Rank-Aware Hyperbolic Alignment for Vision–Language Dataset Distillation

Jongoh Jeong[1], Sun-Kyung Lee[2], and Kuk-Jin Yoon[1]

[1] Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2] Electronics and Telecommunications Research Institute (ETRI), Republic of Korea
{jeong2, kjyoon}@kaist.ac.kr, sklee2014@etri.re.kr Project Page

Abstract. Vision-language dataset distillation (VLDD) compresses a large image-text paired dataset into a small set of synthetic pairs that can efficiently train contrastive vision-language models under strict data and compute budgets. Most existing methods match expert trajectories or cross-modal statistics, yet still enforce full-dimensional alignment in a Euclidean embedding space. This is often overly restrictive due to rank-deficient image–text correlation, with shared semantics concentrated in a low-dimensional range and remaining variation spread across a weakly correlated residual subspace. LoRS relaxes alignment at the similarity level by low-rank factorization, but does not explicitly control dominant alignment capacity and structure in the representation space. We thus propose a rank-aware hyperbolic alignment (RAHA) that combines hierarchical geometry with explicit alignment-capacity control. RAHA lifts multimodal representations to hyperbolic space and optimizes distilled pairs with asymmetric objectives that enforce geodesic alignment in the shared range while regularizing the residual subspace to preserve modality-private diversity and improve transfer robustness. Experiments on benchmarks show that RAHA demonstrates competitive cross-modal retrieval and improved transfer indicators under fixed budgets.

Keywords: Vision–language · Dataset distillation · Hyperbolic geometry

1 Introduction

Vision–language models (VLMs) have become a foundational layer of modern AI, connecting perception with language to support retrieval, captioning, grounding, and multimodal reasoning [1–3, 29, 33, 56, 69]. Their rapid progress is driven by large paired image–text datasets and scalable contrastive objectives, amplified by careful data mixture design, filtering, deduplication, and hard-negative effects in batch training. As paired collections grow from millions to billions of pairs [6, 21, 31, 53, 58, 67], capability improves, but the associated burdens escalate. These burdens include privacy exposure from unconsented personal data [4,57], uncertain licensing across countries and platforms [37], limited provenance traceability and consent management [38], and vulnerability to data poisoning and content drift [7]. These constraints increasingly shape how multimodal models can be built, audited, shared, and deployed.

This tension motivates a practical question: *can we replace web-scale paired data with compact and auditable surrogates that retain multimodal training utility?* Dataset distillation (DD) offers a principled direction by synthesizing a small set of examples that approximates the learning signal of a much larger dataset [28,35, 62,68]. Beyond efficiency, distilled surrogates can reduce the footprint of sensitive data, enable controlled redistribution when raw pairs cannot be shared, and accelerate research iteration through cheaper ablations and faster model selection. In this sense, distillation is not only a systems optimization, but also a practical tool for developing VLMs under real constraints.

Extending DD to vision–language data is harder than the unimodal case as a distilled set must preserve *cross-modal relational structure*, not only unimodal diversity. Table 1 organizes prior VLDD methods by the supervision signal extracted from real data and the spaces in which they distill images and text. Existing approaches broadly fall into three families: (i) Trajectory-based methods [64, 66, 70] match training dynamics but incur high cost and can inherit architectural bias from the teacher; (ii) Generative methods [75] leverage diffusion priors to synthesize pairs, trading direct control of alignment structure for scalability; and (iii) Distribution statistics-based methods [27] match cross-modal moments efficiently, yet they operate in the Euclidean space and still impose a largely uniform alignment constraint across feature directions. Across these families, a common limitation persists: *current methods provide limited explicit control over which representation directions should be aligned and which should remain modality-private, especially under the extreme compression.*

This limitation is consequential because image–text correlation is often effectively *low-rank* [66]. A compact shared subspace captures dominant semantic factors and coarse relations that should align across modalities, while the remaining directions form a weakly correlated *residual* component that absorbs modality-specific cues, annotation artifacts, and nuisance variation. Under tight budgets, enforcing alignment uniformly in this residual component can suppress complementary information and harm transfer. Separately, multimodal semantics are naturally hierarchical, spanning entities, attributes, and relations at different abstraction levels. Euclidean geometry offers limited inductive bias for preserving such nested structure when only a small distilled set must carry the training signal.

We thus address these limitations altogether by proposing **RAHA** (**R**ank-**A**ware **H**yperbolic **A**lignment), a geometry-aware distillation framework that controls *what* is aligned and *how*. As shown in Table 1, RAHA couples hyperbolic contrastive alignment with an explicit range–residual treatment of cross-modal correlation. RAHA estimates an adaptive rank decomposition of real batch coupling, matches synthetic relevance in the shared *range* component, and regularizes the complementary *residual* component so that weak interactions do not dominate under compression. These objectives complement a base hyperbolic contrastive term that maintains pairwise discriminability on the synthetic set. We evaluate RAHA on standard VLDD benchmarks and analyze retrieval, cross-architecture transfer, robustness to perturbations, and qualitative fidelity of synthesized pairs,

Table 1: Comparison in the approach, categorized by source of supervision from real data, latent, image/text spaces each method operates on with the respective focus.

Method	Source of supervision	Latent space	Image space	Text space	Approach/Focus
MTT-VL [64]	Training Trajectory (Expert)	Euclidean	Pixel	Enc. output	Matching Training Trajectory
LoRS [66]			Pixel	Enc. output	+ Low-rank similarity
RepBlend [70]			Pixel	Enc. output	+ Modality collapse mitigation
EDGE [75]	Diffusion Model (SD [51])		DM latent	Caption	Generative
CovMatch [27]	Distribution (Statistics)		Pixel	Enc. input	Image-text cross-covariance
Ours	Distribution (Statistics)	Hyperbolic	Pixel	Enc. input	Explicit subspace decomposition

with the goal of understanding when structured relevance distillation is most beneficial.

Contributions. We make the following main contributions in this paper:

– **Rank-aware hyperbolic formulation for VLDD.** We formulate VLDD as hyperbolic contrastive learning with tangent-space relevance distillation that exposes an explicit range–residual decomposition of cross-modal correlation.
– **Range–residual relevance distillation objective.** We propose an objective that matches real and synthetic relevance in the shared dominant range component while regularizing the residual component to preserve modality-private variability under tight budgets.
– **Evaluation with component ablations.** We evaluate RAHA on standard retrieval benchmarks and provide ablations that isolate the contributions of hyperbolic contrast, range matching, and residual regularization, alongside analyses of cross-architecture transfer, robustness, and qualitative samples.

2 Related Work

Unimodal dataset distillation (DD) synthesizes a small surrogate set that preserves the training utility of a much larger corpus [62]. Early kernel-based formulations [42, 43] offer analytical insight but scale poorly. More practical approaches match training signals or feature statistics: gradient matching [71,73], distribution matching [61, 72, 74], and efficient parameterizations of synthetic data [24,26,36,39,77]. In parallel, trajectory matching aligns full optimization paths [8] and has been extended with memory-efficient scaling [11] and improved stability [16,20,32,76]. Recent work broadens the toolkit further with implicit-gradient views [40], information-theoretic criteria [54], optimal-transport objectives [12,34], and diffusion-based generation [9,55,60]. We further refer to [10,28,68] for comprehensive surveys. These unimodal advances provide the methodological foundation, but the field is transitioning to the multimodal regime, where distilled sets must additionally preserve cross-modal correspondence.

