Multimodal Distribution Matching for Vision-Language Dataset Distillation

Supplementary Material

In this supplementary material, we elaborate on the details of our method and experiments and provide additional results with further analyses. We remark that throughout the manuscript, we intend to work with features denoted as z^v for the image and z^t for the text, the angle between the two displacement vectors as $\angle\phi(\cdot,\cdot)$, and the distance (discrepancy) function as $\phi(\cdot,\cdot)$. The list of contents is as follows:

S1. MDM Formulation

Generalizing Unimodal DM to Multimodal. In our approach, we consider a multimodal dataset $\mathcal{D}_{\text{real}} = \{(x_i, t_i)\}_{i=1}^{B}$ of image-text pairs and a much smaller synthetic dataset $\mathcal{D}_{\text{syn}} = \{(\tilde{x}_j, \tilde{t}_j)\}_{j=1}^{\tilde{B}}$ with $|\mathcal{D}_{\text{syn}}| \ll |\mathcal{D}_{\text{real}}|$. A unified image-text model $\Psi(\cdot,\cdot)$ maps an image-text pair (x,t) into a joint feature space and is composed of a pretrained image encoder θ^v, and a pretrained text encoder θ^t with a projection layer.

Our goal is to acquire an optimal set of distilled set $\mathcal{D}^{\star}_{\text{syn}}$ via the multimodal distribution matching (MDM) objective (Eq. 2):

$$\mathcal{D}^{\star}_{\text{syn}} = \arg\min_{\mathcal{D}_{\text{syn}}} \phi\Big(\underbrace{\mathbb{E}_{(X,T)\sim\mathcal{D}_{\text{real}}}\big[\Psi(X,T)\big]}_{\text{real joint}}, \underbrace{\mathbb{E}_{(\tilde{X},\tilde{T})\sim\mathcal{D}_{\text{syn}}}\big[\Psi(\tilde{X},\tilde{T})\big]}_{\text{synthetic joint}}\Big),$$

where $\phi : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ is a nonnegative discrepancy function measuring the distance between two mean joint feature vectors. Below, we justify this objective as a natural multimodal extension of the standard unimodal distribution matching formulation.

We first interpret the real and synthetic datasets as empirical distributions to ease the understanding of their distributions. The real data $\mathcal{D}_{\text{real}}$ induces the empirical distribution:

$$\widehat{P}^{\text{real}}_{XT} = \frac{1}{B}\sum_{i=1}^{B}\delta_{(x_i,t_i)}, \tag{S1}$$

where $\delta_{(x_i,t_i)}$ is the Dirac measure at the pair (x_i,t_i). Similarly, the synthetic dataset $\mathcal{D}_{\text{syn}}$ induces

$$\widehat{Q}^{\text{syn}}_{XT} = \frac{1}{\tilde{B}}\sum_{j=1}^{\tilde{B}}\delta_{(\tilde{x}_j,\tilde{t}_j)}. \tag{S2}$$

For any fixed joint feature extractor Ψ, the expectations of Ψ under these empirical distributions are simply empirical means:

$$\mathbb{E}_{(X,T)\sim\widehat{P}^{\text{real}}_{XT}}\big[\Psi(X,T)\big] = \frac{1}{B}\sum_{i=1}^{B}\Psi(x_i,t_i), \tag{S3}$$

$$\mathbb{E}_{(\tilde{X},\tilde{T})\sim\widehat{Q}^{\text{syn}}_{XT}}\big[\Psi(\tilde{X},\tilde{T})\big] = \frac{1}{\tilde{B}}\sum_{j=1}^{\tilde{B}}\Psi(\tilde{x}_j,\tilde{t}_j). \tag{S4}$$

Thus, the expectations appearing in Eq.2 can be interpreted as dataset-wise averages of joint features produced by Ψ over the real and synthetic datasets, respectively. We can then define the real and synthetic mean joint features as

$$\mu_{\text{real}} := \mathbb{E}_{(X,T)\sim P^{\text{real}}_{XT}}\big[\Psi(X,T)\big], \tag{S5}$$

$$\mu_{\text{syn}}(\mathcal{D}_{\text{syn}}) := \mathbb{E}_{(\tilde{X},\tilde{T})\sim Q^{\text{syn}}_{XT}}\big[\Psi(\tilde{X},\tilde{T})\big], \tag{S6}$$

where we denote the true data distribution over image-text pairs (X,T) as P^{real}_{XT}, and the distribution represented by $\mathcal{D}_{\text{syn}}$ as Q^{syn}_{XT}. The goal of multimodal DM is then to find a synthetic dataset $\mathcal{D}_{syn}$ whose induced distribution Q^{syn}_{XT} yields a mean feature vector $\mu_{\text{syn}}(\mathcal{D}_{\text{syn}})$ that is as close to the real mean μ_{real} under some discrepancy function $\phi(\cdot,\cdot)$:

$$\mathcal{D}^{\star}_{\text{syn}} \in \arg\min_{\mathcal{D}_{\text{syn}}} \phi\big(\mu_{\text{real}}, \mu_{\text{syn}}(\mathcal{D}_{\text{syn}})\big). \tag{S7}$$

Replacing the population expectations by their empirical counterparts $\widehat{P}^{\text{real}}_{XT}$ and $\widehat{Q}^{\text{syn}}_{XT}$, we obtain

$$\mathbb{E}_{(X,T)\sim\mathcal{D}_{\text{real}}}[\Psi(X,T)] \approx \mu_{\text{real}}, \tag{S8}$$

$$\mathbb{E}_{(\tilde{X},\tilde{T})\sim\mathcal{D}_{\text{syn}}}[\Psi(\tilde{X},\tilde{T})] \approx \mu_{\text{syn}}(\mathcal{D}_{\text{syn}}). \tag{S9}$$

For fixed Ψ and $\mathcal{D}_{\text{real}}$, the real mean μ_{real} is constant and serves as the target, while the synthetic mean $\mu_{\text{syn}}(\mathcal{D}_{\text{syn}})$ depends on the content of $\mathcal{D}_{\text{syn}}$ and is optimized by updating the synthetic pairs. We remark that this is equivalent to the unimodal DM formulation if we replace the joint image-text feature $\Psi(x,t)$ with a unimodal encoder $\theta_0(x)$ as:

$$\mathcal{D}^{\star}_{\text{syn}} = \arg\min_{\mathcal{D}_{\text{syn}}} \phi\Big(\mathbb{E}_{X\sim\mathcal{D}_{\text{real}}}\big[\theta_0(X)\big], \mathbb{E}_{\tilde{X}\sim\mathcal{D}_{\text{syn}}}\big[\theta_0(\tilde{X})\big]\Big), \tag{S10}$$

for some distance function $\phi(\cdot,\cdot)$ for unimodal distributions.

