Model Distribution-Aware Multimodal Dataset Distillation

Jiajun Shen* **Yonglin Wu*** **Ruonan Yu** **Xinchao Wang†**
National University of Singapore
{jiajun.shen,yonglin.wu,ruonan}@u.nus.edu
xinchao@nus.edu.sg

Abstract

Multimodal dataset distillation (MDD) aims to reduce the cost of vision-language pre-training by synthesizing compact image-text datasets. Although recent feature-matching methods have improved distillation efficiency, synthetic data may become specialized to the model representations encountered during optimization, limiting generalization to unseen model states. We propose **Model Distribution-Aware Multimodal Dataset Distillation (MDA-DD)**, which leverages supervision from a dynamic staggered queue of models at different training stages. We further introduce asymmetric dual projection for complementary alignment in visual and linguistic spaces and an uncentered second-order moment objective that retains mean-related cross-modal information. Experiments on Flickr30K and MS-COCO demonstrate consistent retrieval improvements over existing methods across synthetic-data budgets. With only 100 synthetic pairs, MDA-DD improves IR@1 and TR@1 on Flickr30K by 12.9 and 18.2 percentage points, respectively, over the strongest baseline at the same budget, outperforming all compared methods using 500 pairs on both datasets across all six retrieval metrics. Our code is available at https://github.com/Jun-Shen04-Jia/MDA-DD.

1 Introduction

Dataset distillation (DD) [40, 2, 44, 1, 51, 61] aims to compress large datasets into compact synthetic training sets. Contrastive vision-language models such as CLIP [31] and SigLIP [53, 37] rely heavily on web-scale image-text corpora, incurring substantial storage and computational costs. Although DD has proven effective across unimodal domains, extending it to image-text data requires preserving fine-grained semantic alignment between modalities.

Trajectory-based methods such as MTT-VL [43] and LoRS [46] address this issue by matching expert training trajectories to guide cross-modal alignment. RepBlend [54] further improves efficiency by freezing both encoders and matching projection-layer trajectories. However, generating expert trajectories still requires training on the full dataset and storing intermediate checkpoints, incurring substantial computational and storage costs. CovMatch [22] avoids expert trajectory generation by updating both encoders online and

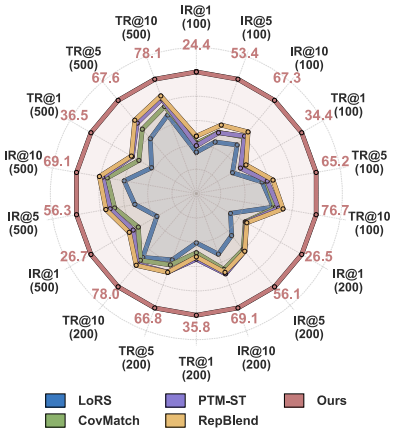


Figure 1: Radar charts comparing MDA-DD against all baselines on Flickr30K across retrieval metrics.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

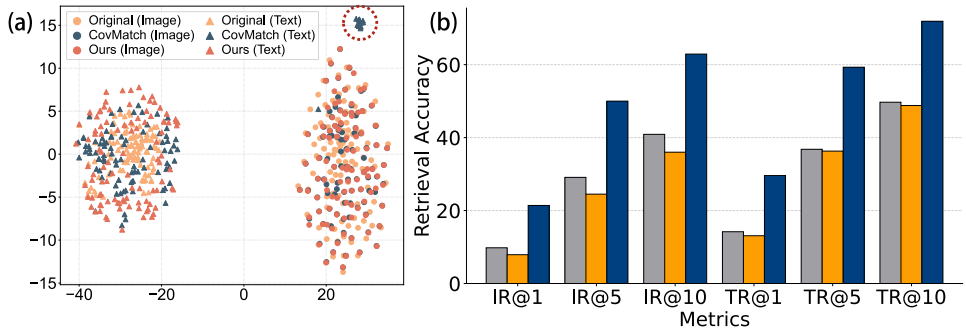


Figure 2: (a) Joint t-SNE visualization of multimodal embeddings under a different model initialization. Circles denote image embeddings and triangles denote text embeddings; colors indicate the source (**yellow:real data, cyan:CovMatch, orange:MDA-DD**). The **red dashed circle** highlights a subset of CovMatch text embeddings that collapses to an isolated region. (b) Retrieval performance of CovMatch variants on Flickr30K with 100 pairs. Removing the image projection **(yellow)** has minimal impact relative to the baseline **(grey)**, while random model initialization **(blue)** yields consistent gains across all metrics.

matching feature distributions between real and synthetic data, substantially reducing the cost of distillation.

Despite this progress, feature matching remains dependent on the model representations used during distillation. CovMatch periodically resets its encoders to the same pretrained weights before updating them online, which may cause synthetic data to specialize to the encountered model states and degrade downstream retrieval performance under different initializations. Figure 2(a) shows a subset of its text embeddings separated from the main clusters under a different model initialization, while Figure 2(b) shows that changing the initialization strategy improves retrieval performance.

Beyond model-state dependence, aligning both modalities in the visual feature space leaves complementary alignment in the linguistic space unexplored, while centered cross-covariance removes mean-related cross-modal information from the matched second-order statistic, potentially overlooking useful global cross-modal information.

We propose **Model Distribution-Aware Multimodal Dataset Distillation (MDA-DD)**, which uses a dynamic staggered model queue to provide supervision from models at different training stages. We further introduce **asymmetric dual projection (ADP)** for complementary alignment in visual and linguistic spaces and a **Second-Order Cross-Modal Moment (SOCM)** objective that retains mean-related cross-modal information without explicit centering.

Experiments on Flickr30K [30] and MS-COCO [24] show consistent improvements across synthetic-data budgets. With only 100 pairs, MDA-DD achieves 24.4 IR@1 and 34.4 TR@1 on Flickr30K, and 7.3 IR@1 and 10.5 TR@1 on MS-COCO. On both datasets, this 100-pair configuration outperforms all compared methods using 500 pairs across all six retrieval metrics.

Our main contributions are summarized as follows:

(1) We introduce a model-distribution-aware distillation framework with a dynamic staggered model queue that provides supervision from models at different training stages.

(2) We develop complementary cross-modal matching through asymmetric dual projection and an uncentered second-order moment objective that retains mean-related cross-modal information.

(3) We demonstrate substantial improvements in retrieval performance and data efficiency, as well as generalization to unseen visual encoder architectures.

2 Preliminaries and Related Work

We formalize multimodal dataset distillation for image-text contrastive learning and review existing distribution-based approaches (see Appendix A.3 for a discussion on trajectory-based methods).

Multimodal Dataset Distillation. Vision-language models embed image-text pairs into a joint representation space. Given a dataset $\mathcal{T} = \{(\mathbf{x}_v^i, \mathbf{x}_l^i)\}_{i=1}^M$, an image encoder f_v and a text encoder f_l, each followed by a linear projection head, produce multimodal embeddings $\mathbf{z}_v, \mathbf{z}_l \in \mathbb{R}^z$. The model is trained with the symmetric InfoNCE [29] loss:

$$\mathcal{L}_{\text{NCE}}(S) = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{B}\exp(S_{ij}/\tau)} + \log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{B}\exp(S_{ji}/\tau)}\right]. \tag{1}$$

where $s_{ij} = \cos(\mathbf{z}_v^i, \mathbf{z}_l^j)$ and $\tau > 0$ is a learnable temperature.

The goal of cross-modal dataset distillation is to compress $\mathcal{T}$ into a compact synthetic set $\mathcal{S} = \{(\tilde{\mathbf{x}}_v^i, \tilde{\mathbf{x}}_l^i)\}_{i=1}^N$ with $N \ll M$. Let $P_{\mathcal{D}}$ denote the distribution of the real data. We seek the synthetic set such that a model trained on $\mathcal{S}$ attains cross-modal alignment comparable to training on the full corpus $\mathcal{T}$, by minimizing the expected alignment gap under $P_{\mathcal{D}}$:

$$\mathcal{S}^* \;=\; \arg\min_{\mathcal{S}} \; \mathbb{E}_{(\mathbf{x}_v,\mathbf{x}_l)\sim P_{\mathcal{D}}}\big\|\, \mathcal{L}(\theta_{\mathcal{T}}) \;-\; \mathcal{L}(\theta_{\mathcal{S}})\,\big\|, \tag{2}$$

where $\theta_{\mathcal{T}}$ and $\theta_{\mathcal{S}}$ denote model parameters obtained by training on $\mathcal{T}$ and $\mathcal{S}$ respectively, and $\mathcal{L}$ is the cross-modal retrieval loss.

Distribution-Based Paradigms. CovMatch replaces costly trajectory unrolling with single-step distribution matching. By interleaving model training with data distillation, CovMatch uses the updated model directly as a feature extractor for both real and synthetic sets. Its distillation objective combines a shared projection design with cross-covariance matching.

Projection Design. The image encoder extracts embeddings $\mathbf{e}_v \in \mathbb{R}^{d_v}$, which are mapped to $\mathbf{z}_v = \mathbf{W}_v^{\updownarrow}\mathbf{e}_v \in \mathbb{R}^{d_v}$ through a projection matrix $\mathbf{W}_v^{\updownarrow} \in \mathbb{R}^{d_v \times d_v}$. Similarly, the text encoder produces embeddings $\mathbf{e}_l \in \mathbb{R}^{d_l}$ and maps them to $\mathbf{z}_l = \mathbf{W}_l^{\uparrow}\mathbf{e}_l \in \mathbb{R}^{d_v}$ through a projection matrix $\mathbf{W}_l^{\uparrow} \in \mathbb{R}^{d_v \times d_l}$. This projection scheme aligns text representations with the visual feature space.

Cross-Covariance Matching. Given a batch of B paired embeddings $\{\mathbf{e}_v^i, \mathbf{e}_l^i\}_{i=1}^B$, the batch means $\boldsymbol{\mu}_v$ and $\boldsymbol{\mu}_l$ are first computed to obtain centered embeddings $\bar{\mathbf{e}}_v^i = \mathbf{e}_v^i - \boldsymbol{\mu}_v$ and $\bar{\mathbf{e}}_l^i = \mathbf{e}_l^i - \boldsymbol{\mu}_l$. The cross-modal covariance matrix is then defined as:

$$\mathbf{C} \;=\; \frac{1}{B}\sum_{i=1}^{B}\bar{\mathbf{e}}_v^i\,(\bar{\mathbf{e}}_l^i)^\top \;=\; \frac{1}{B}\,\bar{\mathbf{E}}_v^\top\,\bar{\mathbf{E}}_l \;\in\; \mathbb{R}^{d_v \times d_l}, \tag{3}$$

where $\bar{\mathbf{E}}_v \in \mathbb{R}^{B\times d_v}$ and $\bar{\mathbf{E}}_l \in \mathbb{R}^{B\times d_l}$ stack the centered embeddings row-wise.

The distillation loss aligns the cross-covariance matrices and the projected means of each modality:

$$\mathcal{L}_{\text{total}} \;=\; \big\|\mathbf{C}_{\mathcal{S}} - \rho\,\mathbf{C}_{\mathcal{T}}\big\|_F^2 \;+\; \lambda\big(\|\bar{\mathbf{z}}_v^{\mathcal{S}} - \bar{\mathbf{z}}_v^{\mathcal{T}}\|_2^2 + \|\bar{\mathbf{z}}_l^{\mathcal{S}} - \bar{\mathbf{z}}_l^{\mathcal{T}}\|_2^2\big), \tag{4}$$

where ρ is a scaling factor, $\bar{\mathbf{z}}$ denotes the projected batch mean, and λ is a balancing weight. This scheme avoids generating and storing expert trajectories.

