is guided by positive pairs $\tilde{\boldsymbol{\tau}}'^{\top}_m \tilde{\boldsymbol{\tau}}'_n$, promoting concentration in similar directions. A detailed derivation is provided in Appendix B. When distilling a large dataset into a compact one, the optimization process tends to be dominated by a few salient features [9, 15, 18, 43]. Once this convergence trend emerges, cross-modal supervision further reinforces it: modality-specific diversity is implicitly suppressed, and intra-modal representations are increasingly aligned toward a limited set of dominant directions. As illustrated in Figure 2 (left), *the intra-modal similarity consistently increases throughout the distillation process*.

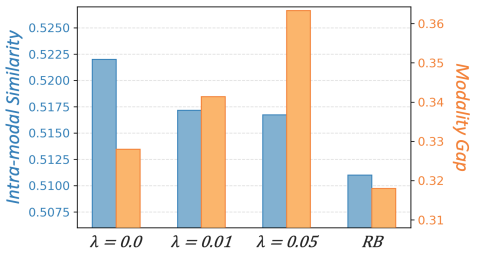

Figure 3: As the noise level λ increases, intra-modal similarity (blue bars) shows a slight decline, while the modality gap (yellow bars) rises markedly. In contrast, our representation blending (RB) leverages in-distribution samples to simultaneously reduce intra-modal similarity and inter-modal gap, effectively mitigating modality collapse during distillation.

In addition to the aggravated intra-modal similarity, modality collapse also exacerbates the cross-modal representation gap, as features from each modality become increasingly centralized within compact regions of the shared embedding space. Consequently, the similarities between non-matching image-text pairs converge toward a uniform distribution. Such behavior undermines the utility of soft label distributions, which are designed to encode fine-grained relational information beyond

[3] In LoRS [55], no image projection head is used.

the binary supervision provided by one-hot labels. As illustrated in Figure 2 (right), *non-diagonal similarity values exhibit a near-uniform pattern*, where image embeddings produce nearly constant similarity scores across all non-matching text embeddings (manifesting as horizontal stripes), and vice versa for text samples (vertical stripes).

3.2 Mitigating Modality Collapse via Representation Blending

As analyzed in Equation 3, modality collapse arises from overly strong cross-modal supervision, which implicitly encourages intra-modal concentration and undermines representational diversity. To alleviate this constraint, one potential approach is to inject directional signals that deviate from $\tilde{\boldsymbol{\tau}}'_m$ and $\tilde{\boldsymbol{\tau}}'_n$. To empirically validate this hypothesis and explore a viable remedy, we conduct a controlled perturbation experiment on Flickr-30K [36]. In particular, we adopt two key metrics following [26]: the intra-modal similarity (`Sim`) and the modality gap (`Gap`), defined as,

$$\texttt{Sim} = \frac{1}{|\mathcal{S}|(|\mathcal{S}|-1)} \sum_{i \neq j}^{|\mathcal{S}|} \tilde{\boldsymbol{x}}_i'^{\top} \tilde{\boldsymbol{x}}'_j, \quad \texttt{Gap} = \frac{1}{|\mathcal{S}|} \| \sum_{i=1}^{|\mathcal{S}|} \tilde{\boldsymbol{x}}'_i - \sum_{j=1}^{|\mathcal{S}|} \tilde{\boldsymbol{\tau}}'_j \|_2. \tag{4}$$

We inject Gaussian noise into the text representations,
$\tilde{\boldsymbol{\tau}}'^{\text{+noise}}_m = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_m + \lambda\vec{\Delta}_m)\right), \quad \tilde{\boldsymbol{\tau}}'^{\text{+noise}}_n = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_n + \lambda\vec{\Delta}_n)\right),$
where $\vec{\Delta}_m$ and $\vec{\Delta}_n$ are independently sampled random noise from $\mathcal{N}(0,1)$, and λ controls the noise level. We evaluate `Sim` and `Gap` under varying levels of λ. As shown in Figure 3, a slight increase in noise reduces intra-modal similarity (blue bars), indicating enhanced modality-specific diversity. These results support our hypothesis that perturbing in the representation space can effectively counteract modality concentration.

However, as noise level continues to grow, the injected perturbation begins to introduce semantically meaningless signals, which hinders cross-modal alignment. This is evidenced by the growing modality gap (yellow bars), accompanied by a performance drop of 1.9% in IR@1 and 2.1% in TR@1 at $\lambda = 0.01$ under 100 distilled pairs on Flickr-30K dataset. To mitigate this issue, we propose replacing the random perturbation with a structure-preserving variant using in-distribution samples. Specifically, we blend representations from different synthetic instances:

$$\tilde{\boldsymbol{\tau}}'^{\text{blend}}_m = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_m + \lambda\tilde{\boldsymbol{\tau}}_i)\right), \quad \tilde{\boldsymbol{\tau}}'^{\text{blend}}_n = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_n + \lambda\tilde{\boldsymbol{\tau}}_j)\right), \tag{5}$$

where $1 \leq i, j \leq |\mathcal{S}|$. This operation resembles the idea of MixUp, but is applied in the representation space. As shown in the last group of Figure 3, we can maintain a low level of intra-modal similarity and small modality gap. Note that although here we illustrate the formulation on text, the same operation is also applied to image side in practice.