Vision–language dataset distillation (VLDD) extends DD to paired image–text data, where a compact synthetic set must maintain both intra-modal diversity and cross-modal alignment. MTT-VL [64] initiated the field with trajectory matching and retrieval-based evaluation. LoRS [66] improves efficiency by distilling low-rank similarity structure within the same paradigm. RepBlend [70] addresses

multimodal-specific pathologies such as modality collapse through representation blending and balanced supervision. Moving beyond expert trajectories, distribution matching approaches align cross-modal statistics via geodesic kernel energy on a unit hypersphere [22] and cross-covariance [27] with within-modality regularization, where [27] is the first to train the text encoder jointly. In contrast, EDGE [75] leverages diffusion priors to generate pairs with correlation and diversity objectives. While these existing methods operate in Euclidean space with uniform alignment pressure, our method is trajectory-free and geometry-aware: we perform *rank-aware hyperbolic alignment* that decomposes cross-modal correlation into a shared range and a residual subspace, selectively enforcing alignment only where it carries semantic content while preserving modality-private diversity.
Hyperbolic geometry for vision–language. Hyperbolic spaces are Riemannian manifolds with constant negative curvature that embed tree-like hierarchies with low distortion [45,47]. MERU [14] maps image–text features onto the Lorentz model and reveals an interpretable radial layout in which generic concepts lie near the origin, suggesting that the modality gap can reflect abstraction-level differences rather than mere misalignment. Ramasinghe *et al.* [50] argue that alignment objectives should respect, not collapse, unimodal hierarchies. On the data-centric side, HDD [30] studies hyperbolic centroid matching for unimodal distillation, and recent hyperbolic VL methods exploit entailment-like asymmetry through hierarchy-aware objectives [25, 46, 48]. We build on these foundations but introduce an explicit *range–residual decomposition* within hyperbolic space, enabling selective alignment along shared semantic directions while suppressing information-deficient residual components to preserve hierarchical cross-modal structure under extreme compression.

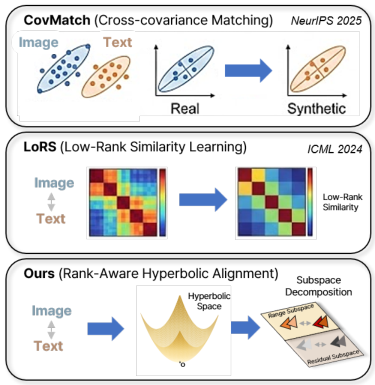
3 Proposed Method

We propose **Rank-Aware Hyperbolic Alignment** (RAHA), a vision–language dataset distillation method that learns a compact synthetic set by optimizing two objectives jointly: (i) a *hyperbolic contrastive* loss on synthetic image–text pairs, and (ii) a *rank-aware relevance distillation* loss that transfers cross-modal rank structure from real to synthetic data through a range–residual decomposition of batch-level statistics, as highlighted in Table 2.

3.1 Preliminaries and Distillation Protocol

Scope and protocol. Dataset distillation seeks a small synthetic set that preserves the training utility of a much larger dataset [8, 62, 72]. In vision–language learning, paired datasets are large and training must preserve cross-modal alignment. Since retrieval-oriented VLMs use contrastive objectives, the distilled set must preserve *bidirectional* ranking behavior (image-to-text and text-to-image). VLDD is evaluated by training a retrieval model on the synthetic set and reporting Recall@K in both directions.

Table 2: Comparison of structural information preserved and missed in the existing methods. Left: matched schematic targets. Right: description of targeted structures versus our proposed rank-aware hyperbolic alignment. *Not in particular order.



(a) Matching target structures

(b) Matching target by method

Method	Matched Target	Structure
CovMatch [27]	Cross-covariance (+ per-modality feature reg.)	• **Preserves:** 2^{nd}-order cross-modal correlation • **Misses:** Explicit pairwise ranking / higher-order structure
LoRS [66]	Explicit similarity matrix $\tilde{S}$ with low-rank factorization	• **Preserves:** Pairwise similarity via low-rank structure • **Misses:** Geometry-/hierarchy-aware structure
Ours	Rank-aware image-text relevance by subspace	• **Preserves:** Relative similarity *order* + hierarchical structure

Vision–language dataset distillation. Let the real paired dataset be $\mathcal{D}_{\text{real}} = \{(x_i, t_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ is an image and $t_i \in \mathcal{T}$ is a caption. We synthesize a compact set $\mathcal{D}_{\text{syn}} = \{(\tilde{x}_j, \tilde{t}_j)\}_{j=1}^{M}$ with $M \ll N$ such that training on $\mathcal{D}_{\text{syn}}$ yields comparable bidirectional retrieval to training on $\mathcal{D}_{\text{real}}$. The retrieval model is $\Psi = \{E_v(\cdot;\theta_v),\, E_t(\cdot;\theta_t),\, \pi_v(\cdot;\omega_v),\, \pi_t(\cdot;\omega_t)\}$, where E_v, E_t are modality-specific encoders and π_v, π_t are linear projection heads that map encoder outputs to a shared d-dimensional space. For a pair (x,t), the projected embeddings are

$$z^v = \pi_v(E_v(x;\theta_v);\omega_v) \in \mathbb{R}^d, \qquad z^t = \pi_t(E_t(t;\theta_t);\omega_t) \in \mathbb{R}^d. \tag{1}$$

RAHA operates on these projected features without cosine normalization.

Synthetic data parameterization. Each synthetic image is a learnable tensor $\tilde{x}_j \in \mathbb{R}^{3\times H\times W}$. Discrete text tokens are not differentiable, so we follow [27] to parameterize each synthetic caption by its input-layer token embeddings and an attention mask:

$$\tilde{t}_j \equiv (\tilde{\mathbf{E}}_j, \tilde{\mathbf{m}}_j), \qquad \text{where } \tilde{\mathbf{E}}_j \in \mathbb{R}^{L\times d_e},\ \ \tilde{\mathbf{m}}_j \in \{0,1\}^L, \tag{2}$$

where L is the maximum sequence length and d_e is the token embedding dimension. The mask $\tilde{\mathbf{m}}_j$ is fixed at initialization, only marking valid-token/padding positions, while $\tilde{\mathbf{E}}_j$ carries learnable linguistic content. This parameterization keeps $\mathcal{D}_{\text{syn}}$ differentiable with respect to both image pixels and text embeddings.

3.2 Hyperbolic Contrastive Alignment

Transition from Euclidean to hyperbolic InfoNCE. Standard contrastive alignment uses InfoNCE with a Euclidean similarity (dot product or cosine) to push matched pairs together and mismatched pairs apart within a batch [49]. Given a batch of paired embeddings $\{(z_i^v, z_i^t)\}_{i=1}^{\tilde{B}}$, Euclidean image–text contrastive loss constructs logits $\ell_{ij} = s(z_i^v, z_j^t)/\tau$ with temperature $\tau > 0$ and similarity function s, then applies symmetric cross-entropy over both directions. RAHA replaces the Euclidean similarity with a geodesic distance on the Lorentz hyperboloid. We base our intuition from the fact that cross-modal semantics that exhibit a hierarchical structure by which captions refine coarse visual concepts into specific attributes, and the hyperbolic geometry accommodates such nested relations through its exponential geometry [18, 44, 45].

