Rank-Aware Hyperbolic Alignment for Vision-Language Dataset Distillation

ECCV 2026



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

[1]Visual Intelligence Lab., KAIST, [2]ETRI

Background: Rise of Data-Centric AI

- **Data quality** is fundamental to training models in ML/CV tasks
- **Key research direction**: how to curate the most valuable set of data for model training?
- **Evolving landscape of data-centric AI toward the dynamic value of data**

Compression Strength

Focus: *performance* ↑

✓ **Data Attribution**
- *Core Question*: Which samples influence model behavior?

Low

✓ **Data Sculpting**
- *Core Question*: How should the training distribution be reshaped?

Low-Medium

Focus: *efficiency* ↑

✓ **Data Pruning / Coreset Selection**
- *Core Question*: Which real samples can represent the entire dataset?

Medium

✓ **Data Quantization**
- *Core Question*: How can real datasets be compacted systematically?

Medium-High

✓ **Data Distillation**
- *Core Question*: Can we synthesize data to replace the original dataset?

High

Shift from "Finding better data" → "Compressing dataset semantics"

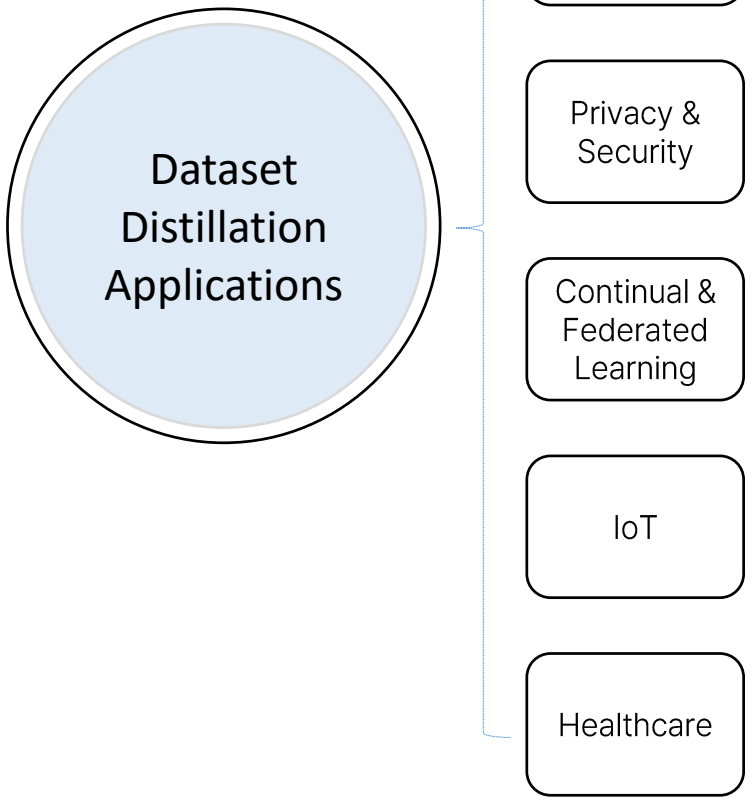
Background: Dataset Distillation

- **Dataset distillation** aims to compress a large training set into a small set of synthetic samples that retain its learning utility.

  ✓ **Why compress the dataset?**
  - Modern training pipelines are often limited by data scale, storage, and repetitive training cost
  - DD relieves this cost by building a compact yet highly informative core, enabling efficient learning under tight budgets

  ✓ **What should a distilled dataset preserve?**
  - The most transferable training signals of the original dataset
  - Strong generalization performance, despite using only a tiny number of synthetic samples

  ✓ **Practical significance**
  - Allows repeated training feasible for resource-constrained and iterative settings
  - Broadly useful for applications including edge device deployment, NAS, privacy-sensitive learning, federated learning, IoT, and healthcare



Background: Unimodal Dataset Distillation

- **(Image) Recent real-world-facing application:** Visual semantic condensation of image datasets
  - ✓ **Raw pixels → task-relevant visual prototypes**
    - Synthesizes compact image that retain the semantics needed for recognition tasks
  - ✓ **Dataset-scale training → efficient surrogate training**
    - Replaces repeated full-dataset training with small distilled sets for rapid and continual model search and retraining
  - ✓ **Full data memory → compact semantic buffer**
    - Uses distilled images as replay/shareable surrogates while reducing storage, transmission, and privacy breach

- **(Text) Recent real world application:** Semantic distillation of text inputs for LLMs
  - ✓ **Raw text → task-relevant abstraction**
    - Extracts entities, intents, relations, and constraints from noisy documents, emails, logs, or prompts
  - ✓ **Surface-form compression, semantic preservation**
    - Removes redundant wording while retaining the information needed for reasoning, retrieval, or action
  - ✓ **Schemas as distilled semantic carriers**
    - Converts unstructured text into compact fields

Common goal: to mine abstract semantics inherent in the original data sufficient for downstream behavior

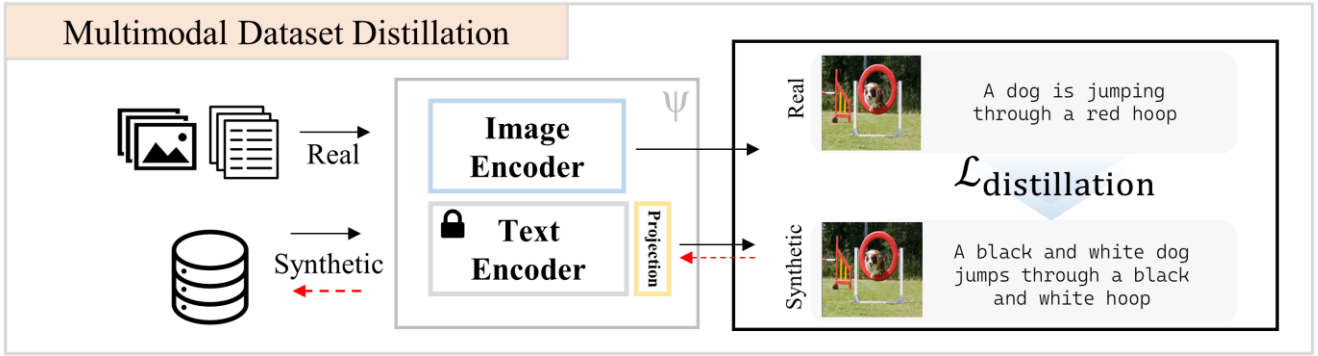
Background: Vision-Language Dataset Distillation

- **Vision-Language Dataset Distillation** compresses a large-scale image-text paired dataset into a smaller set of compact yet representative synthetic pairs.

