TRUST WHAT YOU MATCH: RELIABLE CROSS-MODAL COUPLING MATCHING FOR MULTIMODAL DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal dataset distillation (MDD) aims to compress large-scale corpora into compact synthetic data while preserving their downstream utility. However, existing methods typically match cross-modal couplings estimated from stochastic mini-batches, making supervision highly sensitive to batch composition and consequently impairing preservation of dataset-level multimodal structure. To address this limitation, we propose Reliable Cross-Modal Coupling Matching (ReCoup), which calibrates encoder-space couplings using fluctuations induced by random image–text pairings within each mini-batch. Accordingly, ReCoup assigns higher weights to couplings that are more reproducible across batches and thus better reflect full-dataset statistics. The resulting reliability-aware objective yields more coherent synthetic-data gradients and reduces conflicting optimization updates, while an auxiliary constraint further preserves the diagonal coupling profile in the projected shared space. Experiments on Flickr30K and MS-COCO show consistent improvements over state-of-the-art MDD methods across synthetic-set budgets. With 500 synthetic pairs, ReCoup improves mean Recall by 7.3 points on Flickr30K and 5.1 points on MS-COCO.

1 INTRODUCTION

Training modern multimodal models typically requires massive paired datasets, imposing substantial computational costs on iterative experimentation. Multimodal dataset distillation (MDD) reduces this burden by condensing the original training signal into a compact synthetic set. A faithful distilled set should retain the semantics within each modality while preserving the cross-modal structure essential for multimodal alignment.

Existing multimodal dataset distillation methods mainly follow trajectory or distribution matching. Trajectory- and parameter-matching approaches, such as MTT-VL (Wu et al., 2023) and LoRS (Xu et al.), optimize synthetic data to reproduce the learning dynamics or parameter states induced by real data, but incur substantial computational and memory overhead. Distribution-matching approaches, such as CovMatch (Lee & Chung) and MDM (Jeong et al., 2026a), instead directly align representations or statistics between real and synthetic data. The effectiveness of these methods therefore depends critically on whether the matched cross-modal structure reliably reflects full-dataset statistics.

CLIP-style contrastive learning (Radford et al., 2021) constructs an image–text similarity matrix, whose entries are cosine similarities between image and text embeddings. These sample-wise similarities supervise contrastive training, but cannot be directly matched for distillation: real and synthetic sets differ in size and sample composition, so they have no natural sample-wise correspondence. A dimension-wise cross-modal coupling matrix instead aggregates over paired samples to capture interactions between visual and textual feature dimensions. Its size depends only on the feature dimensions rather than the number of samples, enabling direct matching between real and synthetic data.

In practical distillation, this coupling is estimated from stochastic mini-batches to avoid full-data computation, making individual entries unevenly reliable. Some remain consistent across batches, whereas others change sharply with batch composition. Uniformly matching them transfers these

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.

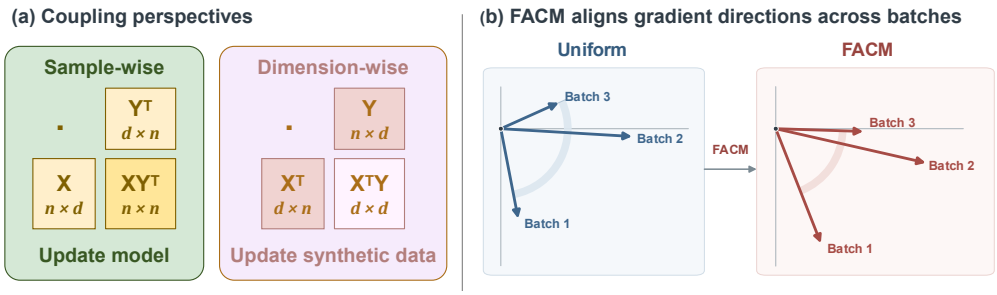


Figure 1: **From sample-wise similarities to reliable cross-modal couplings.** (a) Sample-wise similarities capture relations between image and text samples, whereas dimension-wise cross-modal couplings aggregate paired samples to capture interactions between visual and textual feature dimensions. (b) Uniform coupling matching across stochastic mini-batches can induce inconsistent synthetic gradients. FACM uses batch-level fluctuations to emphasize more reliable couplings.

batch-specific fluctuations to the synthetic data, causing different real mini-batches to induce inconsistent or even conflicting updates on the same synthetic pairs. Projection heads further reshape cross-modal coupling, and the projected representations develop a pronounced diagonal coupling structure in the shared embedding space. This projection-specific organization is not captured by encoder-space matching, motivating an explicit diagonal constraint to preserve it during distillation.

These observations motivate **ReCoup**, a Reliable Cross-Modal Coupling Matching framework for multimodal dataset distillation. **Fluctuation-Aware Cross-Modal Coupling Matching (FACM)** calibrates dense encoder-level coupling supervision against a random-pairing null. **Dimension-Wise Diagonal Coupling Matching (DDCM)** complements FACM with a compact projection-space constraint on corresponding cross-modal dimensions.

Experiments on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014) demonstrate consistent improvements across synthetic-set budgets. With 100, 200, and 500 image-text pairs, ReCoup outperforms prior methods across all retrieval metrics on both datasets. At the 100-pair budget, ReCoup improves mean Recall over the strongest prior method by 19.0% on Flickr30K and 19.6% on MS-COCO, with substantial gains persisting at larger budgets. Mechanism analyses further show that higher FACM scores are associated with stronger cross-batch coupling reproducibility and smaller deviations from full-data couplings. FACM also substantially reduces conflicts among gradient directions during synthetic-data optimization.

Our main contributions are as follows:

- We reveal that cross-modal couplings estimated from stochastic mini-batches are not equally reliable: batch-sensitive couplings are less reproducible and induce more conflicting synthetic-data updates.
- We propose **ReCoup** for reliable cross-modal coupling matching. FACM provides calibrated dense coupling supervision in the encoder space, while DDCM adds a compact dimension-wise constraint in the projected space.
- ReCoup consistently outperforms prior methods on Flickr30K and MS-COCO across synthetic-set budgets. Further analyses link FACM scores to coupling reproducibility and full-data coupling error, and reveal fewer gradient conflicts during optimization.

2 Preliminaries and Related Work

Dataset Distillation. Dataset distillation aims to transfer the training utility of large-scale real datasets to compact synthetic datasets. Existing methods mainly differ in the form of information transferred from real data to synthetic data. Gradient matching methods (Zhao et al.; Zhao & Bilen, 2021) align the optimization signals induced by real and synthetic data, whereas distribution matching methods (Zhao & Bilen, 2023; Zhao et al., 2023) directly align their representations or feature statistics. Trajectory matching methods (Cazenavette et al., 2022; Lee et al., 2022) optimize

synthetic data to reproduce the model states or learning dynamics obtained during training on real data. These paradigms form the primary foundations of modern dataset distillation and have been further extended to multimodal settings.

CLIP and Cross-Modal Representation Learning. CLIP-style vision-language models (Zhai et al., 2023; Tschannen et al., 2025) use modality-specific encoders to process images and text separately and project their outputs into a shared embedding space through modality-specific projection heads. Given an image batch $\mathbf{I}$ and a text batch $\mathbf{T}$, the corresponding encoder representations are

$$\mathbf{E}_v = f_v(\mathbf{I}), \qquad \mathbf{E}_t = f_t(\mathbf{T}), \tag{1}$$

where f_v and f_t denote the vision and text encoders, respectively. The encoder outputs are then passed through the corresponding projection heads and normalized:

$$\mathbf{Z}_v = \mathrm{norm}\left(g_v(\mathbf{E}_v)\right), \qquad \mathbf{Z}_t = \mathrm{norm}\left(g_t(\mathbf{E}_t)\right), \tag{2}$$

where g_v and g_t denote the vision and text projection heads, respectively. Given $\mathbf{Z}_v, \mathbf{Z}_t \in \mathbb{R}^{N\times D}$, CLIP constructs an image-text similarity matrix as

$$\mathbf{S} = \frac{1}{\tau}\mathbf{Z}_v\mathbf{Z}_t^\top \in \mathbb{R}^{N\times N}, \tag{3}$$

where the diagonal entries correspond to matched image-text pairs, while the off-diagonal entries correspond to mismatched image-text pairs. Contrastive learning establishes cross-modal alignment in the shared representation space by increasing the similarity of matched image-text pairs relative to mismatched pairs.

Multimodal Dataset Distillation. Multimodal dataset distillation extends conventional dataset distillation to paired image-text data. MTT-VL introduces trajectory matching into the multimodal setting and jointly distills images and text by reproducing model training dynamics. LoRS further improves the efficiency of this paradigm by using a low-rank parameterization to represent image-text relationships. More recent distribution-based methods directly operate on multimodal representations. CovMatch aligns cross-modal covariance statistics between real and synthetic data, whereas MDM matches the geometric structure of joint multimodal representations.

ReCoup also follows the distribution-matching paradigm but differs in treating the cross-modal interaction matrix as a coupling space with internal structure. It further models the reliability of different couplings and their organization across representation stages.

3 METHOD

ReCoup matches cross-modal structure at two representation stages. FACM matches dense encoder-space couplings with weights calibrated against a random-pairing null, while DDCM constrains the diagonal cross-modal couplings in the projected space.

3.1 FLUCTUATION-AWARE CROSS-MODAL COUPLING MATCHING

Let $\mathbf{X} \in \mathbb{R}^{n\times d_v}$ and $\mathbf{Y} \in \mathbb{R}^{n\times d_t}$ denote the visual and textual encoder representations of a mini-batch containing n paired samples. We define the dimension-wise cross-modal coupling matrix as

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}, \qquad C_{ij} = \frac{1}{n}\mathbf{x}_i^\top\mathbf{y}_j, \tag{4}$$

where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_j \in \mathbb{R}^n$ collect the activations of the i-th visual and j-th textual dimensions across the batch.

Coupling estimates from stochastic mini-batches vary in reliability. FACM therefore constructs a single-batch proxy for coupling reliability by comparing each observed cross-modal association with the fluctuations induced by random image–text pairing.