Lorentz model and lifting. Let $c > 0$ denote the curvature parameter. A point on the Lorentz hyperboloid is $u = (u_0, \bar{u}) \in \mathbb{R}^{d+1}$ satisfying $-u_0^2 + \|\bar{u}\|_2^2 = -1/c$ with $u_0 > 0$. Given a tangent vector $w \in \mathbb{R}^d$ at the origin with norm $r = \|w\|_2$, the exponential map produces the spatial component

$$\mathrm{Exp}_o^c(w) = \frac{\sinh(\sqrt{c}\, r)}{\sqrt{c}\, r}\, w \in \mathbb{R}^d, \tag{3}$$

and the time coordinate is recovered as $u_0 = \sqrt{1/c + \|\mathrm{Exp}_o^c(w)\|_2^2}$. Given a scale parameter $s > 0$, we lift projected embeddings from Eq. 1 to the hyperboloid:

$$h^v = \mathrm{Exp}_o^c(s\, z^v), \qquad h^t = \mathrm{Exp}_o^c(s\, z^t). \tag{4}$$

The geodesic distance between two points $u = (u_0, \bar{u})$ and $v = (v_0, \bar{v})$ on the hyperboloid is

$$d_c(u, v) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\big(-c\, \langle u, v \rangle_{\mathcal{L}}\big), \tag{5}$$

where $\langle u, v \rangle_{\mathcal{L}} = -u_0 v_0 + \bar{u}^\top \bar{v}$ is the Lorentzian inner product.

Hyperbolic image–text contrastive loss (hITC). For a synthetic batch of size $\tilde{B}$, we define logits as the negative geodesic distance scaled by temperature $\tau > 0$ (*i.e.*, 0.07):

$$\ell_{ij} = -\frac{d_c(h_i^v, h_j^t)}{\tau}. \tag{6}$$

The hITC loss applies symmetric cross-entropy:

$$\mathcal{L}_{\mathrm{hITC}} = \frac{1}{2\tilde{B}} \sum_{i=1}^{\tilde{B}} \left[-\log \frac{\exp(\ell_{ii})}{\sum_j \exp(\ell_{ij})} - \log \frac{\exp(\ell_{ii})}{\sum_j \exp(\ell_{ji})} \right]. \tag{7}$$

This loss is computed on *synthetic pairs only* and does not involve real data.

3.3 Distilling Cross-Relevance via Range and Residual Subspaces

$\mathcal{L}_{\mathrm{hITC}}$ enforces alignment within each synthetic pair but does not transfer the *relative* cross-modal ranking structure observed in real data. To distill this cross-modal structure, RAHA decomposes the cross-covariance of real tangent-space features into a low-rank *range* subspace, capturing the adaptive top-k dominant image–text coupling directions, and a complementary *residual* subspace containing the remaining interactions. Relevance distributions in each subspace are then matched from real to synthetic via entropy-regularized optimal transport with Sinkhorn iterations. The final loss groups each subspace's matching term with its regularizer, yielding a clean three-term objective: hITC, range, and residual.

Tangent-space decomposition. Both real and synthetic projected features are lifted to the hyperboloid via Eq. 3 and mapped back to the tangent space at the origin with the logarithmic map:

$$x^v = \mathrm{Log}_o^c(h^v), \qquad x^t = \mathrm{Log}_o^c(h^t). \tag{8}$$

This round-trip (Euclidean → hyperboloid → tangent space) applies a nonlinear warping controlled by c and s that concentrates features according to their hyperbolic norm, so that subsequent linear operations respect the underlying geometry.

Hyperbolic tangent-space cross-covariance. For a real batch $\mathcal{B}$ of size B, let $X^v, X^t \in \mathbb{R}^{B\times d}$ stack the tangent features row-wise. Define batch means $\mu_v = \frac{1}{B}\sum_i x_i^v$, $\mu_t = \frac{1}{B}\sum_i x_i^t$ and the cross-covariance

$$C_{\mathrm{real}} = \frac{1}{B-1}(X^v - \mathbf{1}\mu_v^\top)^\top (X^t - \mathbf{1}\mu_t^\top) \ \in \mathbb{R}^{d\times d}. \tag{9}$$

$C_{\mathrm{syn}} \in \mathbb{R}^{d\times d}$ is computed identically from the synthetic batch. Real features are treated as stop-gradient targets throughout.

Range–residual subspace decomposition. We factorize C_{real} via SVD:

$$C_{\mathrm{real}} = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_d), \quad \sigma_1 \ge \dots \ge \sigma_d \ge 0, \tag{10}$$

where each singular value σ_i measures the strength of cross-modal coupling along that direction. We select the effective rank k as the smallest integer whose top-k singular values capture at least a fraction $\rho \in (0,1)$ of the total squared energy:

$$k = \min\left\{ k' : \frac{\sum_{i=1}^{k'} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2} \ge \rho \right\}. \tag{11}$$

Let $U_k \in \mathbb{R}^{d\times k}$ and $V_k \in \mathbb{R}^{d\times k}$ denote the first k columns of U and V. These define the range basis: U_k spans the image-side coupling directions and V_k spans the text-side directions. Given a tangent feature $x^v \in \mathbb{R}^d$, its range coordinates and residual component are:

$$a^v = x^v U_k \in \mathbb{R}^k, \qquad x_{\mathrm{res}}^v = x^v - a^v U_k^\top \in \mathbb{R}^d, \tag{12}$$

and analogously $a^t = x^t V_k$, $x_{\mathrm{res}}^t = x^t - a^t V_k^\top$ on the text side. The basis (U_k, V_k) is computed from the *real* batch and reused to project both real and synthetic features.

Subspace similarity matrices. Let A^v, A^t stack the range coordinates and $X_{\mathrm{res}}^v, X_{\mathrm{res}}^t$ stack the residual components for a batch. We form separate cross-modal similarity matrices:

$$G^{\mathrm{range}} = A^v (A^t)^\top \in \mathbb{R}^{n\times n}, \qquad G^{\mathrm{res}} = X_{\mathrm{res}}^v (X_{\mathrm{res}}^t)^\top \in \mathbb{R}^{n\times n}, \tag{13}$$

where n is B for real or $\tilde{B}$ for synthetic data.

Row-wise relevance distributions. Each similarity matrix is converted into row-wise probability distributions using a relevance temperature $\tau_r > 0$. The raw similarities in Eq. 13 are divided by τ_r to form logits, and the softmax within the transport loss (below) divides by τ_r again, giving an effective temperature of τ_r^2:

$$P^{\text{range}} = \text{softmax}\left(\frac{G_{\text{real}}^{\text{range}}}{\tau_r^2}\right), \qquad P^{\text{res}} = \text{softmax}\left(\frac{G_{\text{real}}^{\text{res}}}{\tau_r^2}\right), \tag{14}$$

with synthetic counterparts $Q^{\text{range}}, Q^{\text{res}}$ from $G_{\text{syn}}^{\text{range}}$ and $G_{\text{syn}}^{\text{res}}$. Real distributions P are stop-gradient targets.

