NOT EVERY TURN MATTERS: PROGRESSIVE LINEARIZED TRAJECTORY MATCHING FOR MULTIMODAL DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Dataset Distillation (MDD) compresses image–text datasets into synthetic pairs that retain cross-modal retrieval utility. Existing trajectory-based methods use one-epoch expert updates as matching targets. However, each target direction is tied to a sampled intermediate epoch, so a severely compressed synthetic set must reproduce directions that change frequently across training. Across 20 independently trained experts, we observe substantial direction changes across epochs but a consistent cross-epoch trend among experts. We therefore propose Progressive Linearized Trajectory Matching (PLTM), which aligns every target with the overall direction from the selected start state to the end state and progressively increases the matched interval as the synthetic pairs improve. This redesign provides a direction-consistent expert target. Under paired hard supervision, online relational alignment and diversity regularization preserve relations among pairs and sample diversity without optimizing an additional soft similarity matrix. PLTM incurs negligible computational overhead and approximately doubles distillation efficiency. Experiments on Flickr-30K and MS-COCO show that our method consistently outperforms previous state-of-the-art MDD methods under paired hard supervision and optional soft similarity supervision. On MS-COCO with 500 synthetic pairs, PLTM improves IR@10/TR@10 by 7.8%/6.1% without a soft similarity matrix and by 4.2%/4.4% with soft supervision. Our code will be made publicly available to the community.

1 INTRODUCTION

Large-scale image–text datasets advance vision–language representation learning (Radford et al., 2021), but remain expensive to store and train on repeatedly. Multimodal Dataset Distillation (MDD) reduces this cost by synthesizing a small set of pairs with similar training utility, yet extreme compression must preserve both modality-specific representations and cross-modal relations. The central question is which training information such a compact set can transfer reliably.

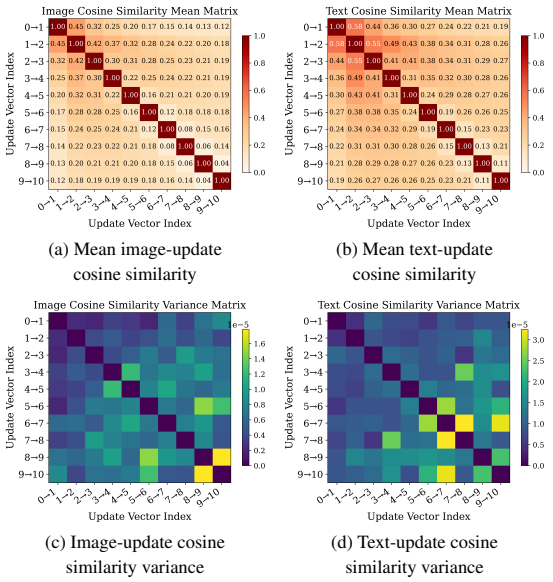
Trajectory matching transfers training information by asking synthetic pairs to reproduce one-epoch parameter updates sampled from expert trajectories trained on the full dataset. MTT-VL (Wu et al., 2024), LoRS (Xu et al., 2024), and RepBlend (Zhang et al., 2025) introduce different forms of cross-modal supervision, but retain these epoch-wise expert updates as matching targets. However, anchoring each target makes its direction conditional on a particular intermediate epoch, so a limited synthetic set must fit an optimization trajectory whose target direction changes frequently.

Given this, must a severely compressed synthetic set reproduce epoch-dependent directions to transfer knowledge from an expert trajectory? To examine this question, we train 20 experts from independent initializations and compute pairwise cosine similarities between their one-epoch update vectors. Figure 1(a,b) shows substantial directional variation across epochs: off-diagonal similarities range from 0.04 to 0.45 for images and from 0.11 to 0.58 for text. Low similarity indicates that the corresponding epochs optimize in different directions. Despite independent initializations, the small variances in Figure 1(c,d) indicate that the changes in their one-epoch update directions follow a consistent cross-epoch trend across experts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Conventional epoch-wise supervision nevertheless asks a severely compressed synthetic set to switch among these directions rather than concentrate on the trajectory's overall start-to-end direction. Such supervision is harder to fit and can require more optimization steps to reproduce the intended training effect.

Given the consistent overall trend across independently trained experts, requiring a compact synthetic set to reproduce every epoch-specific turn is unnecessarily restrictive. Instead, we propose **Progressive Linearized Trajectory Matching (PLTM)** (Figure 2), which defines the overall direction as the vector connecting the first and last selected expert states and replaces the cross-epoch path with the straight line between them. We refer to this replacement as linearization; every target is a positive step along the same vector and is therefore direction-consistent. Once the direction is unified, the remaining question is how much of the expert update to match. At early outer iterations, the synthetic pairs have undergone little optimization, making a long-range expert change difficult to reproduce in a few student steps. Repeated outer updates optimize the pairs against the trajectory-matching loss, motivating longer intervals later in distillation. PLTM therefore follows a short-to-long schedule. Linearization unifies the direction, while progression adapts the target range to the improving synthetic supervision.



(a) Mean image-update cosine similarity

(b) Mean text-update cosine similarity

(c) Image-update cosine similarity variance

(d) Text-update cosine similarity variance

Figure 1: Direction statistics from 20 independently trained experts. For each expert and modality, we obtain ten one-epoch updates as consecutive parameter differences and compute their 10×10 pairwise cosine-similarity matrix. Panels (a,b) report the element-wise mean of these matrices across experts for the image and text branches; panels (c,d) report the corresponding element-wise variance. Rows and columns index the one-epoch updates.

PLTM specifies the expert update that the synthetic set should reproduce. Matching this target depends on the gradients generated when the student is trained on the synthetic image–text pairs, and these gradients are shaped by the supervision assigned to the pairs. Under paired hard supervision, matched pairs have target 1 and all unmatched pairs have target 0. These labels identify paired samples but treat every unmatched pair alike, leaving their relative relations unspecified. Existing methods therefore optimize a soft similarity matrix together with the synthetic data to provide more detailed pairwise targets (Xu et al., 2024; Zhang et al., 2025). Although the soft matrix enriches the training signal, it introduces additional optimized parameters and memory and can obscure the influence of synthetic-data quality. For the paired-hard setting, inspired by these works (Dey et al., 2026; Bo & Wang, 2026), we retain paired hard supervision and obtain the missing relations directly from the current synthetic representations. **Relational alignment** matches image- and text-side Gram matrices computed online. Relational alignment alone does not discourage within-modality redundancy, so **diversity regularization** penalizes excessive similarity. Together, they preserve cross-modal structure and sample diversity without optimizing an additional soft similarity matrix, introducing another network, or requiring additional model passes.

We evaluate paired hard and optional soft similarity supervision in symmetric, asymmetric, and cross-architecture settings. Without a soft similarity matrix, PLTM improves mean recall over RepBlend w/o Sim. by **5.5%, 8.8%, and 12.9%** on Flickr-30K with 100, 200, and 500 pairs. It also outperforms RepBlend with matched soft supervision on both datasets. PLTM obtains these gains in 1,000 rather than 2,000 distillation steps, providing approximately twofold distillation efficiency with negligible computational overhead.