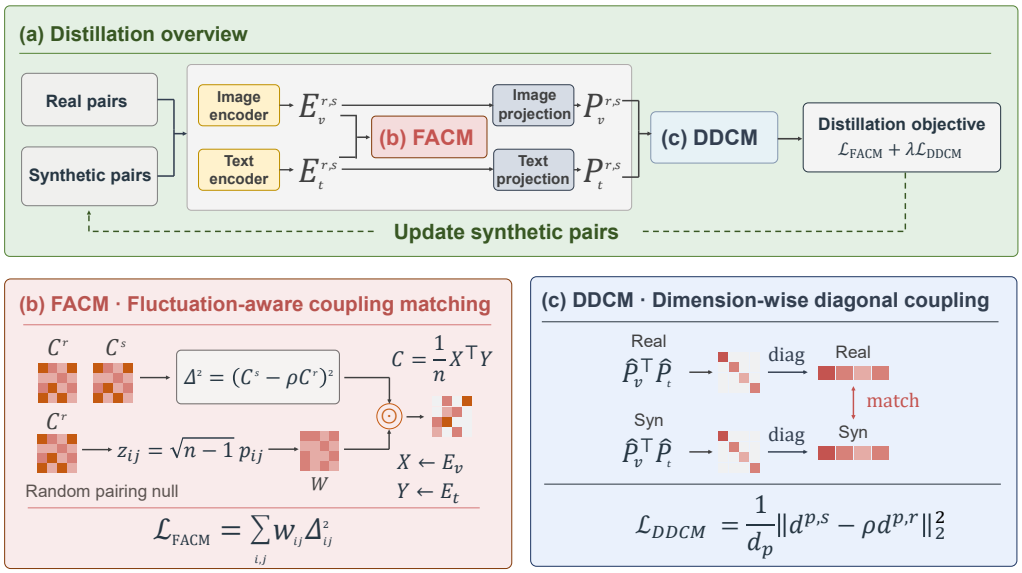


Figure 2: **Overview of ReCoup.** Real and synthetic image–text pairs are processed by the same model. FACM performs reliability-weighted matching of dense encoder-space couplings, and DDCM matches diagonal couplings in the projected space. Their combined objective is used to update the synthetic pairs.

Random-pairing calibration. We first center each activation vector to remove dimension-wise mean activations:

$$\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{x}_i \mathbf{1}, \qquad \widetilde{\mathbf{y}}_j = \mathbf{y}_j - \bar{y}_j \mathbf{1}. \tag{5}$$

For a uniformly sampled permutation π of the textual samples, define

$$R_{ij}^{(\pi)} = \frac{1}{n} \widetilde{\mathbf{x}}_i^\top \mathbf{P}_\pi \widetilde{\mathbf{y}}_j, \tag{6}$$

where $\mathbf{P}_\pi$ is the corresponding permutation matrix. Random re-pairing randomizes image–text correspondence while preserving the marginal activations. Under this null,

$$\mathbb{E}_\pi \left[R_{ij}^{(\pi)} \right] = 0, \qquad \mathrm{Var}_\pi \left[R_{ij}^{(\pi)} \right] = \frac{\|\widetilde{\mathbf{x}}_i\|_2^2 \|\widetilde{\mathbf{y}}_j\|_2^2}{n^2(n-1)}. \tag{7}$$

The corresponding fluctuation scale is

$$\sigma_{ij}^{\mathrm{null}} = \frac{\|\widetilde{\mathbf{x}}_i\|_2 \|\widetilde{\mathbf{y}}_j\|_2}{n\sqrt{n-1}}. \tag{8}$$

For the observed pairing, the centered association is

$$R_{ij} = \frac{1}{n} \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{y}}_j. \tag{9}$$

Standardizing it by the null fluctuation scale gives

$$z_{ij} = \frac{R_{ij}}{\sigma_{ij}^{\mathrm{null}}} = \sqrt{n-1}\, p_{ij}, \tag{10}$$

where p_{ij} is the Pearson correlation between $\mathbf{x}_i$ and $\mathbf{y}_j$. The calibration can therefore be computed in closed form without explicitly sampling permutations.

Reliability-aware coupling matching. We convert the standardized association into the relative reliability score

$$q_{ij} = \frac{|z_{ij}|^\gamma}{\frac{1}{d_v d_t} \sum_{a=1}^{d_v} \sum_{b=1}^{d_t} |z_{ab}|^\gamma}, \tag{11}$$

where γ controls the concentration of the score on stronger pairing-specific associations. For a fixed mini-batch size, the common factor $\sqrt{n-1}$ in equation 10 cancels in equation 11, so q_{ij} can equivalently be computed from $|p_{ij}|^\gamma$.

The coupling weight is

$$w_{ij} = (1-\beta) + \beta q_{ij}, \qquad \beta \in [0,1], \tag{12}$$

where β controls the strength of reliability weighting.

For real and synthetic encoder representations, let $\mathbf{C}^r$ and $\mathbf{C}^s$ denote their cross-modal coupling matrices. FACM minimizes

$$\mathcal{L}_{\text{FACM}} = \frac{1}{d_v d_t} \sum_{i=1}^{d_v} \sum_{j=1}^{d_t} w_{ij} \left(C^s_{ij} - \rho C^r_{ij} \right)^2, \tag{13}$$

where ρ rescales the real coupling target. The weights are computed from the centered associations, while the loss matches the original signed couplings C_{ij}.

3.2 Dimension-Wise Diagonal Coupling Matching

After projection, image and text representations lie in a shared embedding space. During multimodal training, cross-modal coupling becomes increasingly concentrated along the diagonal, as indicated by the rising DOR of paired representations (see Fig. 5). This trend persists under a new random orthogonal rotation at each optimization step, while the DOR remains near its neutral level under random pairing.

Motivated by this emerging diagonal structure, we summarize projection-space cross-modal coupling using its diagonal profile and match this profile between real and synthetic data. This provides a compact $O(d_p)$ statistic that preserves the coordinate-wise organization induced during multimodal training.

Let $\mathbf{P}_v \in \mathbb{R}^{n \times d_p}$ and $\mathbf{P}_t \in \mathbb{R}^{n \times d_p}$ denote the projected visual and textual representations, where n is the batch size and d_p is the projection dimension. We normalize each projected dimension over the batch:

$$\widehat{\mathbf{P}}_{v,:,j} = \frac{\mathbf{P}_{v,:,j}}{\|\mathbf{P}_{v,:,j}\|_2}, \qquad \widehat{\mathbf{P}}_{t,:,j} = \frac{\mathbf{P}_{t,:,j}}{\|\mathbf{P}_{t,:,j}\|_2}, \tag{14}$$

for $j = 1, \ldots, d_p$, where $\mathbf{P}_{v,:,j}$ and $\mathbf{P}_{t,:,j}$ denote the j-th columns of $\mathbf{P}_v$ and $\mathbf{P}_t$, respectively. The projected cross-modal coupling matrix is

$$\mathbf{C}^p = \widehat{\mathbf{P}}_v^\top \widehat{\mathbf{P}}_t. \tag{15}$$

Its diagonal defines the dimension-wise coupling vector:

$$\mathbf{d}^p = \text{diag}(\mathbf{C}^p), \tag{16}$$

where

$$d^p_j = \sum_{k=1}^{n} \widehat{p}_{v,kj}\, \widehat{p}_{t,kj}, \tag{17}$$

and $\widehat{p}_{v,kj}$ and $\widehat{p}_{t,kj}$ are the (k,j)-th entries of $\widehat{\mathbf{P}}_v$ and $\widehat{\mathbf{P}}_t$, respectively.

For real and synthetic batches, let $\mathbf{d}^{p,r}$ and $\mathbf{d}^{p,s}$ denote the corresponding diagonal coupling vectors. DDCM minimizes

$$\mathcal{L}_{\text{DDCM}} = \frac{1}{d_p} \left\| \mathbf{d}^{p,s} - \rho \mathbf{d}^{p,r} \right\|_2^2, \tag{18}$$

where $\rho > 0$ is a scaling coefficient balancing the magnitude of the real and synthetic coupling profiles.

3.3 Overall Distillation Objective

The ReCoup objective combines encoder-level FACM with the projected-space diagonal constraint of DDCM:

$$\mathcal{L}_{\text{ReCoup}} = \mathcal{L}_{\text{FACM}} + \lambda_{\text{DDCM}} \mathcal{L}_{\text{DDCM}}. \tag{19}$$

Table 1: Cross-modal retrieval performance on Flickr30K and MS-COCO under different synthetic pair budgets. Results are averaged over five independent evaluation runs, with standard deviations omitted for clarity. Training on the full dataset yields IR1/5/10 of 41.7/73.7/84.0 and TR1/5/10 of 53.8/80.0/89.4 on Flickr30K, and 22.0/49.7/63.8 and 30.0/57.5/69.4 on MS-COCO, respectively. The best and second-best results for each budget are highlighted in bold and underlined. † denotes a protocol-aligned re-evaluation.