Entropy-regularized set matching. Given a synthetic row distribution q_i (the i-th row of Q) and a real row distribution p_j (the j-th row of P), we measure discrepancy by KL divergence:

$$\mathrm{D}(q_i, p_j) = \sum_m q_{im} \log \frac{q_{im}}{p_{jm}}. \tag{15}$$

Stacking all pairwise divergences yields a cost matrix $\Gamma \in \mathbb{R}^{\tilde{B}\times B}$ with $\Gamma_{ij} = \mathrm{D}(Q_{i,:}, P_{j,:})$. Since there is no prescribed one-to-one correspondence between synthetic and real rows, we compute a soft coupling $T \in \mathbb{R}^{\tilde{B}\times B}$ via entropy-regularization [13]:

$$T = \arg\min_{T\geq 0} \langle T, \Gamma\rangle - \varepsilon\, \mathcal{H}(T) \quad \text{s.t.} \quad T\mathbf{1} = \tfrac{1}{\tilde{B}}\mathbf{1},\ T^\top\mathbf{1} = \tfrac{1}{B}\mathbf{1}, \tag{16}$$

where $\varepsilon > 0$ is the regularization strength and $\mathcal{H}(T) = -\sum_{i,j} T_{ij} \log T_{ij}$. This is solved by Sinkhorn–Knopp iterations [13] on the Gibbs kernel $K_{ij} = \exp(-\Gamma_{ij}/\varepsilon)$. The coupling T is computed under stop-gradient on Γ, so gradients flow only through the KL cost terms. The transport-weighted cost gives a one-directional loss:

$$\mathcal{L}_{\rightarrow}(G_{\text{real}}, G_{\text{syn}}) = \tau_r^2 \sum_{i,j} T_{ij}\, \Gamma_{ij}, \tag{17}$$

where the τ_r^2 factor compensates for the temperature scaling applied during distribution computation. We enforce bidirectionality by averaging the image-to-text and text-to-image directions:

$$\mathcal{L}_{\text{match}}(G_{\text{real}}, G_{\text{syn}}) = \tfrac{1}{2}\left[\mathcal{L}_{\rightarrow}(G_{\text{real}}, G_{\text{syn}}) + \mathcal{L}_{\rightarrow}(G_{\text{real}}^\top, G_{\text{syn}}^\top)\right]. \tag{18}$$

Applying Eq. 18 to range and residual similarities yields $\mathcal{L}_{\text{match}}^{\text{range}}$ and $\mathcal{L}_{\text{match}}^{\text{res}}$.

Range loss. The range loss combines relevance matching with an energy regularizer that prevents the synthetic range component from collapsing. Define the projected cross-covariance blocks $\widehat{C}_{\text{real}}^{\text{range}} = U_k^\top C_{\text{real}} V_k$ and $\widehat{C}_{\text{syn}}^{\text{range}} = U_k^\top C_{\text{syn}} V_k$, both in $\mathbb{R}^{k\times k}$, and their mean squared energies

$$e_{\text{real}} = \frac{1}{k^2}\|\widehat{C}_{\text{real}}^{\text{range}}\|_F^2, \qquad e_{\text{syn}} = \frac{1}{k^2}\|\widehat{C}_{\text{syn}}^{\text{range}}\|_F^2. \tag{19}$$

The regularizer penalizes synthetic energy only when it falls below the real energy:

$$\mathcal{L}_{\text{reg}}^{\text{range}} = \frac{\max(0,\ e_{\text{real}} - e_{\text{syn}})}{e_{\text{real}} + \epsilon}, \tag{20}$$

where $\epsilon > 0$ is a small constant for numerical stability. This one-sided penalty ensures the synthetic data maintains at least as much coupling energy as the real data in the top-k subspace, without penalizing cases where synthetic energy exceeds the real. The grouped range loss is then:

$$\mathcal{L}_{\text{range}} = \mathcal{L}_{\text{match}}^{\text{range}} + \mathcal{L}_{\text{reg}}^{\text{range}}. \tag{21}$$

Residual loss. The residual subspace captures cross-modal interactions outside the top-k range. Matching these interactions can sharpen ranking margins, but if residual energy dominates range energy, the distilled data over-represents weakly correlated directions at the expense of the dominant structure. The residual loss thus pairs a matching term with a compression regularizer. We form the residual cross-covariance of synthetic data by projecting out the range on both sides:

$$C_{\text{syn}}^{\text{res}} = (I - U_k U_k^\top)\, C_{\text{syn}}\, (I - V_k V_k^\top) \ \in \mathbb{R}^{d\times d}, \tag{22}$$

where I is the $d\times d$ identity. We let the residual energy defined as $e_{\text{res}} = \frac{1}{d^2}\|C_{\text{syn}}^{\text{res}}\|_F^2$ and the ratio $r = e_{\text{res}}/(e_{\text{syn}} + \epsilon)$, measuring how much synthetic coupling energy lies outside the range. The compression regularizer is defined as:

$$\mathcal{L}_{\text{reg}}^{\text{residual}} = r + \max(0,\ r - 1). \tag{23}$$

The first term provides a steady gradient shrinking r toward zero, while the second activates an additional penalty when $r > 1$ (residual energy exceeds range energy). We then define the grouped residual loss in Eq. 24 where λ_{comp} controls the relative strength of compression versus matching within the residual subspace.

$$\mathcal{L}_{\text{residual}} = \mathcal{L}_{\text{match}}^{\text{residual}} + \lambda_{\text{comp}}\, \mathcal{L}_{\text{reg}}^{\text{residual}}. \tag{24}$$

3.4 Final Distillation Training Objective

The total distillation objective is then defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hITC}} \ + \ \lambda_{\text{range}}\, \mathcal{L}_{\text{range}} \ + \ \lambda_{\text{residual}}\, \mathcal{L}_{\text{residual}}. \tag{25}$$

Here, λ_{range} and $\lambda_{\text{residual}}$ weight the range and residual subspace losses relative to the contrastive term. Expanding Eq. 21 and Eq. 24, the three scalar hyperparameters $(\lambda_{\text{range}}, \lambda_{\text{residual}}, \lambda_{\text{comp}})$ control four loss components: range matching and its regularizer share λ_{range}, residual matching is scaled by λ_{res}, and residual compression by $\lambda_{\text{residual}}\lambda_{\text{comp}}$. By default, λ_{range}=0.8, $\lambda_{\text{residual}}$=0.4, λ_{comp}=0.1.

Training procedure. Only the synthetic parameters $(\tilde{x}_j, \tilde{\mathbf{E}}_j)$ are updated by gradient descent by Eq. 25. The encoder weights $(\theta_v, \theta_t, \omega_v, \omega_t)$ are initialized from pretrained checkpoints and held fixed during the synthetic update step. Following the online distillation protocol of [27], each distillation iteration consists of multiple outer-loop steps. Within each outer step: (1) a real batch is sampled and its features are computed with stop-gradient, (2) a synthetic sub-batch of size $\tilde{B} \leq M$ is sampled , (3) distillation loss is computed and backpropagated through the synthetic pathway only, and (4) the synthetic parameters are updated by SGD. For each outer-loop step, the network is updated for one inner-loop step on real data, providing a slowly evolving latent feature landscape for enhanced generalization in the next outer step. Alg. 1 summarizes the procedure.

