

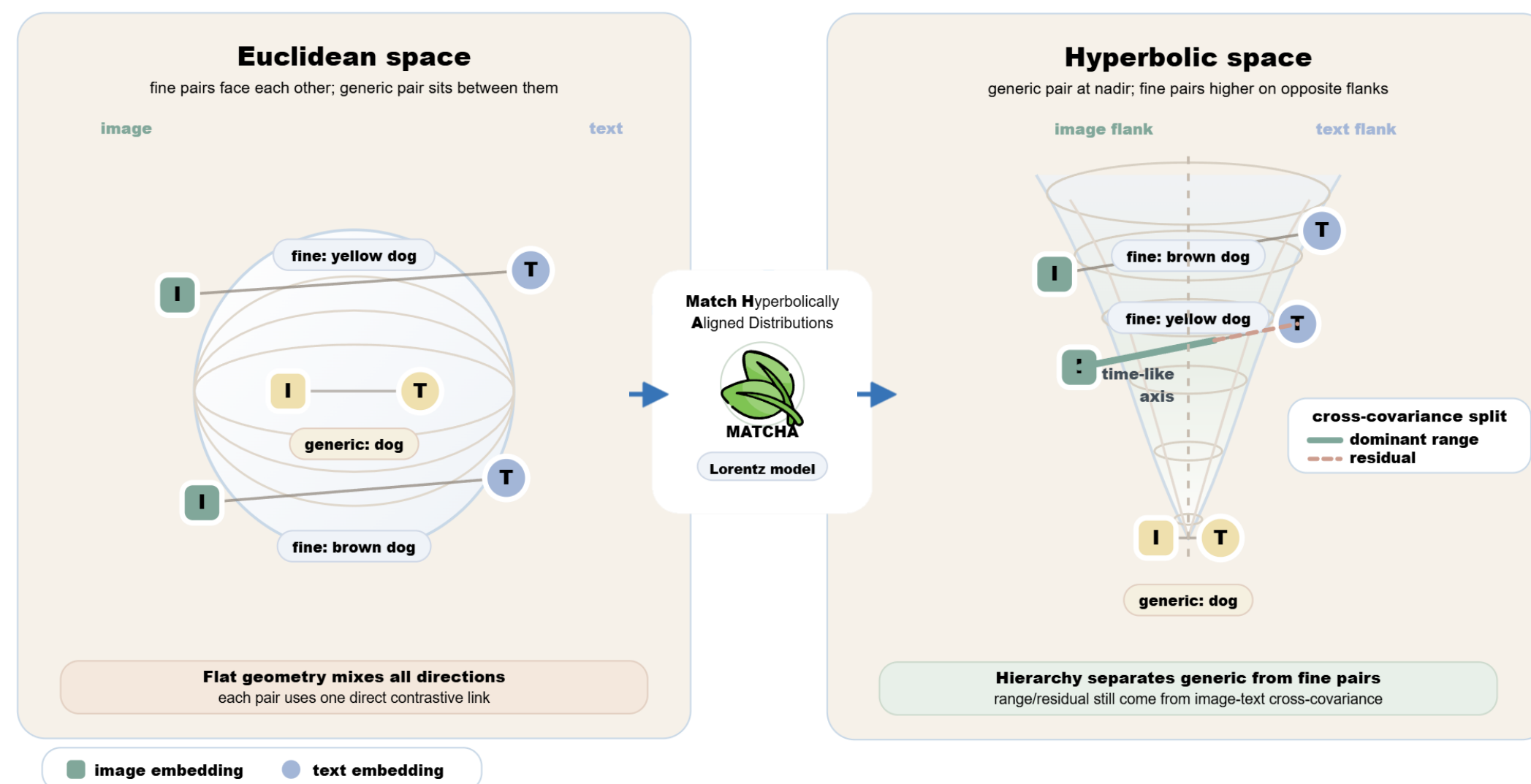


Introduction

- 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?