Pairs	Method	Flickr30K						MS-COCO					
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	Random	1.0	4.0	6.5	1.3	5.9	10.1	0.3	1.3	2.7	0.8	3.0	5.0
	Herding	0.7	2.8	5.3	1.1	4.7	7.9	0.5	1.4	3.5	0.8	2.1	4.9
	K-Center	0.7	3.1	6.1	0.6	5.0	7.6	0.4	1.4	2.5	1.4	3.7	5.5
	Forgetting	0.7	2.4	5.6	1.2	4.2	9.7	0.3	1.5	2.5	0.7	2.6	4.8
	MTT-VL	4.7	15.7	24.6	9.9	28.3	39.1	1.3	5.4	9.5	2.5	10.0	15.7
	LoRS	8.3	24.1	35.1	11.8	35.8	49.2	1.8	7.1	12.2	3.3	12.2	19.6
	RepBlend	11.5	32.0	44.5	16.2	41.7	55.5	4.1	13.9	22.3	5.2	17.9	28.0
	PTM-ST	9.6	28.4	41.5	14.4	38.6	52.7	2.3	9.0	15.6	4.1	13.4	22.0
	AMD	10.4	30.5	43.0	14.4	39.1	52.6	2.8	10.5	17.2	4.1	13.8	21.8
	APM	12.8	34.2	47.1	17.8	43.0	57.2	4.7	16.2	25.8	6.2	20.0	31.1
	CovMatch†	20.0	47.7	61.0	27.9	56.8	70.5	6.0	19.3	30.1	9.0	25.1	36.7
	MDM	8.1	24.7	36.2	11.5	32.6	45.0	1.9	7.6	13.2	3.6	13.7	21.6
	Uniform CCM	20.9	48.7	62.1	29.0	57.7	70.1	6.4	20.1	31.4	9.0	25.0	36.3
	ReCoup	**26.6**	**57.3**	**69.8**	**37.1**	**68.2**	**78.7**	**8.3**	**24.6**	**36.6**	**10.7**	**29.4**	**41.3**
200	Random	1.1	4.8	9.2	2.1	8.7	13.2	0.6	2.3	4.4	1.0	4.0	7.2
	Herding	1.5	5.5	9.3	2.3	8.4	14.4	0.9	2.4	4.1	1.0	3.6	7.7
	K-Center	1.5	5.4	9.9	2.2	8.2	13.5	0.7	2.1	5.8	1.2	3.8	7.5
	Forgetting	1.2	3.1	8.4	1.5	8.4	10.2	0.6	2.8	4.9	1.1	3.5	7.0
	MTT-VL	4.6	16.0	25.5	10.2	28.7	41.9	1.7	6.5	12.3	3.3	11.9	19.4
	LoRS	8.6	25.3	36.6	14.5	38.7	53.4	2.4	9.3	15.5	4.3	14.2	22.6
	RepBlend	12.7	34.7	47.6	18.6	46.0	60.0	6.1	19.3	29.8	6.9	21.8	32.3
	PTM-ST	12.5	34.7	48.5	19.3	45.9	58.9	3.9	13.7	22.2	5.7	18.2	27.8
	AMD	13.1	34.9	47.5	16.9	42.3	56.2	3.8	13.4	21.4	4.6	15.5	24.1
	APM	14.6	38.5	52.0	18.9	47.8	62.2	6.1	19.6	30.4	7.7	23.6	35.3
	CovMatch†	21.2	49.0	62.1	28.7	58.8	71.5	7.4	23.0	34.9	10.1	29.2	42.3
	MDM	9.1	26.7	39.1	13.0	33.7	47.4	2.9	11.1	18.4	4.9	16.2	25.3
	Uniform CCM	21.6	50.6	62.9	30.1	59.3	71.5	7.8	24.0	36.2	11.1	30.2	43.1
	ReCoup	**29.7**	**60.1**	**72.0**	**38.5**	**69.8**	**80.3**	**11.0**	**31.5**	**44.9**	**15.1**	**37.9**	**50.7**
500	Random	2.4	10.5	17.4	5.2	18.3	25.7	1.1	5.0	8.7	1.9	7.5	12.5
	Herding	3.0	10.0	17.0	5.1	16.4	24.3	1.7	5.3	9.9	1.9	7.8	13.7
	K-Center	3.5	10.4	17.3	4.9	16.4	23.3	1.1	6.3	10.5	2.5	8.7	14.3
	Forgetting	1.8	9.0	15.9	3.6	12.3	19.3	0.8	5.8	8.2	2.1	8.2	13.0
	MTT-VL	6.6	20.2	30.0	13.3	32.8	46.8	2.5	8.9	15.8	5.0	17.2	26.0
	LoRS	10.0	28.9	41.6	15.5	39.8	53.7	2.8	9.9	16.5	5.3	18.3	27.9
	RepBlend	17.0	42.5	55.9	22.5	53.2	66.7	6.2	19.9	30.6	7.0	22.0	32.9
	PTM-ST	16.0	40.5	54.0	22.2	51.1	64.6	6.6	20.5	30.7	6.9	20.1	30.0
	AMD	15.8	39.1	53.2	19.3	46.4	60.0	4.2	14.2	22.3	5.7	19.3	28.7
	APM	17.5	43.5	56.8	21.6	52.7	66.4	7.1	21.8	33.3	8.0	24.3	37.1
	CovMatch†	24.6	52.7	66.2	33.3	62.8	74.8	8.9	26.6	39.1	12.8	33.7	46.7
	MDM	10.0	29.3	42.0	13.7	37.0	51.5	3.7	13.6	22.2	5.6	18.4	28.2
	Uniform CCM	25.7	53.7	67.1	32.8	63.7	76.0	9.3	27.4	40.0	12.5	33.9	46.8
	ReCoup	**31.3**	**60.9**	**72.6**	**40.7**	**71.0**	**81.5**	**12.0**	**32.7**	**45.4**	**16.1**	**39.3**	**53.0**

Here, λ_{DDCM} controls the contribution of DDCM. FACM and DDCM thus constrain cross-modal coupling at complementary representation stages.

At each distillation step, a single online vision–language model produces both encoder and projected representations for real and synthetic data, allowing FACM and DDCM to operate under a shared representation state. Gradients of $\mathcal{L}_{\mathrm{ReCoup}}$ are used to update only the synthetic image–text pairs, while the online model is trained on real data and periodically reinitialized after a fixed number of optimization steps.

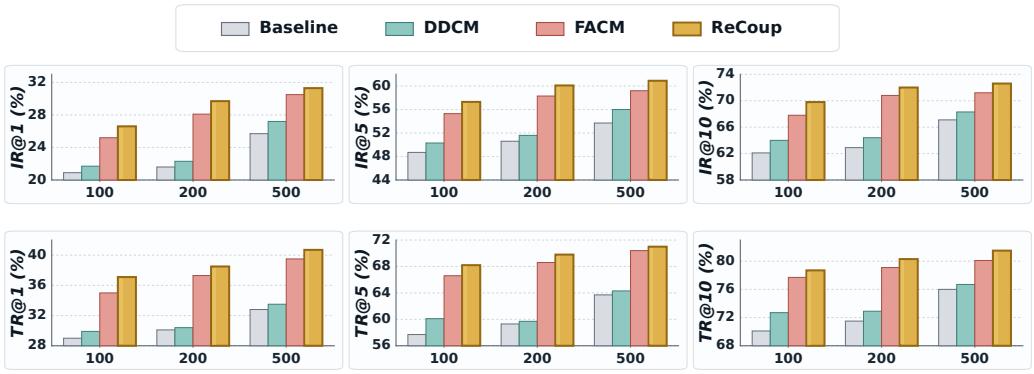


Figure 3: **Ablation of FACM and DDCM.** The baseline uses uniform matching of the cross-modal coupling matrix (CCM). Both FACM and DDCM improve image retrieval (IR) and text retrieval (TR) over the baseline, while combining them yields further gains across all six retrieval metrics.

4 Experiments

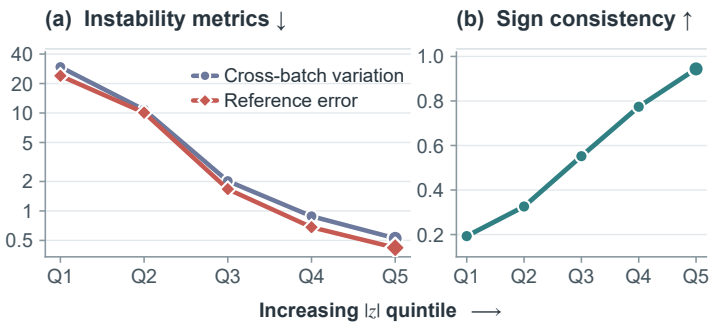
4.1 Experimental Setup

Datasets and evaluation. We evaluate ReCoup on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014) using the standard Karpathy split (Karpathy & Fei-Fei, 2015). Flickr30K contains 29K/1K/1K image–text pairs for training, validation, and testing, while MS-COCO uses a 113K/5K/5K split; all MS-COCO results are reported on the full 5K test set. Retrieval is performed by ranking candidates according to cosine similarity in the shared embedding space. We report Recall@K ($K \in \{1, 5, 10\}$) for image retrieval (IR) and text retrieval (TR), together with mean Recall (mR) averaged over the six metrics. Following prior multimodal dataset distillation work (Wu et al., 2023; Xu et al.), we use an ImageNet-pretrained NFNet (Brock et al., 2021b) as the image encoder and BERT-base (Devlin et al., 2019) as the text encoder, with modality-specific projection heads mapping both modalities into a shared representation space.

Baselines. We compare ReCoup with representative coreset selection and multimodal dataset distillation methods. Coreset baselines include Random Sampling, Herding (Welling, 2009), K-Center (Farahani & Hekmatfar, 2009), and Forgetting (Toneva et al.). For dataset distillation, we include trajectory- or parameter-matching methods, including MTT-VL (Wu et al., 2023), LoRS (Xu et al.), RepBlend (Zhang et al.), PTM-ST (Guo et al., 2026), AMD (Qi et al., 2026), and APM (Bo & Wang, 2026), as well as distribution-based methods CovMatch (Lee & Chung) and MDM (Jeong et al., 2026a). We additionally include Uniform CCM as a controlled baseline that uses the same CCM formulation with uniform allocation. For fair comparison, we align encoder initialization across MDD methods whenever applicable; specifically, we re-evaluate CovMatch with randomized initialization, consistent with the remaining methods, and denote this protocol-aligned result by † in Table 1.

4.2 Main Results

As shown in Table 1, ReCoup consistently outperforms all competing methods on Flickr30K and MS-COCO across synthetic-pair budgets and retrieval metrics. Under the challenging 100-pair budget, ReCoup achieves mR scores of 56.3 and 25.2, outperforming the strongest prior method by 9.0 and 4.1 points, respectively, while exceeding Uniform CCM by 8.2 and 3.8 points. The advantage remains substantial with 500 pairs, with gains of 7.3 and 5.1 points over prior methods and 6.5 and 4.8 points over Uniform CCM. These results establish consistent improvements over both existing methods and the controlled CCM baseline across datasets and compression levels.



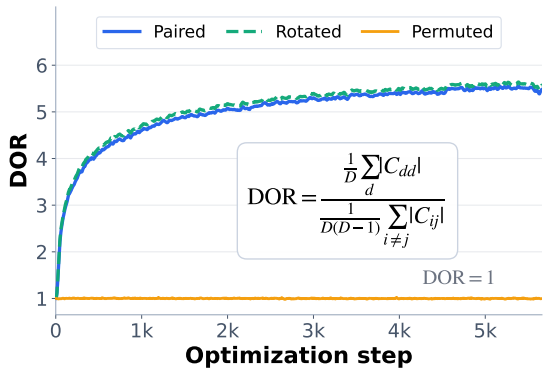
(c) Synthetic-gradient consistency

Metric	Uniform	FACM
$G_{\text{pair}} \uparrow$	0.138	**0.188**
$I_{\text{neg}} \downarrow$	2.53%	**0.23%**

30 real mini-batches

Figure 4: **FACM reliability and gradient consistency.** Larger $|z|$ identifies couplings that are more stable across mini-batches and closer to full-data statistics, while FACM produces more consistent synthetic gradients with substantially fewer conflicting updates.