  ✓ **Aim**
  - Learn a small set of synthetic image-text pairs
  - Retain the training utility of the full real multimodal dataset

  ✓ **Why compress multimodal data?**
  - Reduce storage and training cost for vision-language learning
  - Enable faster experimentation and deployment with compact synthetic data

  ✓ **What must be preserved?**
  - Intra-modal structure within image and text embeddings
  - Cross-modal alignment between paired images and captions

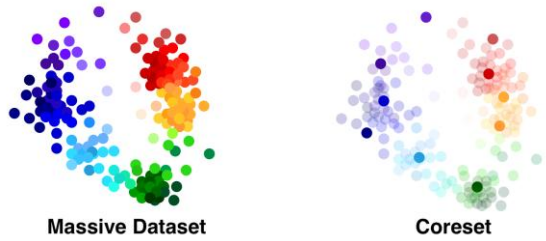


Source: MDM[1]

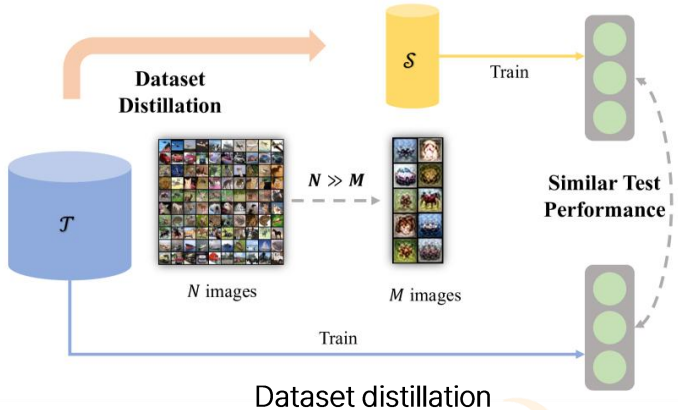
[1] Jeong, Jongoh, et al. "Multimodal Distribution Matching for Vision-Language Dataset Distillation." CVPR 2026.

Limitations: Why Prior Methods Fall Short

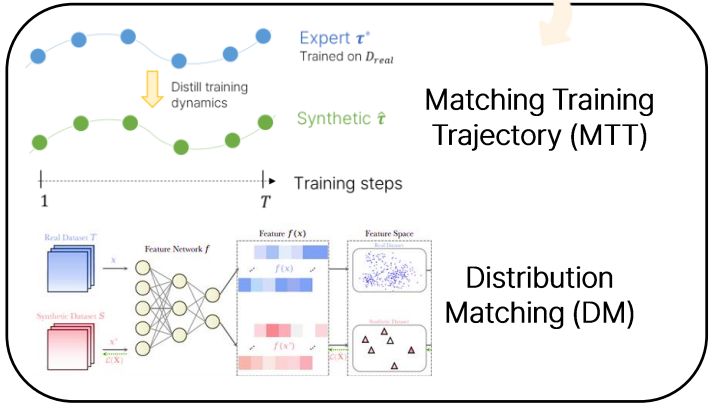
- **Previous approaches**
  - ✓ **Coreset Selection**
    - Selects a subset of real image-text pairs that approximately covers the dataset feature space
    - Efficient and simple, yet essentially a re-sampling strategy
    - Coverage-based heuristics often miss joint semantic modes and weakly preserve cross-modal structure
  - ✓ **Trajectory / Similarity-based Supervision**
    - Prior MDD methods replay training trajectories or preserve low-rank similarity structure
    - While this improves fidelity to the training dynamics, it relies on indirect supervision from particular experts rather than directly matching the multimodal data distribution itself
- **Limitations**
  - ✓ **Reliance on Euclidean geometry without considering inherent text hierarchy**
    - Provides limited inductive bias for preserving naturally hierarchical levels of text captions → capture multimodal semantics on hyperbolic space
  - ✓ **Lack of explicit cross-modal relational structure preservation**
    - Existing methods provide limited control over which representation directions should be aligned and remain modality-specific → explicit subspace decomposition by dominant rank