Algorithm 1 Distilling data with Rank-Aware Hyperbolic Alignment

Require: Real dataset $\mathcal{D}_{\mathrm{real}}$, synthetic budget M, curvature c, scale s,
outer steps N_{out}, inner steps N_{in}, iterations T, batch sizes $B, \tilde{B}$
Initialize $\mathcal{D}_{\mathrm{syn}} = \{(\tilde{x}_j, \tilde{\mathbf{E}}_j, \tilde{\mathbf{m}}_j)\}_{j=1}^{M}$ from real samples
Initialize encoder Ψ from pretrained weights
for $t = 1$ **to** T **do**
  for $\ell = 1$ **to** N_{out} **do**
    Synthetic update:
      Sample real batch $\mathcal{B}$ and synthetic sub-batch $\tilde{\mathcal{B}}$
      Compute projected features (z^v, z^t) for both batches via Ψ {Eq. 1}
      Lift synthetic features to hyperboloid via Exp, compute $\mathcal{L}_{\mathrm{hITC}}$ {Eqs. 4–7}
      Re-map all features to tangent space via Log {Eq. 8}
      Compute $C_{\mathrm{real}}, C_{\mathrm{syn}}$, SVD of C_{real}, and rank k {Eqs. 9–11}
      Project into range and residual, form $G^{\mathrm{range}}, G^{\mathrm{res}}$ {Eqs. 12–13}
      Compute relevance distributions and Sinkhorn coupling {Eqs. 14–16}
      Compute $\mathcal{L}_{\mathrm{range}}$ and $\mathcal{L}_{\mathrm{res}}$ {Eqs. 21, 24}
      Update $\mathcal{D}_{\mathrm{syn}}$ by SGD on $\mathcal{L}_{\mathrm{total}}$ {Eq. 25}
    Model update: Train Ψ for N_{in} steps on real data
  end for
end for
return $\mathcal{D}^{*}_{\mathrm{syn}}$

4 Experiments

4.1 Experimental Setup

Datasets and splits. We evaluate multimodal dataset distillation on three canonical image–caption benchmarks for bidirectional retrieval: Flickr8k [21], Flickr30k [67], and MS COCO [31]. We use the standard retrieval splits popularized by Karpathy *et al.* [23]. Unless otherwise noted, COCO results are reported on the 5k-image test split (rather than the 1k subset), which is the most common setting in recent MDD evaluations. Dataset statistics are summarized in Table 3.

Table 3: Dataset statistics for image–text retrieval. Splits follow the standard retrieval protocol [23]. Each image is annotated with five human-written captions.

Dataset	Year	#Images (Train/Val/Test)	Caps/img	Caption characteristics
Flickr8k [21]	2013	6k / 1k / 1k	5	Short, single-sentence human captions describing everyday scenes; relatively clean and small-scale.
Flickr30k [67]	2014	29k / 1k / 1k	5	Crowd-sourced, concise descriptions with broader visual diversity (people, objects, actions, relations).
MS COCO [31]	2014	113k / 5k / 5k	5	Large-scale, object-centric captions with richer compositionality and vocabulary; the most diverse of the three.

Task and metrics. We use bidirectional image–text retrieval as the main probe of cross-modal alignment. We report Recall@K for $K \in \{1, 5, 10\}$ in both directions: text-to-image (T$\rightarrow$I, denoted **IR@K**) and image-to-text (I$\rightarrow$T, denoted **TR@K**). Retrieval similarities are computed by **dot product** in the shared embedding space. For fair comparison, we evaluate *all* methods using the same similarity function.

Synthetic budget. We distill a compact synthetic set of N image–text queries with $N \in \{100, 200, 500\}$. Each query consists of (i) one image of size $3\times224\times224$ optimized in pixel space and (ii) one *continuous* 768-dimensional text embedding optimized directly in embedding space (i.e., we synthesize embeddings rather than discrete tokens), following the standard embedding-level MDD setup [64, 66]. These budgets correspond to strong compression regimes (e.g., N=100 is below 1% of COCO training images).

Baselines and network architecture. We compare RAHA to (i) coreset selection under the same budget (Random, Herding, K-Center, Forgetting) [17, 59, 63], and (ii) vision–language distillation methods, including trajectory-matching approaches ([64], [11], [66], [70]) and distribution/statistics matching [27]. All methods use the same budget, architecture, and evaluation protocol. We use the network composed from NFNet [5] and BERT [15], each followed by a lightweight linear projection head to a shared d-dimensional embedding space [27, 64, 66, 70].

Distillation protocol. We optimize the synthetic images and synthetic text embeddings while using an *online* surrogate model to compute distillation gradients. Each outer distillation iteration alternates between two stages: (i) an offline synthetic update stage, where we update $\mathcal{D}_{\text{syn}}$ for N_{out} (*e.g.*, 50) gradient steps using the current model state, and (ii) online model update stage for N_{in} (*e.g.*, 1), where we update the trainable retrieval model for one step on real data. This alternating schedule intentionally varies the surrogate model state during distillation, exposing the synthetic set to a shifting encoder geometry and improving the generality of the distilled pairs across latent-space configurations. RAHA computes its relevance matching terms in the hyperbolic representation induced by Lorentz lifting and tangent-space operations, and we allow a structured radial modality gap, consistent with the observation that text concepts are often more generic than images [14]. After distillation, we train a retrieval model from scratch on the distilled set for 100 epochs and evaluate on the real test split.

4.2 Main Results

Image–Text Retrieval. We report bidirectional retrieval performance on Flickr8k, Flickr30k, and COCO for budgets $N \in \{100, 200, 500\}$ in Table 4. Across all datasets and budgets, RAHA consistently outperforms coreset selection under the same budget, highlighting the benefit of optimizing synthetic pairs rather than selecting real pairs. Compared to trajectory-matching baselines (MTT-VL, TESLA, LoRS, RepBlend), RAHA is competitive across budgets while operating in a distribution-matching regime that does not require storing expert trajectories. On Flickr8k, RAHA scales strongly with N and attains the best overall mean recall at N=500, indicating that subspace-conditioned relevance matching preserves retrieval-relevant ranking structure even under extreme compression. On Flickr30k and COCO, RAHA remains comparable to strong trajectory-based baselines and improves with increasing N, suggesting

Table 4: Image-text retrieval for 100, 200, and 500 synthetic pairs using the coreset methods (*left*) and representative distillation methods (*right*). The compression 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, respectively.