Under these choices, the random variable (X, T) effectively reduces to X, and the mean joint features become mean image features. Hence, the proposed multimodal DM objective strictly generalizes the standard image-only DM objective by extending the feature space from $\theta_0(x)$ to the joint representation $\Psi(x, t)$ and by allowing a general discrepancy function.

S2. Distinctions of Our Method

Our MDM method first seeds synthetic image–text pairs by running K-means clustering on the concatenated joint features $[z^v; z^t]$. It then optimizes these synthetic pairs so that the spherical distributions of an agreement vector u and a discrepancy vector g match those of the real data. The agreement u captures shared image–text content and is obtained from a normalized combination of z^v and z^t on the unit hypersphere. The discrepancy g encodes the modality gap between image and text and is obtained from a normalized difference of z^v and z^t on the same hypersphere. Matching the real and synthetic distributions in both u and g encourages the synthetic set to capture *architecture-agnostic joint semantics* rather than reproducing individual training trajectories. Here, interestingly, we observe from the ablation study in Table 5 that matching the distribution of g improves retrieval performance by a larger gap than that from matching u. This indicates that MDM primarily encourages the learning of how captions *deviate* from images through the global structure of the gap distribution, rather than focusing on *shared* semantics.

S2.1. Synthetic Data Initialization

Coreset-based initializations such as K-center and herding select real image–text pairs that approximately cover the encoder feature space under max–min radius or greedy moment-matching criteria, and then reuse these as seeds for optimization under the same MDM objective. However, these heuristics neglect the structure of the image-text agreement and discrepancy, and encourage only marginal coverage, rather than explicitly targeting the joint semantic modes that are crucial for effective cross-modal retrieval. In contrast, our initialization performs K-means clustering directly in the *joint* feature space and assigns the sample nearest the cluster centroid to each synthetic pair, anchoring the synthetic parameters near representative joint centroids. This joint-feature-aware seeding reduces the burden on MDM to relocate poorly placed seeds and instead focuses optimization on fine-grained refinement around already well-positioned prototypes, yielding synthetic datasets that more faithfully approximate the real joint distribution. Empirically, this leads to consistently higher image-text retrieval performance than coreset-seeded variants, shown in Table S1.

Table S1. Performance comparison on Ours seeded with different synthetic data initialization strategies.

	K-center	Herding	Joint-feature K-Means Clustering (Ours)
IR	19.5	19.2	**19.7**
TR	22.1	23.9	**24.2**
Mean	20.8	21.5	**21.9**

S2.2. Model Initialization

Our method employs a dynamic multimodal weight-initialization scheme that merges a pretrained image encoder–text projector pair with N finetuned counterparts randomly sampled from a pool of experts, each updated at every training iteration as with vision-only DM methods [22–24]. Inspired by the angular, layer-wise interpolation strategy of Model Stock [8] (originally proposed for robust *unimodal* classification), we compute layer-wise interpolation coefficients from the angles between the displacement vectors of each finetuned expert anchored on the model composed of a pretrained image encoder and a randomly initialized text projector. Using these coefficients, we then merge them in the weight space, with the goal of shifting the pretrained anchor towards real data distribution, implicitly guided by these experts. To this, we add an additional global weighting factor, α, to further modulate these layer-wise ratios, effectively controlling the trade-off between expert-specific bias and generic structure in a fine-grained manner. Unlike the static, one-shot merge in Model Stock that is used as the final predictor, the merged multimodal encoder in our framework serves as a stochastic, expert-pool–aware initialization that is refreshed at every distillation step and then optimized under the MDM objective. This construction extends [8] from unimodal to multimodal (image encoder and text projector) weight fusion and allows synthetic data to be optimized in a more diverse region of the joint image–text weight space, effectively stabilizing and improving the optimization process.

S2.3. Geodesic Kernel Formulation

Geodesic distance on the sphere measures the shortest path along the sphere with respect to the manifold's curvature, hence yielding the *intrinsic* distance. Our intuition to employ the geodesic distance stems from [13], which highlighted its advantage for understanding the complex geometric structure of multimodal data. Since typical InfoNCE loss and retrieval tasks are primarily driven by angular similarity (*e.g.*, cosine), our *geodesic* perspective using the geodesic kernel energy as formulated in Eq. 9 aligns well with the *angular* distributions of multimodal data.

Kernel variants. For two d-dim feature vectors $a, b \in \mathbb{R}^d$ and a bandwidth $\sigma > 0$, we consider the following radial basis function (RBF) kernels with different angular distance functions that induce the same topology on $\mathbb{S}^{d-1}$, where we demonstrate that the geodesic kernel performs the best

Table S2. Performance by different types of RBF Kernel.

$k(\cdot,\cdot)$	Laplacian	Chordal	Geodesic (Ours)
IR	18.3	19.5	**19.7**
TR	22.4	23.5	**24.2**
Mean	20.3	21.5	**21.9**

in Table S2. Although the image retrieval score improvements appear incremental, our geodesic kernel dramatically increases text retrieval scores.

- **Laplacian RBF kernel.** We also consider a Laplacian RBF kernel based on the L1-distance in the ambient space:

$$k_{\text{laplacian}}(a,b) = \exp\left(-\frac{\|a-b\|_1}{\sigma}\right). \tag{S11}$$

- **Chordal RBF kernel on the unit hypersphere.** Similar to the well-established Euclidean distance, the chordal distance measures the straight-line chord distance between the two unit-normalized points, but restricted to points on a sphere. To force the features onto the sphere, we first L2-normalize the features to lie on the unit hypersphere, $\hat{a} = a/\|a\|_2$, $\hat{b} = b/\|b\|_2$, and define the *chordal distance* between $\hat{a}$ and $\hat{b}$ as

$$d_{\text{chord}}(\hat{a},\hat{b}) = \|\hat{a}-\hat{b}\|_2 = \sqrt{2-2\langle\hat{a},\hat{b}\rangle}, \tag{S12}$$

where $\langle\cdot,\cdot\rangle$ denotes inner product. The corresponding chordal RBF kernel is then:

$$k_{\text{chord}}(a,b) = \exp\left(-\frac{d_{\text{chord}}(\hat{a},\hat{b})^2}{2\sigma^2}\right) = \exp\left(-\frac{2-2\langle\hat{a},\hat{b}\rangle}{2\sigma^2}\right). \tag{S13}$$

Intuitively, chordal distance measures the Euclidean length of the straight-line chord in the ambient space connecting two points on the sphere.