3 Method

3.1 Model-Distribution-Aware Distillation

Motivation. Feature-distribution matching depends on model representations that evolve during training, so synthetic data optimized along one trajectory may match real-data statistics less well under other model states. CovMatch [22] introduces model-state variation through online training and periodic reinitialization, but its encoders repeatedly return to the same pretrained weights. Synthetic-data updates are also guided by one active model at a time, rather than by models at different training stages. This may cause synthetic data to specialize to the states encountered during distillation and generalize poorly to others.

To characterize this limitation, let $\mathcal{T}$ and $\mathcal{S}$ denote the real and synthetic datasets, respectively, and let $\mathcal{L}_{\text{match}}(\theta; \mathcal{S}, \mathcal{T})$ denote their feature-matching discrepancy under model state θ, as defined in Eq. (13). Let $P_{\text{seq}}(\Theta \mid \phi_0)$ represent the model-state distribution covered by sequential online training cycles

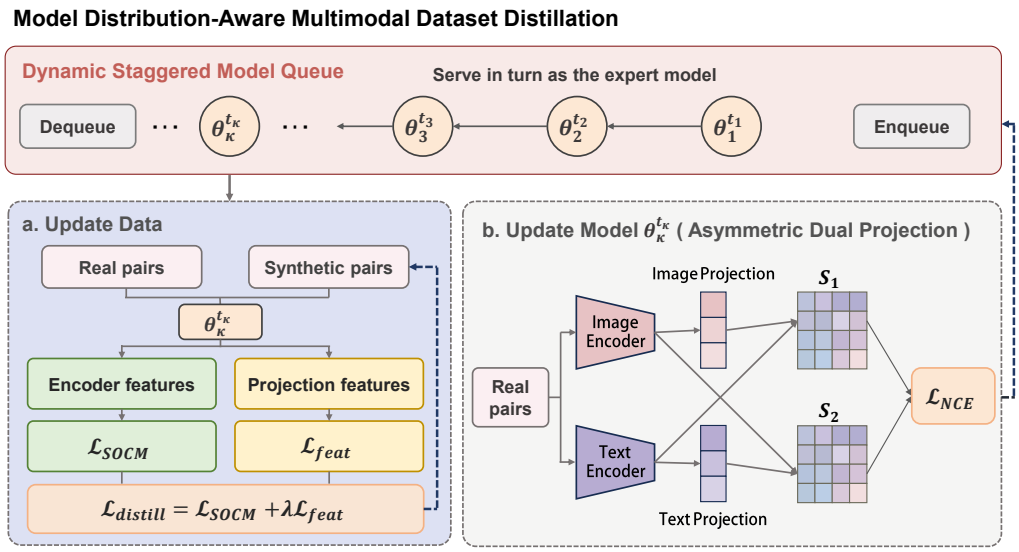


Figure 3: **Overview of MDA-DD.** A dynamic staggered model queue maintains expert models at different training stages. (a) **Update Data:** Synthetic pairs are optimized by matching encoder-level second-order cross-modal moments (SOCM) and projection-level feature means with real pairs. (b) **Update Model:** Each expert is trained on real pairs using asymmetric dual projections for bidirectional contrastive learning.

starting from a common pretrained encoder initialization ϕ_0, including their subsequent parameter updates. Let $P(\Theta)$ denote a broader target distribution of model states. We characterize potential model-state overfitting through the generalization gap

$$\Delta_{\text{state}}(\mathcal{S}) = \mathbb{E}_{\theta\sim P(\Theta)}\left[\mathcal{L}_{\text{match}}(\theta;\mathcal{S},\mathcal{T})\right] - \mathbb{E}_{\theta\sim P_{\text{seq}}(\Theta|\phi_0)}\left[\mathcal{L}_{\text{match}}(\theta;\mathcal{S},\mathcal{T})\right]. \tag{5}$$

A positive Δ_{state} indicates that synthetic data matches real-data statistics more closely on the states covered during distillation than on the broader target distribution. This motivates expanding model-state coverage beyond the repeatedly encountered training trajectories.

Distributional Objective. To reduce this sensitivity, we formulate multimodal dataset distillation as feature matching over a distribution of model states:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \ \mathbb{E}_{\theta\sim P(\Theta)}\Big[\big\|M_\theta(\mathcal{S}) - \rho M_\theta(\mathcal{T})\big\|_F^2 + \lambda\big(\|\bar{z}^{\mathcal{S}}_{v,\theta} - \bar{z}^{\mathcal{T}}_{v,\theta}\|_2^2 + \|\bar{z}^{\mathcal{S}}_{l,\theta} - \bar{z}^{\mathcal{T}}_{l,\theta}\|_2^2\big)\Big]. \tag{6}$$

Here, $M_\theta(\cdot)$ denotes the second-order cross-modal moment defined in Eq. (11), while $\bar{z}^{\mathcal{D}}_{v,\theta}$ and $\bar{z}^{\mathcal{D}}_{l,\theta}$ are the mean projected features of the visual and linguistic modalities for dataset $\mathcal{D}$. The coefficient ρ scales the real-data moment, and λ balances moment matching and feature-mean regularization.

Since the target model-state distribution is generally inaccessible, we approximate its supervision using K concurrently maintained online models at different training stages. At iteration t, the active queue $\mathcal{Q}_t = \{\theta_k\}_{k=1}^{K_t}$ induces the empirical matching objective

$$\widehat{\mathcal{L}}_{\text{distill}}(\mathcal{S};\mathcal{Q}_t) = \frac{1}{K_t}\sum_{k=1}^{K_t}\Big[\big\|M_{\theta_k}(\mathcal{S}) - \rho M_{\theta_k}(\mathcal{T})\big\|_F^2 + \lambda\big(\|\bar{z}^{\mathcal{S}}_{v,k} - \bar{z}^{\mathcal{T}}_{v,k}\|_2^2 + \|\bar{z}^{\mathcal{S}}_{l,k} - \bar{z}^{\mathcal{T}}_{l,k}\|_2^2\big)\Big], \tag{7}$$

where $K_t = |Q_t| \le K$ denotes the number of active models. The queue provides a stage-stratified approximation that covers different training states without treating its models as independent samples from the target distribution.

Dynamic Staggered Model Queue. As illustrated in Figure 3, we maintain a queue of at most K online models, each with a maximum lifetime of $T_{\max}$ training steps on real data. New models share the pretrained encoder initialization but use independently initialized projection heads and separate training randomness. Each model evolves through online contrastive training, with its age defined by the number of real-data updates since initialization.

To maintain coverage across training stages, we introduce new models at intervals of $\Delta t = T_{\max}/K$ and retire the oldest model when it reaches its maximum lifetime. During the initial warm-up period, the queue is gradually populated. Once the queue reaches its steady state, the active models span different training ages. Let $a_{(k)}(t)$ denote the age of the k-th model ordered from youngest to oldest. The steady-state ages satisfy

$$\mathcal{Q}_t = \{\theta_k^{(t)}\}_{k=1}^{K}, \qquad a_{(k)}(t) = (k-1)\Delta t + t', \qquad t' = t \bmod \Delta t, \tag{8}$$

where the iteration index is aligned with the model replacement cycle. Consecutive models are therefore separated by Δt training steps, providing simultaneous supervision from early, intermediate, and later stages of model optimization.

At each distillation iteration, the active models take turns updating the synthetic dataset. For each model θ_k, we first extract encoder-level cross-modal moments and projection-level feature means from real and synthetic image-text batches. Keeping θ_k fixed, we treat the real-data statistics as targets and backpropagate the matching loss in Eq. (13) only through the synthetic branch to update $\mathcal{S}$. The model is then advanced by one contrastive-learning step on real data using the asymmetric dual projection objective (§3.2). After all active models have been processed, the queue is updated by retiring expired models and introducing new ones according to the staggered schedule.

Following Algorithm 1, synthetic-data updates are performed sequentially across the active models, providing an incremental optimization of the multi-state objective in Eq. (7). By exposing synthetic pairs to multiple evolving feature geometries within each distillation iteration, the dynamic queue broadens model-state supervision without requiring stored expert trajectories.

3.2 Asymmetric Dual Projection

Limitations of Existing Projection Design. CovMatch aligns both modalities in the visual feature space $\mathbb{R}^{d_v}$ by projecting image embeddings with $W_v \in \mathbb{R}^{d_v \times d_v}$ and text embeddings with $W_l \in \mathbb{R}^{d_v \times d_l}$. These projections produce a single similarity matrix S for contrastive learning. This design has two limitations: the same-dimensional image projection W_v may be absorbed into the trainable image encoder's final layer, while text embeddings are projected into the visual space without a corresponding projection in the reverse direction. Consequently, cross-modal alignment is restricted to the visual space, leaving complementary alignment in the linguistic space unexplored.

Dual Projection. We replace the redundant same-dimension image projection with two asymmetric cross-dimensional matrices, $\mathbf{W}v^{\downarrow} \in \mathbb{R}^{d_l \times d_v}$ and $\mathbf{W}l^{\uparrow} \in \mathbb{R}^{d_v \times d_l}$, which project image and text embeddings into the feature space of the other modality:

$$\mathbf{z}_v = \mathbf{W}_v^{\downarrow}\,\mathbf{e}_v \in \mathbb{R}^{d_l}, \qquad \mathbf{z}_l = \mathbf{W}_l^{\uparrow}\,\mathbf{e}_l \in \mathbb{R}^{d_v}. \tag{9}$$

Neither projection is redundant: $\mathbf{W}_v^{\downarrow}$ genuinely reduces dimensionality and cannot be absorbed by the image encoder, while $\mathbf{W}_l^{\uparrow}$ performs the complementary expansion.

Dual-Space Similarity. The two projected representations produce two cross-modal similarity matrices in different feature spaces. Within the text space $\mathbb{R}^{d_l}$, projected image embeddings are compared with original text embeddings, whereas within the visual space $\mathbb{R}^{d_v}$, original image embeddings are compared with projected text embeddings:

$$\mathbf{S}_1^{(ij)} = \cos\big(\mathbf{z}_v^i,\ \mathbf{e}_l^j\big), \qquad \mathbf{S}_2^{(ij)} = \cos\big(\mathbf{e}_v^i,\ \mathbf{z}_l^j\big). \tag{10}$$

$\mathbf{S}_1$ quantifies how effectively visual content is represented in the linguistic space, whereas $\mathbf{S}_2$ quantifies how reliably textual content is grounded in the visual space. Together, they provide complementary alignment signals through the two projection directions, rather than relying on a single feature space. We then apply the symmetric InfoNCE loss in Eq. 1 to each similarity matrix and average the resulting losses.

3.3 Second-Order Cross-Modal Moment

From Covariance to Raw Moment. The cross-modal covariance in CovMatch (Eq. 3) requires explicit mean centering before computing the outer product. We instead match the raw second-order cross-modal moment, avoiding explicit centering.