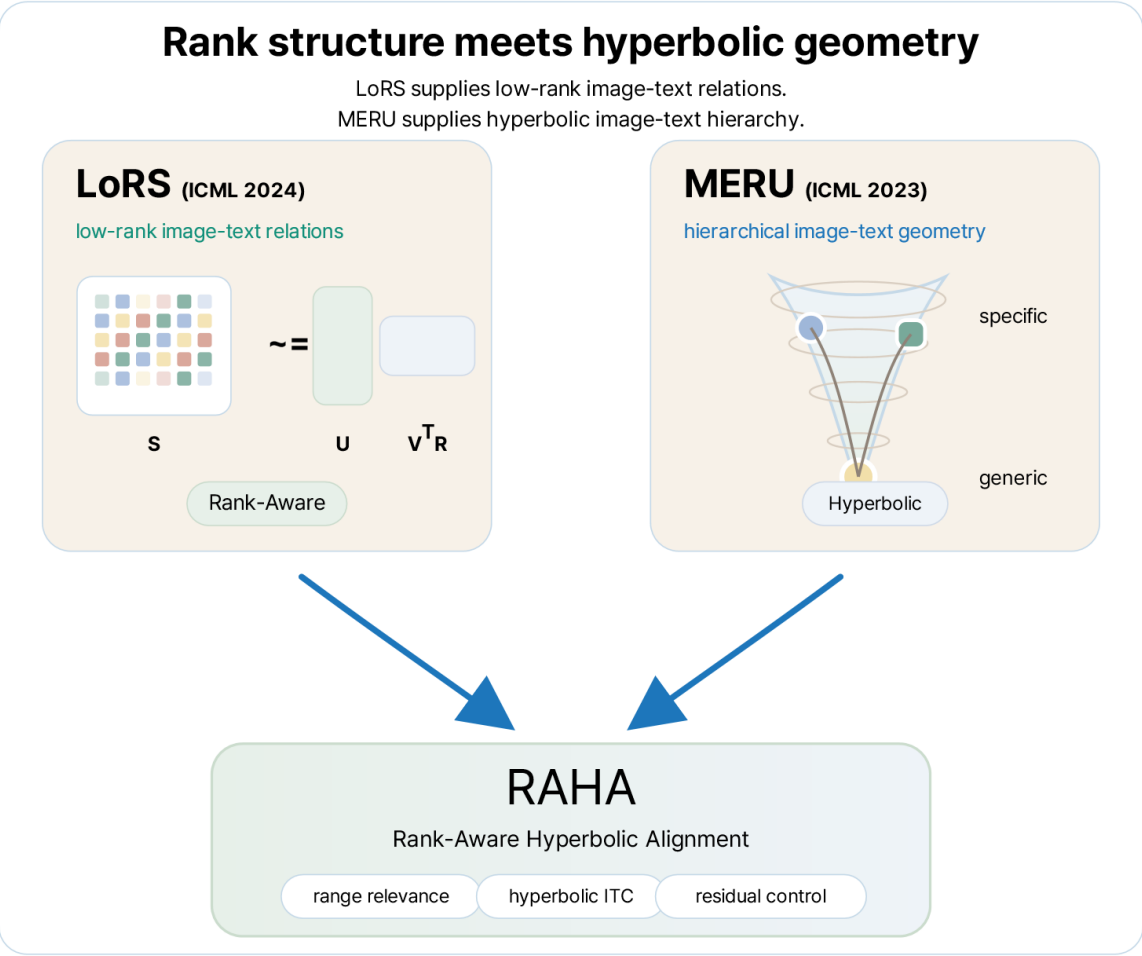
Coreset Selection



Dataset distillation



Motivation: Rank structure meets hyperbolic hierarchy



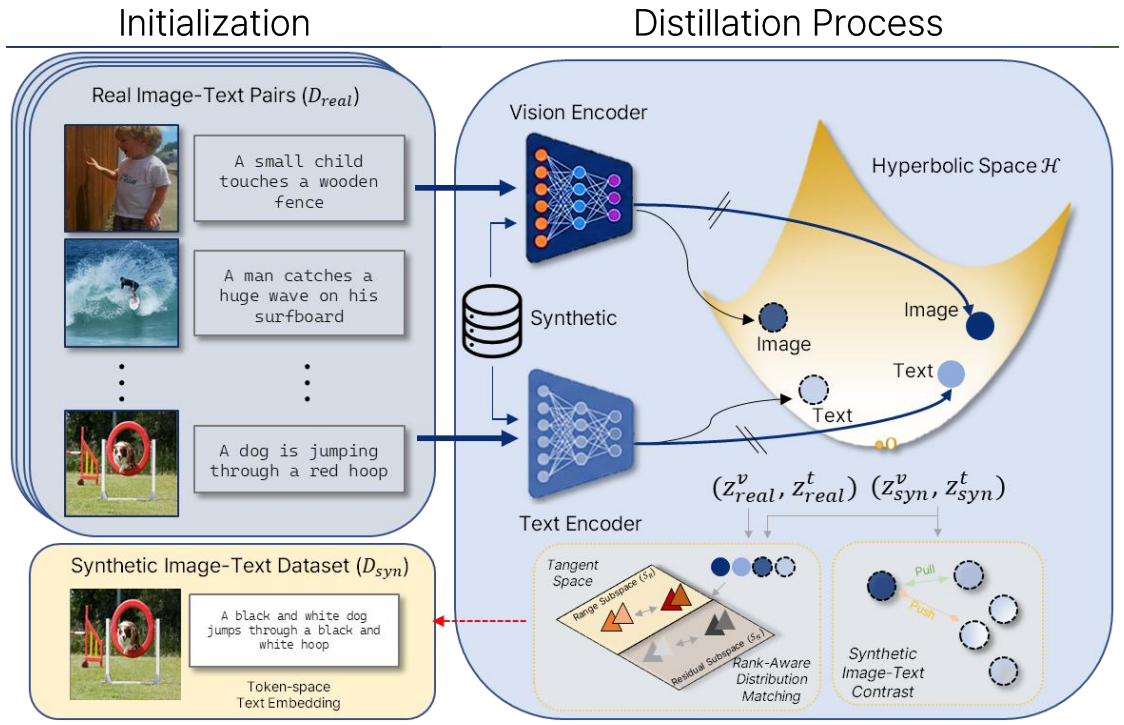
- **Insights from previous literature**
  - ✓ **Key semantics lie in low-rank**
    - LoRS[2] entails the useful semantic information lies in a relatively low-rank structure
  - ✓ **Hyperbolic structure for text hierarchy**
    - MERU[3] and a follow-up work HYPE[4] demonstrate hierarchy-aware hyperbolic embeddings are beneficial for learning image-text associations
  - ▶ **Our proposed approach: RAHA**
    - Enjoy the best of both worlds by exploiting the dominant rank-decomposed structure on the hyperbolic space!

[2] Xu, Yue, et al. "Low-rank similarity mining for multimodal dataset distillation.", ICML 2024.
[3] Desai, Karan, et al. "Hyperbolic image-text representations." ICML 2023.
[4] Kim, Wonjae, et al. "Hype: Hyperbolic entailment filtering for underspecified images and texts." ECCV 2024.

Rank-Aware Hyperbolic Alignment

✓ Key Research Questions

1. **What must V-L data compression preserve?**
   - A distilled image-text set is useful only if a model trained on it learns the same bidirectional ranking behavior as from real data. This makes cross-modal relevance the core information capacity to preserve, beyond visual appearance alone.
2. **How should text hierarchy be retained?**
   - The cross-modal relevance signal is not flat because captions move from generic categories to fine attributes and relations. This calls for geometry that keeps generic text broad while keeping fine-grained image-text pairs distinct.
3. **Where should alignment capacity be spent?**
   - LoRS[2] shows that useful image-text relations are concentrated in low-rank structure. RAHA therefore separates range and residual subspaces, matching dominant shared semantics while regulating weaker residual variation.



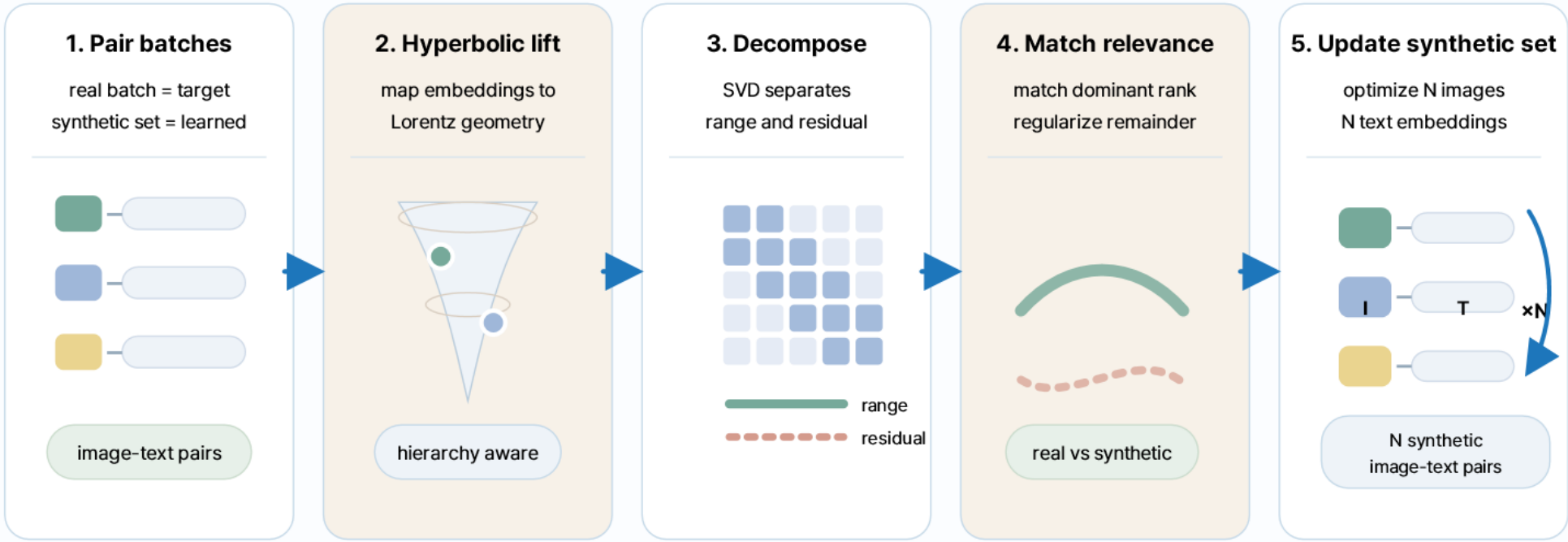
Overview of RAHA.

[2] Xu, Yue, et al. "Low-rank similarity mining for multimodal dataset distillation.", ICML 2024.

RAHA: One-Step Walkthrough

RAHA walkthrough in one distillation step

A real batch defines cross-modal relevance targets, while learned synthetic image-text pairs are the only optimized data.