- **Geodesic RBF kernel on the unit hypersphere.** The intrinsic (geodesic) distance on the unit hypersphere is given by the angular distance along the arc on the surface of the sphere as:

$$d_{\text{geo}}(\hat{a},\hat{b}) = \arccos\big(\langle\hat{a},\hat{b}\rangle\big) \in [0,\pi]. \tag{S14}$$

Using this intrinsic distance, we define the geodesic Gaussian kernel

$$k_{\text{geo}}(a,b) = \exp\left(-\frac{d_{\text{geo}}(\hat{a},\hat{b})^2}{2\sigma^2}\right). \tag{S15}$$

Although this kernel is not guaranteed to be positive definite on $\mathbb{S}^{d-1}$ in general, we use the resulting quantity as a *geodesic kernel energy* to encourage the alignment of real and synthetic feature distributions in the unit hypersphere.

S2.4. Cross-Modal Agreement and Discrepancy

To better capture the structure of multimodal representations, we construct the joint image-text features as a *cross-modal agreement* vector u and a *cross-modal discrepancy* vector g. The agreement component encodes **modality-shared semantics** (objects, actions, coarse scene layout), whereas the discrepancy component models **modality-specific** information, *i.e.*, the inherent "modality gap". This gap is non-negligible in image-text data: a single image can be paired with multiple, diverse captions (as in Flickr8k, Flickr30k, MS-COCO datasets), and the same visual concept can be expressed in many textual forms, leading to a structured one-to-many relationship between the two modalities. As reported in Fig. S1, reducing geodesic kernel energy on each modality independently ("*singles*") or on naïvely concatenated joint features ("*concat*") fails to *fully* account for this structure. In contrast, our joint optimization over u and g ("*joint*") explicitly matches both modality-shared semantics and modality-specific gaps, yielding better preservation of joint image-text relationships and consistently higher retrieval performance.

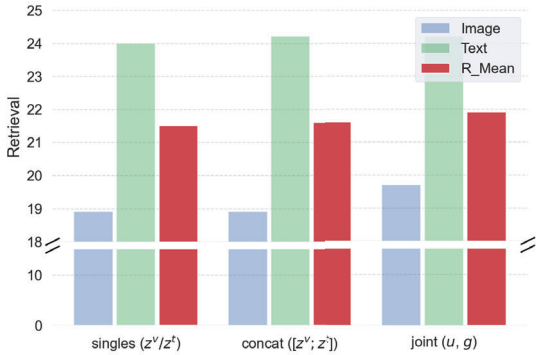

Figure S1. Comparison of matching geodesic kernel energy on (i) each image and text features separately, (ii) concatenated image-text features only, and (iii) our joint cross-modal agreement and discrepancy features.

S3. Experimental Details

S3.1. Deliberate Selection of Datasets

To study how our multimodal dataset distillation behaves as the training corpus grows, we deliberately choose three image–text benchmark datasets as exemplified in Fig. S2 that particularly *differ in scale*: Flickr8k [6] (≈8k pairs), Flickr30k [21] (≈30k pairs) and MS-COCO [11] (≈123k pairs). These datasets of increasing scale allow us to probe whether the same distillation procedure continues to yield meaningful compression and retrieval performance as we move from a small to a substantially larger dataset, while keeping the task and evaluation protocol comparable. Across this entire scale range, our method remains highly competitive, effectively condensing real data into compact synthetic subsets at substantially lower computational resource cost than existing baselines.

Figure S2. Examples of image-text datasets consisting of natural scene images and corresponding captions.

S3.2. Reasons for Underperformance on Flickr30k

In Table 1 of the main paper, we observe that on Flickr30k, a mid-scale dataset with relatively low redundancy, our text retrieval (TR) performance slightly underperforms the LoRS baseline. Each image in Flickr30k has multiple *locally diverse* captions, and many images are visually similar while differing only in subtle relational or attribute-level details. Under MDM, the real gap vectors $\{g_i\}$ associated with such visually and semantically related image–caption pairs are pulled toward a small number of synthetic gap prototypes. This averaging of gap modes reduces the margin between the ground-truth captions and near-duplicate but different captions for a given image. Caption-level discrimination in text retrieval on Flickr30k, therefore, becomes harder than for LoRS [20], which preserves more instance-level detail via low-rank similarity and maintains sharper caption margins.

In contrast, the Flickr8k dataset is about $4\times$ smaller and has a sparser caption space, with fewer near-duplicate captions per image and fewer confusing alternatives at evaluation time. In this low-data regime, the same smoothing of the gap distribution acts mainly as regularization. It suppresses unstable or idiosyncratic gaps instead of collapsing many truly distinct modes, which improves generalization relative to LoRS. On COCO, the largest and most redundant dataset in our evaluation, many caption patterns and gap structures repeat across a large number of images. In this setting, MDM compresses these repeated gap patterns into a compact set of prototypes and achieves more efficient coverage of typical multimodal relations than LoRS. This explains why our method outperforms the baseline on COCO despite operating at a stronger effective compression rate.