3.3 Enhancing Cross-modal Alignment via Symmetric Projection Trajectory Matching

In prior MDD practices, methods such as MTT-VL [53] and LoRS [55] follow a de facto protocol wherein the text encoder is frozen and the image projection layer is omitted. The image encoder and the text projection head are trained to generate expert trajectories for distillation. In this setup, the image encoder is initialized with pretrained weights from ImageNet-1K [8], while the text projection head is trained from scratch. This design is motivated by two key considerations: 1) the prohibitive computational and memory cost of optimizing and storing expert trajectories for large-scale text encoders such as BERT [10]; and 2) the fact that text distillation operates in the representation space, where supervision is applied only through the projection head, thus, matching at the encoder level cannot propagate supervision to the representation space. LoRS [55] minimize the objective in Equation 1 through trajectory matching, which is formulated as follows:

$$\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\tau}}^*, \tilde{\boldsymbol{y}}^* = \arg\min_{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{y}}} \left(\left\| \boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{imgE}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgE}}} \right\|_2^2 + \left\| \boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{textP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}} \right\|_2^2 \right) / \left(\left\| \boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{imgE}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgE}}} \right\|_2^2 + \left\| \boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{textP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}} \right\|_2^2 \right),$$

where $\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{imgE}}}$ and $\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{textP}}}$ denote the T-step finetuned weights of the image encoder and text projection head using $\mathcal{S}$, initialized from $\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{imgE}}}$ and $\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{textP}}}$, respectively. The objective is to align the T-step synthetic trajectory with the M-step real trajectory by minimizing the ℓ_2 distance between their terminal weights, given the same initialization.

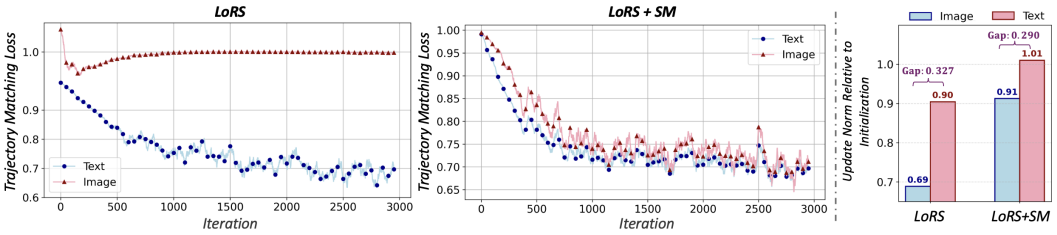

Figure 4: Current MDD methods adopt asymmetric distillation. **Left**: The loss on the image side shows much smaller variation than that of the text side, fluctuating mildly around 1.0 without notable reduction. **Right**: The update norm relative to initialization is significantly lower for the image modality in LoRS (0.69) compared to the text modality (0.90), suggesting insufficient representation transfer. The update norm is computed in the shared representation space for both modalities. After incorporating symmetric matching (SM), both image and text modalities exhibit more balanced and synchronized update dynamics, leading to more effective cross-modal alignment (reduced `Gap`).

However, the aforementioned trajectory matching is asymmetric. As shown in Figure 4 (left), the trajectory matching losses of the image and text modalities exhibit divergent trends: the text-side loss decreases steadily, whereas the image-side loss quickly plateaus and remains relatively high. This is primarily because the image encoder contains significantly more parameters than the text projection head, thus, even small per-parameter errors can accumulate into a large overall mismatch. This imbalance is further evidenced in Figure 4 (right), the norm of updates relative to initialization for the image modality is significantly smaller than that of the text, indicating insufficient distillation on the image side. While the representation blending introduced in Section 3.2 helps narrow the modality gap, its effect is still constrained by the inherently asymmetric distillation. To address this imbalance and further enhance cross-modal alignment, we propose a symmetric distillation strategy by matching trajectories of projection head for both modalities:

$$\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\tau}}^*, \tilde{\boldsymbol{y}}^* = \arg\min_{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{y}}} \left(\left\| \boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{imgP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgP}}} \right\|_2^2 + \left\| \boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{textP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}} \right\|_2^2 \right) / \left(\left\| \boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{imgP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgP}}} \right\|_2^2 + \left\| \boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{textP}}} - \boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}} \right\|_2^2 \right). \quad (6)$$

Here, the image encoder is initialized with ImageNet-1K pretrained weights and kept frozen. While the added image projection head incurs slight computational overhead, it enables projection-based matching that significantly enhances the overall efficiency of the distillation process (as discussed in Section 4.4). As shown in Figure 4, symmetric projection matching leads to a more consistent decrease in loss for both image and text branches. Moreover, the increased magnitude of updates suggests stronger supervision signals across modalities, resulting in a more balanced and effective distillation process. With symmetric distillation, the modality gap is further narrowed from 0.318 (in Figure 3) to 0.290, indicating enhanced cross-modal alignment.