Algorithm 1 MDA-DD: Model-Distribution-Aware Multimodal Dataset Distillation

Require: Real dataset $\mathcal{T}$, synthetic size N, iterations T_{iter}, queue size K, max lifetime $T_{\max}$, learning rates $\eta_{\mathcal{S}}, \eta_\theta$
Ensure: Synthetic dataset $\mathcal{S} = \{(\tilde{\mathbf{x}}_v^i, \tilde{\mathbf{x}}_l^i)\}_{i=1}^N$

```
Initialize S by sampling N pairs from T
Initialize queue Q = {θ_1} with t_1 = 0 {t_k: training steps of θ_k}
for t = 1 to T_iter do
    for each θ_k ∈ Q do
        // Compute real-data statistics (frozen)
        Sample B∼T; compute M_{θ_k}(B) via Eq. 11 and z̄_v^T, z̄_l^T
        // Compute synthetic-data statistics (with gradients)
        Sample B̃∼S; compute M_{θ_k}(B̃) and z̄_v^S, z̄_l^S
        // Update synthetic data
        Compute L_total via Eq. 13;  S ← S − η_S ∇_S L_total
        // Advance model on real data
        Sample B'∼T;  θ_k ← θ_k − η_θ ∇_{θ_k} L_NCE(θ_k, B');  t_k ← t_k + 1
    end for
    // Queue maintenance
    if max_k t_k ≥ T_max then
        Remove oldest θ_k from Q
    end if
    if |Q| < K and t mod ⌊T_max/K⌋ = 0 then
        Enqueue freshly initialized θ_new with t_new = 0
    end if
end for
```

Given a dataset $\mathcal{D} = \{(\mathbf{x}_v^i, \mathbf{x}_l^i)\}_{i=1}^n$ and model parameters θ, let $\mathbf{e}_v^i = f_v(\mathbf{x}_v^i; \theta_v) \in \mathbb{R}^{d_v}$ and $\mathbf{e}_l^i = f_l(\mathbf{x}_l^i; \theta_l) \in \mathbb{R}^{d_l}$ denote the encoder outputs. The Second-Order Cross-Modal Moment (SOCM) is defined as:

$$\mathbf{M}_\theta(\mathcal{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_v^i \, \mathbf{e}_l^{i\top} = \frac{1}{n}\mathbf{E}_v^\top \mathbf{E}_l \in \mathbb{R}^{d_v \times d_l}, \tag{11}$$

where $\mathbf{E}_v \in \mathbb{R}^{n \times d_v}$ and $\mathbf{E}_l \in \mathbb{R}^{n \times d_l}$ stack the encoder outputs row-wise. Its relationship to the cross-modal covariance $\mathbf{C}$ (Eq. 3) is:

$$\mathbf{M}_\theta(\mathcal{D}) = \underbrace{\mathbf{C}}_{\text{covariance}} + \underbrace{\boldsymbol{\mu}_v \boldsymbol{\mu}_l^\top}_{\text{mean outer product}}, \tag{12}$$

where $\boldsymbol{\mu}_v$ and $\boldsymbol{\mu}_l$ denote the batch means. SOCM therefore captures both the centered cross-modal covariance and the outer product of the modality means within a single matrix. Computing the raw moment directly avoids the mean-centering and element-wise subtraction required by covariance matching, reducing the computational cost of the second-order matching term.

Complete Distillation Loss. We pair SOCM matching with a per-modality mean constraint at the projection level. The full objective is:

$$\mathcal{L}_{\text{total}} = \underbrace{\left\|\mathbf{M}_\theta(\mathcal{S}) - \rho\,\mathbf{M}_\theta(\mathcal{T})\right\|_F^2}_{\mathcal{L}_{\text{SOCM}}} + \lambda \underbrace{\left(\left\|\bar{\mathbf{z}}_v^{\mathcal{S}} - \bar{\mathbf{z}}_v^{\mathcal{T}}\right\|_2^2 + \left\|\bar{\mathbf{z}}_l^{\mathcal{S}} - \bar{\mathbf{z}}_l^{\mathcal{T}}\right\|_2^2\right)}_{\mathcal{L}_{\text{feat}}}, \tag{13}$$

where $\bar{\mathbf{z}}_v^{\mathcal{D}}$ and $\bar{\mathbf{z}}_l^{\mathcal{D}}$ are the batch-wise mean projected features, ρ is a scaling coefficient, and λ balances the two terms. $\mathcal{L}_{\text{SOCM}}$ captures second-order cross-modal statistics at the encoder level, while $\mathcal{L}_{\text{feat}}$ matches the mean features of each modality at the projection level.

Table 1: Cross-modal retrieval on **Flickr30K** (NFNet+BERT). The model trained on full dataset achieves: IR@1 = 47.8, IR@5 = 79.0, IR@10 = 87.4, TR@1 = 61.5, TR@5 = 85.1, TR@10 = 92.1.

Pairs	Ratio	Metric	Coreset selection				Dataset distillation				
			Rand	Herd	K-Cent	Forget	LoRS	CovMatch	RepBlend	PTM-ST	**Ours**
100	0.3%	IR@1	1.0	0.7	0.7	0.7	8.3±0.2	10.1±0.2	11.5±0.4	9.6±0.3	**24.4±0.6**
		IR@5	4.0	2.8	3.1	2.4	24.1±0.2	28.6±0.4	32.0±0.7	28.4±0.5	**53.4±0.7**
		IR@10	6.5	5.3	6.1	5.6	35.1±0.3	40.9±0.6	44.5±0.6	41.5±0.5	**67.3±0.7**
		TR@1	1.3	1.1	0.6	1.2	11.8±0.2	14.8±0.9	16.2±0.8	14.4±0.4	**34.4±0.7**
		TR@5	5.9	4.7	5.0	4.2	35.8±0.6	38.0±0.4	41.7±0.9	38.6±0.8	**65.2±0.9**
		TR@10	10.1	7.9	7.6	9.7	49.2±0.5	50.6±0.6	55.5±0.4	52.7±0.7	**76.7±0.7**
200	0.7%	IR@1	1.1	1.5	1.5	1.2	8.6±0.3	12.3±0.4	12.7±0.8	12.5±0.1	**26.5±0.5**
		IR@5	4.8	5.5	5.4	3.1	25.3±0.2	33.6±0.3	34.7±0.6	34.7±0.2	**56.1±0.6**
		IR@10	9.2	9.3	9.9	8.4	36.6±0.3	45.8±0.2	47.6±0.5	48.5±0.2	**69.1±0.7**
		TR@1	2.1	2.3	2.2	1.5	14.5±0.5	17.4±0.5	18.6±0.7	19.3±0.2	**35.8±0.8**
		TR@5	8.7	8.4	8.2	8.4	38.7±0.5	41.7±0.5	46.0±0.8	45.9±0.4	**66.8±0.6**
		TR@10	13.2	14.4	13.5	10.2	53.4±0.5	55.8±0.5	60.0±0.6	58.9±0.5	**78.0±0.8**
500	1.7%	IR@1	2.4	3.0	3.5	1.8	10.0±0.2	14.7±0.3	17.0±0.6	16.0±0.2	**26.7±1.2**
		IR@5	10.5	10.0	10.4	9.0	28.9±0.7	38.4±0.4	42.5±0.5	40.5±0.3	**56.3±1.5**
		IR@10	17.4	17.0	17.3	15.9	41.6±0.6	51.4±0.3	55.9±0.6	54.0±0.3	**69.1±1.2**
		TR@1	5.2	5.1	4.9	3.6	15.5±0.7	19.9±0.6	22.5±0.4	22.2±0.6	**36.5±1.1**
		TR@5	18.3	16.4	16.4	12.3	39.8±0.4	46.7±0.9	53.2±0.3	51.1±0.3	**67.6±1.3**
		TR@10	25.7	24.3	23.3	19.3	53.7±0.3	59.5±0.7	66.7±0.3	64.6±0.3	**78.1±1.0**

4 Experiments

4.1 Experimental Setup

We evaluate on two standard vision-language benchmarks: Flickr30K [30] ($\sim$31K images) and MS-COCO [24] ($\sim$123K images), each with five human-annotated captions per image. Using the Karpathy split [19], Flickr30K follows a 29K/1K/1K train/val/test partition, while MS-COCO follows 113K/5K/5K, with results reported on the full 5K test set. Cross-modal candidates are ranked by cosine similarity within the shared embedding space. We report Recall@K ($K = 1, 5, 10$) for both image-to-text retrieval (TR@K) and text-to-image retrieval (IR@K). Following prior work [43, 46], we use an ImageNet-pretrained NFNet [5] as the image encoder and a pretrained BERT-base [10] as the text encoder. Each encoder is equipped with the asymmetric dual projection layer (§3.2) to map features into two complementary shared spaces for bidirectional alignment.

Baselines. We compare with two families of methods. (i) Coreset selection: Random Sampling, Herding [42], K-Center [15], and Forgetting [36], all adapted to the vision-language setting following prior work. (ii) Dataset distillation: LoRS [46], RepBlend [54], CovMatch [22], and PTM-ST [16].

Implementation Details. We report the mean and standard deviation across 5 independent runs. All experiments are conducted on a single NVIDIA RTX 5090 (32 GB). Detailed hyperparameters and optimization settings are provided in the appendix.

4.2 Main Results

Tables 1 and 2 report cross-modal retrieval results on Flickr30K and MS-COCO under synthetic budgets of 100, 200, and 500 pairs. MDA-DD ranks first on every budget–metric combination on both benchmarks. At 100 pairs, MDA-DD reaches 24.4 IR@1 and 34.4 TR@1 on Flickr30K (0.3% of training data) and 7.3 IR@1 and 10.5 TR@1 on MS-COCO (0.8‰), more than doubling the strongest baseline on Flickr30K and outperforming it by +3.2 IR@1 and +5.3 TR@1 on MS-COCO. On both datasets, this 100-pair configuration already surpasses all baselines operating at 500 pairs, yielding an effective 5$\times$ reduction in synthetic data requirement.

Table 2: Cross-modal retrieval on **MS-COCO** (NFNet+BERT). The model trained on full dataset achieves: IR@1 = 24.9, IR@5 = 53.7, IR@10 = 67.5, TR@1 = 32.5, TR@5 = 61.9, TR@10 = 74.0.