S3.3. Limitations of the Baseline [20]

The LoRS baseline constructs synthetic data by enforcing low-rank similarity between the behavior of real and synthetic examples on a fixed source architecture. This ties the distilled set closely to the geometry and inductive biases of a particular architecture and favors instance-level reproduction of that model's gradients and feature updates. LoRS does not explicitly model the multimodal feature distribution. Instead, it approximates the real data distributions through a low-rank subspace of parameter updates. As a result, LoRS can preserve fine caption-level distinctions that benefit same-architecture text retrieval on Flickr30k. Yet, the synthetic data often remains specialized to the source model's decision boundaries and transfers poorly to different encoders. In contrast, our MDM approach operates directly in a joint feature space and matches distributions over both agreement and discrepancy between real and synthetic pairs. This induces the synthetic set to capture architecture-agnostic structure in the joint image-text manifold rather than reproducing low-rank gradient behavior tailored to a single architecture. The resulting synthetic data provide a more faithful distribution-level approximation of multimodal semantics and modality gaps, which accounts for the superior cross-architecture generalization, even though LoRS retains a small advantage in same-architecture text retrieval on Flickr30k.

S3.4. Further Analysis on Initialization

In Fig. S3, we further inspect how different choices of data and model initialization affect the distilled (synthesized) samples. Across all three settings, the synthesized captions change locally relative to their initial counterparts, making the semantic effect of distillation more apparent on the text side. With random data and a pretrained model (*left*), the generated captions are often clearly misleading. For instance, they hallucinate or misidentify people and scene context, indicating poor image–text alignment. When we keep the model pretrained but initialize the synthetic set using our clustering-based joint-feature seeding (*middle*), the captions become slightly more faithful: core objects such as people are better captured, yet high-level scene understanding (*e.g.*, "mountain" vs. "lake") is still inaccurate. In contrast, using randomly sampled real data together with our mixed model initialization (*right*) yields captions that more reliably track the local objects and actions in the image, although some residual mismatches remain. These qualitative trends align with the quantitative ablation studies reported in Table 4 of the main paper, where combining both our data initialization and model initialization gives the strongest overall image–text matching performance. We refer to Fig. 3 of our main paper for the qualitative results of Ours altogether.

S3.5. Implementation Details

Unless otherwise noted, we conducted all ablation studies on the Flickr8k [6] dataset using 100 pairs for uniformity.

S3.5.1. System Configuration

We carried out all the experiments on a Linux Machine equipped with an Intel Xeon(R) Silver 4210R CPU and a single NVIDIA RTX A6000 GPU, following the hyperparameter settings listed in Table S3. The software stack includes Python 3.10, PyTorch 2.6.0, and Torchvision 0.21.0, with support for CUDA 11.8 and cuDNN 9.1.0.

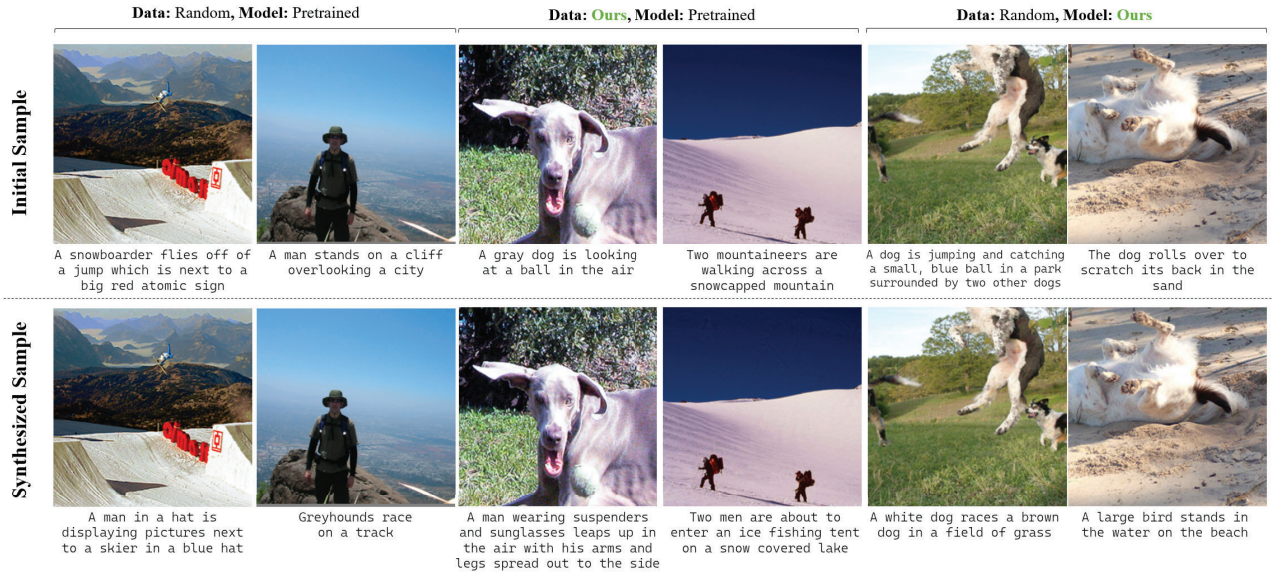

Figure S3. Qualitative comparisons for the ablation studies with different data and model initializations.

Table S3. Hyperparameters for different experiments.

Hyperparam.	Flickr8k [6]			Flickr30k [21]			COCO [11]		
# Pairs	100	200	500	100	200	500	100	200	500
Batch size	64	64	64	64	64	64	64	64	64
$\mathrm{LR}_{\mathrm{img}}$	100	100	1000	100	1000	1000	1000	1000	5000
$\mathrm{LR}_{\mathrm{txt}}$	100	100	1000	100	1000	1000	1000	1000	5000
λ_{agr}	———				0.8		———		
σ_{agr}	———				0.5		———		
λ_{dis}	———				0.8		———		
σ_{dis}	———				0.5		———		
α	———				0.5		———		
min expert epoch	———				1		———		
max expert epoch	———				10		———		

S3.5.2. Coreset Selection

To compare the image-text retrieval performance of our distilled data with traditional coreset selection methods [1–4, 7, 9, 10, 12, 14–18], we selected several benchmarked methods as practiced in [19, 20]. Specifically, we reproduced the results using DeepCore [5] for Herding [18], K-center [4], and Forgetting [17] in Table 1 of our main paper.

Herding [18]. We adapt DeepCore [5]'s herding strategy to the image–text retrieval setting using a CLIP-style encoder. We first warm up the image encoder and text projection for 5 epochs on the full training set with InfoNCE loss, using SGD (learning rate 0.1, batch size 64). After warmup epochs, we extract a joint representation for each pair by concatenating the ℓ_2-normalized image and text features, and compute the global mean feature over all training pairs. Herding then greedily selects pairs whose features make the cumulative sum of selected features closest to $(t+1)$ times the global mean at step t, yielding an ordered coreset of all training pairs. For each budget $\tilde{B} \in \{100, 200, 500\}$, we take the first $\tilde{B}$ pairs in this ordering.