4 Experiments

In this section, we conduct extensive experiments on multiple benchmark datasets to demonstrate the effectiveness of the proposed RepBlend framework. We first present the experimental setup, including the datasets, baseline methods, and implementation details. The main results are summarized in Table 1, Table 2, and Table 3. In addition, we also provide detailed ablation studies to evaluate the individual contribution of each component. All experiments are conducted using two NVIDIA RTX 3090 GPUs and one NVIDIA H100 GPU.

4.1 Experimental Setup

Datasets and Networks. We evaluate our method on two widely-used image captioning datasets: Flickr-30K [36] and MS-COCO [27], which contain approximately 31k and 123k images respectively, with each image paired with five human-annotated captions. For the image encoder, we experiment with NFNet [3], RegNet [38], ResNet-50 [19], and ViT [12]. For the text encoder, we consider both BERT [10] and DistilBERT [40]. To further demonstrate the generalizability of our approach across modalities, we extend our evaluation to the AudioCaps [23] audio-text benchmark, utilizing EfficientAT [41] as the audio encoder. Model performance is primarily evaluated using Recall at K (R@K) in cross-modal retrieval tasks. Given a query from one modality, we retrieve the top-K most similar samples from the other modality and measure the retrieval accuracy. We denote text-to-image retrieval as IR@K, and image-to-text retrieval as TR@K.

Algorithm 1 Blending Representations to Mitigate Modality Collapse in MDD

Require: Original large dataset $\mathcal{D}$; CLIP-style network $\{f^{\text{imgE}}, f^{\text{textE}}, f^{\text{imgP}}, f^{\text{textP}}\}$; real trajectories set $\mathbf{\Theta}_{\mathcal{D}_{\text{imgP}}}$ and $\mathbf{\Theta}_{\mathcal{D}_{\text{textP}}}$, real trajectory matching length M, synthetic trajectory matching length T; total optimization iteration number $Iter$

Initialize $\mathcal{S}$ with $|\mathcal{S}|$ randomly sampled image-text pairs and one-hot groundtruth labels
Load pretrained weights into encoders (frozen); randomly initialize projection heads
for $it = 1$ to $Iter$ **do**
  Sample $\boldsymbol{\theta}^t_{\mathcal{D}_{\text{imgP}}}$, $\boldsymbol{\theta}^t_{\mathcal{D}_{\text{textP}}}$ and $\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgP}}}$, $\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}}$ from $\mathbf{\Theta}_{\mathcal{D}_{\text{imgP}}}$ and $\mathbf{\Theta}_{\mathcal{D}_{\text{textP}}}$
  Initialize $\boldsymbol{\theta}^t_{\mathcal{S}_{\text{imgP}}}$ and $\boldsymbol{\theta}^t_{\mathcal{S}_{\text{textP}}}$ using $\boldsymbol{\theta}^t_{\mathcal{D}_{\text{imgP}}}$ and $\boldsymbol{\theta}^t_{\mathcal{D}_{\text{textP}}}$
  for $i = 1$ to T **do**
    for mini-batch $\mathcal{B} = \{(\tilde{\boldsymbol{x}}_b, \tilde{\boldsymbol{\tau}}_b), \tilde{\boldsymbol{y}}_b\}_{b=1}^{|\mathcal{B}|} \in \mathcal{S}$ **do**
      Calculate image representaion $\{f^{\text{imgE}}(\tilde{\boldsymbol{x}}_b)\}$
      ▷ `Blending in representation space`
      $\{f^{\text{imgE}}(\tilde{\boldsymbol{x}}_b), \tilde{\boldsymbol{\tau}}_b\} = \text{RepBlend}(\{f^{\text{imgE}}(\tilde{\boldsymbol{x}}_b), \tilde{\boldsymbol{\tau}}_b\})$
      Compute loss $\mathcal{L}^{\mathcal{B}}_{\text{wBCE}}$ using Equation 2
      Update projection head weights $\boldsymbol{\theta}^{t+i}_{\mathcal{S}_{\text{imgP}}}$ and $\boldsymbol{\theta}^{t+i}_{\mathcal{S}_{\text{textP}}}$
    end for
    ▷ `Symmetric projection trajectory matching`
    Optimize $\mathcal{S} = \{(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\tau}}_j), \tilde{\boldsymbol{y}}_j\}_{j=1}^{|\mathcal{S}|}$ according to Equation 6
  end for
end for

Ensure: Synthetic dataset $\mathcal{S}$

Baselines. The comparison encompasses a range of SOTA approaches, including coreset selection methods such as Random sampling, Herding [52], K-Center [16], and Forgetting [47], as well as recent advances in dataset distillation tailored for vision-language models, including MTT-VL [53], TESLA-VL [55], and LoRS [55]. A detailed description of these methods can be found in the Appendix E. For fairness, both LoRS [55] and our method synthesize one fewer pair per distillation budget (e.g., 99 pairs for a budget of 100) to account for the additional similarity-matrix overhead.