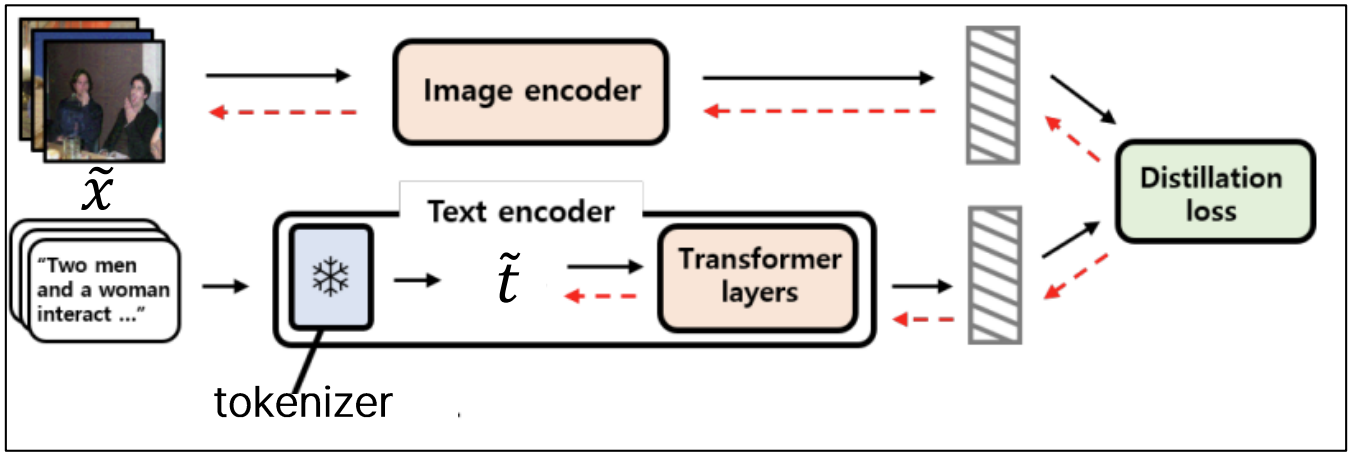


Real batches define target image-text relevance.
Synthetic pairs are updated to reproduce it with far fewer pairs while preserving downstream utility.

$$L_{total} = L_{hITC} + \lambda_{range} \cdot L_{range} + \lambda_{residual} \cdot L_{residual}$$

Synthetic Data Parameterization

- **Insights from previous literature**
  - ✓ CovMatch[4] observed a bottleneck in scaling multi-modal semantic alignment due to frozen text encoder, existing in prior works[2]
  - ✓ **Learnable parameters (D_{syn}):**
    - Image: $\tilde{x} \in \mathbb{R}^{3\times H\times W}$
    - Text embeddings: $\tilde{t} \in \mathbb{R}^{d}$



Distillation Training Protocol[4]

[2] Xu, Yue, et al. "Low-rank similarity mining for multimodal dataset distillation.", ICML 2024.
[4] Lee, Yongmin, and Hye Won Chung. "CovMatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder." NeurIPS 2025.

Hyperbolic Contrastive Alignment

- **Transition from Euclidean space to Hyperbolic space**
  - ✓ Conventional Euclidean contrastive objective (eITC)
  - ✓ Hyperbolic-lifted contrastive objective (hITC)

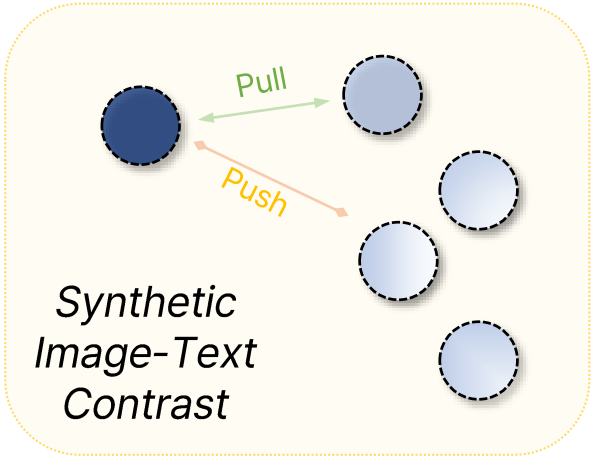
Define Lorentz lifting: (see Sec. 3.2 for details)

$$\boldsymbol{Exp_O^C(w) := \frac{sinh(\sqrt{c}\cdot r)}{(\sqrt{c}\cdot r)} \cdot w \in \mathbb{R}^d,}$$

$where \quad \boldsymbol{r} = ||\boldsymbol{w}||_2$ (norm), c>0 (curvature).

Lift projected embeddings to the hyperboloid as:

$$h^v = Exp_O^C(s \cdot z^v), \quad h^t = Exp_O^C(s \cdot z^t)$$



Define geodesic distance between two points on the hyperboloid as:

$$d_c(u,v) = \frac{1}{\sqrt{c}} arcosh(-c\langle u,v\rangle_L), \quad where \langle u,v\rangle_L = -u_0v_0 + \bar{u}^T v \text{ is the Lorentzian inner product.}$$

Hyperbolic image-text contrastive loss:

$$L_{hITC} = \frac{1}{2\tilde{B}} \sum_{i=1}^{\tilde{B}} \left[-\log \frac{\exp(l_{ii})}{\exp(l_{ij})} - \log \frac{\exp(l_{ii})}{\exp(l_{ji})} \right], \quad \text{where } l_{ij} = -\frac{d_c(h_i^v, h_j^t)}{\tau}$$

Cross-Modal Relevance Matching

- **Rank-Aware Decomposition**
  - ✓ Tangent space projection
  - ✓ Compute hyperbolic tangent-space cross-covariance
  - ✓ Range-residual subspace decomposition
  - ✓ Rank-aware cross-modal relevance matching via optimal transport

Algorithm 1 Distilling data with Rank-Aware Hyperbolic Alignment

Require: Real dataset $\mathcal{D}_{\text{real}}$, synthetic budget M, curvature c, scale s, outer steps N_{out}, inner steps N_{in}, iterations T, batch sizes $B, \tilde{B}$

Initialize $\mathcal{D}_{\text{syn}} = \{(\tilde{x}_j, \tilde{\mathbf{E}}_j, \tilde{\mathbf{m}}_j)\}_{j=1}^{M}$ 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
      Lift synthetic features to hyperboloid via Exp, compute $\mathcal{L}_{\text{hITC}}$
      Re-map all features to tangent space via Log
      Compute $C_{\text{real}}, C_{\text{syn}}$, SVD of C_{real}, and rank k
      Project into range and residual, form $G^{\text{range}}, G^{\text{res}}$
      Compute relevance distributions and Sinkhorn coupling
      Compute $\mathcal{L}_{\text{range}}$ and $\mathcal{L}_{\text{res}}$
      Update $\mathcal{D}_{\text{syn}}$ by SGD on $\mathcal{L}_{\text{total}}$
    Model update: Train Ψ for N_{in} steps on real data
  end for
end for
return $\mathcal{D}^*_{\text{syn}}$

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

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

$$\boldsymbol{L_{total}} = L_{hITC} + \lambda_{range} \cdot L_{range} + \lambda_{residual} \cdot L_{residual}$$

Experimental Setup

- ✓ **Datasets** (scale)
  - Flickr8k (~8k)
  - Flickr30k (~30k)
  - MS COCO (~123k)