K-center [4]. For K-center, we again use the same 5-epoch CLIP warmup (SGD optimizer, learning rate 0.1, batch size 64) and extract pairwise joint sum features. We then run greedy K-center on these embeddings: the first center is chosen uniformly at random, and subsequent centers are added by iteratively selecting the pair with the maximum Euclidean distance to the current selected set (farthest-first traversal) until we reach the maximum budget.

Forgetting [17]. For forgetting-based selection, we train the initial model on the full training set for 10 epochs with InfoNCE loss, using an SGD optimizer. At each training epoch, we compute the CLIP similarity logits for every batch and evaluate whether each image–text pair is correctly retrieved. A pair is marked as "correct" only when its image retrieves its own caption at rank-1 and, simultaneously, the caption retrieves its corresponding image at rank-1. For every pair, we maintain its correctness state over epochs and increment a forgetting counter whenever the pair transitions from correct to incorrect. We also track whether the pair was ever learned (*i.e.*, correct at least once) and whether it remains correct in the final epoch. After 10 epochs of training, we assign each pair a ranking score composed of its number of forgetting events and an additional penalty for pairs that were never learned. Sorting all pairs by this score in ascending order produces a forgetting-based coreset ordering, from which the first $\tilde{B}$ pairs define the selected subset.

S3.6. Performance Sensitivity

S3.6.1. Expert Pool Size

We further investigate how the size of the finetuned expert pool used for model initialization affects image–text retrieval. Specifically, we vary the number of available fine-

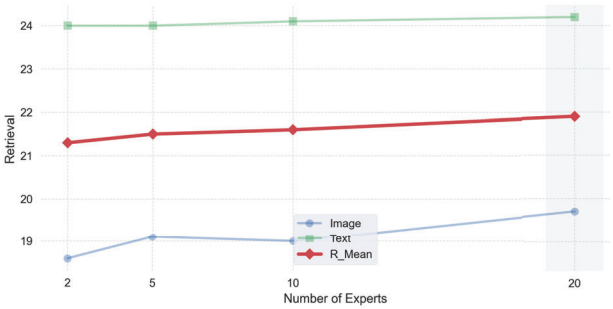

Figure S4. Retrieval performance on Flickr8k with 100 pairs as a function of the number of experts in the randomly sampled pool. Performance remains largely stable, with a slight improvement as the pool size increases.

tuned experts from which models are *randomly sampled* (non-overlapping N experts), and report retrieval performance across different pool sizes in Fig. S4. The results indicate that performance remains consistently strong, with a slight improvement as the expert pool grows. We attribute this behavior to increased stability in random expert sampling when more experts are available, which in turn raises the likelihood of selecting a more diverse set of experts.

S3.6.2. Weighting Factors

On hyperparameters λ_{agr} and λ_{dis}. We tested the sensitivity of the hyperparameter to the weighting factors for the loss of agreement and discrepancy, *i.e.*, λ_{agr}, and λ_{dis} by sweeping the range $[0, 1]$. As shown in Fig. S5, we choose the weighting factors λ_{agr} and λ_{dis} that yield the highest retrieval scores, at (0.8, 0.8). We observe that λ_{dis} contributes more to performance than λ_{agr} as there is more variation in performance with higher λ_{dis} relative to λ_{agr}. Performance improvement is relatively marginal with λ_{agr} in all sweep values, which is consistent with the results of a slightly incremental improvement shown in Table 5 (ablation study).

On hyperparameter α. The hyperparameter α in Eq. 5 of the main paper acts as a global scaling factor on the layer-wise mixing coefficient t_ℓ^m, which determines how strongly each layer of the pretrained anchor $\theta_{0,\ell}^m$ is nudged toward the averaged finetuned updates $\frac{1}{2}(\Delta_{1,\ell}^m + \Delta_{2,\ell}^m)$. When $\alpha = 0$, the initialization collapses to the original pretrained model, whereas larger α values increasingly inject real-data finetuning information along the direction prescribed by t_ℓ^m. As shown in Fig. S6, retrieval performance rapidly improves as we move away from $\alpha = 0$, peaks in an intermediate range, and then slightly degrades when α approaches 1.0, where the model becomes overly biased toward the finetuned solutions. This behavior suggests that too small α underexploits the benefits of real-data finetuning, while too large α over-specializes the distilled initialization. A mid-range value of $\alpha = 0.5$, as adopted in ours, provides a balanced compromise and yields the best mean recall across image-to-text and text-to-image retrieval.

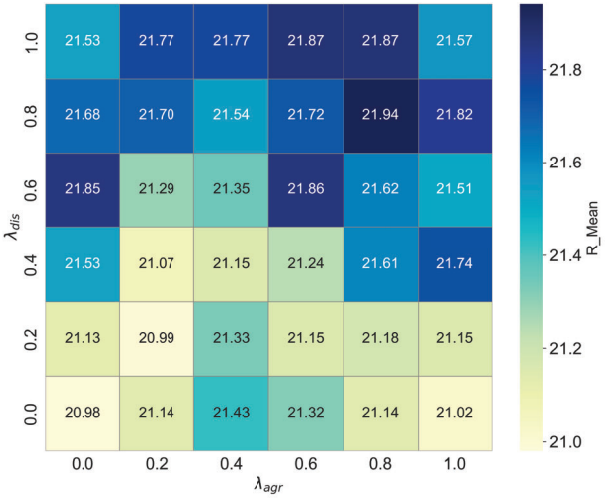

Figure S5. Hyperparameter sensitivity. While all choices of λ_{agr} and λ_{dis} consistently perform high, our selected values return the highest mean recall in image-text retrieval tasks.