Implementation Details. We construct a CLIP-style architecture using the aforementioned image and text encoders. The image encoder is initialized with ImageNet-pretrained weights [8], while the text encoder is initialized with the official pretrained weights provided by the corresponding language model. After feature extraction, the outputs from both branches are passed through separate linear projection layers to obtain the final embeddings. During buffer generation, distillation, and evaluation training, the encoders are frozen and only the projection layers are optimized. We collect 20 expert trajectories, each consisting of 10 training epochs. The hyperparameter settings follow those used in LoRS [55] and can be found in Table 7 and Table 8 in Appendix F.

4.2 Main Results

The results on Flickr-30K [36] and MS-COCO [27] are presented in Table 1 and Table 2, respectively. Our method consistently outperforms all baseline methods, across all distillation budgets and evaluation metrics. Notably, on Flickr-30k, under the extremely low-data regime of 100 training pairs (0.3%), our method achieves an IR@1 of 11.5%, substantially surpassing LoRS (8.3%) and MTT-VL (4.7%). Similarly, our TR@10 reaches 55.5%, a considerable gain over the best baseline LoRS (49.2%). These trends hold consistently across all pair settings. Under the 500-pair scenario (1.7%), our method improves the IR@10 from 41.6% (LoRS) to 55.9% and TR@10 from 53.7% to 66.7%, reflecting a relative gain of over 30%. On MS-COCO, a dataset known for higher complexity and variability, our method continues to exhibit superior performance. Under the 100-pair setting (0.8‰), our approach achieves IR@10 = 22.3% and TR@10 = 28.0%, substantially outperforming LoRS, which attains 12.2% and 19.6%, respectively. At a higher budget of 500 training pairs (4.4‰), our method maintains its advantage, achieving the highest IR@10 (30.6%) and TR@10 (32.9%) among all evaluated methods. Besides, we also extend our method to a larger-scale setting using the LLaVA-cc3m dataset, which serves as the pretraining dataset for LLaVA and consists of 558k image-text pairs. We use approximately 60% of the data (about 334k pairs) for training and reserve a

Table 1: Results on Flickr-30k [36]. Both distillation and validation are performed using NFNet+BERT. The model trained on 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	Ratio	Metric	Coreset Selection				Dataset Distillation			
			Rand	Herd [52]	K-Cent [16]	Forget [47]	MTT-VL [53]	TESLA-VL [55]	LoRS [55]	Ours
100	0.3%	IR@1	1.0	0.7	0.7	0.7	$4.7_{\pm 0.2}$	$0.5_{\pm 0.2}$	$8.3_{\pm 0.2}$	**11.5**$_{\pm 0.4}$
		IR@5	4.0	2.8	3.1	2.4	$15.7_{\pm 0.5}$	$2.3_{\pm 0.2}$	$24.1_{\pm 0.2}$	**32.0**$_{\pm 0.7}$
		IR@10	6.5	5.3	6.1	5.6	$24.6_{\pm 1.0}$	$4.7_{\pm 0.4}$	$35.1_{\pm 0.3}$	**44.5**$_{\pm 0.6}$
		TR@1	1.3	1.1	0.6	1.2	$9.9_{\pm 0.3}$	$5.5_{\pm 0.5}$	$11.8_{\pm 0.2}$	**16.2**$_{\pm 0.8}$
		TR@5	5.9	4.7	5.0	4.2	$28.3_{\pm 0.5}$	$19.5_{\pm 0.9}$	$35.8_{\pm 0.6}$	**41.7**$_{\pm 0.9}$
		TR@10	10.1	7.9	7.6	9.7	$39.1_{\pm 0.7}$	$28.9_{\pm 1.0}$	$49.2_{\pm 0.5}$	**55.5**$_{\pm 0.4}$
200	0.7%	IR@1	1.1	1.5	1.5	1.2	$4.6_{\pm 0.9}$	$0.2_{\pm 0.1}$	$8.6_{\pm 0.3}$	**12.7**$_{\pm 0.8}$
		IR@5	4.8	5.5	5.4	3.1	$16.0_{\pm 1.6}$	$1.3_{\pm 0.2}$	$25.3_{\pm 0.2}$	**34.7**$_{\pm 0.6}$
		IR@10	9.2	9.3	9.9	8.4	$25.5_{\pm 2.6}$	$2.5_{\pm 0.2}$	$36.6_{\pm 0.3}$	**47.6**$_{\pm 0.5}$
		TR@1	2.1	2.3	2.2	1.5	$10.2_{\pm 0.8}$	$2.8_{\pm 0.5}$	$14.5_{\pm 0.5}$	**18.6**$_{\pm 0.7}$
		TR@5	8.7	8.4	8.2	8.4	$28.7_{\pm 1.0}$	$10.4_{\pm 1.5}$	$38.7_{\pm 0.5}$	**46.0**$_{\pm 0.8}$
		TR@10	13.2	14.4	13.5	10.2	$41.9_{\pm 1.9}$	$17.4_{\pm 1.6}$	$53.4_{\pm 0.5}$	**60.0**$_{\pm 0.6}$
500	1.7%	IR@1	2.4	3.0	3.5	1.8	$6.6_{\pm 0.3}$	$1.1_{\pm 0.2}$	$10.0_{\pm 0.2}$	**17.0**$_{\pm 0.6}$
		IR@5	10.5	10.0	10.4	9.0	$20.2_{\pm 1.2}$	$7.3_{\pm 0.4}$	$28.9_{\pm 0.7}$	**42.5**$_{\pm 0.5}$
		IR@10	17.4	17.0	17.3	15.9	$30.0_{\pm 2.1}$	$12.6_{\pm 0.5}$	$41.6_{\pm 0.6}$	**55.9**$_{\pm 0.6}$
		TR@1	5.2	5.1	4.9	3.6	$13.3_{\pm 0.6}$	$5.1_{\pm 0.2}$	$15.5_{\pm 0.7}$	**22.5**$_{\pm 0.4}$
		TR@5	18.3	16.4	16.4	12.3	$32.8_{\pm 1.8}$	$15.3_{\pm 0.5}$	$39.8_{\pm 0.4}$	**53.2**$_{\pm 0.3}$
		TR@10	25.7	24.3	23.3	19.3	$46.8_{\pm 0.8}$	$23.8_{\pm 0.3}$	$53.7_{\pm 0.3}$	**66.7**$_{\pm 0.3}$