Pairs	Ratio	Metric	Coreset selection				Dataset distillation				
			Rand	Herd	K-Cent	Forget	LoRS	CovMatch	RepBlend	PTM-ST	**Ours**
100	0.8‰	IR@1	0.3	0.5	0.4	0.3	1.8±0.1	2.8±0.1	4.1±0.3	2.3±0.1	**7.3**±0.2
		IR@5	1.3	1.4	1.4	1.5	7.1±0.2	10.5±0.2	13.9±0.8	9.0±0.1	**22.6**±0.4
		IR@10	2.7	3.5	2.5	2.5	12.2±0.2	17.7±0.3	22.3±0.5	15.6±0.3	**34.4**±0.6
		TR@1	0.8	0.8	1.4	0.7	3.3±0.2	3.8±0.1	5.2±0.5	4.1±0.2	**10.5**±0.2
		TR@5	3.0	2.1	3.7	2.6	12.2±0.3	13.1±0.3	17.9±0.9	13.4±0.2	**28.6**±0.6
		TR@10	5.0	4.9	5.5	4.8	19.6±0.3	21.1±0.2	28.0±0.3	22.0±0.2	**41.8**±0.7
200	1.7‰	IR@1	0.6	0.9	0.7	0.6	2.4±0.1	3.8±0.1	6.1±0.8	3.9±0.1	**8.3**±0.2
		IR@5	2.3	2.4	2.1	2.8	9.3±0.2	13.4±0.1	19.3±0.7	13.7±0.2	**25.6**±0.4
		IR@10	4.4	4.1	5.8	4.9	15.5±0.2	21.8±0.2	29.8±0.5	22.2±0.2	**38.4**±0.5
		TR@1	1.0	1.0	1.2	1.1	4.3±0.1	5.3±0.2	6.9±0.6	5.7±0.2	**12.1**±0.3
		TR@5	4.0	3.6	3.8	3.5	14.2±0.3	17.3±0.2	21.8±0.9	18.2±0.3	**32.6**±0.5
		TR@10	7.2	7.7	7.5	7.0	22.6±0.2	27.0±0.2	32.3±0.7	27.8±0.3	**45.6**±0.8
500	4.4‰	IR@1	1.1	1.7	1.1	0.8	2.8±0.2	5.4±0.1	6.2±0.1	6.6±0.1	**9.5**±0.4
		IR@5	5.0	5.3	6.3	5.8	9.9±0.5	18.0±0.1	19.9±0.3	20.5±0.2	**27.5**±0.7
		IR@10	8.7	9.9	10.5	8.2	16.5±0.7	28.2±0.1	30.6±0.1	30.7±0.2	**40.3**±0.8
		TR@1	1.9	1.9	2.5	2.1	5.3±0.5	8.1±0.3	7.0±0.2	6.9±0.3	**13.0**±0.6
		TR@5	7.5	7.8	8.7	8.2	18.3±1.5	23.5±0.3	22.0±0.3	20.1±0.2	**34.7**±1.1
		TR@10	12.5	13.7	14.3	13.0	27.9±1.4	34.6±0.6	32.9±0.6	30.0±0.3	**48.1**±1.1

Table 3: Scaling performance on Flickr30K and MS-COCO with larger synthetic-data budgets (NFNet+BERT). The real-data baseline uses equally sized subsets sampled from the original training data. mR is the mean of six retrieval metrics (%).

Dataset	Pairs	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	mR
Flickr30K	5K	Real Data	26.6	57.4	69.6	34.3	64.3	76.5	54.8
		CovMatch	24.0	53.9	64.9	32.1	61.7	73.3	51.7
		MDA-DD	**31.7**	**61.7**	**73.0**	**43.2**	**73.5**	**82.1**	**60.9**
MS-COCO	5K	Real Data	11.9	32.2	45.4	15.3	38.3	51.0	32.4
		CovMatch	10.8	29.7	42.8	15.1	36.8	50.4	30.9
		MDA-DD	**12.9**	**34.6**	**47.6**	**19.3**	**44.7**	**58.5**	**36.3**

4.3 Scaling with Synthetic Data Budget

We evaluate MDA-DD with 5,000 synthetic pairs on Flickr30K and MS-COCO to examine its effectiveness at larger data budgets. As shown in Table 3, increasing the budget from 500 to 5,000 pairs improves mR by 5.2 and 7.4 points, respectively. However, CovMatch remains below equally sized real-data subsets across all six retrieval metrics on both datasets. This suggests that increasing the synthetic budget alone may not overcome the limitations of feature matching along repeatedly sampled model trajectories. In contrast, MDA-DD incorporates supervision from models at different training stages through the dynamic staggered queue, while ADP and SOCM provide complementary cross-modal matching constraints. At 5,000 pairs, MDA-DD surpasses the real-data baselines across all six metrics, achieving 6.1 and 3.9 higher mR points on Flickr30K and MS-COCO, respectively.

4.4 Ablation Study

Component-Wise Contribution. Figure 4 evaluates ADP, DQ, and SOCM on Flickr30K across three synthetic-data budgets. DQ provides the largest retrieval gains, with ADP yielding further improvements. With DQ enabled and ADP disabled, replacing centered covariance (COV) with SOCM changes mean Recall (mR) by $+0.10$, -0.23, and -0.13 points at 100, 200, and 500 pairs, respectively. In a separate timing comparison with DQ and ADP fixed, SOCM reduces moment-matching computation by 1.8% and full iteration time by 0.3% at 500 pairs, while maintaining comparable retrieval performance.

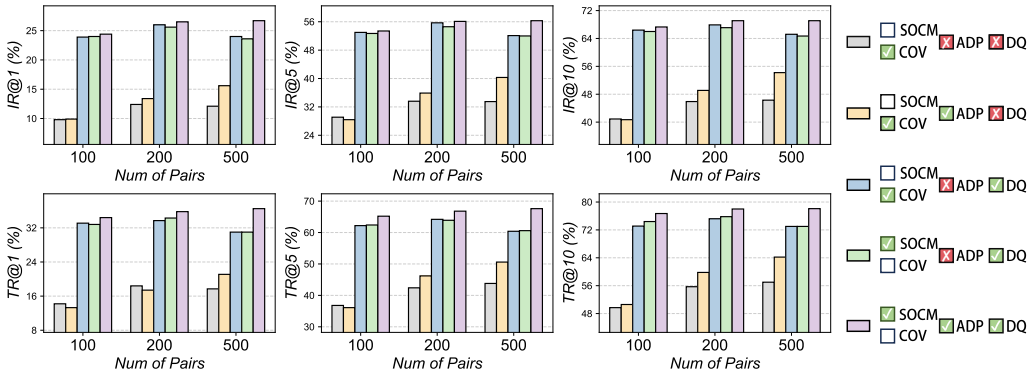


Figure 4: Component-wise ablation on Flickr30K with NFNet+BERT. We compare centered covariance (COV) and SOCM under different configurations of asymmetric dual projection (ADP) and dynamic model queue (DQ).

Table 4: Effects of asymmetric dual projection (ADP) during distillation and downstream training on Flickr30K with 500 synthetic pairs. Original denotes the CovMatch projection design. Results use the controlled ablation setting, and mR averages the six retrieval metrics.

Distillation	Downstream	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	mR
Original	Original	12.1	33.5	46.3	17.7	43.8	57.0	35.1
Original	ADP	15.3	39.1	52.1	18.3	49.3	62.5	39.4
ADP	Original	15.4	37.7	49.8	20.8	45.9	58.2	38.0
ADP	ADP	**15.6**	**40.3**	**54.2**	**21.1**	**50.6**	**64.2**	**41.0**

Distillation-Stage Effect of ADP. To distinguish improvements in synthetic-data quality from downstream architectural gains, we independently vary the projection designs during distillation and downstream training. As shown in Table 4, introducing ADP during distillation improves mR from 35.1 to 38.0 and from 39.4 to 41.0 under the two fixed downstream architectures, respectively. These gains across all six retrieval metrics indicate that ADP benefits the distillation process itself, rather than merely providing a stronger downstream model. By introducing alignment in both visual and linguistic spaces, ADP changes the online model training and projected feature-mean constraints that guide synthetic-data optimization. Moreover, under the same ADP architecture, MDA-DD achieves 55.7 mR compared with 41.0 for CovMatch, showing that its performance advantage extends beyond the projection design.

4.5 Cross-Architecture Generalization.

Distribution-matching methods align feature-level statistics rather than architecture-specific parameter trajectories, making distilled data less tied to the encoder architecture. Following prior work, we distill with NFNet+BERT and evaluate on alternative image encoders, including ResNet-50 [4] and RegNet [45], with the text encoder held fixed. As detailed in Table 6, MDA-DD consistently outperforms the compared baselines across both architectures.

4.6 Computational Efficiency

Table 5 compares the computational costs of trajectory-based and distribution-matching methods. RepBlend incurs substantial overhead from generating and storing expert trajectories, whereas CovMatch and MDA-DD avoid this checkpoint storage. Notably, despite maintaining multiple online models at different training stages, MDA-DD achieves shorter distillation time than CovMatch. Thus, broader model-state supervision need not increase distillation cost.

Table 5: Computational cost on MS-COCO with 100 synthetic pairs (NVIDIA RTX 5090, 32 GB).

Method	Expert		Distill
	storage	time	time
RepBlend	16.6 GB	78.6 h	1.5 h
CovMatch	0	2.0 h	2.8 h
MDA-DD	0	2.0 h	1.3 h

Table 6: Cross-modal retrieval performance under alternative image encoder architectures on Flickr30K with 500 synthetic pairs. Results are reported as the mean and standard deviation over 5 independent runs.

Backbone	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet+BERT	LoRS	3.3±0.2	12.7±0.3	20.4±0.2	6.8±0.2	19.6±1.3	31.1±0.3
	RepBlend	4.2±0.2	14.1±0.2	23.6±0.6	8.4±0.2	23.1±0.8	35.0±1.3
	CovMatch	10.6±0.8	29.0±1.3	41.0±1.2	13.4±1.3	34.2±0.9	46.2±1.4
	Ours	**14.5**±2.2	**36.4**±4.6	**49.1**±5.2	**17.8**±2.2	**42.2**±4.5	**54.1**±5.3
RegNet+BERT	LoRS	3.5±0.1	12.6±0.3	21.1±0.4	6.8±0.3	20.8±0.3	30.2±0.3
	RepBlend	3.9±0.2	13.9±0.3	24.0±0.6	7.9±0.3	24.2±0.3	36.2±1.1
	CovMatch	10.6±0.3	29.5±0.7	41.0±0.8	13.9±0.7	33.5±1.0	44.8±0.5
	Ours	**16.1**±2.1	**40.6**±3.2	**53.8**±3.4	**21.3**±2.6	**46.5**±2.8	**59.6**±2.9

Table 7: Cross-modal retrieval on LLaVA-CC3M with NFNet+BERT (%). The full-data reference achieves IR@1/5/10 = 23.77/50.95/62.84 and TR@1/5/10 = 24.66/50.82/63.58.

Pairs	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
200	LoRS	1.56	5.33	8.77	1.01	4.04	6.81
	RepBlend	3.53	12.38	19.33	4.39	13.45	20.18
	CovMatch	2.04	6.95	12.31	2.67	9.46	14.93
	MDA-DD	**7.92**	**22.29**	**32.13**	**10.21**	**28.54**	**39.62**
500	LoRS	1.96	6.72	10.55	1.41	5.11	8.51
	RepBlend	4.45	15.09	22.31	5.14	14.89	22.97
	CovMatch	3.00	10.06	16.07	3.83	13.04	20.14
	MDA-DD	**8.88**	**24.76**	**35.09**	**10.23**	**28.99**	**40.06**

4.7 Beyond Curated Benchmarks

We further evaluate MDA-DD on LLaVA-CC3M, using 334K image-text pairs for distillation and a disjoint 10K-pair evaluation set, with identical backbones and optimization settings across methods. As shown in Table 7, MDA-DD outperforms all baselines across six retrieval metrics at both 200 and 500 pairs. Notably, with only 200 pairs, MDA-DD surpasses all 500-pair baselines, suggesting that its multi-state supervision and cross-modal matching design can better utilize limited synthetic data capacity on larger web-sourced corpora.

5 Conclusion

We propose MDA-DD, which extends feature matching across models at different training stages through a dynamic staggered queue. Combined with asymmetric dual projection and uncentered second-order cross-modal moment matching, MDA-DD outperforms existing methods on Flickr30K and MS-COCO across synthetic-data budgets. The distilled datasets also achieve improved retrieval performance under alternative visual encoder architectures.

Limitations and Future Work. Despite the strong performance of MDA-DD, current MDD frameworks, including our framework, rely heavily on CLIP-style dual-encoder architectures, which limits their direct applicability to emerging native MLLMs. The dependence on explicit visual encoders also hinders the transferability of distilled datasets to unified autoregressive generation paradigms. In future work, we plan to explore architecture-agnostic distillation strategies for native MLLMs.