Our contributions are summarized as follows:

- We identify the changing directions imposed by epoch-wise matching and propose PLTM, which uses a shared start-to-end direction with a progressively increasing matching interval.
- We introduce online relational alignment and diversity regularization to preserve cross-modal relations and sample distinction under paired hard supervision, without optimizing an additional soft similarity matrix.
- Experiments across datasets, synthetic-set budgets, supervision settings, and architectures demonstrate the effectiveness, efficiency, and generality of our method.

2 Related Work

2.1 Unimodal Dataset Distillation

Coreset selection compresses a unimodal dataset by retaining representative real samples, using criteria such as feature-space coverage, centroid proximity, and forgetting events (Welling, 2009; Farahani & Hekmatfar, 2009; Toneva et al., 2019). Dataset distillation instead optimizes synthetic samples so that training on them approximates training on the full dataset (Wang et al., 2018). Representative objectives match gradients (Zhao et al., 2021), improve synthetic optimization through differentiable augmentation (Zhao & Bilen, 2021), align feature distributions (Wang et al., 2022; Zhao & Bilen, 2023), or transfer training trajectories (Cazenavette et al., 2022).

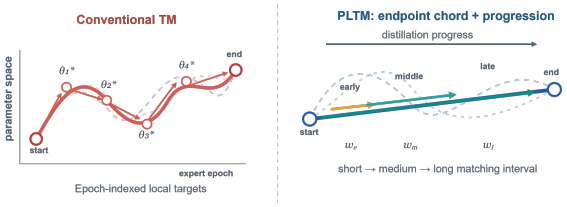


Figure 2: Comparison of conventional epoch-wise trajectory matching and our PLTM. Conventional matching follows different one-epoch directions along the expert trajectory (left), whereas PLTM aligns all targets with the expert trajectory's start-to-end direction and gradually increases the matched interval from short to long (right).

Sequential subset matching decomposes trajectory learning into ordered stages (Du et al., 2023), difficulty-aligned matching adapts supervision to the synthetic learning stage (Guo et al., 2024), and MCT uses convex interpolation between recorded expert states for continuous sampling and reduced storage (Zhong et al., 2025). Its targets still depend on intermediate expert states. PLTM instead removes these states from target construction and replaces directions tied to individual epochs with a shared start-to-end direction.

2.2 Multimodal Dataset Distillation

MTT-VL (Wu et al., 2024) establishes bi-trajectory co-distillation, while LoRS (Xu et al., 2024) augments it with mined low-rank relations. RepBlend (Zhang et al., 2025) uses representation blending and symmetric projection matching to counter modality collapse and optimization bias. PTM-ST (Guo et al., 2026) stages teachers and constructs smooth shortcuts to improve absorption of later-stage knowledge, whereas AMD (Qi et al., 2026) addresses feature-space misalignment and asynchronous dynamics through decoupled starting points and joint-space prototypes.

Other MDD lines preserve multimodal structure without primarily matching expert trajectories. CovMatch (Lee & Chung, 2025) combines cross-covariance matching with intra-modal regularization, and MDM (Jeong et al., 2026) matches real and synthetic distributions in a geometry-aware joint space. MDW (Dang et al., 2025) focuses on noisy web pairs. Generative approaches instead exploit pretrained synthesis priors: scalable data synthesis has been explored with generative models (Zhao et al., 2025), while PDS (Choi et al., 2026) extracts CLIP prototypes and synthesizes images with unCLIP without iterative dataset distillation.

3 METHOD

3.1 MULTIMODAL DATASET DISTILLATION VIA TRAJECTORY MATCHING

Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a real image–text dataset containing N paired samples, and let $\mathcal{S} = \{(\widetilde{x}_j, \widetilde{y}_j)\}_{j=1}^{M}$ be its learnable synthetic counterpart, where $M \ll N$. We use *w/o Sim.* and *w/ Sim.* to denote settings without and with a learnable soft similarity matrix, respectively. The former optimizes only the synthetic pairs, with fixed diagonal positives and off-diagonal negatives; the latter also optimizes an $M \times M$ matrix of pairwise targets. For each modality $m \in \mathcal{M} = \{\mathrm{img}, \mathrm{txt}\}$, $f_m(\cdot; \Theta_m)$ denotes its representation function, where Θ_m contains only the trajectory-matched module and fixed components are omitted. Given a paired dataset $\mathcal{D}$, $\mathcal{L}_{\mathrm{NCE}}(\Theta; \mathcal{D})$ denotes the symmetric image–text contrastive loss based on noise-contrastive estimation (NCE) (van den Oord et al., 2018), and $\mathcal{L}_{\mathrm{WBCE}}(\Theta; \mathcal{D})$ is the weighted binary cross-entropy loss introduced by LoRS (Xu et al., 2024). The student loss is

$$\mathcal{L}_{\mathrm{student}} = \begin{cases} \mathcal{L}_{\mathrm{NCE}}, & \text{w/o Sim.}, \\ \mathcal{L}_{\mathrm{WBCE}}, & \text{w/ Sim.}, \end{cases}$$

For each modality, let $\{\Theta^*_{m,t}\}_{t=0}^{H}$ be the expert trajectory over H epochs on $\mathcal{T}$. The conventional protocol studied here uses $R = 1$, so each sampled target is a one-epoch update. Standard bi-trajectory matching initializes the student at $\widehat{\Theta}^{(0)}_m = \Theta^*_{m,s}$, updates it on $\mathcal{S}$ for R' steps to obtain $\widehat{\Theta}^{(R')}_m$, and minimizes

$$\mathcal{L}_{\mathrm{bi}} = \sum_{m \in \mathcal{M}} \frac{\left\| \widehat{\Theta}^{(R')}_m - \Theta^*_{m,s+R} \right\|_2^2}{\left\| \Theta^*_{m,s} - \Theta^*_{m,s+R} \right\|_2^2}, \tag{1}$$

where s is the sampled expert starting epoch, R is the expert horizon, and R' is the number of student updates. Let $\widehat{\Theta}^{(r)} = (\widehat{\Theta}^{(r)}_{\mathrm{img}}, \widehat{\Theta}^{(r)}_{\mathrm{txt}})$ collect the matched student parameters after update r, and let η denote the student step size. MDD can then be written as the bilevel optimization

$$\begin{aligned} \mathcal{S}^* &= \arg\min_{\mathcal{S}} \mathcal{L}_{\mathrm{bi}}, \\ \widehat{\Theta}^{(r+1)} &= \widehat{\Theta}^{(r)} - \eta \nabla_{\widehat{\Theta}} \mathcal{L}_{\mathrm{student}} \left(\widehat{\Theta}^{(r)}; \mathcal{S} \right). \end{aligned} \tag{2}$$

3.2 DIRECTION-CONSISTENT PROGRESSIVE TRAJECTORY MATCHING

Overall start–end update versus epoch-wise matching. For modality m, define the one-epoch expert update as $d^*_{m,t} = \Theta^*_{m,t+1} - \Theta^*_{m,t}$. The overall update across R epochs beginning at s is

$$\Delta^*_{m,s:R} = \Theta^*_{m,s+R} - \Theta^*_{m,s} = \sum_{t=s}^{s+R-1} d^*_{m,t}. \tag{3}$$