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

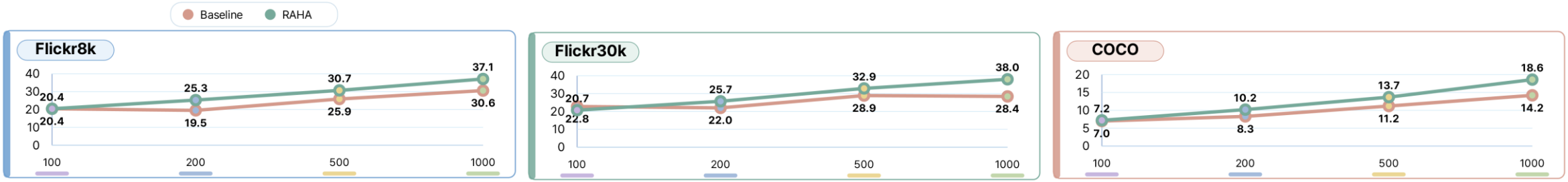
- ✓ **Task**
  - Image-to-Text Retrieval at $K \in \{1,5,10\}$
  - Text-to-Image Retrieval at $K \in \{1,5,10\}$
  - Prompted Classification

- ✓ **Synthetic Budget**
  - N $\in \{100, 200,500\}$
  - Each pair contains $3 \times 224 \times 224$ synthetic image and 768-dimensional synthetic text embedding

- ✓ **Baselines and Architectures**
  - Coreset methods (Random, herding, K-center, forgetting)
  - VLDD baselines (MTT-VL, $\text{TESLA}_{\text{WBCE}}$, LoRS, CovMatch)
  - Architectures: NFNet (Image), BERT (Text) with lightweight projection heads

Results: Image-Text Retrieval

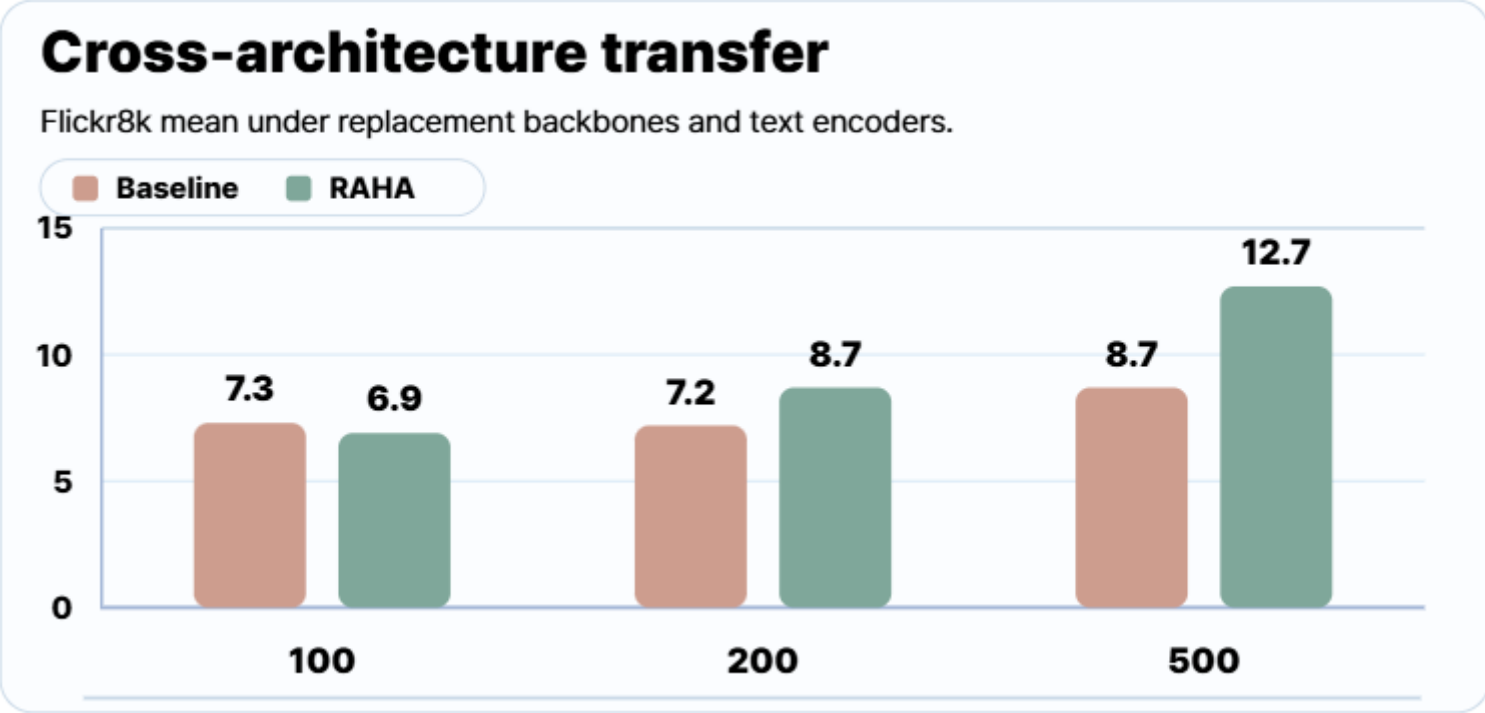
✓ Increasing performance gains with synthetic image-text pair budget



✓ RAHA keeps stronger image-text agreement while reducing high-frequency artifacts