		Coreset sampling												**Distillation methods**															
														Matching Training Trajectories (MTT)									Distribution Matching (DM)						
		Random			Herding [63]			K-Center [17]			Forgetting [59]			MTT-VL [64]			$\text{TESLA}_{\text{WBCE}}$ [11]			LoRS [66]			CovMatch [27]			Ours		
Data	# Pairs	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**	IR	TR	**Mean**
Flickr8k	100	3.5	7.8	5.7	4.7	8.3	6.5	5.0	8.6	6.8	4.1	4.9	5.1	3.9	6.2	5.1	4.6	14.1	9.4	17.3	21.5	19.4	18.4	22.4	20.4	19.0	21.9	20.4
	200	7.8	11.8	9.8	5.5	11.9	8.7	8.8	12.8	10.8	6.6	9.4	8.0	7.0	10.1	8.6	4.8	18.5	11.7	19.5	24.7	22.1	17.4	20.3	18.8	22.7	27.9	25.3
	500	12.6	18.1	15.4	12.0	16.3	14.1	12.8	17.9	15.4	15.1	19.7	17.4	12.7	17.3	15.0	8.5	18.5	13.5	20.7	29.2	25.0	23.8	28.0	25.9	27.7	33.6	30.7
Flickr30k	100	7.4	9.9	8.6	7.9	9.5	8.7	7.5	9.6	8.6	2.9	5.0	4.0	15.0	25.8	20.4	2.5	18.0	10.2	22.5	32.3	27.4	20.9	24.6	22.8	18.7	22.7	20.7
	200	11.1	15.1	13.1	10.9	14.3	12.6	10.7	14.6	12.6	4.2	6.7	5.5	15.4	26.9	21.2	1.3	10.2	5.8	23.5	35.5	29.5	20.2	23.7	22.0	23.5	27.9	25.7
	500	19.7	25.6	22.6	19.5	24.2	21.8	20.4	26.9	23.7	8.9	11.7	10.3	18.9	31.0	25.0	7.0	14.7	10.9	26.8	36.3	31.6	26.3	31.5	28.9	30.0	35.9	32.9
COCO	100	2.9	4.0	3.4	3.0	4.0	3.5	3.3	4.2	3.8	1.4	2.7	2.1	5.4	9.4	7.4	1.0	7.7	4.4	7.0	11.7	9.4	6.5	7.4	7.0	6.6	7.7	7.2
	200	4.7	6.4	5.6	4.8	6.5	5.6	5.1	6.7	5.9	2.8	3.9	3.3	6.8	11.5	9.2	0.3	3.0	1.7	9.1	13.7	11.4	7.4	9.2	8.3	9.3	11.0	10.2
	500	8.6	11.5	10.1	8.6	11.0	9.8	8.9	12.0	10.5	4.9	7.8	6.4	9.1	16.1	12.6	3.7	5.9	4.8	9.7	17.2	13.5	9.9	8.3	11.2	12.6	14.9	13.7

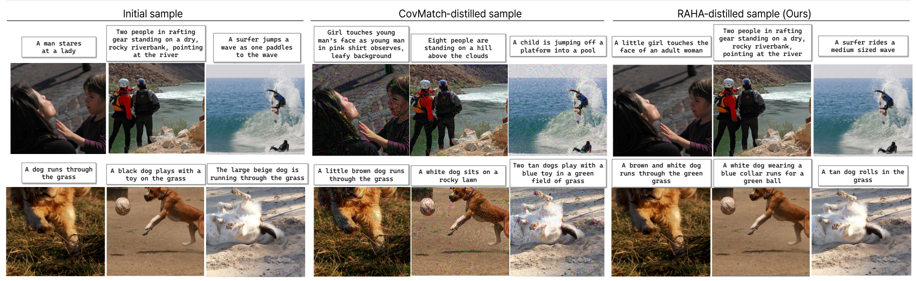


Fig. 1: Qualitative synthesized pairs. Representative samples at initialization (*left*), after CovMatch (*middle*), and after RAHA distillation (*right*). Please zoom in for details and view in color.

that relevance distillation benefits from additional synthetic capacity on more diverse datasets. See comparison with EDGE [75] in Appendix.

Relative to the strongest distribution/statistics matching baseline, CovMatch [27], RAHA trades some absolute retrieval performance in the smallest-budget regime for a different inductive bias. CovMatch directly aligns Euclidean cross-covariances and feature-level statistics, which is highly effective when N is extremely small. RAHA instead matches *row-wise relevance distributions* after extracting a rank-adaptive range/residual decomposition from real batch structure, explicitly controlling how alignment capacity is allocated across coupled and weakly-coupled directions. This distinction is reflected most clearly in cross-architecture transfer and robustness (Table 5), and is further supported by qualitative differences in the synthesized pairs (Fig. 1).

Qualitative comparison. Fig. 1 contrasts representative synthesized pairs. CovMatch outputs often retain visually salient *residual artifacts* such as high-frequency speckling and banding, even when coarse layout is plausible. Such artifacts are consistent with satisfying global second-order alignment through structured noise that perturbs features without improving perceptual realism. In contrast, RAHA yields cleaner textures and more natural edges, suggesting that range–residual relevance matching suppresses spurious degrees of freedom that manifest as visible artifacts. RAHA also better preserves fine-grained image–text

Table 5: Comparison on Flickr8k [21]: (i) **Cross-architecture generalization** averaged over IR/TR@K={1,5,10}. Source-model results marked with '*' are not averaged; best average results are in **boldface**, and runner-up in underline. (a)–(d): NFNet, NF-ResNet, NF-RegNet, ViT-B. (ii) **Robustness to noise (δ) by modality**: (a)–(c): JPEG [19] (75%)/Bit Quantization [65] (4-bit), Additive White Gaussian Noise (AWGN) ($\sigma = 0.01$), 10-PGD [41] (step size=2).

# Pairs		**(i) Cross-architecture generalization**									**(ii) Robustness to noise**							
	Text ε	BERT [15]				DistilBERT [52]					Image-side δ				Text-side δ			
	Image ε	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	Mean	(a)	(b)	(c)	Mean	(a)	(b)	(c)	Mean
100	CovMatch [27]	20.4*	6.4	6.3	7.7	9.6	6.8	6.3	7.8	7.3	6.0	6.0	1.9	4.6	3.8	10.1	0.0	4.6
	Ours	20.4*	5.1	5.4	6.3	9.9	6.7	6.7	8.0	6.9	4.9	5.6	1.8	4.1	3.1	8.1	0.0	3.7
200	CovMatch [27]	18.8*	7.2	6.3	6.7	9.3	7.2	6.7	7.2	7.2	5.0	5.7	1.9	4.2	3.3	8.7	0.0	4.0
	Ours	25.3*	7.4	7.2	9.1	10.5	8.0	7.9	10.8	8.7	5.8	6.8	2.4	5.0	3.6	9.8	0.0	4.5
500	CovMatch [27]	25.9*	8.1	7.3	8.9	12.3	8.4	7.3	9.1	8.7	8.0	8.3	3.0	6.4	3.9	11.5	0.0	5.1
	Ours	30.7*	5.4	11.3	14.6	15.4	12.5	12.8	16.9	12.7	8.0	9.5	3.0	6.8	4.4	15.0	0.0	6.4

semantics in several cases, where CovMatch drifts to mismatched captions (e.g., incorrect scenes/actions), while RAHA retains the correct relational content and discriminative attributes. Overall, the qualitative evidence aligns with RAHA's design goal: preserve dominant coupled structure while explicitly controlling weakly-coupled residual interactions.

4.3 Architectural Transfer and Robustness

To test whether distilled pairs capture transferable cross-modal relevance rather than overfitting the source encoders used during distillation, we follow a cross-architecture protocol. We distill with a fixed source configuration and then retrain/evaluate retrieval models that replace either the image backbone or the text encoder while keeping the distilled set unchanged. We also evaluate robustness via the degradation metric δ under common perturbations applied to either modality.