Proposition 1 (epoch-wise matching is sufficient but not necessary). Consider any sequence of updates $\{\widehat{d}_{m,t}\}_{t=s}^{s+R-1}$ produced by a candidate synthetic set, and let $\widehat{\Delta}_{m,s:R} = \sum_{t=s}^{s+R-1} \widehat{d}_{m,t}$. Matching every epoch-wise update, $\widehat{d}_{m,t} = d^*_{m,t}$ for all t, is sufficient to obtain $\widehat{\Delta}_{m,s:R} = \Delta^*_{m,s:R}$. It is not necessary when $R > 1$: equality of the overall updates only requires

$$\sum_{t=s}^{s+R-1} \left(\widehat{d}_{m,t} - d^*_{m,t} \right) = 0, \tag{4}$$

which can hold even when individual epoch-wise updates differ. Moreover, by the Cauchy–Schwarz inequality,

$$\left\| \widehat{\Delta}_{m,s:R} - \Delta^*_{m,s:R} \right\|_2^2 \le R \sum_{t=s}^{s+R-1} \left\| \widehat{d}_{m,t} - d^*_{m,t} \right\|_2^2. \tag{5}$$

Proof. Summing the epoch-wise equalities gives equality of the overall updates. Conversely, Equation 4 constrains only the sum of the epoch-wise errors; for example, two nonzero errors with equal

magnitudes and opposite directions cancel. The bound follows by applying the Cauchy–Schwarz inequality to their sum.

Proposition 1 shows that reproducing every intermediate update imposes more conditions than reproducing the overall start-to-end update. The triangle inequality further gives

$$\left\|\Delta^*_{m,s:R}\right\|_2 \leq \sum_{t=s}^{s+R-1} \left\|d^*_{m,t}\right\|_2, \tag{6}$$

with equality only when the nonzero epoch-wise updates share the same direction. Figure 1 shows changing epoch-wise directions but a consistent cross-epoch trend among experts. PLTM therefore retains $\Delta^*_{m,s:R}$ and constructs targets from the selected start and end states. This result applies to parameter updates under a fixed model parameterization and does not imply functional equivalence.

At outer distillation iteration k, we first determine the normalized progress and matching-interval length as

$$q_k = \min\left(1, \frac{k}{K}\right), \qquad w_k = w_{\min} + (w_{\max} - w_{\min})q_k, \tag{7}$$

where I is the number of outer updates, $k \in \{0, \ldots, I-1\}$, $K = \max(1, I)$, ρ is the matching-range factor, $w_{\min} = 1/\rho$, and $w_{\max}$ is the maximum interval. We assume $0 < w_{\min} \leq w_{\max} \leq 1$.

To place the interval on the linearized trajectory, we sample its starting coefficient $a_k \sim \mathcal{U}(0, 1-w_k)$ and set $b_k = a_k + w_k$. For each modality m, the interpolated start and target are

$$\begin{aligned} \overline{\Theta}^0_{m,k} &= \Theta^*_{m,s} + a_k \Delta^*_{m,s:R}, \\ \overline{\Theta}^1_{m,k} &= \Theta^*_{m,s} + b_k \Delta^*_{m,s:R}. \end{aligned} \tag{8}$$

The target update is $\overline{\Theta}^1_{m,k} - \overline{\Theta}^0_{m,k} = w_k \Delta^*_{m,s:R}$. In the implementation, $s = 0$ and R equals `expert_epochs`. We initialize the student as $\widehat{\Theta}^{(0)}_{m,k} = \overline{\Theta}^0_{m,k}$ and update it on $\mathcal{S}$ for R' steps to obtain $\widehat{\Theta}^{(R')}_{m,k}$. The loss is

$$\mathcal{L}^{(k)}_{\mathrm{PLTM}} = \sum_{m \in \mathcal{M}} \frac{\left\|\widehat{\Theta}^{(R')}_{m,k} - \overline{\Theta}^1_{m,k}\right\|_2^2}{\left\|\overline{\Theta}^0_{m,k} - \overline{\Theta}^1_{m,k}\right\|_2^2}. \tag{9}$$

Algorithm 1 Progressive Linearized Trajectory Matching

Input: real set $\mathcal{T}$; synthetic set $\mathcal{S}$; expert trajectories $\{\Theta^*_{m,t}\}$; supervision setting; horizons R, R'; outer updates I; window bounds $w_{\min}, w_{\max}$
Output: optimized synthetic set $\mathcal{S}$
Set $s \leftarrow 0$ and $K \leftarrow \max(1, I)$
for $k = 0$ to $I - 1$ **do**
  Compute q_k and w_k using Equation 7
  Sample $a_k \sim \mathcal{U}(0, 1 - w_k)$; set $b_k \leftarrow a_k + w_k$
  for $m \in \mathcal{M}$ **do**
    $\Delta^*_{m,s:R} \leftarrow \Theta^*_{m,s+R} - \Theta^*_{m,s}$
    Construct $\overline{\Theta}^0_{m,k}$ and $\overline{\Theta}^1_{m,k}$ using Equation 8
    Initialize $\widehat{\Theta}^{(0)}_{m,k} \leftarrow \overline{\Theta}^0_{m,k}$
  end for
  for $r = 0$ to $R' - 1$ **do**
    Sample a mini-batch $\mathcal{B} \subset \mathcal{S}$
    Update both modalities using $\mathcal{L}_{\mathrm{NCE}}$ (w/o Sim.) or $\mathcal{L}_{\mathrm{WBCE}}$ (w/ Sim.) on $\mathcal{B}$
  end for
  Compute $\mathcal{L}^{(k)}_{\mathrm{PLTM}}$ using Equation 9
  For w/o Sim., compute $\mathcal{L}_{\mathrm{reg}}$ on $\mathcal{B}$; for w/ Sim., set $\mathcal{L}_{\mathrm{reg}} \leftarrow 0$
  Update $\mathcal{S}$ using $\mathcal{L}^{(k)}_{\mathrm{total}}$
end for

3.3 Cross-Modal Relational Alignment and Diversity Regularization

Paired hard labels identify the matched image and text in each pair but do not describe relations among different pairs. We therefore obtain these relations directly from the current synthetic mini-batch. Consider B paired synthetic samples. Let $z_i^m \in \mathbb{R}^d$ be the ℓ_2-normalized representation of sample i from modality $m \in \mathcal{M}$, and stack these row-wise into $Z_m \in \mathbb{R}^{B\times d}$. We compute a within-modality Gram matrix from the current synthetic representations,

$$G_m = Z_m Z_m^\mathsf{T}, \qquad (G_m)_{ij} = (z_i^m)^\mathsf{T} z_j^m, \tag{10}$$

where $(G_m)_{ij}$ is the cosine similarity between samples i and j in modality m. Rather than optimizing an additional soft similarity matrix or directly forcing paired image and text embeddings to coincide, we align their pairwise structures through

$$\mathcal{L}_{\text{align}} = \frac{1}{B^2} \left\| G_{\text{img}} - G_{\text{txt}} \right\|_F^2, \tag{11}$$

where B is the synthetic mini-batch size and $\|\cdot\|_F$ denotes the Frobenius norm.