(a) Diagonal structure emerges during optimization



(b) Projection-space statistic ablation

Statistic	Stats	Cost	mR
None	0	—	54.6
Trace	$O(1)$	$O(nd_p)$	54.8
Random-D	$O(d_p)$	$O(nd_p)$	55.1
Top-D offdiag	$O(d_p)$	$O(nd_p^2)$	55.0
Full matrix	$O(d_p^2)$	$O(nd_p^2)$	55.7
Diagonal(DDCM)	$\mathbf{O(d_p)}$	$\mathbf{O(nd_p)}$	**56.3**

Figure 5: **Diagonal structure in projection space and structural ablation.** The DOR of paired and rotated representations progressively increases during optimization, indicating increasingly pronounced diagonal structure, whereas randomly permuted pairs remain near the neutral level of $\text{DOR} = 1$. Preserving the diagonal profile yields the best performance with only $O(d_p)$ statistics and $O(nd_p)$ computation.

4.3 Ablation and Mechanistic Analysis

Component contributions of FACM and DDCM. Figure 3 evaluates FACM and DDCM across different synthetic-data budgets, using a protocol-matched uniform CCM baseline. Both components improve retrieval performance across the evaluated settings, while their combination yields the strongest results at 100, 200, and 500 synthetic pairs, increasing mean Recall from 48.1/49.3/53.2 to 56.3/58.4/59.7, respectively. The consistent gains from combining FACM and DDCM suggest that the two components provide complementary supervision.

Reliable couplings provide consistent supervision. The fluctuation-calibrated score is strongly aligned with empirical coupling reliability: higher $|z|$ consistently corresponds to lower cross-batch variation and reference error, together with higher sign consistency. This reliability also carries over to optimization, where FACM increases agreement among synthetic gradients while sharply reducing conflicting update directions. Together, these results support the central premise of FACM: calibrating coupling supervision by its susceptibility to batch-level fluctuations produces more stable and coherent synthetic-data updates.

Structure of dimension-wise coupling. The DOR of paired representations increases steadily during multimodal training, and this trend persists under independently sampled random orthogonal rotations of the shared representation basis, while the permutation control remains near its neutral level. This suggests that the observed concentration reflects paired multimodal structure rather than

Table 2: **Cross-architecture generalization on Flickr30K with 500 synthetic pairs.** Synthetic data are distilled with NFNet+BERT and evaluated on unseen ResNet+BERT and RegNet+BERT architectures. ReCoup achieves the strongest transfer performance across all retrieval metrics.

Evaluation	Method	Image Retrieval			Text Retrieval		
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet+BERT	LoRS	3.3	12.7	20.4	6.8	19.6	31.1
	RepBlend	4.2	14.1	23.6	8.4	23.1	35.0
	AMD	4.1	14.2	22.6	7.6	22.3	33.2
	APM	6.9	21.2	31.2	8.7	24.5	35.9
	CovMatch	10.6	29.0	41.0	13.4	34.2	46.2
	Ours	**21.7**	**47.7**	**61.7**	**28.0**	**52.7**	**66.6**
RegNet+BERT	LoRS	3.5	12.6	21.1	6.8	20.8	30.2
	RepBlend	3.9	13.9	24.0	7.9	24.2	36.2
	AMD	4.0	14.4	23.1	7.5	22.8	32.7
	APM	5.4	16.7	25.3	7.9	22.2	32.1
	CovMatch	10.6	29.5	41.0	13.9	33.5	44.8
	Ours	**20.9**	**46.5**	**59.4**	**25.9**	**52.9**	**64.9**

a property specific to a fixed projection basis. The structural ablation further shows that the diagonal profile outperforms trace, random, off-diagonal, and full-matrix alternatives using only $O(d_p)$ statistics and $O(nd_p)$ computation. These results motivate DDCM as a compact structural constraint.

4.4 Cross-Architecture Generalization

Synthetic data distilled with NFNet+BERT are further evaluated on unseen ResNet (Brock et al., 2021a)+BERT and RegNet (Xu et al., 2022)+BERT architectures. Methods tied more closely to training dynamics or model parameters transfer less effectively under architecture shift, whereas distribution-level matching shows substantially stronger generalization. ReCoup achieves mean Recall scores of 46.4 and 45.1 on the two target architectures, compared with 29.1 and 28.9 for CovMatch, indicating that the preserved cross-modal coupling structure is less dependent on the source architecture.

5 Conclusion

We presented ReCoup, a reliable cross-modal coupling matching framework for multimodal dataset distillation. Instead of treating all observed cross-modal statistics as equally informative, ReCoup distinguishes couplings by both their reliability and their dimension-wise organization. FACM calibrates encoder-level couplings against random-pairing fluctuations to reduce unstable supervision, while DDCM preserves compact dimension-wise diagonal structure in the projection space. Experiments on Flickr30K and MS-COCO show consistent improvements across compression ratios, together with stronger coupling reproducibility, more coherent synthetic-data updates, and improved cross-architecture transfer. These results suggest that effective multimodal distillation depends not simply on matching more statistics, but on preserving cross-modal structure that remains reliable under compression.

Limitations Our study focuses on paired image–text data and models cross-modal structure between two modalities. It therefore remains unclear whether the same coupling formulation extends directly to other modality pairs, such as audio–text or video–text, where representation characteristics and alignment mechanisms can differ substantially. More importantly, settings involving three or more modalities introduce interactions that cannot always be reduced to independent pairwise couplings. Extending ReCoup to such settings may require modeling higher-order or multiway coupling structure, together with reliability estimation across multiple modality combinations. Exploring these broader multimodal settings is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability, grammar, and language of the manuscript. We did not use generative AI tools to develop the methodology, formulate mathematical claims or proofs, design or conduct experiments, analyze data, or interpret results. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and results.

REFERENCES

Deyu Bo and Xinchao Wang. Efficient multi-modal dataset distillation via analytic parameter matching. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *arXiv preprint arXiv:2101.08692*, 2021a.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pp. 1059–1071. PMLR, 2021b.

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*, 2022.

Junhyeok Choi, Sangwoo Mo, and Minwoo Chae. Multimodal dataset distillation made simple by prototype-guided data synthesis. *arXiv preprint arXiv:2602.19756*, 2026.

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*, pp. 6565–6590. PMLR, 2023.

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

Zhuohang Dang, Minnan Luo, Chengyou Jia, Hangwei Qian, Xinyu Zhang, Xiaojun Chang, and Ivor Tsang. Correspondence coverage matters for multi-modal dataset distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20693–20701, 2026.

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, 2019.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3758. IEEE, 2023a.

Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. *Advances in Neural Information Processing Systems*, 36:67487–67504, 2023b.

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

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

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 (CVPR)*, 2026a.

Jongoh Jeong, Sun-Kyung Lee, and Kuk-Jin Yoon. Rank-aware hyperbolic alignment for vision-language dataset distillation. *arXiv preprint arXiv:2606.29464*, 2026b.

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*, pp. 4904–4916. PMLR, 2021.

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*, pp. 3128–3137, 2015.

Saksham Singh Kushwaha, Siva Sai Nagender Vasireddy, Kai Wang, and Yapeng Tian. Audio-visual dataset distillation. *Transactions on Machine Learning Research*, 2024.

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

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*.

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*, pp. 740–755. Springer, 2014.

Yue Min, Shaobo Wang, Jiaze Li, Tianle Niu, Junxin Fan, Yongliang Miao, Lijin Yang, and Linfeng Zhang. Imagebinddc: Compressing multi-modal data with imagebind-based condensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15537–15545, 2026.

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

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 Cai Zhao. Asynchronous matching with dynamic sampling for multimodal dataset distillation. In *International Conference on Learning Representations*, 2026.

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*, pp. 8748–8763. PmLR, 2021.

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*, pp. 17097–17107, 2023.

Haoru Tan, Wang Wang, Sitong Wu, Xiuzhe Wu, Yang-Tian Sun, Chirui Chang, Shaofeng Zhang, and Xiaojuan Qi. Dataset distillation by influence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19654–19664, 2026.

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*.

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.

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. *arXiv preprint arXiv:1811.10959*, 2018.

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

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

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.

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*.

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*, pp. 11975–11986, 2023.

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*.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 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*.

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)*, pp. 7856–7865. IEEE, 2023.

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. *arXiv preprint arXiv:2509.15472*, 2025.

A Additional Related Work

Multimodal dataset distillation builds on the broader development of dataset distillation, from bi-level synthetic-data optimization (Wang et al., 2018) and gradient matching (Zhao et al.; Zhao & Bilen, 2021; Lee et al., 2022) to trajectory matching (Cazenavette et al., 2022; Cui et al., 2023; Du et al., 2023a;b) and feature/distribution matching (Wang et al., 2022; Zhao & Bilen, 2023; Zhao et al., 2023; Sajedi et al., 2023). In vision–language learning, most existing dataset-distillation methods operate on paired image–text data under the dual-encoder paradigm popularized by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), typically with contrastive objectives such as InfoNCE (Oord et al., 2018).

Optimization-Dynamics and Parameter-Based Multimodal Distillation. Early vision–language dataset-distillation methods transfer supervision through model optimization dynamics. MTT-VL (Wu et al., 2023) extends trajectory matching to vision–language learning by jointly distilling image–text pairs under contrastive training, while using low-rank adaptation to improve tractability for large multimodal models. LoRS (Xu et al.) further incorporates cross-modal correspondence by jointly distilling a soft image–text similarity matrix with a low-rank parameterization.

Subsequent work improves trajectory-based supervision along complementary directions. RepBlend (Zhang et al.) addresses modality collapse through representation blending and symmetric projection trajectory matching. MDW (Dang et al., 2025) targets noisy web-scale data by learning fine-grained correspondences from partially mismatched image–text pairs. ProCo (Dang et al., 2026) instead improves correspondence coverage through retrieval-distribution-based clustering and coverage-aware regularization. PTM-ST (Guo et al., 2026) models teacher evolution across training stages using stage-aware supervision and shortcut trajectories, whereas AMD (Qi et al., 2026) accounts for asynchronous optimization between modalities and introduces semantics-aware prototype initialization.

Beyond trajectory matching, APM (Bo & Wang, 2026) derives analytic solutions for modality projectors and directly matches real and synthetic parameters, avoiding dense expert checkpoints and iterative unrolling. Inf-Match (Tan et al., 2026) instead matches the influence of real and synthetic data on converged model parameters and extends this formulation to vision–language distillation. These methods therefore transfer supervision through optimization trajectories, parameters, correspondence variables, or their effects on learned parameters, retaining varying degrees of dependence on the underlying model and training procedure.