Across budgets. At N=100, RAHA is competitive with CovMatch on cross-architecture transfer while improving robustness to both image- and text-side perturbations, reducing the average degradation δ on both modalities. At N=200, RAHA yields a clear gain in mean transfer (from 7.2 to 8.7), with improvements that are consistent across both BERT and DistilBERT targets and across vision backbones. At N=500, transfer and robustness become less capacity-limited, and both methods generally improve; importantly, RAHA remains competitive in this higher-budget regime, indicating that the relevance structure distilled by RAHA does not rely on a single encoder geometry and continues to generalize as synthetic capacity grows. Taken together, transfer improves strongly with N, while robustness does not improve uniformly at the highest budget. These trends are consistent with the intended role of RAHA: distilling relevance in a rank-adaptive shared range while regulating residual interactions, which can improve stability under perturbations and reduce overfitting to a particular encoder configuration.

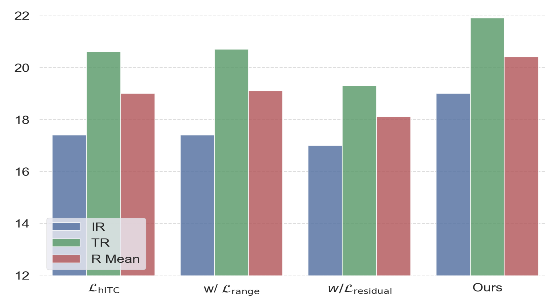


Fig. 2: Ablation study for Flickr8k N=100 setting with each component added, demonstrating the synergy of the two subspace losses.

5 Auxiliary Discussion

Ablation study. Using only hyperbolic contrast $\mathcal{L}_{\mathrm{hITC}}$ provides a strong baseline, confirming that geodesic InfoNCE on synthetic pairs already yields retrieval-relevant alignment (Fig. 2). Adding the range relevance term further improves mean retrieval, supporting the role of the range component as the dominant carrier of cross-modal coupling distilled from real data. Using the residual term alone is weaker, consistent with residual interactions being harder to match reliably without anchoring to the dominant coupled structure. Combining range and residual relevance matching with their subspace regularizers yields the best performance, indicating that residual matching is most effective when explicitly controlled rather than optimized in isolation.

Distillation cost. For completeness, we report a coarse per-iter wall-clock breakdown on an RTX A6000 into (i) data loading, (ii) model init., and (iii) the synthetic update (distillation step), as a complementary metric that contextualizes the computational profile of each objective. With identical data (0.67s) and model (0.3s) initialization, the main difference arises in the distillation step: at a batch size of 1, both are similar at $\approx 25\,$s, while the difference grows with batch size due to lifting and decomposition particularly (*e.g.*, at 64, $\approx 55\,$s vs. $\approx 400\,$s), reflecting the stronger batch-scaling cost of RAHA's structure-aware update. We do not use runtime as the primary comparison axis, since our evaluation focuses on retrieval accuracy, transfer, and robustness under fixed synthetic budgets.

In sum, RAHA distills structured image–text relevance through a rank-aware shared subspace. It does not claim uniform dominance in every extreme-compression cell. RAHA is competitive at `100` pairs and outperforms [27] at `200`/`500`. Here, RAHA brings to the surface an important difficulty in VLDD: dataset-scale semantics can be abstracted into decomposed range/residual bases, but a very small synthetic set may not have enough capacity to materialize those semantic modes. At extreme rates, *e.g.*, on Flickr30k/MS COCO, the selected basis can be stable, but too few pairs may not populate the decomposed structure. Thus, MTT-style [66] can remain strong at the smallest budgets. As # pairs $\uparrow$, RAHA better materializes these modes. Structurally, RAHA uses the conceptual text hierarchy [14, 25] as an inductive bias for VLDD, since hyperbolic geometry can keep generic concepts close to many descendants while keeping specific instances separable.

On transfer and robustness comparison. [70] builds on the MTT-based [66] line by addressing modality collapse, whereas RAHA follows the distribution-matching line of [27], with hyperbolic range–residual relevance modeling in a methodologically complementary, rather than mutually exclusive, direction.

# Pairs	Method	Trainable Text Enc.	Cross-arch (Mean)	Image-δ (Mean)	Text-δ (Mean)
100/200/500	LoRS [66]	✗	10.2 / 10.8 / 11.0	0.5 / 0.5 / 0.5	0.5 / 0.5 / 0.6
	RepBlend [70]	✗	8.1 / 8.9 / 10.7	0.6 / 0.5 / 0.5	0.5 / 0.6 / 0.5
	CovMatch [27]	✓	7.3 / 7.2 / 8.7	4.6 / 4.2 / 6.4	4.6 / 4.0 / 5.1
	Ours	✓	6.9 / 8.7 / 12.7	4.1 / 5.0 / 6.8	3.7 / 4.5 / 6.4

Rank selection & subspace stability. RAHA selects the effective rank not manually but based on the smallest k whose cumulative squared singular-value energy of $C_{\text{real}}{=}U\Sigma V^\top$ reaches ρ, with numerical-rank clipping and $k_{\max}$ (Eq.11, §A3.2). This defines the range as the minimal energy-bearing image–text coupling subspace and treats the complement as residual structure. Fig. 3(a) shows that larger real batches improve retrieval by stabilizing the cross-covariance estimate, rather than causing k to simply follow the algebraic ceiling $\texttt{rank}(C_{\text{real}}) \leq B{-}1$. Fig. 3(b) shows that $\rho{=}0.95$ is the best energy threshold, while 1.0 degrades by absorbing the low-energy residual tail. Fig. 3(c) further shows that the selected rank remains stably within a dataset-specific bound over iterations, supporting stable batch-wise SVD in practice.

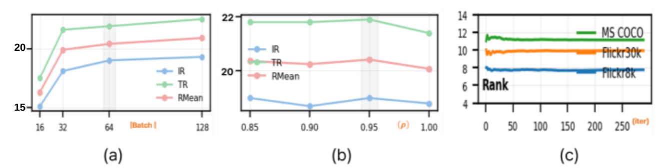


Fig. 3: Batch, hyperparameter ρ ablations (a,b) and rank by dataset (c).

6 Conclusion

In this work, we present a hyperbolic geometry-aware objective for VLDD, termed **RAHA**, that lifts image–text representations to *hyperbolic space* and enforces *rank-consistent* cross-modal relevance matching, improving stability and cross-architecture transfer under tight compression budgets.

Limitations. As with [22, 27, 64, 66, 70], RAHA remains bounded by the expressivity of chosen teacher encoders, and performance may degrade under domain-shifted or noisy captions. RAHA is driven by inherent hierarchical structures, and thus datasets with weak hierarchy may see limited gains (§A5).