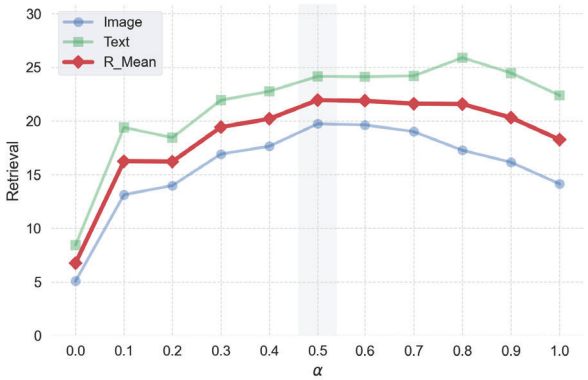

Figure S6. Hyperparameter sensitivity. While all choices of λ_{agr} and λ_{dis} consistently perform high, our selected values return the highest mean recall in image-text retrieval tasks.

S3.6.3. Additional Insights

Weight mixing for large heterogeneous encoders. We agree weight-space update-agreement should be scale-robust. [8] provides encouraging evidence in CLIP-scale models that fine-tuning updates exhibit highly structured, layer-wise regularities: displacement magnitudes and directions remain consistent across a wide range of fine-tuning conditions, and similar patterns are observed across backbone families (*e.g.*, ViT, ResNet, ConvNeXt). Given these coherent patterns, using directional agreement as a conservative mixing criterion is well-motivated, since it explicitly checks the compatibility of updates rather than relying on unstructured averaging, which can introduce interference.

Comparison to a concurrent work. In comparison to a concurrent work, EDGE [25], on the total cost, our pipeline includes a one-time expert training (∼91.7h), whereas EDGE

Table S4. **Full Image-text retrieval results** for 100, 200, and 500 synthetic pairs using the coreset methods and distillation method, including standard deviation over five random evaluations. The condensation rate for {Flickr8k, Flickr30k, and COCO} datasets are approximately {1.7%, 0.3%, 0.8‰}, {3.3%, 0.7%, 1.7‰}, {8.3%, 1.7%, 4.4‰} for 100, 200, and 500 pairs. Best and runner-up results are indicated in **boldface** and underline, respectively.