Representation and Distribution Matching for Multimodal Distillation. A complementary family directly aligns multimodal representation statistics. CovMatch (Lee & Chung) matches cross-covariance matrices between real and synthetic image–text representations together with modality-wise feature statistics, establishing cross-modal second-order structure as a distillation target while allowing both encoders to remain trainable. MDM (Jeong et al., 2026a) performs distribution matching in the joint embedding space and exploits its unit-hypersphere geometry through cross-modal agreement and discrepancy directions. RAHA (Jeong et al., 2026b) models the low-rank structure of cross-modal correlation by separating a dominant shared semantic subspace from a weakly correlated residual subspace and combining rank-aware matching with hyperbolic alignment.

Beyond conventional vision–language dual encoders, ImageBindDC (Min et al., 2026) performs multimodal condensation in the unified ImageBind embedding space using characteristic-function-based objectives to preserve uni-modal, cross-modal, and joint-modal distributions. Audio-Visual Dataset Distillation (AVDD) (Kushwaha et al., 2024) extends distribution matching to paired audio and visual signals by jointly preserving modality-wise information and cross-modal semantic associations. ReCoup belongs to this representation-matching family but instead focuses on the reliability and internal organization of cross-modal statistics estimated under stochastic mini-batching, rather than primarily expanding their order, geometry, or modality coverage.

Generative and Learning-Free Alternatives. An orthogonal direction constructs compact multimodal training data without iterative real–synthetic model matching. EDGE (Zhao et al., 2025) uses generative models to synthesize multimodal training pairs while promoting cross-modal consistency and sample diversity. PDS (Choi et al., 2026) follows a learning-free route by extracting aligned representations with a pretrained vision–language model, constructing cross-modal proto-

Algorithm 1 ReCoup Distillation

Require: real dataset $\mathcal{D}$; synthetic budget m; iterations T; reset interval R; $\rho, \gamma, \beta, \lambda_{\mathrm{DDCM}}$
Ensure: distilled image–text set $\mathcal{S}$

Initialize $\mathcal{S} = \{(\mathbf{I}_k^s, \mathbf{T}_k^s)\}_{k=1}^m$ from sampled training pairs
Initialize online vision–language model θ
$a \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  Sample real batch $\mathcal{B}^r \sim \mathcal{D}$
  Extract $(\mathbf{E}_v^r, \mathbf{E}_t^r, \mathbf{P}_v^r, \mathbf{P}_t^r) \leftarrow f_\theta(\mathcal{B}^r)$
  Compute $\mathbf{C}^r = |\mathcal{B}^r|^{-1}(\mathbf{E}_v^r)^\top \mathbf{E}_t^r$
  Compute FACM weights $\mathbf{W} \leftarrow \mathrm{ReliabilityWeight}(\mathbf{E}_v^r, \mathbf{E}_t^r; \gamma, \beta)$
  Sample synthetic batch $\mathcal{B}^s \subseteq \mathcal{S}$
  Extract $(\mathbf{E}_v^s, \mathbf{E}_t^s, \mathbf{P}_v^s, \mathbf{P}_t^s) \leftarrow f_\theta(\mathcal{B}^s)$
  Compute $\mathbf{C}^s = |\mathcal{B}^s|^{-1}(\mathbf{E}_v^s)^\top \mathbf{E}_t^s$
  $\mathcal{L}_{\mathrm{FACM}} \leftarrow \mathrm{FACM}(\mathbf{C}^s, \mathbf{C}^r, \mathbf{W}; \rho)$
  $\mathcal{L}_{\mathrm{DDCM}} \leftarrow \mathrm{DDCM}(\mathbf{P}_v^s, \mathbf{P}_t^s, \mathbf{P}_v^r, \mathbf{P}_t^r; \rho)$
  $\mathcal{L}_{\mathrm{ReCoup}} \leftarrow \mathcal{L}_{\mathrm{FACM}} + \lambda_{\mathrm{DDCM}} \mathcal{L}_{\mathrm{DDCM}}$
  Update synthetic pairs: $\mathcal{S} \leftarrow \mathcal{S} - \eta_s \nabla_{\mathcal{S}} \mathcal{L}_{\mathrm{ReCoup}}$
  Sample independent real batch $\widetilde{\mathcal{B}}^r \sim \mathcal{D}$
  Update online model: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\mathrm{CLIP}}(\widetilde{\mathcal{B}}^r; \theta)$
  $a \leftarrow a + 1$
  if $a = R$ **then**
    Reinitialize online model θ
    $a \leftarrow 0$
  end if
end for
return $\mathcal{S}$

types through modality-specific clustering and matching, and using them to guide data synthesis. These approaches complement representation-matching methods by reducing reliance on iterative optimization-based distillation.

B Additional Method Details and Derivations

This section provides the optimization procedure and additional derivations underlying ReCoup. We derive the random-pairing calibration used by FACM, clarify its connection to Pearson correlation, characterize the resulting reliability weighting, and summarize the efficient computation of DDCM.

B.1 Optimization Procedure

Algorithm 1 summarizes ReCoup distillation. A single online vision–language model provides the encoder and projection spaces used to construct the real and synthetic statistics. At each iteration, ReCoup first updates the synthetic image–text pairs using FACM and DDCM and subsequently updates the online model using an independent real mini-batch. Periodic model reinitialization exposes the synthetic set to multiple model states rather than a single optimization trajectory.

B.2 Derivation of Random-Pairing Calibration

FACM calibrates an observed cross-modal association against the fluctuation induced by randomly re-pairing samples within the same mini-batch. We derive the corresponding permutation mean and variance below.

For a visual dimension i and textual dimension j, let

$$\mathbf{a} = \widetilde{\mathbf{x}}_i, \qquad \mathbf{b} = \widetilde{\mathbf{y}}_j,$$

where $\mathbf{1}^\top \mathbf{a} = \mathbf{1}^\top \mathbf{b} = 0$. For a uniformly sampled permutation π of the n textual samples,

$$R^{(\pi)} = \frac{1}{n} \sum_{k=1}^{n} a_k b_{\pi(k)}. \tag{20}$$

Permutation mean. Since each entry of $\mathbf{b}$ is equally likely to occupy position k,

$$\mathbb{E}_\pi[b_{\pi(k)}] = \frac{1}{n} \sum_{r=1}^{n} b_r = 0,$$

and therefore

$$\mathbb{E}_\pi[R^{(\pi)}] = 0. \tag{21}$$

Permutation variance. Because the permutation mean is zero, the variance follows from the second moment:

$$\begin{aligned} &\mathbb{E}_\pi \left[\left(\sum_{k=1}^{n} a_k b_{\pi(k)} \right)^2 \right] \\ &= \sum_{k=1}^{n} a_k^2 \mathbb{E}_\pi[b_{\pi(k)}^2] + \sum_{k \neq \ell} a_k a_\ell \mathbb{E}_\pi[b_{\pi(k)} b_{\pi(\ell)}]. \end{aligned} \tag{22}$$

For each k,

$$\mathbb{E}_\pi[b_{\pi(k)}^2] = \frac{\|\mathbf{b}\|_2^2}{n}.$$

For $k \neq \ell$,

$$\begin{aligned} \mathbb{E}_\pi[b_{\pi(k)} b_{\pi(\ell)}] &= \frac{\sum_{r \neq s} b_r b_s}{n(n-1)} \\ &= -\frac{\|\mathbf{b}\|_2^2}{n(n-1)}, \end{aligned} \tag{23}$$

where the second equality follows from $\sum_r b_r = 0$. Similarly,

$$\sum_{k \neq \ell} a_k a_\ell = -\|\mathbf{a}\|_2^2.$$

Substitution gives

$$\mathbb{E}_\pi \left[\left(\sum_{k=1}^{n} a_k b_{\pi(k)} \right)^2 \right] = \frac{\|\mathbf{a}\|_2^2 \|\mathbf{b}\|_2^2}{n-1}. \tag{24}$$

Accounting for the factor $1/n$ in $R^{(\pi)}$ yields

$$\mathrm{Var}_\pi[R^{(\pi)}] = \frac{\|\mathbf{a}\|_2^2 \|\mathbf{b}\|_2^2}{n^2(n-1)}. \tag{25}$$

Substituting $\mathbf{a} = \widetilde{\mathbf{x}}_i$ and $\mathbf{b} = \widetilde{\mathbf{y}}_j$ recovers the random-pairing variance used in the main text.

Connection to Pearson correlation. The Pearson correlation between $\mathbf{x}_i$ and $\mathbf{y}_j$ is

$$p_{ij} = \frac{\widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{y}}_j}{\|\widetilde{\mathbf{x}}_i\|_2 \|\widetilde{\mathbf{y}}_j\|_2}. \tag{26}$$

Since

$$R_{ij} = \frac{1}{n} \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{y}}_j, \qquad \sigma_{ij}^{\mathrm{null}} = \frac{\|\widetilde{\mathbf{x}}_i\|_2 \|\widetilde{\mathbf{y}}_j\|_2}{n\sqrt{n-1}},$$

the standardized association becomes

$$z_{ij} = \frac{R_{ij}}{\sigma_{ij}^{\mathrm{null}}} = \sqrt{n-1}\, p_{ij}. \tag{27}$$

The random-pairing calibration therefore admits a closed-form computation and does not require explicit permutation sampling.

B.3 Properties of Reliability Weighting

FACM converts the standardized association into

$$q_{ij} = \frac{|z_{ij}|^\gamma}{\frac{1}{d_v d_t}\sum_{a=1}^{d_v}\sum_{b=1}^{d_t}|z_{ab}|^\gamma}, \tag{28}$$

and uses

$$w_{ij} = (1-\beta) + \beta q_{ij}. \tag{29}$$

Let $M = d_v d_t$. By construction,

$$\frac{1}{M}\sum_{i,j} q_{ij} = 1, \tag{30}$$

and therefore

$$\frac{1}{M}\sum_{i,j} w_{ij} = 1. \tag{31}$$

FACM thus redistributes emphasis across couplings without changing the average loss weight.

The interpolation parameter satisfies

$$\beta = 0 \Rightarrow w_{ij} = 1, \qquad \beta = 1 \Rightarrow w_{ij} = q_{ij},$$

and for intermediate values

$$w_{ij} \geq 1 - \beta. \tag{32}$$

FACM therefore retains a uniform contribution for every coupling.

Finally, for any common scale $c > 0$,

$$\frac{|cz_{ij}|^\gamma}{\frac{1}{M}\sum_{a,b}|cz_{ab}|^\gamma} = q_{ij}. \tag{33}$$

Hence the relative weighting depends on the organization of associations within the batch rather than their common scale. In particular, for a fixed mini-batch size, the factor $\sqrt{n-1}$ in $z_{ij} = \sqrt{n-1}p_{ij}$ cancels from q_{ij}.