Relational alignment alone can admit a solution in which many samples have similar representations. We therefore penalize excessive off-diagonal similarity within each modality:

$$\mathcal{L}_{\text{div}}^m = \frac{1}{B(B-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{B} \big((G_m)_{ij}\big)^2, \qquad m \in \mathcal{M}. \tag{12}$$

The diagonal terms are excluded since normalized representations satisfy $(G_m)_{ii} = 1$. The combined regularizer is

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{div}} \sum_{m\in\mathcal{M}} \mathcal{L}_{\text{div}}^m, \tag{13}$$

where $\lambda_{\text{align}} \geq 0$ and $\lambda_{\text{div}} \geq 0$ control relational alignment and diversity, respectively.

3.4 Overall Objective and Training Algorithm

At each outer iteration k, we optimize the synthetic image–text pairs using

$$\mathcal{L}_{\text{total}}^{(k)} = \begin{cases} \mathcal{L}_{\text{PLTM}}^{(k)} + \mathcal{L}_{\text{reg}}, & \text{w/o Sim.}, \\ \mathcal{L}_{\text{PLTM}}^{(k)}, & \text{w/ Sim.}. \end{cases} \tag{14}$$

The regularizers directly constrain the synthetic images and texts, but not the soft similarity matrix. Adding them under w/ Sim. may therefore change the data representations without corresponding changes to the learned pairwise targets. Since the soft matrix already provides relations among samples, we omit the regularizers in this setting. Algorithm 1 summarizes the complete optimization procedure.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets and Metrics.

We conduct experiments on Flickr-30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014), two standard image–text retrieval benchmarks containing approximately 31K and 123K images, respectively, with five captions associated with each image. We evaluate bidirectional retrieval using Recall@K for $K \in \{1, 5, 10\}$. Following the notation in prior MDD work, IR@K denotes text-to-image retrieval and TR@K denotes image-to-text retrieval. The distillation budgets are 100, 200, and 500 synthetic image–text pairs; the corresponding compression ratios are reported in each table. Following prior work, we adopt NFNet (Brock et al., 2021), RegNet (Radosavovic et al., 2020), and ResNet-50 (He et al., 2016) as image encoders and BERT (Devlin et al., 2019) as the text encoder. NFNet+BERT is used in the primary protocol, while RegNet+BERT and ResNet-50+BERT are used for cross-architecture evaluation.

Table 1: Flickr-30K retrieval results (Plummer et al., 2015) with NFNet+BERT for distillation and evaluation. Full-data IR@1/5/10 and TR@1/5/10 are 23.2/54.0/66.6 and 33.8/65.7/76.9, respectively. Gray columns denote our variants; bold indicates the better result within each matched supervision setting.

Pairs	Ratio	Metric	Coreset Selection				MDD Baselines			w/ Sim.		w/o Sim.	
			Rand	Herd	K-Cent	Forget	MTT-VL TMLR'24	LoRS ICML'24	AMD ICLR'26	RepBlend NeurIPS'25	Ours –	RepBlend NeurIPS'25	Ours –
100	0.3%	IR@1	1.0	0.7	0.7	0.7	4.7±0.2	8.3±0.2	10.4±0.3	**11.5±0.4**	11.4±0.4	10.3±0.3	**12.5±0.2**
		IR@5	4.0	2.8	3.1	2.4	15.7±0.5	24.1±0.2	30.5±0.7	32.0±0.7	**32.0±0.9**	29.8±0.1	**35.6±0.1**
		IR@10	6.5	5.3	6.1	5.6	24.6±1.0	35.1±0.3	43.0±0.6	44.5±0.6	**44.8±1.1**	41.2±0.6	**48.6±0.1**
		TR@1	1.3	1.1	0.6	1.2	9.9±0.3	11.8±0.2	14.4±0.5	16.2±0.8	**17.4±0.9**	12.4±0.4	**16.1±0.5**
		TR@5	5.9	4.7	5.0	4.2	28.3±0.5	35.8±0.6	39.1±0.6	41.7±0.9	**43.5±1.0**	36.1±0.5	**42.6±0.9**
		TR@10	10.1	7.9	7.6	9.7	39.1±0.7	49.2±0.5	52.6±0.6	55.5±0.4	**57.9±0.8**	50.4±0.4	**57.8±1.1**
200	0.7%	IR@1	1.1	1.5	1.5	1.2	4.6±0.9	8.6±0.3	13.1±0.3	12.7±0.8	**14.2±0.4**	10.4±0.3	**14.5±0.3**
		IR@5	4.8	5.5	5.4	3.1	16.0±1.6	25.3±0.2	34.9±0.6	34.7±0.6	**37.1±0.7**	29.3±0.4	**38.6±0.4**
		IR@10	9.2	9.3	9.9	8.4	25.5±2.6	36.6±0.3	47.5±0.7	47.6±0.5	**50.2±0.5**	41.0±0.5	**51.5±0.4**
		TR@1	2.1	2.3	2.2	1.5	10.2±0.8	14.5±0.5	16.9±0.4	18.6±0.7	**21.3±0.2**	12.8±0.5	**19.6±0.8**
		TR@5	8.7	8.4	8.2	8.4	28.7±1.0	38.7±0.5	42.3±0.6	46.0±0.8	**50.9±0.6**	36.7±0.5	**46.9±0.6**
		TR@10	13.2	14.4	13.5	10.2	41.9±1.9	53.4±0.5	56.2±0.8	60.0±0.6	**64.4±0.8**	49.9±0.9	**61.5±0.6**
500	1.7%	IR@1	2.4	3.0	3.5	1.8	6.6±0.3	10.0±0.2	15.8±0.4	17.0±0.6	**17.3±0.2**	9.5±0.2	**17.7±0.1**
		IR@5	10.5	10.0	10.4	9.0	20.2±1.2	28.9±0.7	39.8±0.4	42.5±0.5	**43.0±0.3**	27.4±0.4	**43.5±0.3**
		IR@10	17.4	17.0	17.3	15.9	30.0±2.1	41.6±0.6	53.2±0.5	55.9±0.6	**56.7±0.5**	37.7±0.4	**56.3±0.4**
		TR@1	5.2	5.1	4.9	3.6	13.3±0.6	15.5±0.7	19.3±0.5	**22.5±0.4**	22.2±0.6	16.0±0.7	**24.0±0.3**
		TR@5	18.3	16.4	16.4	12.3	32.8±1.8	39.8±0.4	46.4±0.4	**53.2±0.3**	52.9±0.6	42.5±0.4	**55.9±0.9**
		TR@10	25.7	24.3	23.3	19.3	46.8±0.8	53.7±0.3	60.0±0.6	66.7±0.3	**67.5±0.6**	57.1±0.7	**70.0±0.3**

Table 2: MS-COCO retrieval results (Lin et al., 2014) with NFNet+BERT for distillation and evaluation. Full-data IR@1/5/10 and TR@1/5/10 are 14.6/38.9/53.2 and 20.6/46.8/61.3, respectively. Gray columns denote our variants; bold indicates the better result within each matched supervision setting.