Acknowledgement

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00), the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002).

References

[1] Deyu Bo, Songhua Liu, and Xinchao Wang. Understanding Dataset Distillation via Spectral Filtering. In *International Conference on Learning Representations*, 2026.

[2] Deyu Bo and Xinchao Wang. Beyond soft label: Dataset distillation via orthogonal gradient matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[3] Deyu Bo and Xinchao Wang. Efficient multi-modal dataset distillation via analytic parameter matching. In *International Conference on Machine Learning*, 2026.

[4] Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. In *International Conference on Learning Representations*.

[5] 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.

[6] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4750–4759, 2022.

[7] Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. DMax: Aggressive Parallel Decoding for dLLMs. In *Neural Information Processing Systems*, 2026.

[8] Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. dParallel: Learnable Parallel Decoding for dLLMs. In *International Conference on Learning Representations*, 2026.

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

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[11] Guodong Ding, Rongyu Chen, and Angela Yao. Condensing action segmentation datasets via generative network inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17733–17742, 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 *International Conference on Learning Representations*.

[13] Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3758, 2023.

[14] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

[16] Shengbin Guo, Hang Zhao, Senqiao Yang, Chenyang Jiang, Yuhang Cheng, Xiangru Peng, Rui Shao, and Zhuotao Tian. Multimodal dataset distillation via phased teacher models. In *The Fourteenth International Conference on Learning Representations*.

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[18] Wei Jin, Lingxiao Zhao, Shichang Zhang, Yozen Liu, Jiliang Tang, and Neil Shah. Graph condensation for graph neural networks. In *International Conference on Learning Representations*.

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[20] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.

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

[22] Yongmin Lee and Hye Won Chung. Covmatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[23] Zhe Li, Hadrien Reynaud, Mischa Dombrowski, Sarah Cechnicka, Franciskus Xaverius Erick, and Bernhard Kainz. Video dataset condensation with diffusion models. 2025.

[24] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[26] Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Unidetox: Universal detoxification of large language models via dataset distillation. *arXiv preprint arXiv:2504.20500*, 2025.

[27] Aru Maekawa, Naoki Kobayashi, Kotaro Funakoshi, and Manabu Okumura. Dataset distillation with attention labels for fine-tuning bert. *Journal of Natural Language Processing*, 32(1):283–299, 2025.

[28] 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.

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

[30] 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.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[32] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17097–17107, 2023.

[33] Seungjae Shin, Heesun Bae, Donghyeok Shin, Weonyoung Joo, and Il-Chul Moon. Loss-curvature matching for dataset selection and condensation. In *International Conference on Artificial Intelligence and Statistics*, pages 8606–8628. PMLR, 2023.

[34] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9390–9399, 2024.

[35] Yefan Tao, Luyang Kong, Andrey Kan, and Laurent Callot. Textual dataset distillation via language model embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12557–12569, 2024.

[36] 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. In *International Conference on Learning Representations*.

[37] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022.

[40] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation.

[41] 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.

[42] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pages 1121–1128, 2009.

[43] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transactions on Machine Learning Research*.

[44] Lingao Xiao, Songhua Liu, Yang He, and Xinchao Wang. Unifying dataset pruning and distillation for efficient large-scale compression. In *International Conference on Machine Learning*, 2026.

[45] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification. *IEEE transactions on neural networks and learning systems*, 34(11):9562–9567, 2022.

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

[47] Beining Yang, Kai Wang, Qingyun Sun, Cheng Ji, Xingcheng Fu, Hao Tang, Yang You, and Jianxin Li. Does graph distillation see like vision dataset counterpart? *Advances in Neural Information Processing Systems*, 36:53201–53226, 2023.

[48] Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*.

[49] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36:73582–73603, 2023.

[50] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE transactions on pattern analysis and machine intelligence*, 46(1):150–170, 2023.

[51] Ruonan Yu, Songhua Liu, and Xinchao Wang. Flash-dd: An ultra parameter-efficient approach to dataset distillation. In *European Conference on Computer Vision*, 2026.

[52] Ruonan Yu, Songhua Liu, Jingwen Ye, and Xinchao Wang. Teddy: Efficient large-scale dataset distillation via taylor-approximated matching. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[53] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[54] Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representation blending for multimodal dataset distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[55] 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*.

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

[57] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6514–6523, 2023.

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

[59] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7856–7865. IEEE Computer Society, 2023.

[60] Yinjie Zhao, Heng Zhao, Bihan Wen, Yew-Soon Ong, and Joey Tianyi Zhou. Video set distillation: Information diversification and temporal densification. *arXiv preprint arXiv:2412.00111*, 2024.

[61] Letian Zhou, Songhua Liu, and Xinchao Wang. CoDA: From Text-to-Image Diffusion Models to Training-Free Dataset Distillation. In *International Conference on Learning Representations*, 2026.

A More Related Work

A.1 Vision-Language Pre-Training

Our work distills the training data of dual-encoder vision-language models, a paradigm popularized by CLIP [31] and ALIGN [17]. In this architecture, an image encoder and a text encoder independently produce embeddings that are aligned via the InfoNCE objective [29] in a shared space. SigLIP [53] replaces the softmax-based contrastive loss with a per-pair sigmoid loss, improving batch efficiency. All existing multimodal dataset distillation methods [43, 46, 54, 22, 16] operate under this dual-encoder setting.

A.2 Dataset Distillation

Dataset distillation synthesizes a compact training set whose learning signal approximates that of the full dataset [50]. Early meta-learning formulations [40] optimize synthetic pixels through bi-level optimization but are computationally prohibitive.

Gradient matching (DC [58], DCC [21], DSA [56], LCMat [33]) sidesteps the inner loop by aligning single-step gradients, while trajectory matching (MTT [6], TESLA [9], FTD [13], SeqMatch [14]) extends this to multi-step parameter trajectories at the cost of storing dense expert checkpoints. Teddy [52] further improves distillation efficiency through Taylor-approximated matching and a pre-cached pool of weak models. Distribution matching (DM [57], DataDAM [32], CAFE [39], IDM [59], RDED [34], SRe^{2}L [49], CDA [48]) takes a different route, directly minimizing the discrepancy between real and synthetic feature distributions. Beyond image-centric settings, dataset distillation has also been studied for text [26, 28], video [41, 11], and graph data [25, 55], where recent studies synthesize compact data by exploiting domain-specific structures, including linguistic generation [35, 27, 8, 7], temporal dynamics [60, 23], and graph spectra [47, 18]. These advances indicate that dataset distillation is evolving from a vision-oriented technique into a general data compression framework, further motivating its extension to multimodal learning. Among existing multimodal efforts, MTT-VL [43] adapts trajectory matching to vision-language data; by contrast, distribution matching provides a more favorable efficiency–quality trade-off and serves as the methodological basis of CovMatch [22] and our method. We follow this paradigm but address challenges specific to multimodal settings: cross-modal alignment between heterogeneous feature spaces, dimensional asymmetry between encoders, and the lack of discrete class labels that can serve as alignment anchors.

A.3 Trajectory-Based Multimodal Distillation

Trajectory-based methods extend MTT to vision-language models by matching expert parameter trajectories across both encoder branches. Given a real dataset $\mathcal{T}$, an expert trajectory $\tau^* = \{\Theta_t^*\}_{t=0}^{T}$ is pre-computed by training a vision-language model with the symmetric InfoNCE loss (Eq. 1), where $\Theta_t^* = (\Theta_{\text{img},t}^*, \Theta_{\text{txt},t}^*)$. A student model initialized from an expert checkpoint $\hat{\Theta}_s = \Theta_s^*$ is then trained on the synthetic data $\mathcal{S}$ for R' inner steps, and the synthetic data is updated by minimizing the bi-modal trajectory discrepancy:

$$\mathcal{L}_{\text{bi-traj}} = \frac{\|\hat{\Theta}_{\text{img},s+R'} - \Theta_{\text{img},s+R}^*\|^2}{\|\Theta_{\text{img},s}^* - \Theta_{\text{img},s+R}^*\|^2} + \frac{\|\hat{\Theta}_{\text{txt},s+R'} - \Theta_{\text{txt},s+R}^*\|^2}{\|\Theta_{\text{txt},s}^* - \Theta_{\text{txt},s+R}^*\|^2}, \tag{14}$$

where R is the expert matching horizon and R' is the student unrolling horizon. Representative methods include MTT-VL, which provides the first baseline; LoRS [46], which jointly optimizes a soft-label similarity matrix alongside the synthetic image-text pairs and applies low-rank decomposition to this matrix to reduce computational cost; and RepBlend [54], which freezes both encoders and blends real and synthetic representations. Despite these refinements, all trajectory-based methods require a full pre-training run to produce the expert trajectory, which is precisely the cost that distillation aims to avoid. As shown in Table 5, RepBlend incurs 16.6 GB of expert storage and 78.6 hours of expert training, compared with zero additional storage and 2.0 hours for distribution-based methods including ours. More recently, APM [3] proposed an efficient multimodal dataset distillation framework that matches analytically estimated projection parameters, reducing the computational and storage overhead associated with conventional trajectory-based methods.

B t-SNE Visualization

This section describes the procedure used to generate the t-SNE visualization in Figure 2 [38].

Feature Extraction. We instantiate a CLIPModel (NFNet image encoder + BERT text encoder) with pre-trained weights under a random seed distinct from those used during distillation; all parameters are frozen. For each method (Real Data, CovMatch, MDA-DD), we extract image embeddings $\mathbf{e}_v^i = f_v(\tilde{\mathbf{x}}_v^i; \theta_v) \in \mathbb{R}^{d_v}$ without projection or normalization, and text embeddings $\mathbf{z}_l^i = \mathbf{W}_l^{\uparrow} \cdot f_l(\tilde{\mathbf{x}}_l^i; \theta_l) \in \mathbb{R}^{d_v}$, projected into the visual space for direct comparison.

Joint t-SNE Reduction. All image and text embeddings from the three configurations are concatenated into a single matrix $\mathbf{X} \in \mathbb{R}^{6N \times d_v}$ (2 modalities $\times$ 3 methods $\times$ N pairs) and jointly reduced by a single t-SNE instance to $\mathbf{X}_{\text{2D}} \in \mathbb{R}^{6N \times 2}$. Joint reduction is essential because all embeddings are extracted by the same model and reside in the same high-dimensional space; reducing them simultaneously preserves their relative proximity. We set the perplexity to 50 and use PCA initialization [20].

Observations. As shown in Figure 2(a), real-data embeddings form two well-spread, spatially interleaved clusters with strong cross-modal alignment. CovMatch exhibits two pathologies: a subset of text embeddings collapses into an isolated region far from both main clusters, consistent with the initialization overfitting analyzed in Section 3.1. The embeddings produced by MDA-DD distribute similarly to real data, with both modalities well-spread and interleaved, confirming that initialization overfitting is effectively alleviated.

C Proofs and Derivations

C.1 Matrix Form Equivalence of SOCM (Eq. 11)

Let $\mathbf{E}_v \in \mathbb{R}^{n \times d_v}$ and $\mathbf{E}_l \in \mathbb{R}^{n \times d_l}$ be obtained by stacking $\mathbf{e}_v^i$ and $\mathbf{e}_l^i$ row-wise. Then $\frac{1}{n}\sum_{i=1}^n \mathbf{e}_v^i (\mathbf{e}_l^i)^\top = \frac{1}{n}\mathbf{E}_v^\top \mathbf{E}_l$.

