Rank-Aware Hyperbolic Alignment for Vision-Language Dataset Distillation

Jongoh Jeong[1] · Sun-Kyung Lee[2] · Kuk-Jin Yoon[1,†] †Corresponding author

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

KAIST Korea Advanced Institute of Science and Technology · ETRI Electronics and Telecommunications Research Institute

Project Page

ECCV 2026 | MALMÖ | SEPT 8–12

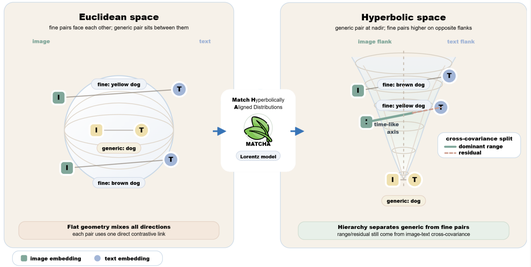
Introduction

- Demand for multimodal datasets for efficient training of VLMs is growing!
- Existing vision-language dataset distillations preserve *cross-correlation* structure, yet miss explicit pairwise structure and does not consider the hierarchy in text

→ RAHA: hierarchical structure + relative similarity order

Motivation

Transition to Hyperbolic Space:



What must data compression preserve?

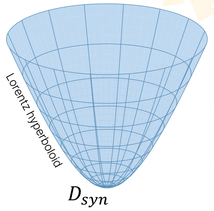
- A distilled image-text set is useful only if a model trained on it learns the same bidirectional ranking behavior as from real data. This makes **cross-modal relevance** the core information capacity to preserve, beyond visual appearance alone.

How should text hierarchy be retained?

- Since captions move from generic categories to fine attributes and relations, a geometry that keeps **generic text broad**, and **fine-grained distinct** on a hyperboloid is beneficial

Where should alignment capacity be spent?

- Previous study has shown that **useful image-text relations** are concentrated in **low-rank** structure. In RAHA, we separates range and residual subspaces, matching dominant shared semantics while regulating weaker residual variation.

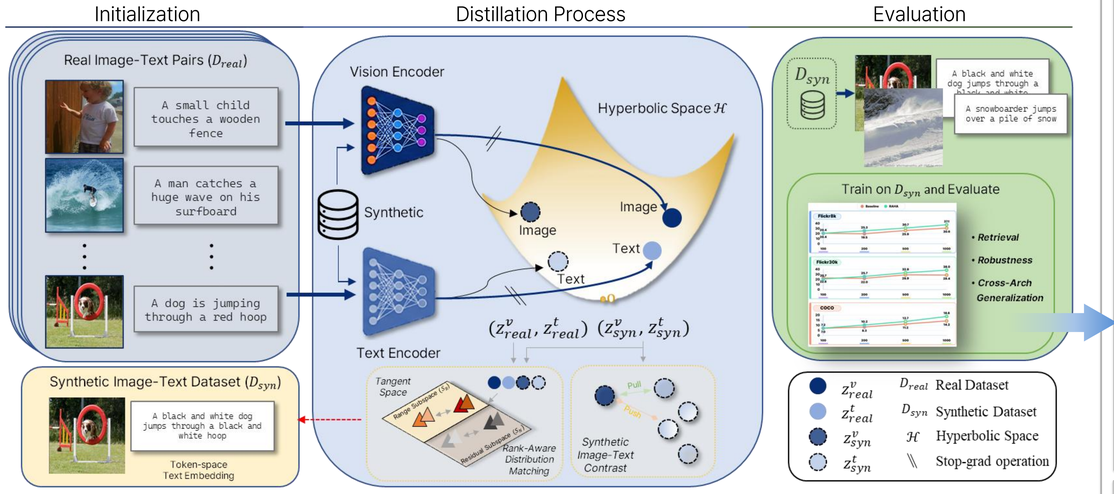


Advantages of hyperbolic embeddings

- Can map more generic data closer to the origin, more specific data farther away from the origin,
- Conceptually, distance from the origin ∝ uncertainty of inputs from inherent noisy image-text pairs

Proposed Method: RAHA

Match Hyperbolically Aligned Distributions!

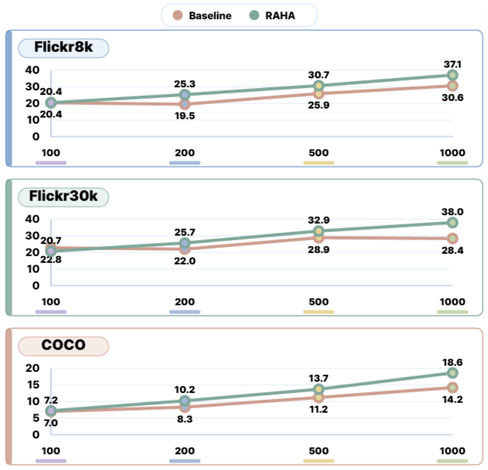


Key Contributions

1. Rank-aware hyperbolic formulation for vision-language dataset distillation
2. Range-residual subspace-decomposition for cross-modal relevance distillation with optimal transport
3. Comprehensive evaluation on retrieval, cross-arch. transfer, robustness, prompted classification

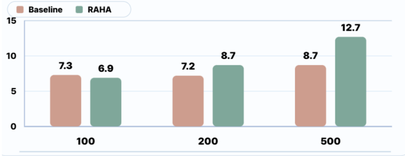
Qualitative Comparison

Experiments



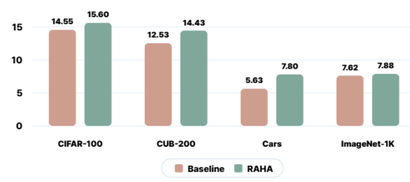
DATASET	PAIRS	BASELINE MEAN	RAHA MEAN	Δ
Flickr8k	100	20.4	20.4	+0.0
	200	19.5	25.3	+5.8
	500	25.9	30.7	+4.8
	1000	30.6	37.1	+6.5
Flickr30k	100	22.8	20.7	-2.1
	200	22.0	25.7	+3.7
	500	28.9	32.9	+4.0
	1000	28.4	38.0	+9.6
COCO	100	7.0	7.2	+0.2
	200	8.3	10.2	+1.9
	500	11.2	13.7	+2.5
	1000	14.2	18.6	+4.4

Cross-Architecture Transfer

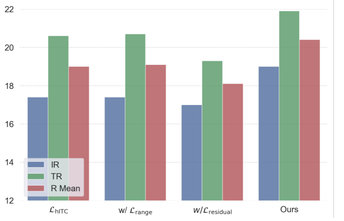


*Averaged over {NF-ResNet, NF-RegNet, ViT-B}, {BERT, DistilBERT} w/o source model

Prompted Image Classification

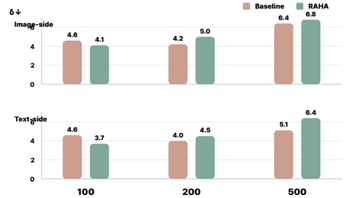


Component Analysis



Robustness to Noise

(a) JPEG (75%)/Bit Quantization (4-bit)
(b) Additive White Gaussian Noise (AWGN) (σ = 0.01)
(c) 10-PGD (step size=2)



*Averaged over (a—c)