Results: Generalization to Encoder Architecture

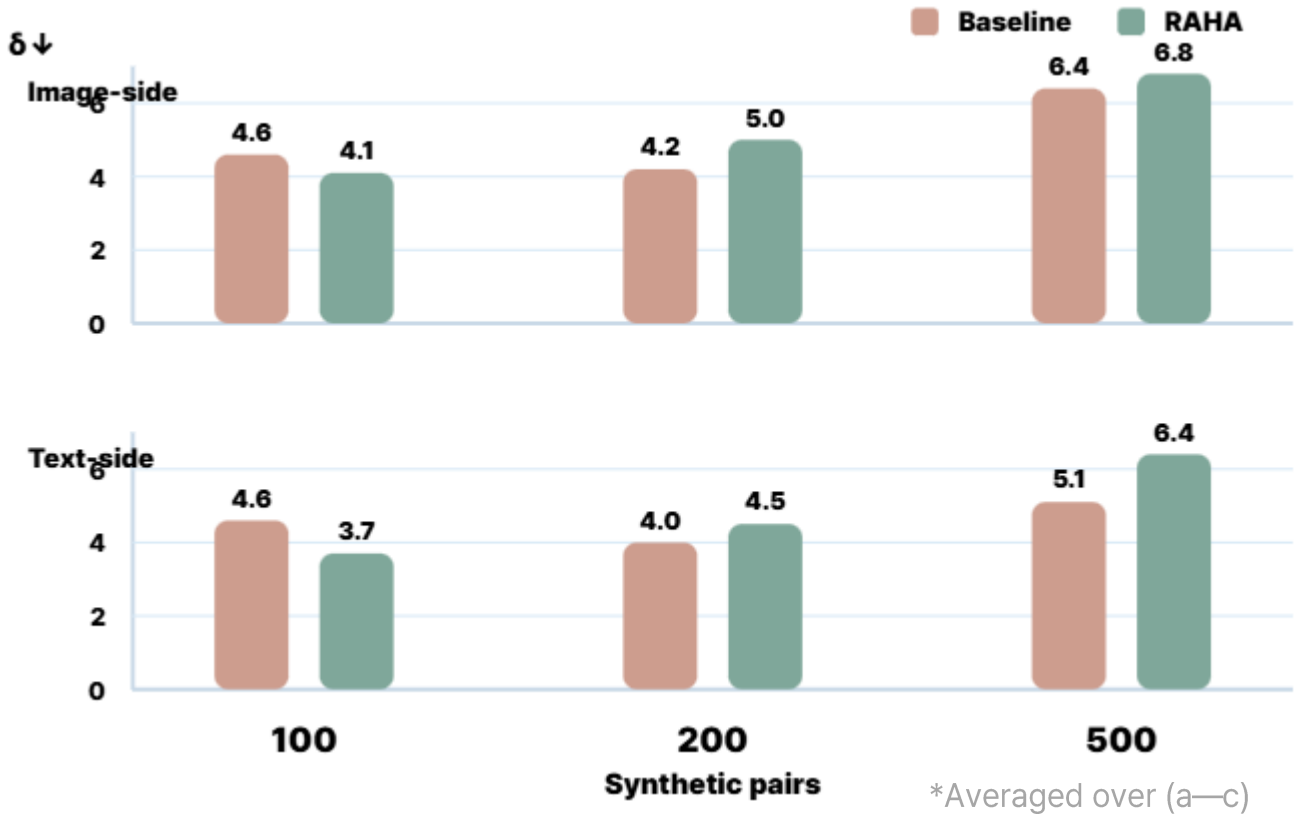
✓ **Relevance distillation survives encoder architecture changes**



*Averaged over {NF-ResNet, NF-RegNet, ViT-B}, {BERT, DistilBERT} w/o source model

Results: Robustness to Noise/Perturbations

✓ Improvements in robustness to perturbations



*Averaged over (a—c)

(a)

JPEG compression on images and bit quantization on text. Included in each modality mean from Table 5.

(b)

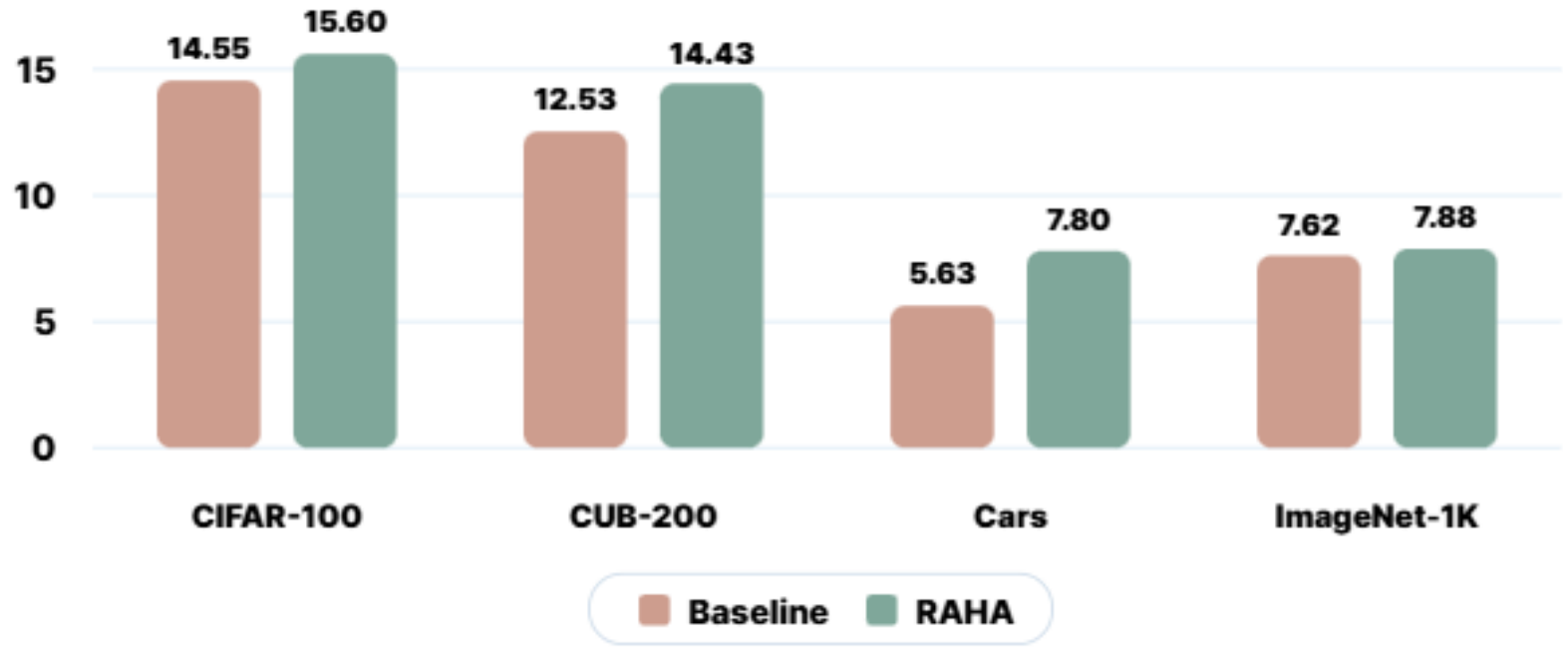
AWGN for both image and text. Included in the image-side and text-side mean δ columns.

(c)

PGD-10 for both image and text. Included in the mean-column robustness summary, with lower δ indicating less degradation.

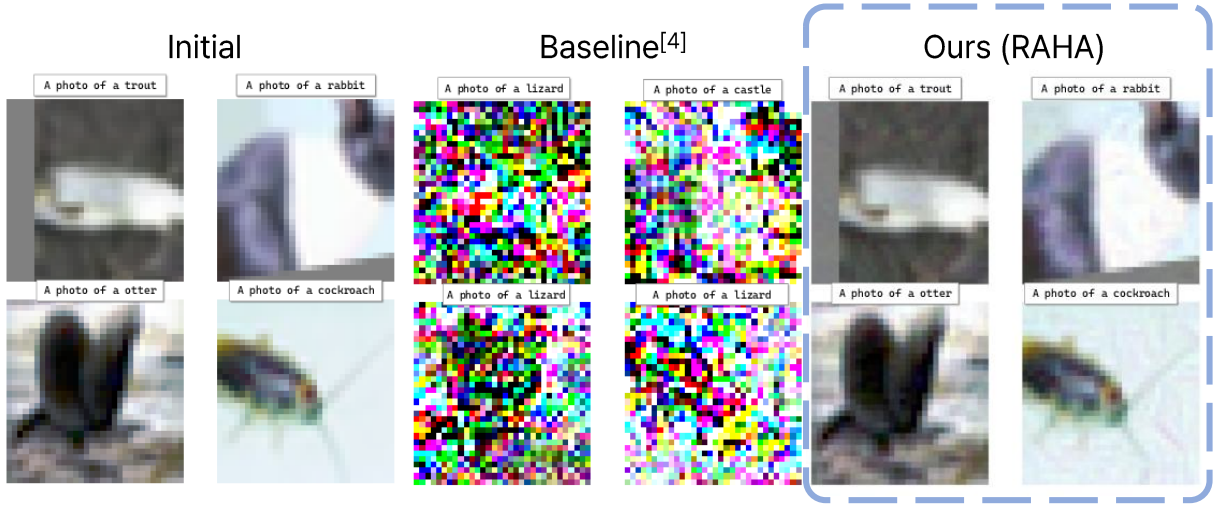
Results: Prompted Image Classification

✓ **RAHA is effective consistently across multiple classification datasets** (CLIP-style prompts, e.g., a photo of [CLS])