The (p, q)-th entry of $\mathbf{E}_v^\top \mathbf{E}_l$ is $\sum_{i=1}^n [\mathbf{e}_v^i]_p [\mathbf{e}_l^i]_q$, which equals the (p, q)-th entry of $\sum_{i=1}^n \mathbf{e}_v^i (\mathbf{e}_l^i)^\top$. Since this holds for all (p, q), dividing by n yields the result.

C.2 SOCM Decomposition (Eq. 12)

Let $\boldsymbol{\mu}_v = \frac{1}{n}\sum_i \mathbf{e}_v^i$, $\boldsymbol{\mu}_l = \frac{1}{n}\sum_i \mathbf{e}_l^i$, and $\mathbf{C} = \frac{1}{n}\sum_i (\mathbf{e}_v^i - \boldsymbol{\mu}_v)(\mathbf{e}_l^i - \boldsymbol{\mu}_l)^\top$. Then $\mathbf{M}_\theta(\mathcal{D}) = \mathbf{C} + \boldsymbol{\mu}_v \boldsymbol{\mu}_l^\top$.

Substituting $\mathbf{e}_v^i = \bar{\mathbf{e}}_v^i + \boldsymbol{\mu}_v$ and $\mathbf{e}_l^i = \bar{\mathbf{e}}_l^i + \boldsymbol{\mu}_l$ into the SOCM definition and expanding:

$$\begin{aligned}
\mathbf{M}_\theta(\mathcal{D}) &= \frac{1}{n}\sum_{i=1}^n (\bar{\mathbf{e}}_v^i + \boldsymbol{\mu}_v)(\bar{\mathbf{e}}_l^i + \boldsymbol{\mu}_l)^\top \\
&= \underbrace{\frac{1}{n}\sum_i \bar{\mathbf{e}}_v^i (\bar{\mathbf{e}}_l^i)^\top}_{\mathbf{C}} + \underbrace{\frac{1}{n}\Big(\sum_i \bar{\mathbf{e}}_v^i\Big)\boldsymbol{\mu}_l^\top}_{=\mathbf{0}} + \underbrace{\boldsymbol{\mu}_v \frac{1}{n}\Big(\sum_i \bar{\mathbf{e}}_l^i\Big)^\top}_{=\mathbf{0}} + \boldsymbol{\mu}_v \boldsymbol{\mu}_l^\top .
\end{aligned} \tag{15}$$

The cross terms vanish because $\sum_i \bar{\mathbf{e}}_v^i = \sum_i \bar{\mathbf{e}}_l^i = \mathbf{0}$ by definition of centering.

This decomposition shows that SOCM incorporates both the centered cross-modal covariance and the mean outer product within a single matrix. Matching the raw moment therefore preserves their combined contribution without requiring explicit mean-centering.

C.3 Approximation Analysis of the Staggered Model Queue (Eq. 6→Eq. 7)

We analyze how the staggered queue approximates a model-state matching objective. The analysis first considers a fixed synthetic dataset and then distinguishes ideal stratified sampling from the deterministic training ages used by the actual queue.

Loss and boundedness. For a fixed synthetic dataset $\mathcal{S}$, define the matching loss under model state θ as

$$g_{\mathcal{S}}(\theta) = \left\|\mathbf{M}_\theta(\mathcal{S}) - \rho\mathbf{M}_\theta(\mathcal{T})\right\|_F^2 + \lambda\Big(\left\|\bar{\mathbf{z}}_{v,\theta}^{\mathcal{S}} - \bar{\mathbf{z}}_{v,\theta}^{\mathcal{T}}\right\|_2^2 + \left\|\bar{\mathbf{z}}_{l,\theta}^{\mathcal{S}} - \bar{\mathbf{z}}_{l,\theta}^{\mathcal{T}}\right\|_2^2\Big). \tag{16}$$

Assume that, over the model states under consideration, encoder features satisfy $\|\mathbf{e}_v\|_2 \le R_v$ and $\|\mathbf{e}_l\|_2 \le R_l$, while projected features satisfy $\|\mathbf{z}_v\|_2 \le Z_v$ and $\|\mathbf{z}_l\|_2 \le Z_l$. For $\lambda \ge 0$, these bounds imply

$$0 \le g_{\mathcal{S}}(\theta) \le B, \tag{17}$$

where one valid upper bound is

$$B = (1 + |\rho|)^2 R_v^2 R_l^2 + 4\lambda(Z_v^2 + Z_l^2). \tag{18}$$

This boundedness condition is an explicit assumption of the analysis, rather than a consequence of network architecture alone.

Training-stage distribution. Let Θ_a denote the random model state after a real-data training updates, starting from the prescribed pretrained encoder weights. The randomness includes independently initialized projection heads and stochastic training updates. Write P_a for the distribution of Θ_a and

$$m(a) = \mathbb{E}_{\theta\sim P_a}[g_{\mathcal{S}}(\theta)]. \tag{19}$$

For integer training ages $a \in \{0, \dots, T_{\max} - 1\}$, consider the stage-averaged objective

$$\mu_{\text{stage}} = \frac{1}{T_{\max}} \sum_{a=0}^{T_{\max}-1} m(a). \tag{20}$$

This corresponds to a model-state distribution obtained by sampling the training age uniformly and then drawing a model state from P_a. It is a specific stage-based distribution and need not coincide with the general target distribution $P(\Theta)$ in Eq. 6.

C.3.1 I.i.d. baseline

Suppose $\theta_1, \dots, \theta_K$ are drawn independently from the stage-based distribution. Define

$$\bar{g}_K^{\text{MC}} = \frac{1}{K} \sum_{k=1}^{K} g_{\mathcal{S}}(\theta_k). \tag{21}$$

Then $\mathbb{E}[\bar{g}_K^{\text{MC}}] = \mu_{\text{stage}}$. Since $g_{\mathcal{S}}(\theta_k) \in [0, B]$, Popoviciu's inequality gives

$$\text{Var}[\bar{g}_K^{\text{MC}}] \le \frac{B^2}{4K}. \tag{22}$$

Consequently, by Cauchy–Schwarz,

$$\mathbb{E}\big[|\bar{g}_K^{\text{MC}} - \mu_{\text{stage}}|\big] \le \frac{B}{2\sqrt{K}}. \tag{23}$$

Hoeffding's inequality further yields

$$\Pr\big(|\bar{g}_K^{\text{MC}} - \mu_{\text{stage}}| \ge \epsilon\big) \le 2\exp\left(-\frac{2K\epsilon^2}{B^2}\right). \tag{24}$$

C.3.2 Ideal stratified sampling

Assume that K divides $T_{\max}$ and let $\Delta t = T_{\max}/K$. Partition the training ages into strata

$$I_k = \{(k-1)\Delta t, \dots, k\Delta t - 1\}. \tag{25}$$

Let H denote the stratum index of a uniformly sampled training age. Each stratum has probability $\Pr(H = k) = 1/K$.

Consider an ideal stratified estimator that independently draws one model state from the conditional distribution associated with each stratum. Let

$$\mu_k = \mathbb{E}[g_{\mathcal{S}}(\theta) \mid H = k], \qquad \sigma_k^2 = \mathrm{Var}[g_{\mathcal{S}}(\theta) \mid H = k]. \tag{26}$$

The estimator

$$\bar{g}_K^{\mathrm{strat}} = \frac{1}{K} \sum_{k=1}^{K} g_{\mathcal{S}}(\theta_k) \tag{27}$$

is unbiased for μ_{stage}. Because the draws are independent across strata,

$$\mathrm{Var}[\bar{g}_K^{\mathrm{strat}}] = \frac{1}{K^2} \sum_{k=1}^{K} \sigma_k^2. \tag{28}$$

By the law of total variance,

$$\begin{aligned} \mathrm{Var}[g_{\mathcal{S}}(\theta)] = \frac{1}{K} \sum_{k=1}^{K} \sigma_k^2 \\ + \frac{1}{K} \sum_{k=1}^{K} (\mu_k - \mu_{\mathrm{stage}})^2. \end{aligned} \tag{29}$$

The between-stratum term is non-negative. Therefore,

$$\mathrm{Var}[\bar{g}_K^{\mathrm{strat}}] \leq \frac{\mathrm{Var}[g_{\mathcal{S}}(\theta)]}{K} = \mathrm{Var}[\bar{g}_K^{\mathrm{MC}}]. \tag{30}$$

The inequality is strict when the stratum means differ. Thus, ideal stratification reduces sampling variance by accounting for variation across training stages.

C.3.3 Approximation by staggered ages

The actual queue differs from ideal stratified sampling because its model ages are determined by the replacement schedule. At steady state, Eq. 7 uses one model from each training-age stratum. For a fixed replacement-cycle offset $r \in \{0, \ldots, \Delta t - 1\}$, the ages are

$$a_k(r) = (k-1)\Delta t + r, \qquad k = 1, \ldots, K. \tag{31}$$

Let $\theta_k(r)$ denote the model state at age $a_k(r)$, and define

$$\widehat{\mu}_K(r) = \frac{1}{K} \sum_{k=1}^{K} g_{\mathcal{S}}(\theta_k(r)). \tag{32}$$

For a fixed $\mathcal{S}$, suppose the model trajectories use independent training randomness, so that the states $\theta_k(r)$ are independent across models. They need not be identically distributed because their ages differ. The estimator has expectation

$$\mu_r = \mathbb{E}[\widehat{\mu}_K(r)] = \frac{1}{K} \sum_{k=1}^{K} m(a_k(r)). \tag{33}$$

Assume that the expected matching loss varies smoothly across training ages:

$$|m(a) - m(b)| \leq L_a |a - b| \tag{34}$$

for some constant $L_a \geq 0$. The difference between the staggered-age expectation and the stage-averaged objective satisfies

$$\begin{aligned} |\mu_r - \mu_{\mathrm{stage}}| &\leq \frac{L_a}{K\Delta t} \sum_{k=1}^{K} \sum_{a \in I_k} |a_k(r) - a| \\ &\leq \frac{L_a(\Delta t - 1)}{2}. \end{aligned} \tag{35}$$

Since the model states are independent under the stated assumption and their losses lie in $[0, B]$, Hoeffding's inequality for independent, not necessarily identically distributed, random variables gives

$$\Pr\left(|\widehat{\mu}_K(r) - \mu_r| \geq \epsilon\right) \leq 2\exp\left(-\frac{2K\epsilon^2}{B^2}\right). \tag{36}$$

Combining this concentration bound with the stage-discretization error yields

$$\mathbb{E}\big[|\widehat{\mu}_K(r) - \mu_{\text{stage}}|\big] \leq \frac{B}{2\sqrt{K}} + \frac{L_a(\Delta t - 1)}{2}. \tag{37}$$

The first term represents stochastic variation across independently evolving models, while the second captures the difference between the staggered ages and the stage-averaged objective. Moreover, averaging the expected staggered estimator over a complete replacement cycle gives

$$\frac{1}{\Delta t}\sum_{r=0}^{\Delta t-1} \mu_r = \mu_{\text{stage}}. \tag{38}$$

Thus, the staggered schedule covers all training ages across a complete cycle, although its expectation at any single offset need not equal the stage average.