# Pairs	Dataset	Flickr8k [6]							Flickr30k [21]							COCO [11]						
	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean
100	Random	1.2	5.6	9.6	2.7	8.0	12.6	6.6	0.9	4.2	7.3	2.0	7.9	12.1	5.7	0.4	1.7	3.0	0.8	3.2	5.4	2.4
	Herding [18]	1.2	4.4	8.5	2.2	8.5	14.2	6.5	0.9	3.5	6.5	2.0	6.9	11.1	5.1	0.3	1.4	2.6	0.8	3.0	5.5	2.3
	K-Center [4]	1.2	4.9	9.0	2.7	9.3	13.9	6.8	1.1	4.9	8.7	3.0	9.1	14.3	6.8	0.5	1.9	3.6	1.1	4.2	7.6	3.2
	Forgetting [17]	1.2	4.1	7.1	1.5	4.8	8.4	4.5	0.8	3.6	6.2	1.2	5.4	9.1	4.4	0.2	1.0	1.9	0.2	1.2	2.6	1.2
	MTT-VL [19]	0.8	4.1	7.0	1.2	6.4	11.5	5.1	4.7	15.7	24.6	9.9	28.3	39.1	20.4	1.3	5.4	9.5	2.5	10.0	15.7	7.4
		± 0.0	± 0.2	± 0.3	± 0.2	± 0.6	± 0.9	± 0.3	± 0.2	± 0.5	± 1.0	± 0.3	± 0.5	± 0.7	± 0.5	± 0.1	± 0.3	± 0.5	± 0.3	± 0.5	± 0.4	± 0.4
	TESLA$_{\text{WBCE}}$	0.8	3.8	7.0	4.7	16.1	25.9	9.7	0.5	2.3	4.7	5.5	19.5	28.9	10.2	0.3	1.0	1.8	2.0	7.7	13.5	4.4
		± 0.1	± 0.4	± 0.3	± 0.3	± 0.6	± 0.8	± 0.3	± 0.2	± 0.2	± 0.4	± 0.5	± 0.9	± 1.0	± 0.5	± 0.2	± 0.4	± 0.5	± 0.2	± 0.3	± 0.3	± 0.4
	LoRS [20]	4.9	18.0	29.0	7.0	22.8	34.8	19.4	8.3	24.1	35.1	11.8	35.8	49.2	**27.4**	1.8	7.1	12.2	3.3	12.2	19.6	9.4
		± 0.3	± 0.8	± 1.2	± 0.3	± 0.4	± 0.8	± 0.5	± 0.2	± 0.2	± 0.3	± 0.2	± 0.6	± 0.5	± 0.3	± 0.1	± 0.2	± 2.0	± 0.2	± 0.3	± 0.3	± 0.5
	Ours	6.0	20.8	32.4	7.9	26.5	38.1	**21.9**	8.1	24.7	36.2	11.5	32.6	45.0	26.4	1.9	7.6	13.2	3.6	13.7	21.6	**10.3**
		± 0.4	± 0.4	± 0.7	± 0.3	± 0.5	± 0.6	± 0.3	± 0.3	± 0.8	± 0.7	± 0.7	± 0.9	± 1.0	± 0.3	± 0.1	± 0.1	± 0.2	± 0.2	± 0.3	± 0.4	± 0.2
200	Random	2.0	7.8	13.7	3.3	12.5	19.5	9.8	1.9	7.1	12.3	1.9	10.3	18.2	8.6	0.6	2.7	4.9	1.3	5.3	9.0	4.0
	Herding [18]	2.0	7.6	14.0	3.2	12.5	19.9	9.9	1.4	5.9	10.5	3.1	9.4	15.5	7.6	0.6	2.5	4.6	1.1	4.6	8.4	3.6
	K-Center [4]	2.3	9.1	15.0	3.8	13.7	20.9	10.8	2.2	8.1	13.3	4.2	13.1	21.2	10.3	0.9	3.4	5.9	2.1	7.0	11.6	5.1
	Forgetting [17]	1.7	6.5	11.5	3.1	9.7	15.4	8.0	1.6	6.6	10.8	2.5	9.0	14.9	7.6	0.4	1.6	3.0	0.7	2.8	5.1	2.3
	MTT-VL [19]	1.8	7.0	12.2	2.8	10.3	17.3	8.6	4.6	16.0	25.5	10.2	28.7	41.9	21.2	1.7	6.5	12.3	3.3	11.9	19.4	9.2
		± 0.2	± 0.2	± 0.2	± 0.3	± 0.7	± 0.7	± 0.3	± 0.9	± 1.6	± 2.6	± 0.8	± 1.0	± 1.9	± 1.5	± 0.1	± 0.4	± 0.8	± 0.2	± 0.6	± 1.2	± 0.6
	TESLA$_{\text{WBCE}}$	1.2	4.7	8.4	6.6	19.5	29.5	11.7	0.2	1.3	2.5	2.8	10.4	17.4	5.8	0.1	0.2	0.5	0.7	3.1	5.3	1.7
		± 0.2	± 0.5	± 0.6	± 0.3	± 1.1	± 1.5	± 0.6	± 0.1	± 0.2	± 0.2	± 0.5	± 1.5	± 1.6	± 0.7	± 0.1	± 0.1	± 0.1	± 0.2	± 0.5	± 0.8	± 0.3
	LoRS [20]	6.3	20.5	31.6	9.5	26.3	38.2	22.1	8.6	25.3	36.6	14.5	38.7	53.4	**29.5**	2.4	9.3	15.5	4.3	14.2	22.6	11.4
		± 0.4	± 0.7	± 0.7	± 0.4	± 0.4	± 0.9	± 0.4	± 0.3	± 0.2	± 0.3	± 0.5	± 0.5	± 0.5	± 0.4	± 0.1	± 0.2	± 0.2	± 0.1	± 0.3	± 0.2	± 0.2
	Ours	7.1	23.2	35.1	9.9	29.0	41.6	**24.3**	9.1	26.7	39.1	13.0	33.7	47.4	28.2	2.9	11.1	18.4	4.9	16.2	25.3	**13.1**
		± 0.2	± 0.4	± 0.6	± 0.2	± 0.7	± 0.6	± 0.3	± 0.2	± 0.4	± 0.4	± 0.5	± 0.9	± 0.5	± 0.2	± 0.1	± 0.2	± 0.3	± 0.1	± 0.3	± 0.3	± 0.1
500	Random	3.7	13.0	21.2	6.0	19.4	28.8	15.3	3.2	11.5	18.9	5.2	18.3	27.4	14.1	1.2	5.2	9.2	2.5	8.7	14.9	7.0
	Herding [18]	3.7	12.5	19.8	4.9	17.5	26.4	14.1	2.7	10.6	17.0	4.1	14.9	24.0	12.2	1.3	5.0	8.8	2.0	7.9	13.6	6.4
	K-Center [4]	4.0	13.4	21.1	5.9	18.9	29.0	15.4	3.4	11.8	18.7	6.7	18.0	30.6	14.9	1.5	5.7	9.7	3.0	9.9	16.2	7.7
	Forgetting [17]	4.6	16.2	24.5	5.8	21.7	31.7	17.4	3.6	12.7	20.6	6.1	18.7	29.5	15.2	1.1	4.3	7.6	2.0	7.3	11.3	5.6
	MTT-VL [19]	3.7	13.3	21.3	5.8	18.0	28.2	15.1	6.6	20.2	30.0	13.3	32.8	46.8	25.0	2.5	8.9	15.8	5.0	17.2	26.0	12.6
		± 0.0	± 0.3	± 0.5	± 0.3	± 0.6	± 0.7	± 0.3	± 0.3	± 1.2	± 2.1	± 0.6	± 1.8	± 0.8	± 1.1	± 0.5	± 0.7	± 1.5	± 0.4	± 1.3	± 1.9	± 1.1
	TESLA$_{\text{WBCE}}$	2.5	8.8	14.1	6.9	19.6	29.0	13.5	1.1	7.3	12.6	5.1	15.3	23.8	10.9	0.8	3.6	6.7	1.7	5.9	10.2	4.8
		± 0.2	± 0.3	± 0.2	± 0.4	± 0.6	± 0.6	± 0.2	± 0.2	± 0.4	± 0.5	± 0.2	± 0.5	± 0.3	± 0.4	± 0.2	± 0.6	± 0.9	± 0.4	± 0.8	± 1.0	± 0.7
	LoRS [20]	6.9	22.0	33.1	10.9	31.0	45.8	25.0	10.0	28.9	41.6	15.5	29.8	53.7	**31.6**	2.8	9.9	16.5	5.3	18.3	27.9	13.5
		± 0.4	± 1.2	± 1.6	± 0.3	± 1.0	± 1.2	± 0.7	± 0.2	± 0.7	± 0.6	± 0.7	± 0.4	± 0.3	± 0.5	± 0.2	± 0.5	± 0.7	± 0.5	± 1.5	± 1.4	± 0.8
	Ours	7.4	25.0	37.1	11.2	32.4	44.2	**26.2**	10.0	29.3	42.0	13.7	37.0	51.5	30.6	3.7	13.6	22.2	5.6	18.4	28.2	**15.3**
		± 0.4	± 0.4	± 0.7	± 0.7	± 0.7	± 0.5	± 0.3	± 0.5	± 0.5	± 0.7	± 0.5	± 0.6	± 0.9	± 0.4	± 0.1	± 0.2	± 0.5	± 0.3	± 0.2	± 0.4	± 0.2
	Full Dataset	25.5	56.1	69.2	32.7	64.5	74.5	53.8	28.1	57.9	70.3	34.9	65.4	77.6	55.7	17.3	42.8	56.7	20.4	47.7	62.1	41.2

trains SDv1.5 (∼49.0h) with additional caption generation time (*e.g.*, ∼248.14s for 100 pairs); *denotes measurement on an RTX A5000 [25], otherwise on an RTX A6000. Our end-to-end GPUh is slightly lower on Flickr30k but becomes higher on COCO as scale increases due to expert training, highlighting the limitation of expert-based methods.

500 Pairs	Method	V-L Model (GB)	Expert train (h)	Data init (s)	Distillation (s)	Total GPUh	Perf.
F30k	[25]	2.58	12.5 (1 SD expert)	-	48,960*	26.1	25.8
	MDM	0.60	25.7 (20 experts)	219	291	25.8	30.6
COCO	[25]	2.58	49.0 (1 SD expert)	-	27,720*	56.7	7.9
	MDM	0.60	91.7 (20 experts)	953	2,765	92.7	15.3

S3.7. Full Retrieval Results

We provide complete image-text retrieval results with standard deviations in five random evaluations in Table S4.