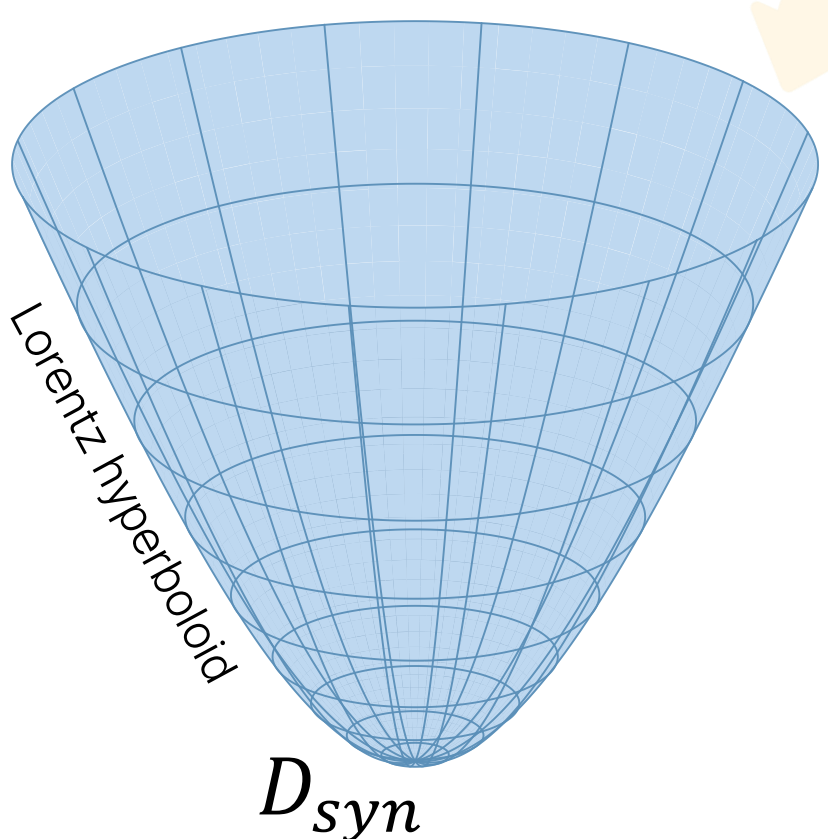
- Since captions move from generic categories to fine attributes and relations, a geometry that keeps **generic text broad**, and **fine-grained distinct** on a hyperboloid is beneficial

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.