B.4 Efficient DDCM Computation and Complexity

The projected cross-modal coupling matrix is

$$\mathbf{C}^p = \widehat{\mathbf{P}}_v^\top \widehat{\mathbf{P}}_t.$$

DDCM requires only its diagonal and therefore does not need to materialize the full $d_p \times d_p$ matrix.

Let $\widehat{\mathbf{p}}_{v,k}, \widehat{\mathbf{p}}_{t,k} \in \mathbb{R}^{d_p}$ denote the k-th rows of the column-normalized projected representations. The diagonal profile can be computed directly as

$$\mathbf{d}^p = \sum_{k=1}^{n} \widehat{\mathbf{p}}_{v,k} \odot \widehat{\mathbf{p}}_{t,k}, \tag{34}$$

where $\odot$ denotes element-wise multiplication.

FACM computes a dense encoder-space coupling $\mathbf{X}^\top\mathbf{Y}$ and therefore requires $O(nd_v d_t)$ computation and $O(d_v d_t)$ coupling statistics. The centered associations used to construct reliability scores have the same asymptotic matrix-multiplication cost, with only constant-factor additional overhead for centering and norm computation.

DDCM requires only $O(nd_p)$ computation and $O(d_p)$ coupling statistics.

Table 3: **Complexity of ReCoup cross-modal statistics.** Forward and backward passes shared across objectives are excluded.

Component	Computation	Stored statistics
FACM	$O(nd_vd_t)$	$O(d_vd_t)$
DDCM	$O(nd_p)$	$O(d_p)$

Table 4: **Distillation hyperparameters used by ReCoup.** Optimization hyperparameters shared with CovMatch are inherited from its configuration without method-specific retuning. ReCoup-specific coefficients are fixed across datasets and synthetic-set budgets; only the batch sizes vary with the synthetic-set budget as listed.

Hyperparameter	100 pairs	200 pairs	500 pairs
Projection dimension d_p	2304	2304	2304
Distillation iterations	8000	8000	8000
Online-model steps / iteration	1	1	1
Real batch size	128	256	300
Synthetic batch size	100	200	300
Synthetic LR (image / text)	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0
Synthetic momentum	0.5	0.5	0.5
Encoder LR (image / text)	0.01 / 0.01	0.01 / 0.01	0.01 / 0.01
Projection LR (image / text)	0.1 / 0.1	0.1 / 0.1	0.1 / 0.1
Online-model momentum	0.9	0.9	0.9
Online-model weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}
Target scaling ρ	1.0	1.0	1.0
FACM exponent γ	2.0	2.0	2.0
FACM mixing coefficient β	$10/11$	$10/11$	$10/11$
DDCM weight λ_{DDCM}	0.1	0.1	0.1

C Experimental Details

C.1 Datasets, Models, and Initialization

Datasets and evaluation. We evaluate ReCoup on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014) using the standard Karpathy split (Karpathy & Fei-Fei, 2015). Flickr30K contains 29K/1K/1K training, validation, and test samples, while MS-COCO uses a 113K/5K/5K split. All MS-COCO retrieval results are reported on the full 5K-image test set. Candidates are ranked according to cosine similarity in the shared embedding space. We report Recall@K, $K \in \{1, 5, 10\}$, for image retrieval (IR) and text retrieval (TR), and compute mean Recall (mR) over the six retrieval metrics.

Model configuration. Following the encoder and projection-head configuration used by CovMatch (Lee & Chung), we use an ImageNet-pretrained NFNet (Brock et al., 2021b) as the visual encoder and BERT-base (Devlin et al., 2019) as the textual encoder. Modality-specific projection heads map both modalities into a shared $d_p = 2304$ dimensional representation space. The architecture is fixed across datasets and synthetic-set budgets. All experiments are conducted on a single NVIDIA H100 GPU.

Synthetic-pair initialization. For a synthetic budget of m pairs, we sample m real training pairs to initialize the distilled set. Synthetic images are initialized from the corresponding real images. Captions are tokenized with the pretrained text encoder and initialized from their token embeddings; the resulting continuous text representations are optimized jointly with the synthetic images while their attention masks remain fixed.

C.2 Distillation and Reporting Protocol

Distillation and reporting protocol. To minimize method-specific optimization tuning, we retain the optimization hyperparameters of CovMatch for settings shared by the two methods. In particular, the learning rates for the synthetic image/text variables, the image/text encoders, and the projection heads are kept unchanged from CovMatch. We do not perform dataset- or budget-specific retuning of these shared hyperparameters.

ReCoup is optimized for 8,000 iterations for every dataset and synthetic budget. At each iteration, one real mini-batch provides the statistics used to update the synthetic pairs. An independent real mini-batch is then used for one contrastive-learning update of the online model. Synthetic image and text variables are optimized using SGD with learning rate 1.0 and momentum 0.5. The online model is optimized with SGD using momentum 0.9 and weight decay 5×10^{-4}. The encoder learning rates are 0.01 for both modalities, and the projection-head learning rates are 0.1.

The real/synthetic batch sizes are 128/100, 256/200, and 300/300 for 100, 200, and 500 synthetic pairs, respectively. The ReCoup-specific coefficients are fixed to $\rho = 1.0$, $\gamma = 2.0$, $\alpha = 10$ ($\beta = 10/11$), and $\lambda_{\mathrm{DDCM}} = 0.1$ throughout all experiments. The same values are used on Flickr30K and MS-COCO and across all synthetic-set budgets, without dataset- or budget-specific tuning.

For baselines, we retain standard deviations whenever they are available from the corresponding reported results and otherwise report the available mean value.

Our implementation parameterizes the FACM interpolation using $\alpha > 0$:

$$w_{ij} = \frac{1 + \alpha q_{ij}}{1 + \alpha}. \tag{35}$$

This is equivalent to $w_{ij} = (1 - \beta) + \beta q_{ij}$ under

$$\beta = \frac{\alpha}{1 + \alpha}. \tag{36}$$

We use $\alpha = 10$, corresponding to $\beta = 10/11 \approx 0.909$.

D Additional Experiments and Analyses

D.1 Protocol-Aligned Baseline Evaluation

Most compared multimodal dataset distillation methods use randomized encoder initialization when the online model is refreshed. The original CovMatch configuration differs in this respect: its encoder is restored from a fixed initialization, while its projection layers are reinitialized. Since encoder initialization has a substantial effect on absolute retrieval performance, we additionally re-evaluate CovMatch using the randomized encoder-initialization protocol adopted in our experiments, while keeping the remaining CovMatch configuration unchanged.

Table 5 summarizes the effect of this protocol alignment. Randomized encoder initialization substantially strengthens CovMatch across both datasets and all synthetic-set budgets. We therefore use this protocol-aligned variant, denoted CovMatch†, in the main comparison to avoid attributing gains caused by different initialization protocols to the proposed objective. For reference, the original-protocol CovMatch results are reported in Tables 6 and 7, but are not used for the main performance comparison.

The protocol-aligned CovMatch is close to Uniform CCM, indicating that a substantial part of the gap to the originally reported CovMatch result is explained by the encoder-initialization protocol. Importantly, ReCoup retains a clear advantage over both protocol-aligned baselines. At 500 synthetic pairs, ReCoup improves mR over CovMatch† by 7.3 points on Flickr30K and 5.1 points on MS-COCO, and over Uniform CCM by 6.5 and 4.8 points, respectively. This controlled comparison isolates the gains from reliability-aware coupling weighting and projection-space structural matching from those caused by the training protocol.

Table 5: **Effect of protocol alignment on Mean Recall (mR).** CovMatch† changes only the encoder-initialization protocol of CovMatch to match the randomized setting used in our experiments.

	Flickr30K			MS-COCO		
Method	100	200	500	100	200	500
CovMatch (original protocol)	30.5	34.4	38.4	11.5	14.8	19.6
CovMatch†	47.3	48.6	52.4	21.0	24.5	28.0
Uniform CCM	48.1	49.3	53.2	21.4	25.4	28.3
ReCoup	**56.3**	**58.4**	**59.7**	**25.2**	**31.9**	**33.1**

D.2 Complete Retrieval Results with Standard Deviations

Tables 6 and 7 report the complete retrieval results on Flickr30K and MS-COCO. Results are reported as mean ± standard deviation whenever the corresponding uncertainty is available. CovMatch† denotes the protocol-aligned re-evaluation described in Appendix D.1. Uniform CCM follows the same distillation and evaluation protocol as ReCoup while uniformly matching the encoder-space cross-modal coupling matrix.

D.3 Coupling Reliability and Gradient Consistency

FACM computes its reliability score from a single stochastic mini-batch. We examine whether this score predicts empirical reproducibility across independently sampled real mini-batches and whether the resulting weighting produces more coherent synthetic-data updates.

Empirical coupling reliability. Let $\mathbf{C}^{(b)} \in \mathbb{R}^{d_v \times d_t}$, $b = 1, \ldots, B$, denote coupling matrices computed from B independently sampled real mini-batches using a fixed model state, and let $\mathbf{C}^{\mathrm{ref}}$ denote the coupling computed using the full-data reference. For each coupling (i, j), define

$$\bar{C}_{ij} = \frac{1}{B} \sum_{b=1}^{B} C_{ij}^{(b)}. \tag{37}$$

Cross-batch variation is measured by

$$V_{ij} = \frac{\sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(C_{ij}^{(b)} - \bar{C}_{ij} \right)^2}}{|\bar{C}_{ij}| + \epsilon}, \tag{38}$$

where smaller values indicate greater reproducibility under mini-batch resampling.

We additionally measure deviation from the full-data reference:

$$E_{ij}^{\mathrm{ref}} = \frac{1}{B} \sum_{b=1}^{B} \frac{\left| C_{ij}^{(b)} - C_{ij}^{\mathrm{ref}} \right|}{|C_{ij}^{\mathrm{ref}}| + \epsilon}. \tag{39}$$

Finally, sign consistency is defined as

$$S_{ij} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ \mathrm{sign}(C_{ij}^{(b)}) = \mathrm{sign}(C_{ij}^{\mathrm{ref}}) \right\}. \tag{40}$$

We group coupling observations into quintiles according to the magnitude of their FACM score $|z|$. As shown in Fig. 4, larger $|z|$ is associated with lower cross-batch variation and reference error and with higher sign consistency. The score therefore identifies couplings that are empirically more reproducible under stochastic mini-batching.

Table 6: Complete retrieval results on Flickr30K. Results are reported as mean ± standard deviation whenever available. Entries without error bars retain the available mean values. CovMatch† denotes the protocol-aligned re-evaluation.