For the more general target distribution $P(\Theta)$ in Eq. 6, define the distributional mismatch

$$\delta_P(\mathcal{S}) = \left|\mu_{\text{stage}} - \mathbb{E}_{\theta\sim P(\Theta)}[g_{\mathcal{S}}(\theta)]\right|. \tag{39}$$

The corresponding expected approximation error is bounded by

$$\begin{aligned}&\mathbb{E}\left[\left|\widehat{\mu}_K(r) - \mathbb{E}_{\theta\sim P(\Theta)}[g_{\mathcal{S}}(\theta)]\right|\right]\\ &\qquad\leq \frac{B}{2\sqrt{K}} + \frac{L_a(\Delta t - 1)}{2} + \delta_P(\mathcal{S}).\end{aligned} \tag{40}$$

The mismatch term vanishes when the target distribution is the stage-based distribution defined above.

Dependence and scope. The preceding concentration bounds hold for a fixed synthetic dataset and independently evolving model trajectories. They characterize model-state objective approximation, not convergence of the joint optimization algorithm.

During actual distillation, the synthetic dataset is updated using the active queue. Consequently, the losses evaluated at the adaptively optimized synthetic dataset may be statistically dependent, even when the underlying model trajectories use independent randomness.

More generally, for dependent losses $Y_k = g_{\mathcal{S}}(\theta_k)$, the estimator variance satisfies

$$\begin{aligned}\text{Var}\left[\frac{1}{K}\sum_{k=1}^{K} Y_k\right] &= \frac{1}{K^2}\sum_{k=1}^{K}\text{Var}[Y_k]\\ &\quad + \frac{2}{K^2}\sum_{i<j}\text{Cov}(Y_i, Y_j).\end{aligned} \tag{41}$$

Thus, a comparable variance rate under dependence requires additional control of the cross-model covariance terms. Such control is not implied solely by staggered replacement or weight decay.

The analysis isolates the benefit of training-stage coverage under explicit sampling and regularity assumptions. The empirical ablations in Section 4.4 and Appendix F further evaluate the effectiveness of the staggered queue in the actual distillation procedure.

D Implementation of Asymmetric Dual Projection

D.1 Training and Evaluation Procedures

Algorithm 2 and Algorithm 3 describe the training and evaluation procedures, respectively. During training, the forward pass computes contrastive losses in two complementary subspaces: one pairs raw image features with projected text representations in $\mathbb{R}^{d_v}$, while the other pairs projected image features with raw text representations in $\mathbb{R}^{d_l}$. At evaluation time, the model trained on synthetic data constructs a fused similarity matrix by averaging scores from both projection directions, and retrieval metrics are computed from this fused matrix through top-k selection.

Algorithm 2 Forward Pass with Asymmetric Dual Projection

Require: Image batch $\{\mathbf{x}_v^i\}_{i=1}^B$, text batch $\{(\mathbf{x}_l^i, \mathbf{m}^i)\}_{i=1}^B$, learnable logit-scale parameter ℓ
Ensure: Contrastive loss $\mathcal{L}$

$\mathbf{e}_v^i \leftarrow f_v(\mathbf{x}_v^i; \theta_v) \in \mathbb{R}^{d_v}, \quad \forall\, i$
$\mathbf{e}_l^i \leftarrow f_l(\mathbf{x}_l^i, \mathbf{m}^i; \theta_l) \in \mathbb{R}^{d_l}, \quad \forall\, i$
$\mathbf{z}_l^i \leftarrow \mathbf{W}_l^{\uparrow} \mathbf{e}_l^i \in \mathbb{R}^{d_v}$ {text $\rightarrow$ visual space}
$\mathbf{z}_v^i \leftarrow \mathbf{W}_v^{\downarrow} \mathbf{e}_v^i \in \mathbb{R}^{d_l}$ {image $\rightarrow$ linguistic space}
$\hat{\mathbf{e}}_v^i \leftarrow \mathbf{e}_v^i / \|\mathbf{e}_v^i\|_2, \quad \hat{\mathbf{z}}_l^i \leftarrow \mathbf{z}_l^i / \|\mathbf{z}_l^i\|_2$
$\hat{\mathbf{z}}_v^i \leftarrow \mathbf{z}_v^i / \|\mathbf{z}_v^i\|_2, \quad \hat{\mathbf{e}}_l^i \leftarrow \mathbf{e}_l^i / \|\mathbf{e}_l^i\|_2$
$\mathbf{S}_{1,ij} \leftarrow \hat{\mathbf{z}}_v^{i\top} \hat{\mathbf{e}}_l^j$ {linguistic space}
$\mathbf{S}_{2,ij} \leftarrow \hat{\mathbf{e}}_v^{i\top} \hat{\mathbf{z}}_l^j$ {visual space}
$\tau \leftarrow \exp(-\ell)$
$\mathcal{L}_1 \leftarrow \mathcal{L}_{\mathrm{NCE}}(\mathbf{S}_1; \tau)$
$\mathcal{L}_2 \leftarrow \mathcal{L}_{\mathrm{NCE}}(\mathbf{S}_2; \tau)$
$\mathcal{L} \leftarrow \frac{1}{2}(\mathcal{L}_1 + \mathcal{L}_2)$
return $\mathcal{L}$

Algorithm 3 Similarity Matrix Construction with Asymmetric Dual Projection at Evaluation

Require: Test images $\{\mathbf{x}_v^i\}_{i=1}^{N_I}$, test texts $\{\mathbf{x}_l^j\}_{j=1}^{N_T}$, trained model $(f_v, f_l, \mathbf{W}_l^{\uparrow}, \mathbf{W}_v^{\downarrow}, \tau)$
Ensure: Score matrices $\mathbf{S}^{i2t} \in \mathbb{R}^{N_I \times N_T}$, $\mathbf{S}^{t2i} \in \mathbb{R}^{N_T \times N_I}$

$\hat{\mathbf{e}}_v^i,\ \hat{\mathbf{z}}_v^i \leftarrow \mathrm{Norm}(f_v(\mathbf{x}_v^i)),\ \mathrm{Norm}(\mathbf{W}_v^{\downarrow} f_v(\mathbf{x}_v^i)), \quad \forall\, i$
$\hat{\mathbf{e}}_l^j,\ \hat{\mathbf{z}}_l^j \leftarrow \mathrm{Norm}(f_l(\mathbf{x}_l^j)),\ \mathrm{Norm}(\mathbf{W}_l^{\uparrow} f_l(\mathbf{x}_l^j)), \quad \forall\, j$
$\mathbf{R}_{ij} \leftarrow \frac{1}{2\tau}\left(\hat{\mathbf{e}}_v^{i\top} \hat{\mathbf{z}}_l^j + \hat{\mathbf{z}}_v^{i\top} \hat{\mathbf{e}}_l^j\right)$
$\mathbf{S}^{i2t} \leftarrow \mathrm{TopKMask}(\mathbf{R}, k)$
$\mathbf{S}^{t2i} \leftarrow \mathrm{TopKMask}(\mathbf{R}^{\top}, k)$
return $\mathbf{S}^{i2t}, \mathbf{S}^{t2i}$

E Comparison Methods

Coreset Selection. Random (Rand) uniformly samples image-text pairs i.i.d. from the full dataset. This unbiased baseline is computationally cheap but ignores sample informativeness, which limits performance at low compression ratios. Herding (Herd) [42] is a greedy algorithm that iteratively selects the sample whose inclusion best approximates the full dataset's feature mean, minimizing first-order distributional discrepancy. K-Center (K-Cent) [15] is a coverage-based strategy that iteratively picks the point farthest from the current subset, solving an approximate k-center problem to maximize geometric coverage in the feature space. Forgetting (Forget) [36] counts how often each sample transitions from correctly to incorrectly matched during training; low-forgetting (easy) samples are discarded first, retaining the most informative hard examples.

Dataset Distillation. LoRS [46] is a trajectory-matching multimodal distillation method that applies low-rank factorization to the cross-modal similarity matrix, reducing memory and computation while preserving key alignment information. CovMatch [22] is a distribution-matching method that projects both modalities into a shared space and aligns the cross-modal covariance between real and synthetic data, relying on a fixed model initialization for optimization stability. RepBlend [54] is a trajectory-matching method that freezes pretrained encoders and symmetrically matches the trajectories of image and text projection heads, while blending synthetic representations to mitigate modality collapse and enhance intra-modal diversity. PTM-ST [16] is a staged distillation framework with two components. PTM partitions the distillation process into multiple phases with dynamically shifted trajectory-matching intervals, enabling progressive cross-modal knowledge absorption. ST replaces the original high-variance expert trajectory with a smoothed linear interpolation between endpoints, weighted by cumulative distance. Combined with EMA updates, this suppresses gradient oscillation and bridges inter-stage knowledge gaps, improving both distillation stability and retrieval performance.

Table 8: Hyperparameter configurations of MDA-DD on Flickr30K and MS-COCO under varying synthetic budget sizes.

Hyperparameter	Flickr30K			MS-COCO		
	100	200	500	100	200	500
Scaling factor ρ	2	2	1	2	2	1
Feature matching weight λ	0.1	0.1	0.5	0.1	0.1	0.5
Batch size (real)	128	128	128	128	128	128
Batch size (synthetic)	–	–	200	–	–	200
Learning rate (image)	1	1	1	1	1	1
Learning rate (text)	1	1	1	1	1	1
Iterations	1000	1000	1000	1000	1000	1000
Number of models in queue K	8	8	8	8	8	8
Maximum model lifetime $T_{\max}$	200	200	200	200	200	200

F Additional Ablation Studies

F.1 Component-Wise Contribution

Baseline reproduction. To ensure a fair comparison, we re-run CovMatch under identical hardware and optimization settings as MDA-DD. All experiments in this ablation use a single NVIDIA RTX 5090 (32 GB). We set the following CovMatch hyperparameters to match those of our method and its ablated variants: inner-loop iterations (50), outer-loop iterations (200), synthetic batch size, feature-matching weight λ, and scaling factor ρ. The remaining architecture-level settings (image encoder, text encoder, image resolution, and synthetic budget N) are also kept consistent with the main experiments (Table 8). This controlled setup isolates the effect of each proposed component, namely the model-distribution-aware objective (§3.1), asymmetric dual projection (§3.2), and the SOCM statistic (§3.3), from confounding factors such as hardware differences or hyperparameter tuning.

F.2 Cross-Architecture Generalization

Network architectures. Table 9 lists all models used in our experiments. The image encoders are NFNet, NF-ResNet, NF-RegNet, and ViT; the text encoder is BERT throughout.

Table 9: Network architectures and parameter counts.

Architecture	Model Variant	Parameter Count
NFNet [5]	nfnet_l0	26.4M
NF-ResNet [4]	nf_resnet50	25.6M
NF-RegNet [45]	nf_regnet_b1	21.0M
ViT [12]	vit_base_patch16_224	85.8M
BERT [10]	bert-base-uncased	110.0M

Cross-architecture evaluation. Table 10 reports the full cross-architecture results on Flickr30K with 100 synthetic pairs. All methods perform distillation using NFNet+BERT and are evaluated on unseen image encoders (NF-ResNet50, NF-RegNet-B1, ViT-B/16) while keeping the text encoder fixed. Trajectory-based methods (MTT-VL, LoRS, RepBlend) generally perform below random selection on image-retrieval metrics, indicating limited transferability across visual encoders. In contrast, MDA-DD consistently outperforms the compared methods across all six retrieval metrics, achieving approximately 2.7–2.8$\times$ the IR@1 of CovMatch across the three unseen architectures. These results demonstrate improved cross-architecture retrieval performance with synthetic data distilled using MDA-DD.

Table 10: Cross-architecture evaluation on Flickr30K with $N = 100$ synthetic pairs. All methods are distilled with NFNet+BERT and evaluated on unseen image encoders.