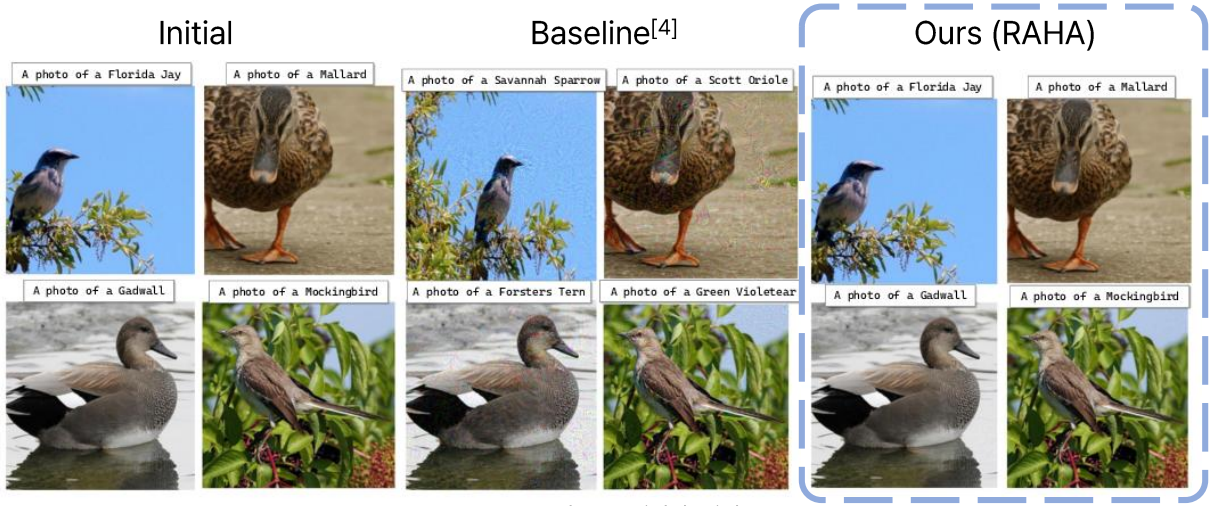


Results: Prompted Image Classification

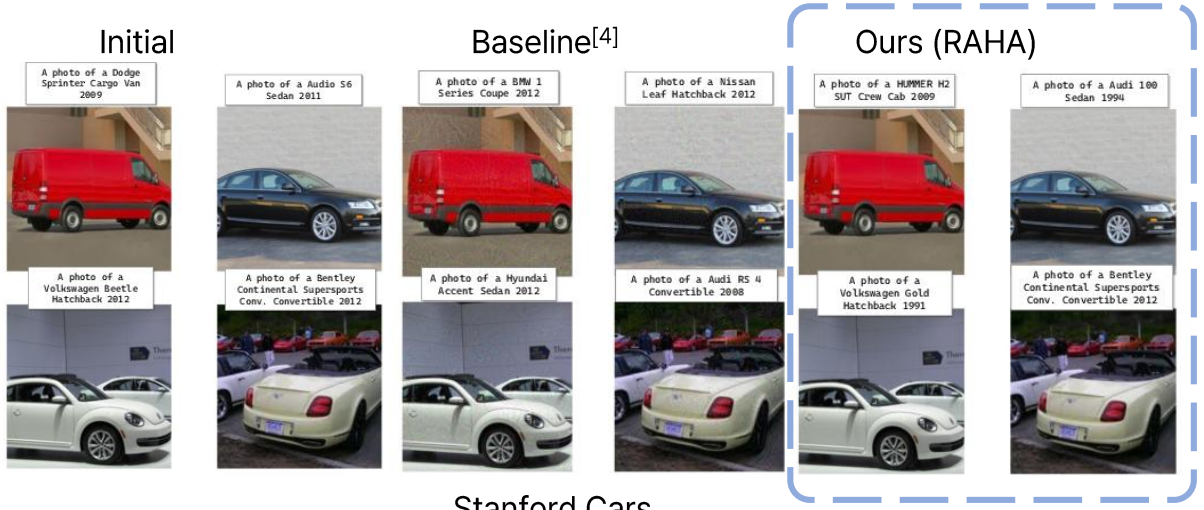
✓ **RAHA is effective consistently across multiple classification datasets** (CLIP-style prompts, e.g., a photo of [CLS])