Table 2: Results on MS-COCO [27]. Both distillation and validation are performed using NFNet+BERT. The model trained on 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	Ratio	Metric	Coreset Selection				Dataset Distillation			
			Rand	Herd [52]	K-Cent [16]	Forget [47]	MTT-VL [53]	TESLA-VL [55]	LoRS [55]	Ours
100	0.8‰	IR@1	0.3	0.5	0.4	0.3	$1.3_{\pm 0.1}$	$0.3_{\pm 0.2}$	$1.8_{\pm 0.1}$	**4.1**$_{\pm 0.3}$
		IR@5	1.3	1.4	1.4	1.5	$5.4_{\pm 0.3}$	$1.0_{\pm 0.4}$	$7.1_{\pm 0.2}$	**13.9**$_{\pm 0.8}$
		IR@10	2.7	3.5	2.5	2.5	$9.5_{\pm 0.5}$	$1.8_{\pm 0.5}$	$12.2_{\pm 0.2}$	**22.3**$_{\pm 0.5}$
		TR@1	0.8	0.8	1.4	0.7	$2.5_{\pm 0.3}$	$2.0_{\pm 0.2}$	$3.3_{\pm 0.2}$	**5.2**$_{\pm 0.5}$
		TR@5	3.0	2.1	3.7	2.6	$10.0_{\pm 0.5}$	$7.7_{\pm 0.5}$	$12.2_{\pm 0.3}$	**17.9**$_{\pm 0.9}$
		TR@10	5.0	4.9	5.5	4.8	$15.7_{\pm 0.4}$	$13.5_{\pm 0.3}$	$19.6_{\pm 0.3}$	**28.0**$_{\pm 0.3}$
200	1.7‰	IR@1	0.6	0.9	0.7	0.6	$1.7_{\pm 0.1}$	$0.1_{\pm 0.1}$	$2.4_{\pm 0.1}$	**6.1**$_{\pm 0.8}$
		IR@5	2.3	2.4	2.1	2.8	$6.5_{\pm 0.4}$	$0.2_{\pm 0.1}$	$9.3_{\pm 0.2}$	**19.3**$_{\pm 0.7}$
		IR@10	4.4	4.1	5.8	4.9	$12.3_{\pm 0.8}$	$0.5_{\pm 0.1}$	$15.5_{\pm 0.2}$	**29.8**$_{\pm 0.5}$
		TR@1	1.0	1.0	1.2	1.1	$3.3_{\pm 0.2}$	$0.7_{\pm 0.2}$	$4.3_{\pm 0.1}$	**6.9**$_{\pm 0.6}$
		TR@5	4.0	3.6	3.8	3.5	$11.9_{\pm 0.6}$	$3.1_{\pm 0.5}$	$14.2_{\pm 0.3}$	**21.8**$_{\pm 0.9}$
		TR@10	7.2	7.7	7.5	7.0	$19.4_{\pm 1.2}$	$5.3_{\pm 0.8}$	$22.6_{\pm 0.2}$	**32.3**$_{\pm 0.7}$
500	4.4‰	IR@1	1.1	1.7	1.1	0.8	$2.5_{\pm 0.5}$	$0.8_{\pm 0.2}$	$2.8_{\pm 0.2}$	**6.2**$_{\pm 0.1}$
		IR@5	5.0	5.3	6.3	5.8	$8.9_{\pm 0.7}$	$3.6_{\pm 0.6}$	$9.9_{\pm 0.5}$	**19.9**$_{\pm 0.3}$
		IR@10	8.7	9.9	10.5	8.2	$15.8_{\pm 1.5}$	$6.7_{\pm 0.9}$	$16.5_{\pm 0.7}$	**30.6**$_{\pm 0.1}$
		TR@1	1.9	1.9	2.5	2.1	$5.0_{\pm 0.4}$	$1.7_{\pm 0.4}$	$5.3_{\pm 0.5}$	**7.0**$_{\pm 0.2}$
		TR@5	7.5	7.8	8.7	8.2	$17.2_{\pm 1.3}$	$5.9_{\pm 0.8}$	$18.3_{\pm 1.5}$	**22.0**$_{\pm 0.3}$
		TR@10	12.5	13.7	14.3	13.0	$26.0_{\pm 1.9}$	$10.2_{\pm 1.0}$	$27.9_{\pm 1.4}$	**32.9**$_{\pm 0.6}$