Pairs	Ratio	Metric	Coreset Selection				MDD Baselines			w/ Sim.		w/o Sim.	
			Rand	Herd	K-Cent	Forget	MTT-VL TMLR'24	LoRS ICML'24	AMD ICLR'26	RepBlend NeurIPS'25	Ours –	RepBlend NeurIPS'25	Ours –
100	0.08%	IR@1	0.3	0.5	0.4	0.3	1.3±0.1	1.8±0.1	2.8±0.2	4.1±0.3	**6.5±0.1**	4.7±0.2	**5.0±0.1**
		IR@5	1.3	1.4	1.4	1.5	5.4±0.3	7.1±0.2	10.5±0.2	13.9±0.8	**20.9±0.2**	15.8±0.5	**16.8±0.2**
		IR@10	2.7	3.5	2.5	2.5	9.5±0.5	12.2±0.2	17.2±0.4	22.3±0.5	**31.8±0.2**	25.3±0.6	**26.6±0.3**
		TR@1	0.8	0.8	1.4	0.7	2.5±0.3	3.3±0.2	4.1±0.3	5.2±0.5	**6.1±0.1**	4.9±0.1	**5.3±0.1**
		TR@5	3.0	2.1	3.7	2.6	10.0±0.5	12.2±0.3	13.8±0.3	17.9±0.9	**20.5±0.1**	16.9±0.2	**17.9±0.2**
		TR@10	5.0	4.9	5.5	4.8	15.7±0.4	19.6±0.3	21.8±0.4	28.0±0.3	**30.9±0.3**	26.5±0.5	**27.7±0.1**
200	0.17%	IR@1	0.6	0.9	0.7	0.6	1.7±0.1	2.4±0.1	3.8±0.2	6.1±0.8	**6.8±0.2**	4.2±0.2	**5.6±0.2**
		IR@5	2.3	2.4	2.1	2.8	6.5±0.4	9.3±0.2	13.4±0.3	19.3±0.7	**21.3±0.5**	14.6±0.2	**18.4±0.1**
		IR@10	4.4	4.1	5.8	4.9	12.3±0.8	15.5±0.2	21.4±0.4	29.8±0.5	**32.4±0.7**	23.6±0.3	**28.6±0.3**
		TR@1	1.0	1.0	1.2	1.1	3.3±0.2	4.3±0.1	4.6±0.2	6.9±0.6	**7.0±0.3**	5.2±0.2	**6.5±0.5**
		TR@5	4.0	3.6	3.8	3.5	11.9±0.6	14.2±0.3	15.5±0.6	21.8±0.9	**22.1±0.5**	17.9±0.1	**21.2±0.3**
		TR@10	7.2	7.7	7.5	7.0	19.4±1.2	22.6±0.2	24.1±0.5	32.3±0.7	**33.6±0.5**	28.3±0.4	**32.0±0.5**
500	0.44%	IR@1	1.1	1.7	1.1	0.8	2.5±0.5	2.8±0.2	4.2±0.2	6.2±0.1	**7.3±0.1**	4.3±0.2	**6.3±0.1**
		IR@5	5.0	5.3	6.3	5.8	8.9±0.7	9.9±0.5	14.2±0.5	19.9±0.3	**23.1±0.2**	14.5±0.3	**20.1±0.4**
		IR@10	8.7	9.9	10.5	8.2	15.8±1.5	16.5±0.7	22.3±0.4	30.6±0.1	**34.8±0.3**	23.1±0.5	**30.9±0.4**
		TR@1	1.9	1.9	2.5	2.1	5.0±0.4	5.3±0.5	5.7±0.6	7.0±0.2	**8.3±0.4**	6.9±0.1	**7.9±0.2**
		TR@5	7.5	7.8	8.7	8.2	17.2±1.3	18.3±1.5	19.3±1.2	22.0±0.3	**25.4±0.5**	21.4±0.3	**26.2±0.4**
		TR@10	12.5	13.7	14.3	13.0	26.0±1.9	27.9±1.4	28.7±1.0	32.9±0.6	**37.3±0.6**	32.6±0.7	**38.7±0.4**

4.1.2 Implementation Details.

Following prior vision–language dataset-distillation protocols (Wu et al., 2024; Zhang et al., 2025), we use a CLIP-style dual encoder built from the backbones above. Each image backbone is ini-

Table 3: Component ablation on Flickr-30K with 500 synthetic pairs. mR averages R@1/5/10, and ΔmR is the absolute gain over the baseline.

Method	R@1	R@5	R@10	mR	ΔmR
Baseline	12.6	34.9	47.9	31.8	–
LTM w/ Sim.	18.8	47.3	61.1	42.4	+10.6
LTM	18.3	45.0	59.4	40.9	+9.1
PLTM	20.1	48.2	61.6	43.3	+11.5
PLTM+div	20.8	48.8	62.5	44.0	+12.2
PLTM+align	20.0	47.9	62.5	43.5	+11.7
Ours	**20.9**	**49.9**	**63.8**	**44.8**	**+13.0**

Table 5: Flickr-30K asymmetric-architecture results with paired hard targets and no soft similarity matrix.

Pairs	Method	R@1	R@5	R@10	mR
100	Baseline	9.5	27.0	39.1	25.2
	LoRS	10.1	30.0	42.2	27.4
	+PLTM	**10.6**	**31.4**	**44.5**	**28.8**
200	Baseline	10.1	30.1	43.1	27.7
	LoRS	11.6	32.0	45.0	29.5
	+PLTM	**12.1**	**33.0**	**45.8**	**30.3**
500	Baseline	13.7	37.6	50.7	34.0
	LoRS	12.8	34.4	47.7	31.6
	+PLTM	**16.8**	**39.9**	**53.5**	**36.7**

Table 4: Cross-architecture Flickr-30K retrieval with 500 NFNet+BERT-distilled hard-target pairs. The best result for each architecture is bold.

Evaluate Model	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
	AMD	4.1±0.3	14.2±0.5	22.6±0.4	7.6±0.4	22.3±0.4	33.2±0.6
	Ours	**5.6±0.2**	**18.9±0.5**	**28.4±0.5**	**10.9±0.8**	**28.7±0.4**	**40.6±0.4**
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**
	AMD	4.0±0.2	14.4±0.3	23.1±0.5	7.5±0.3	22.8±0.5	32.7±0.4
	Ours	**5.0±0.2**	**15.2±0.3**	23.3±0.3	**8.3±0.7**	**24.7±0.6**	35.3±0.3

tialized from ImageNet-pretrained weights, while BERT uses its official pretrained weights. In the symmetric setting, trajectory matching is applied only to the independently initialized projection modules appended to the image and text encoders. In the asymmetric setting, the text branch retains a projection module whereas the image branch has no projection module, so the recorded trajectories correspond to the text projection and the image encoder, respectively. Before distillation, we train 20 expert models on the full real dataset following the same expert-training protocol as prior MDD methods and record their 10-epoch parameter trajectories. We follow RepBlend for shared hyperparameters; PLTM-specific settings are provided in Tables 6–8 of Appendix B. For a fair comparison, all methods use only early expert trajectories and receive no information from later expert states. PLTM constructs targets from the start and end states within this early range, while each baseline follows its published epoch-sampling protocol. Unless stated otherwise, all experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.1.3 Baselines.