Pairs	Method	Image Retrieval			Text Retrieval		
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	Random	1.0	4.0	6.5	1.3	5.9	10.1
	Herding	0.7	2.8	5.3	1.1	4.7	7.9
	K-Center	0.7	3.1	6.1	0.6	5.0	7.6
	Forgetting	0.7	2.4	5.6	1.2	4.2	9.7
	MTT-VL	4.7	15.7	24.6	9.9	28.3	39.1
	LoRS	8.3±0.2	24.1±0.2	35.1±0.3	11.8±0.2	35.8±0.6	49.2±0.5
	RepBlend	11.5±0.4	32.0±0.7	44.5±0.6	16.2±0.8	41.7±0.9	55.5±0.4
	PTM-ST	9.6±0.3	28.4±0.5	41.5±0.5	14.4±0.4	38.6±0.8	52.7±0.7
	AMD	10.4±0.3	30.5±0.7	43.0±0.6	14.4±0.5	39.1±0.6	52.6±0.6
	APM	12.8	34.2	47.1	17.8	43.0	57.2
	CovMatch†	20.0±1.4	47.7±2.1	61.0±2.0	27.9±2.0	56.8±3.0	70.5±2.0
	MDM	8.1	24.7	36.2	11.5	32.6	45.0
	Uniform CCM	20.9±0.5	48.7±0.8	62.1±0.7	29.0±0.7	57.7±0.4	70.1±0.5
	ReCoup	26.6±0.4	57.3±0.8	69.8±0.7	37.1±0.5	68.2±0.5	78.7±0.4
200	Random	1.1	4.8	9.2	2.1	8.7	13.2
	Herding	1.5	5.5	9.3	2.3	8.4	14.4
	K-Center	1.5	5.4	9.9	2.2	8.2	13.5
	Forgetting	1.2	3.1	8.4	1.5	8.4	10.2
	MTT-VL	4.6	16.0	25.5	10.2	28.7	41.9
	LoRS	8.6±0.3	25.3±0.2	36.6±0.3	14.5±0.5	38.7±0.5	53.4±0.5
	RepBlend	12.7±0.8	34.7±0.6	47.6±0.5	18.6±0.7	46.0±0.8	60.0±0.6
	PTM-ST	12.5±0.1	34.7±0.2	48.5±0.2	19.3±0.2	45.9±0.4	58.9±0.5
	AMD	13.1±0.3	34.9±0.6	47.5±0.7	16.9±0.4	42.3±0.6	56.2±0.8
	APM	14.6	38.5	52.0	18.9	47.8	62.2
	CovMatch†	21.2±1.0	49.0±1.5	62.1±1.8	28.7±2.0	58.8±1.7	71.5±1.8
	MDM	9.1	26.7	39.1	13.0	33.7	47.4
	Uniform CCM	21.6±0.7	50.6±1.2	62.9±1.1	30.1±0.9	59.3±1.1	71.5±0.8
	ReCoup	29.7±0.8	60.1±1.0	72.0±1.0	38.5±1.4	69.8±1.3	80.3±1.2
500	Random	2.4	10.5	17.4	5.2	18.3	25.7
	Herding	3.0	10.0	17.0	5.1	16.4	24.3
	K-Center	3.5	10.4	17.3	4.9	16.4	23.3
	Forgetting	1.8	9.0	15.9	3.6	12.3	19.3
	MTT-VL	6.6	20.2	30.0	13.3	32.8	46.8
	LoRS	10.0±0.2	28.9±0.7	41.6±0.6	15.5±0.7	39.8±0.4	53.7±0.3
	RepBlend	17.0±0.6	42.5±0.5	55.9±0.6	22.5±0.4	53.2±0.3	66.7±0.3
	PTM-ST	16.0±0.2	40.5±0.3	54.0±0.3	22.2±0.6	51.1±0.3	64.6±0.3
	AMD	15.8±0.4	39.1±0.4	53.2±0.5	19.3±0.5	46.4±0.4	60.0±0.6
	APM	17.5	43.5	56.8	21.6	52.7	66.4
	CovMatch†	24.6±1.2	52.7±1.7	66.2±1.6	33.3±1.5	62.8±1.6	74.8±1.2
	MDM	10.0	29.3	42.0	13.7	37.0	51.5
	Uniform CCM	25.7±0.4	53.7±0.5	67.1±0.4	32.8±0.9	63.7±1.5	76.0±0.4
	ReCoup	31.3±1.0	60.9±0.9	72.6±0.8	40.7±1.1	71.0±1.4	81.5±1.0

Synthetic-gradient consistency. For a fixed synthetic set $\mathcal{S}$, let

$$\mathbf{g}^{(b)} = \nabla_{\mathcal{S}} \mathcal{L}(\mathcal{S}, \mathcal{B}^{(b)}) \tag{41}$$

denote the synthetic-data gradient induced by the b-th real mini-batch. Gradients with respect to all optimized synthetic variables are flattened before comparison.

Table 7: Complete retrieval results on MS-COCO using the full 5K-image test split. Results are reported as mean ± standard deviation whenever available. Entries without error bars retain the available mean values. CovMatch† denotes the protocol-aligned re-evaluation.

Pairs	Method	Image Retrieval			Text Retrieval		
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	Random	0.3	1.3	2.7	0.8	3.0	5.0
	Herding	0.5	1.4	3.5	0.8	2.1	4.9
	K-Center	0.4	1.4	2.5	1.4	3.7	5.5
	Forgetting	0.3	1.5	2.5	0.7	2.6	4.8
	MTT-VL	1.3	5.4	9.5	2.5	10.0	15.7
	LoRS	1.8±0.1	7.1±0.2	12.2±0.2	3.3±0.2	12.2±0.3	19.6±0.3
	RepBlend	4.1±0.3	13.9±0.8	22.3±0.5	5.2±0.5	17.9±0.9	28.0±0.3
	PTM-ST	2.3±0.1	9.0±0.1	15.6±0.3	4.1±0.2	13.4±0.2	22.0±0.2
	AMD	2.8±0.2	10.5±0.2	17.2±0.4	4.1±0.3	13.8±0.3	21.8±0.4
	APM	4.7	16.2	25.8	6.2	20.0	31.1
	CovMatch†	6.0±0.4	19.3±1.1	30.1±1.6	9.0±0.8	25.1±1.7	36.7±2.2
	MDM	1.9	7.6	13.2	3.6	13.7	21.6
	Uniform CCM	6.4±0.2	20.1±0.5	31.4±0.6	9.0±0.1	25.0±0.7	36.3±0.5
	ReCoup	8.3±0.1	24.6±0.5	36.6±0.6	10.7±0.3	29.4±0.8	41.3±0.5
200	Random	0.6	2.3	4.4	1.0	4.0	7.2
	Herding	0.9	2.4	4.1	1.0	3.6	7.7
	K-Center	0.7	2.1	5.8	1.2	3.8	7.5
	Forgetting	0.6	2.8	4.9	1.1	3.5	7.0
	MTT-VL	1.7	6.5	12.3	3.3	11.9	19.4
	LoRS	2.4±0.1	9.3±0.2	15.5±0.2	4.3±0.1	14.2±0.3	22.6±0.2
	RepBlend	6.1±0.8	19.3±0.7	29.8±0.5	6.9±0.6	21.8±0.9	32.3±0.7
	PTM-ST	3.9±0.1	13.7±0.2	22.2±0.2	5.7±0.2	18.2±0.3	27.8±0.3
	AMD	3.8±0.2	13.4±0.3	21.4±0.4	4.6±0.2	15.5±0.6	24.1±0.5
	APM	6.1	19.6	30.4	7.7	23.6	35.3
	CovMatch†	7.4±0.7	23.0±1.9	34.9±2.4	10.1±1.2	29.2±2.4	42.3±3.1
	MDM	2.9	11.1	18.4	4.9	16.2	25.3
	Uniform CCM	7.8±0.4	24.0±0.7	36.2±0.9	11.1±0.3	30.2±1.1	43.1±1.1
	ReCoup	11.0±0.3	31.5±0.6	44.9±0.7	15.1±0.4	37.9±1.2	50.7±1.0
500	Random	1.1	5.0	8.7	1.9	7.5	12.5
	Herding	1.7	5.3	9.9	1.9	7.8	13.7
	K-Center	1.1	6.3	10.5	2.5	8.7	14.3
	Forgetting	0.8	5.8	8.2	2.1	8.2	13.0
	MTT-VL	2.5	8.9	15.8	5.0	17.2	26.0
	LoRS	2.8±0.2	9.9±0.5	16.5±0.7	5.3±0.5	18.3±1.5	27.9±1.4
	RepBlend	6.2±0.1	19.9±0.3	30.6±0.1	7.0±0.2	22.0±0.3	32.9±0.6
	PTM-ST	6.6±0.1	20.5±0.2	30.7±0.2	6.9±0.3	20.1±0.2	30.0±0.3
	AMD	4.2±0.2	14.2±0.5	22.3±0.4	5.7±0.6	19.3±1.2	28.7±1.0
	APM	7.1	21.8	33.3	8.0	24.3	37.1
	CovMatch†	8.9±0.6	26.6±1.6	39.1±1.8	12.8±1.3	33.7±2.2	46.7±1.8
	MDM	3.7	13.6	22.2	5.6	18.4	28.2
	Uniform CCM	9.3±0.2	27.4±0.4	40.0±0.6	12.5±0.6	33.9±0.6	46.8±0.3
	ReCoup	12.0±0.4	32.7±0.7	45.4±0.8	16.1±0.4	39.3±0.8	53.0±0.9

For mini-batches b and b', define their gradient cosine similarity as

$$A_{bb'} = \frac{\langle \mathbf{g}^{(b)}, \mathbf{g}^{(b')} \rangle}{\|\mathbf{g}^{(b)}\|_2 \|\mathbf{g}^{(b')}\|_2}. \tag{42}$$

The average pairwise agreement is

$$G_{\text{pair}} = \frac{2}{B(B-1)} \sum_{1 \le b < b' \le B} A_{bb'}, \tag{43}$$

and the fraction of conflicting gradient pairs is

$$I_{\text{neg}} = \frac{2}{B(B-1)} \sum_{1 \le b < b' \le B} \mathbf{1}\{A_{bb'} < 0\}. \tag{44}$$

Using $B = 30$ independently sampled real mini-batches, uniform matching yields $G_{\text{pair}} = 0.138$ and $I_{\text{neg}} = 2.53\%$. FACM increases pairwise agreement to 0.188 while reducing the conflict rate to 0.23%. Reliability-aware matching therefore produces substantially more coherent optimization signals for the same synthetic set.