non-overlapping set of 10k pairs for validation[4]. In addition, we evaluate our approach with more powerful encoders, including DiNo-v2 [34] (85,798,656 parameters) for vision and BGE-1.5 [54] (109,482,240 parameters) for text. The results (shown in Table 3) demonstrate that our method remains effective when scaling both model capacity and training data, and it significantly outperforms the SOTA competitor. Moreover, our method also demonstrates strong generalizability to other multimodal settings, such as audio-text benchmark. See Appendix H for details.

4.3 Ablation Study

Representation Blending & Symmetric Matching. We conduct an ablation study on the Flickr-30K dataset using NFNet+BERT to evaluate the individual and combined contributions of the proposed components: Representation Blending (RB) and Symmetric Projection Trajectory Matching (SM). As shown in Figure 5, removing either module leads to consistent performance degradation across all retrieval metrics (IR@1/5/10 and TR@1/5/10) and distillation budgets (100, 200, 500 pairs). RB contributes by mitigating intra-modal collapse; as illustrated in Figure 3, it effectively reduces

[4]see https://huggingface.co/xinxin66/RepBlend/tree/main/datasets/cc3m.

Table 3: Results when scaling to larger dataset and model. NFNet + BERT on LLaVA-cc3m: the model trained on the full dataset performs: IR@1=9.13, IR@5=25.94, IR@10=36.34, TR@1=9.49, TR@5=26.08, TR@10=37.07. DiNo-v2 + BGE-1.5 on MS-COCO: the model trained on the full dataset performs: IR@1=22.70, IR@5=51.13, IR@10=65.26, TR@1=31.04, TR@5=61.96, TR@10=74.1.

Pairs	Methods	*NFNet + BERT on LLaVA-cc3m*						*DiNo-v2 + BGE-1.5 on MS-COCO*					
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
200	LoRS [55]	1.56	5.33	8.77	1.01	4.04	6.81	1.59	6.12	10.59	1.62	6.15	10.47
	Ours	**3.53**	**12.38**	**19.33**	**4.39**	**13.45**	**20.18**	**12.52**	**31.96**	**44.39**	**16.06**	**36.84**	**49.34**
500	LoRS [55]	1.96	6.72	10.55	1.41	5.11	8.51	2.48	8.46	13.42	3.18	10.36	16.37
	Ours	**4.45**	**15.09**	**22.31**	**5.14**	**14.89**	**22.97**	**13.37**	**33.09**	**45.69**	**16.90**	**39.38**	**52.14**
800	LoRS [55]	1.67	5.87	9.69	1.68	6.11	10.25	2.95	9.90	15.69	4.56	13.66	20.63
	Ours	**5.26**	**16.31**	**24.23**	**5.42**	**16.13**	**24.16**	**13.68**	**33.58**	**45.93**	**17.14**	**39.76**	**52.92**

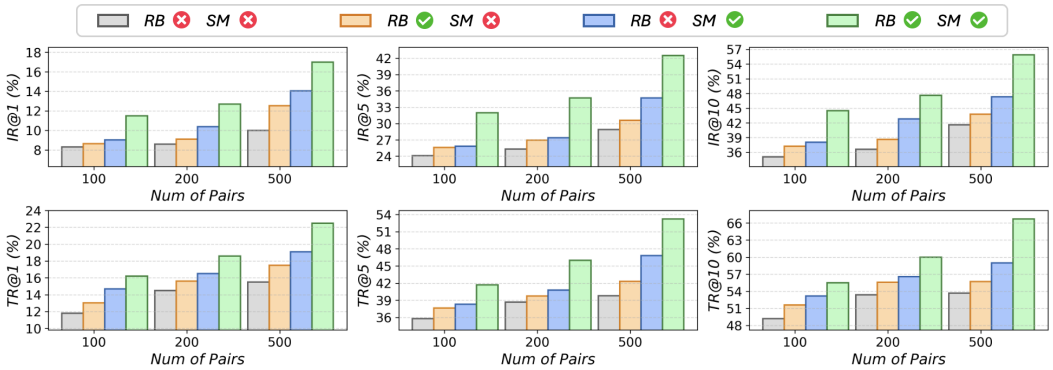

Figure 5: Ablation study of Representation Blending (RB) and Symmetric Projection Trajectory Matching (SM) on Flickr-30K with NFNet+BERT.

intra-modal similarity and enhances representational diversity. SM further balances the learning dynamics across modalities and improves cross-modal alignment, as evidenced in Figure 4. When combined, RB and SM achieve the best overall performance, highlighting their complementary roles in enhancing intra-modal diversity and cross-modal alignment.