S3.8. Full Algorithm

We outline our multimodal distribution matching algorithm in Alg. 1.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020. 5

[2] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

[3] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

[4] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media, 2009. 5, 7

[5] Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A

Algorithm 1 Multimodal Distribution Matching

Input: Real image-text dataset $\mathcal{D}_{\text{real}} = \{(x_i, t_i)\}_{i=1}^{B}$;
Unified pretrained image-text model $\Psi_0 := \{\theta_0^v, \theta_0^t\}$ with frozen text encoder;
Buffer of finetuned experts $\mathcal{B} = \{(\theta_{real}^{v,(i)}, \theta_{real}^{t,(i)})\}_i^{N_{\text{experts}}}$;
Number of synthetic pairs $\tilde{B}$; Maximum distillation iterations $T_{\max}$;

Output: Optimized synthetic set $\mathcal{D}_{\text{syn}}^{\star} = \{(\tilde{x}_j, \tilde{t}_j)\}_{j=1}^{\tilde{B}}$.

Synthetic data initialization

Compute joint real embeddings with Ψ_0: $(z_i^v, z_i^t) = \Psi_0(x_i, t_i) \in \mathbb{R}^d \quad \text{for } i = 1, \ldots, B$

Run k-means clustering on joint embeddings with $K = \tilde{B}$: $\{c_k\}_{k=1}^{\tilde{B}} \leftarrow \texttt{KMeans}(\{z_i^v; z_i^t)\}_i^B)$,

For each cluster c, select the real index i_c whose joint feature $\{z^v; z^t\}_{i_c}$ is closest to the centroid

Initialize the synthetic dataset by selecting at these indices from the real dataset: $\mathcal{D}_{\text{syn}} \leftarrow \{(x_{i_c}, \theta_0^t(t_{i_c}))\}_{c=1}^{\tilde{B}}$,

while $t < T_{\max}$ **do**

Model initialization

Sample N expert checkpoints: $\{\theta_{\text{real}}^{v,(1,\cdots,N)}, \theta_{\text{real}}^{t,(1,\cdots,N)}\} \subset \mathcal{B}$.

For each trainable layer ℓ of image encoder and text projector, merge weights using Eq. 5 (for $N = 2$):
$\theta_{*,\ell}^m = \theta_{0,\ell}^m + \alpha\, t_\ell^m \cdot \frac{1}{2}\big(\Delta_{1,\ell}^m + \Delta_{2,\ell}^m\big), \quad \text{for } m \in \{v, t\},$

Instantiate trainable image-text model $\Psi_t \leftarrow (\theta_*^v, \theta_*^t)$ with text encoder kept frozen

Optimize synthetic data

Sample a real minibatch: $\{(x_i, t_i)\}_{i=1}^{B_r} \subset \mathcal{D}_{\text{real}}$

Encode joint features for real and synthetic:
$(z_{r,i}^v, z_{r,i}^t) \leftarrow \texttt{stop\_grad}(\Psi_t(x_i, \theta_t^t(t_i))), \quad (z_{s,i}^v, z_{s,i}^t) \leftarrow \Psi_t(\mathcal{D}_{\text{syn}})$

Construct cross-modal agreement and discrepancy vectors after ℓ_2 normalization:
$u_{\text{r},i} = \text{norm}(z_{\text{r},i}^v + z_{\text{r},i}^t), \quad g_{\text{r},i} = \text{norm}(z_{\text{r},i}^v - z_{\text{r},i}^t), \quad u_{\text{s},i} = \text{norm}(z_{\text{s},i}^v + z_{\text{s},i}^t), \quad g_{\text{s},i} = \text{norm}(z_{\text{s},i}^v - z_{\text{s},i}^t)$

Compute bidirectional InfoNCE loss on synthetic embeddings using Eq. 3:
$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2}\{\mathcal{L}_{\text{i2t}}(z_{\text{s}}^v, z_{\text{s}}^t) + \mathcal{L}_{\text{t2i}}(z_{\text{s}}^t, z_{\text{s}}^v)\}$

Compute geodesic kernel energies on u and g using Eq. 9:
$\mathcal{L}_{\text{agr}} = \mathsf{GKE}\big(\{u_{r,i}\}_{i=1}^{B_r}, \{u_{s,i}\}_{i=1}^{\tilde{B}}\big), \quad \mathcal{L}_{\text{dis}} = \mathsf{GKE}\big(\{g_{r,i}\}_{i=1}^{B_r}, \{g_{s,i}\}_{i=1}^{\tilde{B}}\big),$

Update synthetic data by the total loss objective using Eq. 11:
$\mathcal{L}_{\text{MDM}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{agr}} \cdot \mathcal{L}_{\text{agr}} + \lambda_{\text{dis}} \cdot \mathcal{L}_{\text{dis}}.$

Update only the synthetic parameters by gradient descent: $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} - \eta \nabla_{\mathcal{D}_{\text{syn}}} \mathcal{L}_{\text{MDM}}$

$t \leftarrow t + 1$

end while

return $\mathcal{D}_{\text{syn}}^{\star}$

comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022. 5

[6] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 3, 4, 5, 7

[7] Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pages 722–754. PMLR, 2021. 5

[8] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision*, pages 207–223. Springer, 2024. 2, 6

[9] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021. 5

[10] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8110–8118, 2021. 5

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 5, 7

[12] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and

Nikolaos Aletras. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*, 2021. 5

[13] Shibin Mei, Hang Wang, and Bingbing Ni. Geomm: On geodesic perspective for multi-modal learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4776–4786, 2025. 2

[14] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 5

[15] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

[16] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[17] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 5, 7

[18] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009. 5, 7

[19] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation, 2024. TMLR 2024. 5, 7

[20] 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*, pages 55144–55161. PMLR, 2024. 4, 5, 7

[21] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 3, 5, 7

[22] Hansong Zhang, Shikun Li, Fanzhao Lin, Weiping Wang, Zhenxing Qian, and Shiming Ge. Dance: Dual-view distribution alignment for dataset condensation. *arXiv preprint arXiv:2406.01063*, 2024. 2

[23] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, 2023.

[24] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 2

[25] 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. 6, 7