We compare against four coreset-selection methods—Random sampling, Herding (Welling, 2009), K-Center (Farahani & Hekmatfar, 2009), and Forgetting (Toneva et al., 2019)—and four multimodal dataset-distillation methods: MTT-VL (Wu et al., 2024), LoRS (Xu et al., 2024), RepBlend (Zhang et al., 2025), and AMD (Qi et al., 2026).

4.2 Comparison with State-of-the-Art Methods

Tables 1 and 2 evaluate PLTM under two forms of cross-modal supervision. *Ours w/o Sim.* uses paired hard supervision with $\mathcal{L}_{\mathrm{NCE}}$, PLTM, relational alignment, and diversity regularization. *Ours w/ Sim.* instead trains student updates with the LoRS $\mathcal{L}_{\mathrm{WBCE}}$ loss and a learnable soft similarity matrix; $\mathcal{L}_{\mathrm{reg}}$ is omitted. For a controlled comparison, *RepBlend w/o Sim.* is the corresponding paired-hard-supervision baseline. Averaged over all six retrieval metrics, the gains are **5.5%, 8.8%,**

and 12.9% on Flickr-30K and **0.9%, 3.1%, and 4.6%** on MS-COCO for 100, 200, and 500 pairs, respectively.

With the same soft similarity supervision as RepBlend, our method improves mean recall by **0.9%, 3.1%, and 0.3%** on Flickr-30K and **4.2%, 1.2%, and 2.9%** on MS-COCO. Across both settings, our method attains the **highest score for every Flickr-30K metric and budget**. Figure 3 further shows that the optimized images retain the main objects and layouts of their sources while introducing fine-grained textures and fine-scale appearance variations.

4.3 Ablation Studies

4.3.1 Component Contributions.

Table 3 isolates progression and the two regularizers on Flickr-30K with 500 pairs. We report $\mathrm{R@}K = (\mathrm{IR@}K + \mathrm{TR@}K)/2$ and their mean mR. LTM uses the same straight start-to-end path as PLTM but matches a fixed long interval from the beginning of distillation, whereas PLTM increases the interval from short to long as the synthetic pairs improve. Except for "LTM w/ Sim.," all variants use paired hard supervision. Replacing LTM with PLTM raises mR from 40.9 to 43.3, a **2.4% gain**, directly supporting the progressive matching strategy.

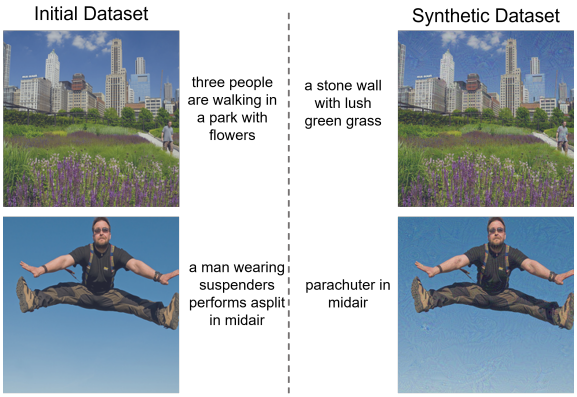


Figure 3: Representative Flickr-30K source images (left) and their optimized synthetic counterparts (right).

Adding diversity or alignment to PLTM raises mR to 44.0 and 43.5, respectively, while their joint use reaches **44.8 mR**. The complete method therefore improves over the baseline and LTM by **13.0%** and **3.9%**, respectively, and exceeds similarity-supervised LTM by **2.4%** without requiring a soft similarity matrix.

4.3.2 Supervision and Architecture Settings.

To assess cross-architecture transfer, we train unseen image encoders on the NFNet+BERT-distilled pairs. As shown in Table 4, PLTM consistently maintains strong retrieval performance across different image-encoder architectures, indicating that its gains are not confined to the distillation backbone. Table 5 further tests a fixed asymmetric architecture throughout distillation and evaluation. PLTM improves mR over the baseline by **3.6%, 2.6%, and 2.7%** at 100, 200, and 500 pairs, respectively. At 500 pairs, it raises R@1/5/10 by **3.1%, 2.3%, and 2.8%**, whereas LoRS falls below the baseline at the same budget. These results indicate that direction-consistent progressive matching remains effective when the two modalities follow different optimization dynamics.

5 Conclusion

PLTM replaces changing epoch-specific targets with a shared start-to-end direction and progressively increasing matching intervals. Under paired hard supervision, relational alignment recovers cross-modal structure and diversity regularization reduces within-modality redundancy, strengthening supervision without a soft similarity matrix or extra model passes. Experiments on Flickr-30K and MS-COCO show consistent retrieval gains at nearly twice the distillation efficiency.

AI USE STATEMENT

Generative AI tools assisted with language editing, literature organization, and analysis and interpretation of experimental results. They also assisted with manuscript assessment and LaTeX formatting, including template migration. The authors take responsibility for the final content of this work, including its text, citations, scientific claims, and reported results.

REPRODUCIBILITY STATEMENT

Section 3 specifies the optimization objectives and provides the proof of Proposition 1 and the training procedure in Algorithm 1. Section 4.1 describes the datasets, evaluation metrics, architecture settings, and expert-training protocol. The appendices document supervision and matched modules in Appendix A, expert-buffer and budget-specific hyperparameters in Appendix B (Tables 6–8), regularization sensitivity in Appendix C, additional 500-pair comparisons in Appendix D, and the loss-adaptive extension in Appendix E.

REFERENCES

Deyu Bo and Xinchao Wang. Beyond soft label: Dataset distillation via orthogonal gradient matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5596–5605, 2026.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1059–1071. PMLR, 2021.

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*, pp. 10718–10727, 2022.

Junhyeok Choi, Sangwoo Mo, and Minwoo Chae. Multimodal dataset distillation made simple by prototype-guided data synthesis. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bl3drImevi.

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

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)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Priyam Dey, Aditya Sahdev, Sunny Bhati, Konda Reddy Mopuri, and Venkatesh Babu Radhakrishnan. Rethinking dataset distillation: Hard truths about soft labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 178–187, 2026.

Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67487–67504. Curran Associates, Inc., 2023. doi: 10.52202/075280-2950.

Reza Zanjirani Farahani and Masoud Hekmatfar (eds.). *Facility Location: Concepts, Models, Algorithms and Case Studies*. Contributions to Management Science. Physica-Verlag, Heidelberg, 2009. doi: 10.1007/978-3-7908-2151-2.

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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Me4AON8160.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rTBL8OhdhH.

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

Jongoh Jeong, Hoyong Kwon, Minseok Kim, and Kuk-Jin Yoon. Multimodal distribution matching for vision-language dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23072–23082, 2026.