Cross-Architecture Generalization. We further validate the generalization capability of RepBlend across diverse architectures. Following the protocol of LoRS [55], we keep the text encoder fixed and evaluate the dataset distilled with NFNet+BERT using alternative image encoders, including ResNet-50 and RegNet. As shown in Table 4, RepBlend consistently maintains strong performance across different encoder architectures. Moreover, we extend the evaluation to a broader set of architecture combinations, such as ResNet-50+BERT, ViT+BERT, RegNet+BERT, and NFNet+DistilBERT, as illustrated in Figure 6 and Figure 7 in Appendix I. Across all architectures, datasets, and distillation budgets, RepBlend consistently outperforms the sota baseline, demonstrating its robustness and architectural adaptability.

Zero-Shot Generalization. To further validate the effectiveness of our distilled dataset, we further evaluate zero-shot ImageNet [8] classification and OCR-relevant retrieval on TextCaps [45]. Specifically, we randomly select 10 classes from ImageNet-1K and report Top-1 and Top-5 zero-shot accuracies. For TextCaps, we measure retrieval performance on 3,166 validation samples. The results, summarized in Table 5, show that under the same budget, models trained on our distilled dataset outperform LoRS and narrow the performance gap to the full-dataset baseline.

4.4 Computational Efficiency

In the proposed method, the training trajectories of image and text projection layers are used for matching optimization. Although we introduce an additional image projection, it incurs negligible computational overhead. In fact, as shown in Table 6, our method achieves significantly better computational efficiency compared to prior work. Specifically, the time required to construct expert trajectories is reduced from 70 minutes to 40 minutes per trajectory (1.75× speedup), and the corresponding memory footprint decreases from 1.63 GB to 0.73 GB (2.23× reduction). During the distillation phase, our method accelerates training iterations from 11.5 seconds to 1.71 seconds per iteration, yielding a 6.7× speedup. Moreover, it lowers the peak GPU memory usage from 21.78 GB

Table 4: Cross-architecture generalization. The distilled data are synthesized with NFNet+BERT and evaluated across architectures on Flickr-30K under the 500-pair setting.

Evaluate Model	Methods	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet+BERT	TESLA-VL [55]	$3.0_{\pm 0.2}$	$10.8_{\pm 0.5}$	$17.0_{\pm 0.8}$	$6.0_{\pm 0.9}$	$18.8_{\pm 0.7}$	$27.7_{\pm 1.2}$
	LoRS [55]	$3.3_{\pm 0.2}$	$12.7_{\pm 0.3}$	$20.4_{\pm 0.2}$	$6.8_{\pm 0.2}$	$19.6_{\pm 1.3}$	$31.1_{\pm 0.3}$
	Ours	$\mathbf{4.2}_{\pm 0.2}$	$\mathbf{14.1}_{\pm 0.2}$	$\mathbf{23.6}_{\pm 0.6}$	$\mathbf{8.4}_{\pm 0.2}$	$\mathbf{23.1}_{\pm 0.8}$	$\mathbf{35.0}_{\pm 1.3}$
RegNet+BERT	TESLA-VL [55]	$3.2_{\pm 0.8}$	$11.1_{\pm 1.8}$	$17.5_{\pm 1.3}$	$5.8_{\pm 0.1}$	$18.6_{\pm 0.6}$	$28.1_{\pm 1.0}$
	LoRS [55]	$3.5_{\pm 0.1}$	$12.6_{\pm 0.3}$	$21.1_{\pm 0.4}$	$6.8_{\pm 0.3}$	$20.8_{\pm 0.3}$	$30.2_{\pm 0.3}$
	Ours	$\mathbf{3.9}_{\pm 0.2}$	$\mathbf{13.9}_{\pm 0.3}$	$\mathbf{24.0}_{\pm 0.6}$	$\mathbf{7.9}_{\pm 0.3}$	$\mathbf{24.2}_{\pm 0.3}$	$\mathbf{36.2}_{\pm 1.1}$

Table 5: Zero-Shot Generalization. Models trained on the distilled MS-COCO dataset under the 500-pair setting are evaluated on zero-shot ImageNet classification and TextCaps retrieval tasks.

Methods	*ImageNet-10 Classification*		*TextCaps Retrieval*					
	ACC@1	ACC@5	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
LoRS [55]	21.4	74.4	1.7	5.1	8.4	0.4	1.7	3.1
Ours	**27.6**	**76.2**	**3.1**	**9.4**	**14.5**	**1.9**	**6.2**	**10.3**

to 10.17 GB (2.14× reduction). These results show that our projection-based design not only enables effective multimodal distillation, but also leads to substantially improved computational efficiency.

5 Conclusion

In this work, we investigate the underexplored challenge of modality collapse in multimodal dataset distillation (MDD), where intra-modal similarity is excessively amplified and inter-modal alignment is degraded. Through theoretical analysis and empirical evidence, we attribute this phenomenon to the inherent over-compression behavior of dataset distillation and its interplay with cross-modal contrastive supervision. To mitigate these issues, we propose RepBlend, a novel MDD framework incorporating two key components: Representation Blending for enhancing intra-modal diversity and Symmetric Projection Trajectory Matching for achieving balanced and effective supervision across modalities. Extensive experiments on Flickr-30K and MS-COCO confirm the superiority of RepBlend in both retrieval performance and distillation efficiency.