D.4 Projection-Space Structure and Statistic Ablations

We analyze the projection-space structure motivating DDCM using the projected coupling matrix

$$\mathbf{C}^p = \widehat{\mathbf{P}}_v^\top \widehat{\mathbf{P}}_t \in \mathbb{R}^{d_p \times d_p}. \tag{45}$$

Diagonal-to-off-diagonal ratio. We quantify diagonal concentration using

$$\text{DOR}(\mathbf{C}^p) = \frac{\frac{1}{d_p}\sum_{j=1}^{d_p} |C^p_{jj}|}{\frac{1}{d_p(d_p-1)}\sum_{i \ne j} |C^p_{ij}|}. \tag{46}$$

DOR $= 1$ indicates equal average magnitude on and off the diagonal, while larger values indicate stronger diagonal concentration.

We consider two controls to distinguish the effects of image–text correspondence and representation basis. First, for the pairing control, we randomly permute the textual samples:

$$\mathbf{C}^{p,(\pi)} = \widehat{\mathbf{P}}_v^\top \mathbf{P}_\pi \widehat{\mathbf{P}}_t, \tag{47}$$

where $\mathbf{P}_\pi$ is the permutation matrix associated with a random permutation π. This operation preserves the marginal projected representations while removing the original image–text correspondence.

Second, for the rotation control, at each optimization step t we independently sample a random orthogonal matrix $\mathbf{Q}_t \in \mathbb{R}^{d_p \times d_p}$ satisfying $\mathbf{Q}_t \mathbf{Q}_t^\top = \mathbf{I}$, and apply the same $\mathbf{Q}_t$ to both modalities. The rotated coupling is

$$\begin{aligned} \mathbf{C}_t^{p,(\text{rot})} &= (\widehat{\mathbf{P}}_{v,t}\mathbf{Q}_t)^\top (\widehat{\mathbf{P}}_{t,t}\mathbf{Q}_t) \\ &= \mathbf{Q}_t^\top \mathbf{C}_t^p \mathbf{Q}_t. \end{aligned} \tag{48}$$

This transformation preserves the image–text correspondence and applies a common change of basis to the two modalities, while the basis itself is resampled independently across optimization steps.

As shown in Fig. 5, the DOR of paired representations increases during multimodal optimization, and the same trend persists under random orthogonal rotations, whereas the permutation control remains close to the neutral level of DOR $= 1$. This indicates that the observed concentration is associated with paired multimodal structure rather than being specific to a fixed projection basis.

DOR is used only as an analysis diagnostic and is not part of the ReCoup training objective. DDCM requires only the diagonal profile and therefore does not construct the off-diagonal entries during distillation.

Alternative projection-space statistics. We compare DDCM with several alternative summaries while keeping the remaining distillation procedure fixed.

The *trace* retains only

$$s_{\text{trace}} = \text{tr}(\mathbf{C}^p), \tag{49}$$

Table 8: **Zero-shot image classification on CIFAR-100.** Models are trained using either 200 real image–text pairs or 200 synthetic pairs and evaluated without task-specific fine-tuning. The full-data result is included as a reference.

Pairs	Training Data	Top-1 (%)	Top-5 (%)
200	Real Subset	9.8	24.7
200	APM	14.5	34.8
200	CovMatch	14.1	33.6
200	**ReCoup**	**22.4**	**45.4**
Full Data	Real Data	25.7	53.3

compressing the entire diagonal profile into a single scalar.

For *random dimension pairing*, we sample a fixed permutation π over projection dimensions and retain

$$\mathbf{d}_{\text{rand}} = [C^p_{1,\pi(1)}, \ldots, C^p_{d_p,\pi(d_p)}], \tag{50}$$

which preserves $O(d_p)$ statistics but discards the learned same-index organization.

For *top off-diagonal matching*, we retain the d_p largest-magnitude off-diagonal entries,

$$\mathcal{I}_{\text{top}} = \text{TopK}_{(i,j):i\neq j}\left(|C^p_{ij}|, d_p\right), \tag{51}$$

which requires constructing the full coupling matrix and therefore costs $O(nd_p^2)$.

The *full-matrix* alternative minimizes

$$\mathcal{L}_{\text{full}} = \frac{1}{d_p^2}\left\|\mathbf{C}^{p,s} - \rho\mathbf{C}^{p,r}\right\|_F^2, \tag{52}$$

retaining $O(d_p^2)$ statistics with $O(nd_p^2)$ computation.

DDCM instead retains

$$\mathbf{d}^p = \text{diag}(\mathbf{C}^p), \tag{53}$$

requiring only $O(d_p)$ statistics and $O(nd_p)$ computation. The structural ablation in Fig. 5 shows that the diagonal profile outperforms the trace, random dimension pairing, top off-diagonal interactions, and full-matrix matching. Together with the DOR dynamics, these results support DDCM as a compact structural constraint on the coordinate organization that emerges during multimodal training, rather than merely a low-cost approximation to full-matrix matching.

D.5 Downstream Zero-Shot Transfer

We further examine whether the representation learned from the distilled image–text pairs transfers beyond retrieval. We evaluate zero-shot image classification on CIFAR-100, where the vision–language model is trained using the corresponding real or distilled image–text set but receives no CIFAR-100 training examples or task-specific fine-tuning.

For each class c, we construct a textual class prompt $t(c)$ and obtain its normalized text representation $\mathbf{z}_t(c)$. Given a test image $\mathbf{x}$ with normalized visual representation $\mathbf{z}_v(\mathbf{x})$, prediction is performed as

$$\hat{y}(\mathbf{x}) = \arg\max_c \ \mathbf{z}_v(\mathbf{x})^\top \mathbf{z}_t(c). \tag{54}$$

The same prompt construction is used for every compared method. No CIFAR-100 labels are used for model adaptation.

ReCoup achieves 22.4% Top-1 and 45.4% Top-5 accuracy using only 200 distilled image–text pairs, substantially exceeding APM (14.5/34.8), CovMatch (14.1/33.6), and a size-matched subset of 200 real pairs (9.8/24.7). The remaining gap to full-data training is 3.3 points in Top-1 accuracy and 7.9 points in Top-5 accuracy. These results provide evidence that the cross-modal structure preserved by ReCoup transfers beyond image–text retrieval to downstream recognition without task-specific adaptation.

E VISUALIZATION OF DISTILLED DATA



Figure 6: Synthetic image–text pairs on Flickr30K before and after distillation. Green (left): initial pairs before distillation. Orange (right): pairs after ReCoup distillation.

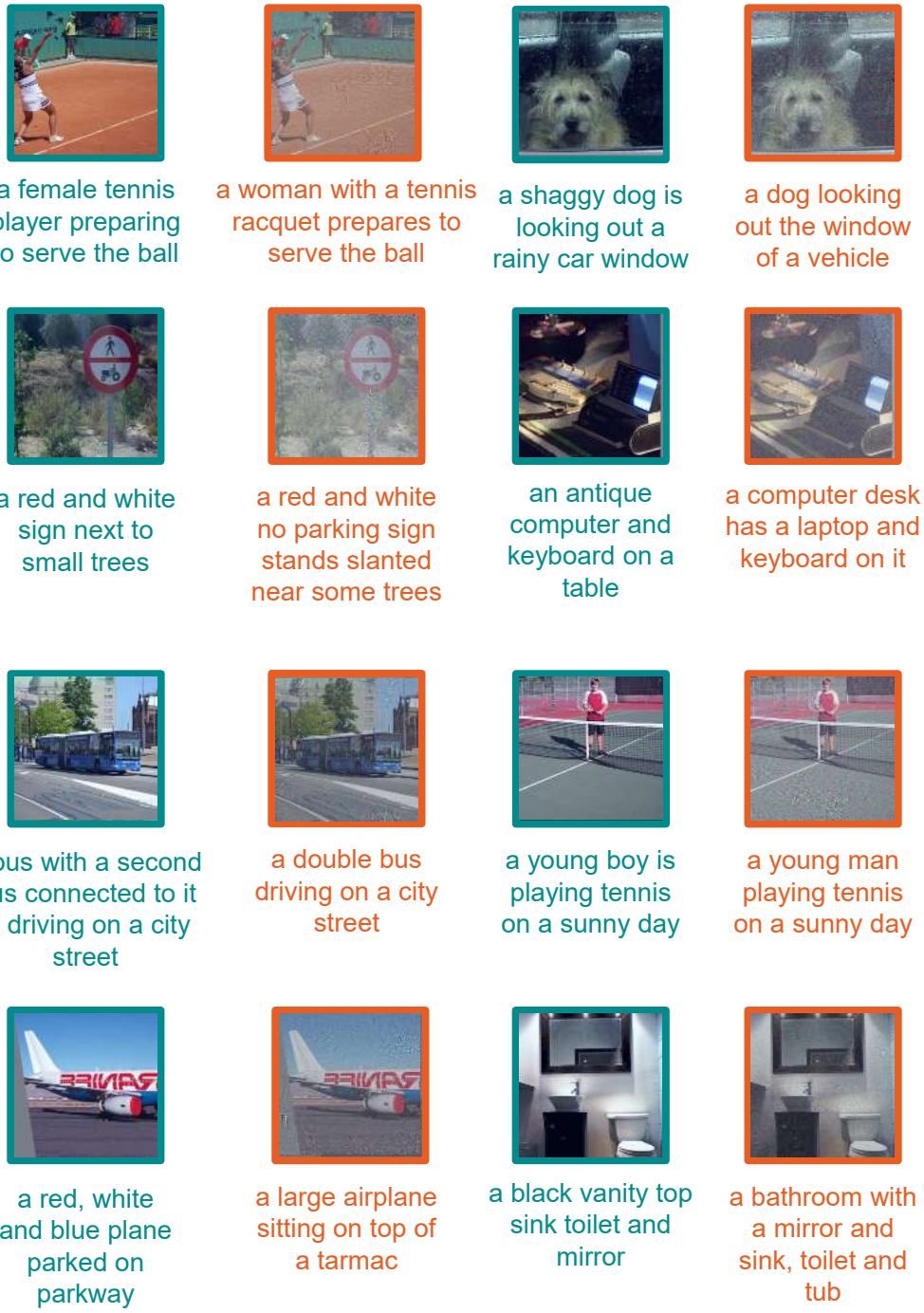


Figure 7: Synthetic image–text pairs on MS-COCO before and after distillation. Green (left): initial pairs before distillation. Orange (right): pairs after ReCoup distillation.