Yongmin Lee and Hye Won Chung. CovMatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. In *Advances in Neural Information Processing Systems*, volume 38, pp. 101490–101518. Curran Associates, Inc., 2025. doi: 10.52202/085713-3398.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, Cham, 2014. Springer. doi: 10.1007/978-3-319-10602-1_48.

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*, pp. 2641–2649, 2015.

Ding Qi, Jian Li, Shuguang Dou, Zifan Song, Junyao Gao, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Asynchronous matching with dynamic sampling for multimodal dataset distillation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/0cdc6c37e7048387fe506c50c23225e7e3ca5f29.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

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*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748. arXiv preprint arXiv:1807.03748.

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*, pp. 12196–12205, 2022.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2018. URL https://arxiv.org/abs/1811.10959. arXiv preprint arXiv:1811.10959.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128. ACM, 2009. doi: 10.1145/1553374.1553517.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=6L6cD65pot.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55144–55161. PMLR, 2024.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representation blending for multimodal dataset distillation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114809–114829. Curran Associates, Inc., 2025. doi: 10.52202/085713-3827.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12674–12685. PMLR, 2021.

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

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf/f6babe705b5d43a6a88ee3b553c209341b63a5e0.pdf.

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 67912–67933. Curran Associates, Inc., 2025. doi: 10.52202/085713-2280.

Wenliang Zhong, Haoyu Tang, Qinghai Zheng, Mingzhu Xu, Yupeng Hu, and Weili Guan. Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25581–25589, 2025.

A Supervision and Architecture Protocols

Supervision settings. The labels w/o Sim. and w/ Sim. denote settings without and with a learnable soft similarity matrix, respectively. In w/o Sim., the synthetic variables are the paired images and texts, student updates use the symmetric contrastive loss based on noise-contrastive estimation (NCE), $\mathcal{L}_{\mathrm{NCE}}$, and both regularizers are enabled. In w/ Sim., the soft matrix is optimized together with the paired images and texts, and student updates use the LoRS weighted binary cross-entropy loss, $\mathcal{L}_{\mathrm{WBCE}}$. We omit the regularizers because they directly constrain the data but not the soft matrix, which may leave changes in the representations and learned targets inconsistent. The soft matrix also already provides relations among samples.

Matched modules. In the symmetric setting, the image and text encoders use pretrained weights and only their independently initialized projection modules are matched. In the asymmetric setting, the image branch has no projection module; the matched modules are therefore the image encoder and the text projection. Both settings follow the expert-training protocols of the corresponding MDD baselines.

B Implementation Details and Hyperparameters

Table 6 reports the shared expert-buffer settings, and Tables 7–8 report the budget-specific distillation settings. Soft-similarity parameters apply only to w/ Sim., whereas `lambda_align` and `lambda_div` apply only to w/o Sim. Each expert buffer contains 20 independently trained trajectories of 10 epochs. Both PLTM and RepBlend use eight student updates.

Table 6: Expert-buffer settings for the primary symmetric protocol.

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

Table 7: Flickr-30K distillation settings for 100/200/500 synthetic pairs.

Setting	100	200	500
`syn_steps` (R')	8	8	8
`iteration`	1000	1000	1000
`lr_img` / `lr_txt`	100	1000	1000
`lr_lr`	10^{-2}	10^{-2}	10^{-2}
`mini_batch_size`	20	20	40
`start_epoch`	0	0	0
`end_epoch`	2	3	4
`lr_sim`	10	10	100
`sim_rank`	10	5	20
`sim_alpha`	3.0	1.0	0.01
`range_factor_rho`	4.0	4.0	3.0
`max_win_len`	0.7	0.7	0.6
`lambda_div`	0.05	0.05	0.05
`lambda_align`	0.05	0.05	0.1

For w/ Sim., `lr_sim`, `sim_rank`, and `sim_alpha` follow the corresponding RepBlend settings. All trajectory-matching methods use only early expert states. PLTM uses the `start_epoch` and `end_epoch` reported in Tables 7–8 and receives no later expert states.

Table 8: MS-COCO distillation settings for 100/200/500 synthetic pairs.

Setting	100	200	500
`syn_steps` (R')	8	8	8
`iteration`	1000	1000	1000
`lr_img` / `lr_txt`	1000	1000	5000
`lr_lr`	10^{-2}	10^{-2}	10^{-2}
`mini_batch_size`	20	20	30
`start_epoch`	0	0	0
`end_epoch`	1	2	3
`lr_sim`	5.0	50.0	500.0
`sim_rank`	10	20	40
`sim_alpha`	1.0	1.0	1.0
`range_factor_rho`	5.0	4.5	3.0
`max_win_len`	0.8	0.9	0.9
`lambda_div`	0.025	0.001	0.01
`lambda_align`	0.05	0.01	0.01

C Regularization-Coefficient Sensitivity

Table 9 presents a one-at-a-time ablation of the two regularization coefficients on MS-COCO with 200 synthetic pairs. With $\lambda_{\text{div}} = 0.001$, changing λ_{align} from 0.005 to 0.010 changes mR from 18.9 to 19.1. With $\lambda_{\text{align}} = 0.010$, changing λ_{div} from 0.001 to 0.005 changes mR from 19.1 to 18.6. These results describe sensitivity only over the tested coefficient values at this dataset and budget; they do not assess the window bounds or the progression schedule.

Table 9: Regularization-coefficient sensitivity on MS-COCO with 200 synthetic pairs. Each block varies one setting while fixing the other.

Varied setting	Value	mR (%)
Fixed: `lambda_div`$= 0.001$		
`lambda_align`	0.005	18.9
`lambda_align`	0.010	**19.1**
Fixed: `lambda_align`$= 0.010$		
`lambda_div`	0.001	**19.1**
`lambda_div`	0.005	18.6

D Additional 500-Pair Comparison

Tables 10 and 11 extend the 500-pair comparison to CovMatch, EDGE, PTM-ST, and AMD. Ours w/ Sim. denotes the soft-similarity setting reported in the main paper. Standard deviations are omitted because they are unavailable for several baselines; mR averages the six retrieval recalls.

Table 10: Additional Flickr-30K retrieval comparison with 500 synthetic pairs. Recall columns list R@1/5/10; all results are percentages, and the best result is bold.

Method	IR@1/5/10	TR@1/5/10	mR
Random	2.4/10.5/17.4	5.2/18.3/25.7	13.3
LoRS	10.0/28.9/41.6	15.5/39.8/53.7	31.6
RepBlend	17.0/42.5/55.9	**22.5/53.2**/66.7	43.0
CovMatch	14.7/38.4/51.4	19.9/46.7/59.5	38.4
EDGE	6.7/21.0/30.5	13.3/35.6/47.5	25.8
PTM-ST	16.0/40.5/54.0	22.2/51.1/64.6	41.4
AMD	15.8/39.8/53.2	19.3/46.4/60.0	39.1
Ours w/ Sim.	**17.3/43.0/56.7**	22.2/52.9/**67.5**	**43.3**

Table 11: Additional MS-COCO retrieval comparison with 500 synthetic pairs. Recall columns list R@1/5/10; all results are percentages, and the best result is bold.