Disclosure of LLM Use. LLM was used only for editing (*e.g.*, formatting) and did not contribute to the underlying technical ideas or experimental results.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25473963). We thank Jaehyuk Jang for providing key insights.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Birhane, A., Prabhu, V., Han, S., Boddeti, V.N., Luccioni, A.S.: Into the LAIONs den: Investigating hate in multimodal datasets. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023)
5. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: International conference on machine learning. pp. 1059–1071. PMLR (2021)
6. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
7. Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., Tramèr, F.: Poisoning web-scale training datasets is practical. In: IEEE Symposium on Security and Privacy (SP) (2024)
8. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: CVPR (2022)
9. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Generalizing dataset distillation via deep generative prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3739–3748 (2023)
10. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Dc-bench: Dataset condensation benchmark. Advances in Neural Information Processing Systems **35**, 810–822 (2022)
11. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: International Conference on Machine Learning. pp. 6565–6590. PMLR (2023)
12. Cui, X., Qin, Y., Zhou, W., Li, H., Li, H.: Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15245–15254 (2025)
13. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
14. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, R.: Hyperbolic image-text representations. In: ICML (2023)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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)
16. Du, J., Jiang, Y., Tan, V.Y., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3758 (2023)
17. Farahani, R.Z., Hekmatfar, M.: Facility location: concepts, models, algorithms and case studies. Springer Science & Business Media (2009)

18. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic neural networks. Advances in neural information processing systems **31** (2018)
19. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
20. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching. arXiv preprint arXiv:2310.05773 (2023)
21. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
22. Jeong, J., Kwon, H., Kim, M., Yoon, K.J.: Multimodal distribution matching for vision-language dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
23. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
24. Kim, J.H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: ICML (2022)
25. Kim, W., Chun, S., Kim, T., Han, D., Yun, S.: Hype: Hyperbolic entailment filtering for underspecified images and texts. In: European Conference on Computer Vision (ECCV) (2024). `https://doi.org/10.48550/arXiv.2404.17507`
26. Lee, H.B., Lee, D.B., Hwang, S.J.: Dataset condensation with latent space knowledge factorization and sharing. arXiv preprint arXiv:2208.10494 (2022)
27. Lee, Y., Chung, H.W.: Covmatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. arXiv preprint arXiv:2510.18583 (2025)
28. Lei, S., Tao, D.: A comprehensive survey of dataset distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 17–32 (2023)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
30. Li, W., Li, G., Maeda, K., Ogawa, T., Haseyama, M.: Hyperbolic dataset distillation. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), poster
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
32. Liu, D., Gu, J., Cao, H., Trinitis, C., Schulz, M.: Dataset distillation by automatic training trajectories. In: European Conference on Computer Vision. pp. 334–351. Springer (2024)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
34. Liu, H., Li, Y., Xing, T., Wang, P., Dalal, V., Li, L., He, J., Wang, H.: Dataset distillation via the wasserstein metric. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1205–1215 (2025)
35. Liu, P., Du, J.: The evolution of dataset distillation: Toward scalable and generalizable solutions. arXiv preprint arXiv:2502.05673 (2025)
36. Liu, S., Wang, K., Yang, X., Ye, J., Wang, X.: Dataset distillation via factorization. In: NeurIPS (2022)
37. Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al.: The data provenance

initiative: A large scale audit of dataset licensing & attribution in AI. arXiv preprint arXiv:2310.16787 (2023)
38. Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., et al.: Consent in crisis: The rapid decline of the AI data commons. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
39. Loo, N., Hasani, R., Amini, A., Rus, D.: Efficient dataset distillation using random feature approximation. In: NeurIPS (2022)
40. Loo, N., Hasani, R., Lechner, M., Rus, D.: Dataset distillation with convexified implicit gradients. arXiv preprint arXiv:2302.06755 (2023)
41. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
42. Nguyen, T., Chen, Z., Lee, J.: Dataset meta-learning from kernel ridge-regression. arXiv preprint arXiv:2011.00050 (2020)
43. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. In: NeurIPS (2021)
44. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. arXiv preprint arXiv:1705.08039 (2017)
45. Nickel, M., Kiela, D.: Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research (PMLR), vol. 80, pp. 3779–3788 (2018)
46. Pal, A., van Spengler, M., D'Amely di Melendugno, G.M., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. In: International Conference on Learning Representations (ICLR) (2025), oral
47. Peng, W., Varanka, T., Mostafa, A., Shi, H., Zhao, G.: Hyperbolic deep neural networks: A survey. arXiv preprint arXiv:2101.04562 (2021)
48. Poppi, T., Kasarla, T., Mettes, P., Baraldi, L., Cucchiara, R.: Hyperbolic safety-aware vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025). `https://doi.org/10.48550/arXiv.2503.12127`
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
50. Ramasinghe, S., Shevchenko, V., Avraham, G., Thalaiyasingam, A.: Accept the modality gap: An exploration in the hyperbolic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27263–27272 (June 2024)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
52. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
53. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
54. Shang, Y., Yuan, Z., Yan, Y.: Mim4dd: Mutual information maximization for dataset distillation. arXiv preprint arXiv:2312.16627 (2023)

55. Su, D., Hou, J., Gao, W., Tian, Y., Tang, B.: D^4m: Dataset distillation via disentangled diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5809–5818 (2024)
56. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
57. Thiel, D.: Identifying and eliminating CSAM in generative ML training data and models. Tech. rep., Stanford Internet Observatory (2023). https://doi.org/10.25740/kh752sm9123, https://purl.stanford.edu/kh752sm9123
58. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
59. Toneva, M., Sordoni, A., Combes, R.T.d., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. arXiv preprint arXiv:1812.05159 (2018)
60. Wang, H., Zhao, Z., Wu, J., Shang, Y., Liu, G., Yan, Y.: Cao_2: Rectifying inconsistencies in diffusion-based dataset distillation (2025), https://arxiv.org/abs/2506.22637
61. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: Cafe: Learning to condense dataset by aligning features. In: CVPR (2022)
62. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
63. Welling, M.: Herding dynamical weights to learn. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1121–1128 (2009)
64. Wu, X., Zhang, B., Deng, Z., Russakovsky, O.: Vision-language dataset distillation (2024), https://openreview.net/forum?id=2y8XnaIiB8, tMLR 2024
65. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: NDSS (2018). https://doi.org/10.14722/ndss.2018.23295, https://www.ndss-symposium.org/ndss-paper/feature-squeezing-detecting-adversarial-examples-in-deep-neural-networks/
66. Xu, Y., Lin, Z., Qiu, Y., Lu, C., Li, Y.L.: Low-rank similarity mining for multimodal dataset distillation. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 55144–55161. PMLR (2024), https://proceedings.mlr.press/v235/xu24q.html
67. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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)
68. Yu, R., Liu, S., Wang, X.: Dataset distillation: A comprehensive review. IEEE transactions on pattern analysis and machine intelligence **46**(1), 150–170 (2023)
69. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pretraining. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
70. Zhang, X., Zhang, Z., Du, J., Liu, Z., Zhou, J.T.: Beyond modality collapse: Representations blending for multimodal dataset distillation. arXiv arXiv:2505.14705 (2025)
71. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: ICML (2021)
72. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: WACV (2023)

73. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)
74. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7856–7865 (2023)
75. Zhao, Z., Wang, H., Wu, J., Shang, Y., Liu, G., Yan, Y.: Efficient multimodal dataset distillation via generative models. arXiv preprint arXiv:2509.15472 (2025)
76. Zhong, W., Tang, H., Zheng, Q., Xu, M., Hu, Y., Guan, W.: Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25581–25589 (2025)
77. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. arXiv preprint arXiv:2206.00719 (2022)