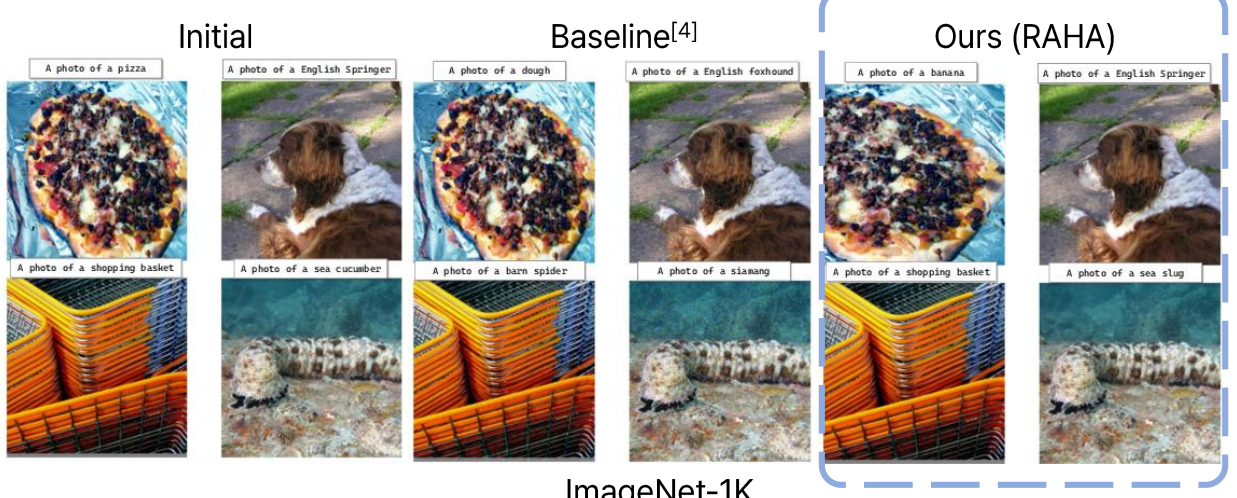
CIFAR-100



CUB-200-2011



Stanford Cars



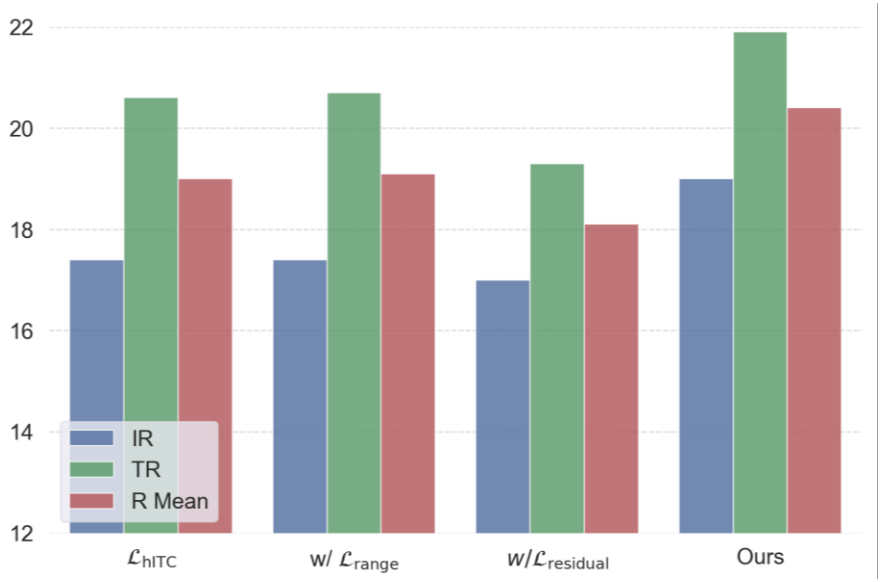
ImageNet-1K

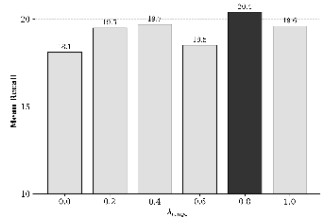
[4] Lee, Yongmin, and Hye Won Chung. "CovMatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder." NeurIPS 2025.

Empirical Findings

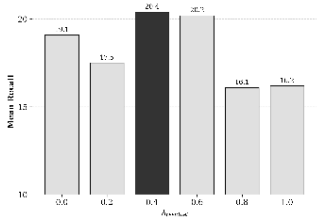
- **Component ablations**
  - We find that range-subspace contributes more strongly than the residual complement

- **Subspace stability analysis**
  - While the dominant rank (top-*k*) is selected adaptively, the rank remains relatively stable across iterations, uniquely for each dataset scale

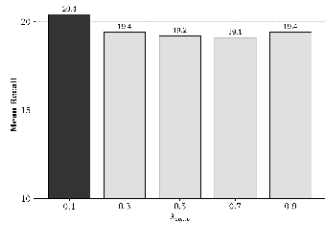




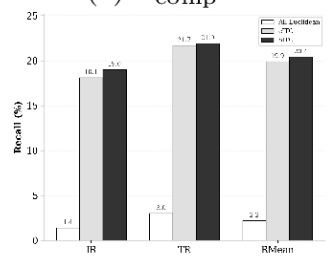
(a) λ_{range}



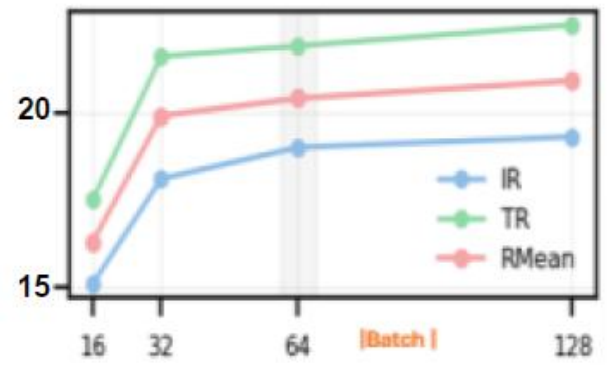
(b) $\lambda_{\text{residual}}$

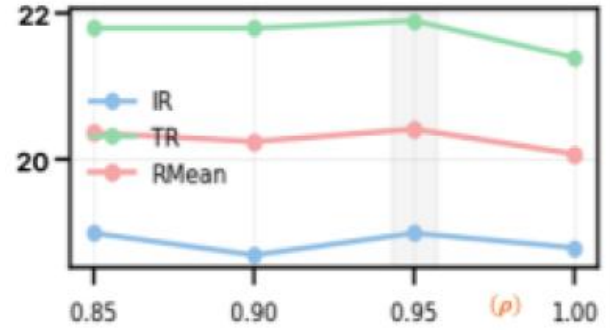


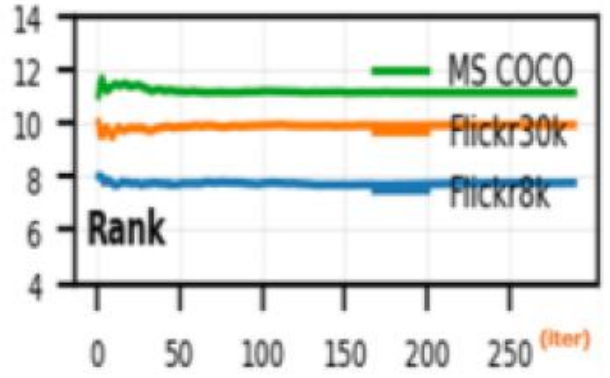
(c) λ_{comp}



(d) Euclidean vs. Hyperbolic space







Key Takeaway

- **Exploiting the right alignment subspace**
  - Image-text correspondence is often low-rank, so distilled datasets should strongly align the dominant shared semantic directions while suppressing over-aligning weak residual directions

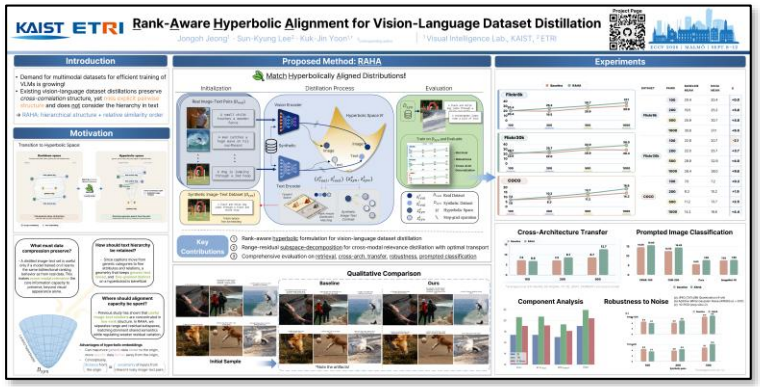
- **Hyperbolic geometry improves structured vision-language distillation**
  - RAHA lifts image and text features into hyperbolic space to better preserve hierarchical semantics, then performs rank-aware range/residual relevance matching in tangent space

- **Performance gains are strongest in transfer, robustness, and larger budgets**
  - RAHA is competitive at very small synthetic budgets and improves more clearly as budget increases, with reported gains in cross-architecture transfer, robustness, and qualitative fidelity of synthesized pairs

KAIST

ECCV 2026 | MALMÖ | SEPT 8–12

Rank-Aware Hyperbolic Alignment for Vision-Language Dataset Distillation



Project Page

Thank you!

Poster Session # (#000):
000, Sep 0, 2026 | 00:00 0M – 00:00 0M | Location