Method	IR@1/5/10	TR@1/5/10	mR
Random	1.1/5.0/8.7	1.9/7.5/12.5	6.1
LoRS	2.8/9.9/16.5	5.3/18.3/27.9	13.5
RepBlend	6.2/19.9/30.6	7.0/22.0/32.9	19.8
CovMatch	5.4/18.0/28.2	8.1/23.5/34.6	19.6
EDGE	1.8/6.5/11.2	2.9/9.5/15.7	7.9
PTM-ST	6.6/20.5/30.7	6.9/20.1/30.0	19.1
AMD	4.2/14.2/22.3	5.7/19.3/28.7	15.7
Ours w/ Sim.	**7.3/23.1/34.8**	**8.3/25.4/37.3**	**22.7**

E Loss-Adaptive Extension

E.1 Expert Dynamics under Asymmetric Architectures

Figure 4 reports image-side directional similarity and modality-specific update magnitudes for 20 independently trained expert trajectories under the asymmetric architecture protocol. The small off-diagonal similarities in Figure 4(a) show that one-epoch image-side directions can differ markedly. Their across-expert variances remain close to 10^{-6} in Figure 4(b), indicating that independently trained experts nevertheless exhibit a consistent cross-epoch trend. This observation supports using a shared start-to-end direction instead of matching each intermediate epoch-specific direction.

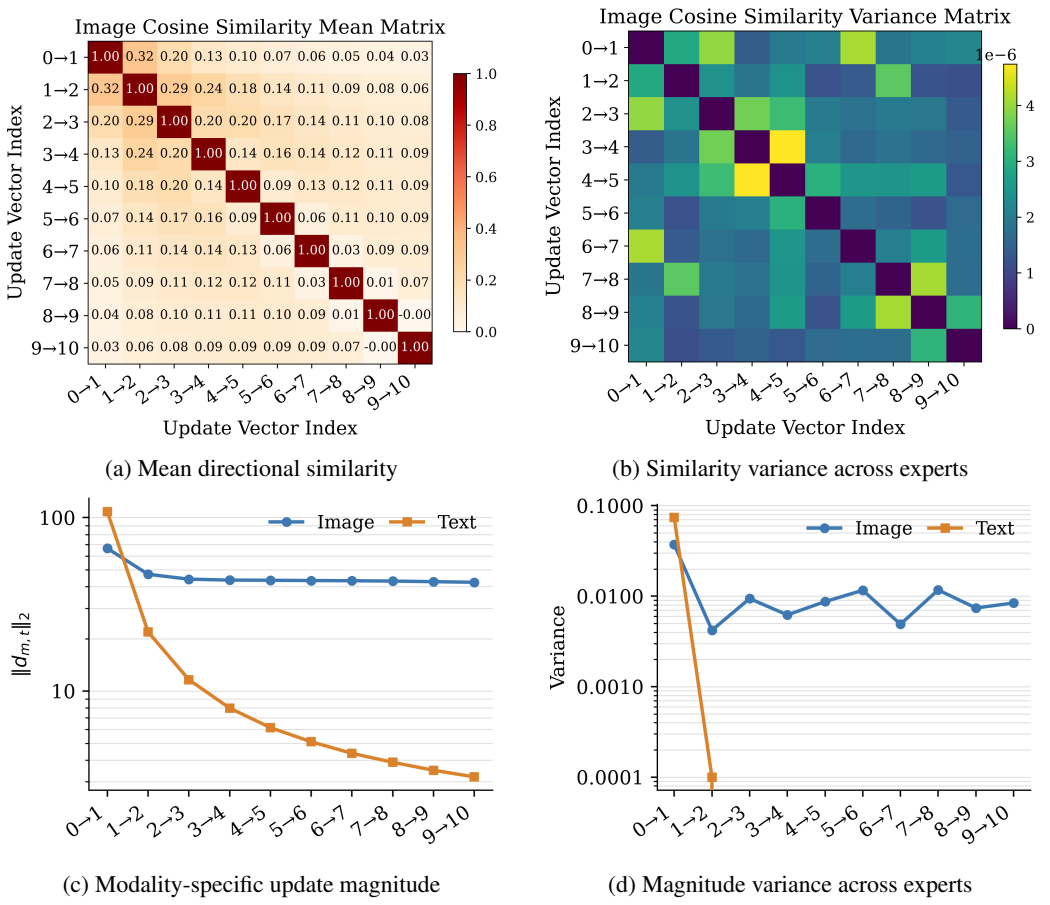
Table 12: Loss-adaptive modality-specific intervals on Flickr-30K with 500 synthetic pairs under the asymmetric architecture protocol. All results are percentages, and the best result is bold.

Temperature τ	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	mR
0.10	10.9	31.4	45.2	21.5	49.5	63.5	37.0
0.25	12.0	33.5	46.8	22.5	52.3	66.1	38.9
0.50	12.1	35.2	48.3	22.3	50.7	**66.4**	**39.2**
0.75	12.2	**35.9**	**48.9**	**23.1**	50.3	63.5	39.0
1.00	**12.4**	34.9	47.9	20.7	51.6	65.7	38.9

Figures 4(c,d) further show that the two modalities can progress at different rates. This motivates the following optional extension, which adapts the matched interval separately for each modality. The extension is evaluated independently and is not required by the core PLTM formulation.

E.2 Loss-Adaptive Modality-Specific Window Lengths

Let $\ell_m^{(k)}$ be the trajectory-matching loss of modality $m \in \{\mathrm{img}, \mathrm{txt}\}$ at outer iteration k, and let w_k be the shared PLTM interval fraction determined by the progression schedule. This shared value is



(a) Mean directional similarity

(b) Similarity variance across experts

(c) Modality-specific update magnitude

(d) Magnitude variance across experts

Figure 4: Expert-trajectory statistics over 20 independently trained trajectories under the asymmetric architecture protocol. (a) Mean cosine similarity between image-side one-epoch update directions; (b) variance of these similarities across experts; (c) mean image- and text-side update magnitudes; and (d) their across-expert variance.

then adapted to each modality as follows.

$$p_{m,k}(\tau) = \frac{\exp\left(\ell_m^{(k)}/\tau\right)}{\sum_{n\in\{\text{img,txt}\}} \exp\left(\ell_n^{(k)}/\tau\right)}, \qquad (15)$$
$$w_{m,k}^{\text{dyn}} = 2w_k p_{m,k}(\tau),$$

where τ is the temperature. The factor of two preserves the average interval fraction across the two modalities. Equal losses recover the shared interval w_k, whereas a modality with a larger matching loss receives a longer interval.

Table 12 ablates the temperature that controls how the matched interval is allocated between modalities. Moderate temperatures yield similar overall retrieval performance, whereas an overly small temperature is less effective. All tested settings improve mean recall over the shared-interval result reported in the main paper, supporting the general value of loss-adaptive allocation rather than a metric-specific temperature choice.

The extension changes only how the matched interval is allocated between the two modalities; it retains PLTM's shared start-to-end direction and progression schedule. These results indicate that modality-specific intervals can be useful when the modalities progress at different rates. The observed association between expert statistics and the preferred allocation is not a causal result and warrants further study.