Table 6: Study of computational efficiency.

Methods	LoRS [55]	Ours
(IR@1, TR@1) (%)	(8.3, 11.8)	**(11.5, 16.2)**
Buffer		
Speed (min/traj)	70	**40**
Memory (GB/traj)	1.63	**0.73**
Distillation		
Speed (s/iter)	11.5	**1.71**
Peak GPU VRAM (GB)	21.78	**10.17**

Limitations and Future work. Despite the promising results of RepBlend, current MDD frameworks, including ours, remain limited to pair-level modeling, which restricts fine-grained alignment between text tokens and visual objects. Additionally, insufficient cross-instance interaction hampers representation expressiveness and limits further gains in compression. In the future, we will explore instance-aware, relation-enhanced strategies to overcome these challenges.

6 Acknowledgement

This research is supported by Xin Zhang's A*STAR Career Development Fund (CDF) (Project No. C243512009), and Jiawei Du's A*STAR Career Development Fund (CDF) (Project No. C233312004). This research is also supported by the Japan Science and Technology Agency (JST) and the Agency for Science, Technology and Research (A*STAR) under the Japan-Singapore Joint Call (Project No. R24I6IR133).

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations*, 2022.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

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

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] 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.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Guodong Ding, Rongyu Chen, and Angela Yao. Condensing action segmentation datasets via generative network inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

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

[14] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *NeurIPS*, 2023.

[15] Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

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

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[18] Ziyao Guo, Kai Wang, George Cazenavette, HUI LI, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *ICLR*, 2024.

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

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[21] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[22] Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. Get more for less: Principled data selection for warming up fine-tuning in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

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

[25] Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *ICML*, 2024.

[26] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[27] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[29] Yang Liu, Deyu Bo, and Chuan Shi. Graph distillation with eigenbasis matching. In *Forty-first International Conference on Machine Learning*, 2024.

[30] Huimin LU, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Unidetox: Universal detoxification of large language models via dataset distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Yao Lu, Jukka Corander, and Zhirong Yang. Doubly stochastic neighbor embedding on spheres. *Pattern Recognition Letters*, 125:581–587, 2019.

[32] Aru Maekawa, Satoshi Kosugi, Kotaro Funakoshi, and Manabu Okumura. Dilm: Distilling dataset into language model for text-level dataset distillation. *Journal of Natural Language Processing*, 32(1):252–282, 2025.

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

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[35] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

[38] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

[39] 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.

[40] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[41] Florian Schmid, Khaled Koutini, and Gerhard Widmer. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *ICASSP 2023-2023 IEEE international Conference on acoustics, Speech and signal processing (ICASSP)*, pages 1–5. IEEE, 2023.

[42] 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.

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

[44] Seungjae Shin, Heesun Bae, Donghyeok Shin, Weonyoung Joo, and Il-Chul Moon. Loss-curvature matching for dataset selection and condensation. In *AISTAS*, 2023.

[45] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] Ziyu Wang, Yue Xu, Cewu Lu, and Yong-Lu Li. Dancing with still images: video distillation via static-dynamic disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6296–6304, 2024.

[52] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009.

[53] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. In *TMLR*, 2024.

[54] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[55] Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *ICML*, 2024.

[56] Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. 2023.

[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] Yuchen Zhang, Tianle Zhang, Kai Wang, Ziyao Guo, Yuxuan Liang, Xavier Bresson, Wei Jin, and Yang You. Navigating complexity: Toward lossless graph condensation via expanding window matching. In *Forty-first International Conference on Machine Learning*, 2024.

[59] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, 2021.

[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. In *ICLR*, 2021.

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 abstract and introduction clearly state the proposed method and its core contribution.

   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: The paper discusses the limitations of the work.

   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: [Yes]

Justification: The paper provides the full set of assumptions and a complete and correct proof.

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 paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   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: This paper provides implementation details and algorithm descriptions for reproduction. We also release our codes for reproduction in camera-ready version.

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: The paper specifies all the training and test details necessary to reproduce the results, including datasets, hyperparameters, optimizer type, and how they were chosen.

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: [Yes]

Justification: Our experiments were rigorously conducted with five repetitions each, and we meticulously reported both the mean values and standard deviations for each experimental trial.

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: The paper comprehensively details compute resources in both the experiments section and supplementary materials, covering GPU type, memory, and storage specifics.

   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 aligns with the NeurIPS Code of Ethics, ensuring ethical standards are upheld throughout the study.

   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: [NA]

    Justification: Our work is foundational research, and therefore, it does not have direct societal impacts to discuss.

    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 does not involve the release of data or models that have a high risk for misuse.

    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: The paper provides proper credit to asset creators, citing relevant papers and explicitly mentioning license and terms of use. URLs are included where possible, and all licenses are respected.

    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: [Yes]

    Justification: The new assets introduced in the paper are well-documented, providing comprehensive details alongside the assets, including training procedures, licenses, limitations, and consent processes, ensuring transparency and reproducibility.

    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 or 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: [NA]

    Justification: LLMs were not used in the development of the core methods; their usage was limited to language polishing and did not influence the scientific contributions of this work.

    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.