Backbone	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet+BERT	Random	2.2	8.0	13.4	3.0	10.1	16.4
	MTT-VL	1.8	7.0	11.9	2.9	10.5	16.4
	LoRS	1.6	7.0	12.1	2.9	11.1	18.0
	RepBlend	1.2	5.3	9.7	4.1	11.2	18.9
	CovMatch	4.7	15.4	23.4	5.5	17.3	26.6
	Ours	**13.3**	**35.4**	**47.3**	**13.6**	**33.8**	**47.0**
RegNet+BERT	Random	2.2	7.6	12.4	2.6	9.7	15.3
	MTT-VL	1.5	6.0	10.6	2.9	9.2	15.0
	LoRS	1.6	6.4	11.2	2.5	11.0	17.7
	RepBlend	1.8	6.8	11.1	3.2	14.0	21.0
	CovMatch	3.9	13.8	21.7	5.6	16.8	26.0
	Ours	**10.8**	**29.6**	**42.7**	**12.2**	**29.9**	**43.1**
ViT+BERT	Random	2.8	9.3	15.1	4.2	13.3	19.8
	MTT-VL	2.1	7.8	13.1	3.9	12.2	18.6
	LoRS	2.0	7.3	12.7	3.2	11.7	19.0
	RepBlend	2.4	9.5	15.3	4.8	16.6	23.9
	CovMatch	3.7	13.5	21.7	5.2	18.6	28.0
	Ours	**9.9**	**30.6**	**43.8**	**13.0**	**33.7**	**45.9**

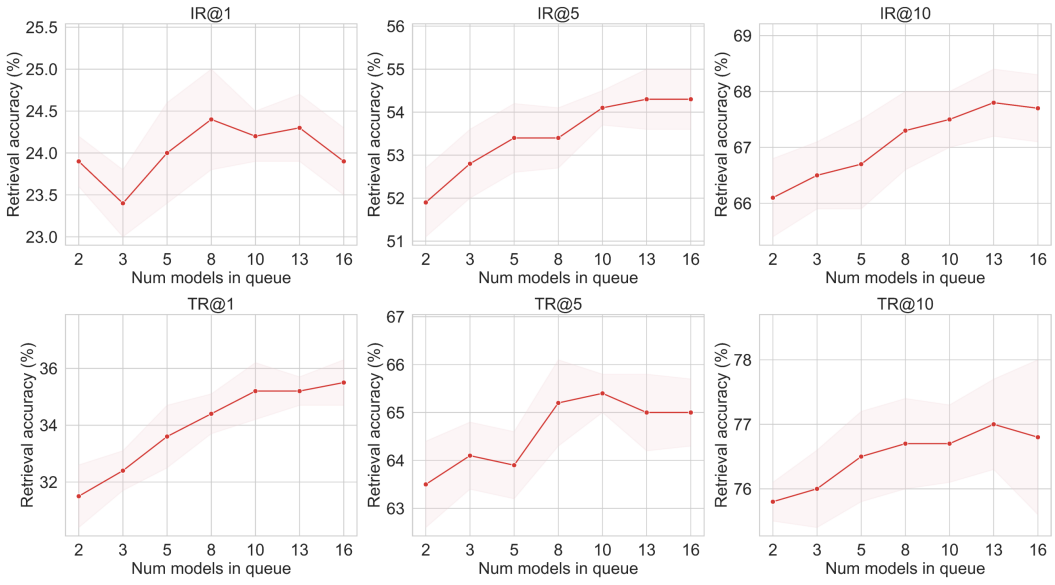


Figure 5: Effect of queue size K on cross-modal retrieval performance (Flickr30K, 100 synthetic pairs). Performance improves steadily up to $K = 8$ and saturates beyond $K = 10$.

F.3 Sensitivity Analysis

Effect of queue size K. Figure 5 reports retrieval performance on Flickr30K with 100 synthetic pairs as a function of the queue size ($K \in \{2, 3, 5, 8, 10, 13, 16\}$), with all other hyperparameters held fixed. All six metrics improve consistently as K increases from 2 to 8, with further marginal gains at $K = 10$ on certain metrics such as IR@5 and TR@1. This trend agrees with the theoretical expectation from Appendix C.3: a larger queue improves coverage of the model-state distribution $P(\Theta)$, thereby narrowing the approximation gap between the particle objective (Eq. 7) and the distributional objective (Eq. 6). Beyond $K = 10$, performance saturates or slightly decreases on several metrics, indicating diminishing returns from additional particles. Since $K = 8$ and $K = 10$ yield comparable accuracy while larger queues incur proportionally higher GPU memory and per-iteration cost, we adopt $K = 8$ as the default to balance performance and efficiency.

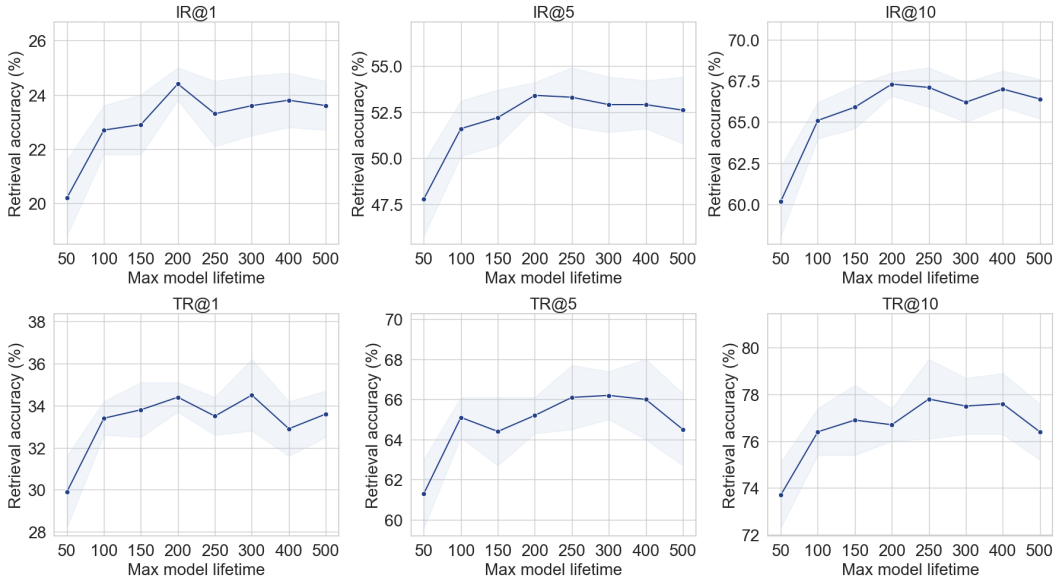


Figure 6: Effect of model lifetime $T_{\max}$ on cross-modal retrieval performance (Flickr30K, 100 synthetic pairs). The optimum lies around $T_{\max} = 200$–250.

Effect of model lifetime $T_{\max}$. Figure 6 shows retrieval performance on Flickr30K with 100 synthetic pairs as $T_{\max}$ varies from 50 to 500. Performance rises sharply from 50 to 200, peaks around $T_{\max} = 200$–250 across most metrics (e.g., IR@1 reaches 24.5% at 200), and plateaus or slightly declines beyond 300. Too short a lifetime limits the training-stage coverage of each particle, reducing the diversity of model states in the queue. Too long a lifetime causes particles to drift into late-training regimes where the feature geometry changes slowly, diminishing the informational gain from queue refreshes. We set $T_{\max} = 200$ as the default, as it provides sufficient training-stage diversity while maintaining frequent queue renewal.

G Distillation under Alternative Budget Settings

Extreme compression with 50 synthetic pairs. Table 11 evaluates performance under an extreme compression setting using only 50 synthetic image-text pairs (0.2% of Flickr30K, 0.4‰ of MS-COCO). At this budget, both RepBlend and CovMatch produce comparable yet modest results on Flickr30K (7–8 IR@1). MDA-DD, by contrast, reaches 20.5 IR@1 and 30.6 TR@1, outperforming the strongest baseline by roughly $2.6\times$ and $2.4\times$, respectively. On MS-COCO, where the compression ratio is even more extreme, MDA-DD achieves 1.67× and 2.17× the IR@1 of RepBlend and CovMatch, respectively. The widening gap at lower budgets suggests that the model-distribution-aware objective is particularly effective when each synthetic sample must carry a larger share of the overall training signal.

Table 11: Cross-modal retrieval performance with 50 synthetic pairs (0.2% of Flickr30K, 0.4‰ of MS-COCO). Results are reported as mean ± standard deviation over 5 independent runs.

Dataset	Pairs	Ratio	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Flickr30K	50	0.2%	RepBlend	7.5±0.5	24.8±1.1	36.0±1.7	12.5±0.5	34.7±0.7	48.4±0.8
			CovMatch	7.8±0.2	23.8±0.5	35.0±0.3	11.5±0.3	31.8±1.3	44.2±1.1
			Ours	**20.5**±1.0	**49.4**±0.8	**63.8**±0.6	**30.6**±0.5	**60.0**±0.7	**71.8**±1.0
MS-COCO	50	0.4‰	RepBlend	3.0±0.2	10.9±0.4	18.6±0.6	4.4±0.2	15.6±0.3	24.7±0.4
			CovMatch	2.3±0.1	8.6±0.2	14.8±0.4	2.8±0.2	10.3±0.4	16.8±0.6
			Ours	**5.0**±0.4	**16.5**±0.8	**26.0**±1.1	**7.7**±0.3	**22.7**±0.9	**33.8**±1.0

H Visualization of Distilled Data



Figure 7: Synthetic image-text pairs on Flickr30K before and after distillation. Red (left): original initialized pairs. Blue (right): distilled pairs produced by MDA-DD.



Figure 8: Synthetic image-text pairs on MS-COCO before and after distillation. Red (left): original initialized pairs. Blue (right): distilled pairs produced by MDA-DD.

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 [N/A] 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 [N/A] 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 limitations of our work are discussed in Section 5.

   Guidelines:

   - The answer [N/A] 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 complete proofs for the theoretical results.

Guidelines:

- The answer [N/A] 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 includes all information required to reproduce the main experimental results.

   Guidelines:

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

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

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

Answer:[Yes]

Justification: The paper provides implementation details and algorithmic descriptions sufficient for reproduction. The source code is publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides the necessary training and testing details, including datasets, hyperparameters, optimizer type, and the criteria for their selection.

   Guidelines:

   - The answer [N/A] 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: Each experiment was repeated five times, and we report the mean and standard deviation to indicate statistical significance.

   Guidelines:

   - The answer [N/A] 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 reports the compute resources required to reproduce the experiments, including GPU type, memory, storage, and execution time.

   Guidelines:

   - The answer [N/A] 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 conforms to the NeurIPS Code of Ethics and adheres to the required ethical standards.

   Guidelines:

   - The answer [N/A] 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:[N/A]

    Justification: This work is primarily methodological and does not directly involve societal impacts; therefore, this item is not applicable.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve the release of high-risk data or models; therefore, this item is not applicable.

    Guidelines:

    - The answer [N/A] 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: We cite the original sources of the code packages, datasets, and models used in this work and comply with their licenses and terms of use.

    Guidelines:

    - The answer [N/A] 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: We provide the code and the distilled dataset introduced in this work in the supplementary material, together with the necessary documentation.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve research with human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer:[N/A]

    Justification: LLMs were used only for language polishing and were not involved in the core methodology; therefore, a declaration of LLM usage is not applicable.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.