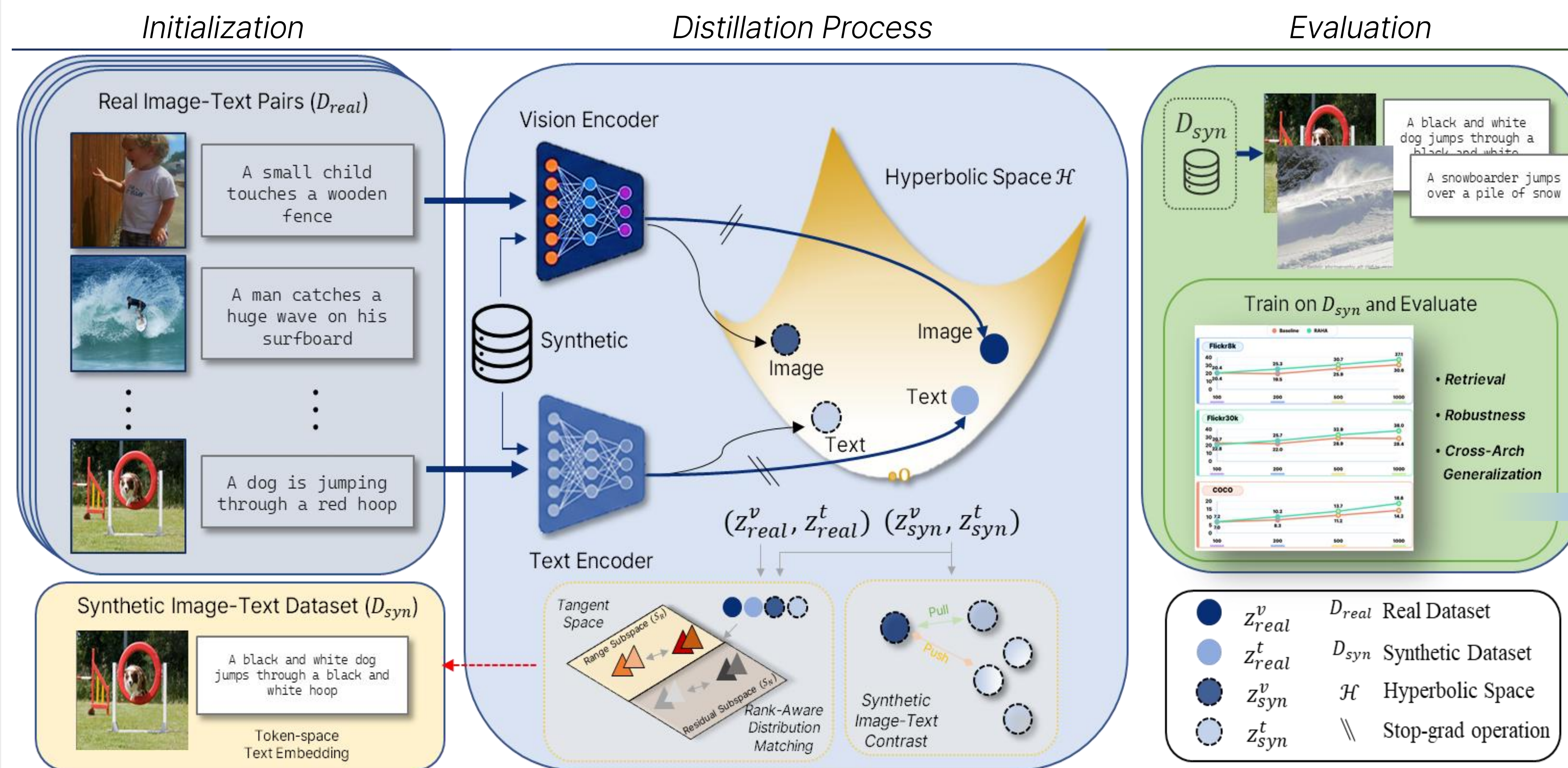
Advantages of hyperbolic embeddings

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



Proposed Method: RAHA

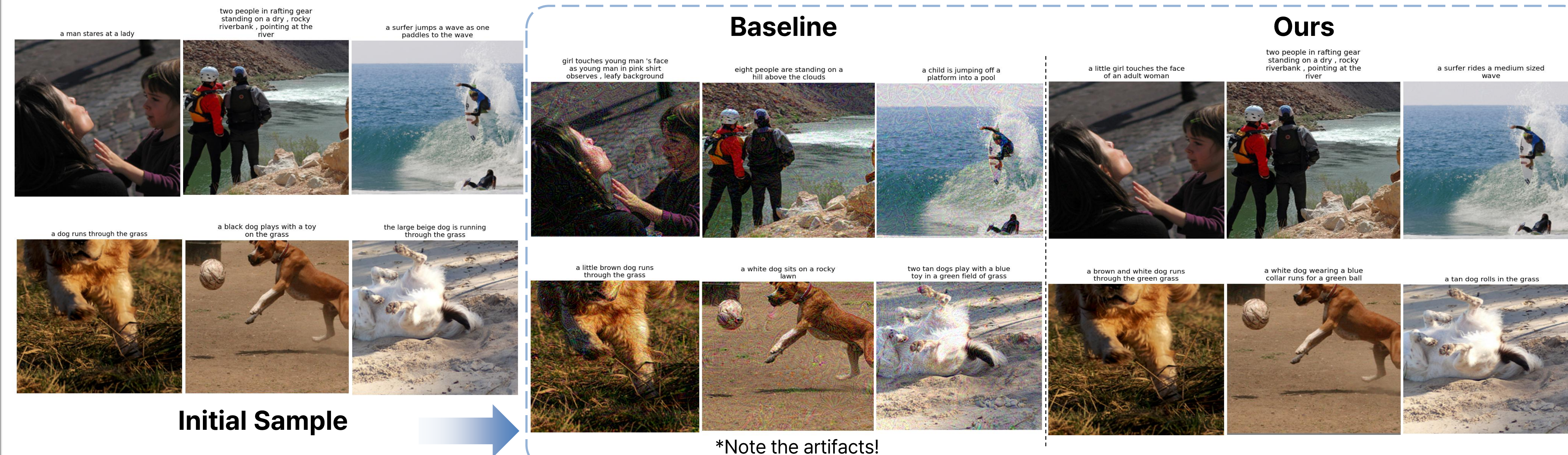
Match Hyperbolically Aligned Distributions!



Key Contributions

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

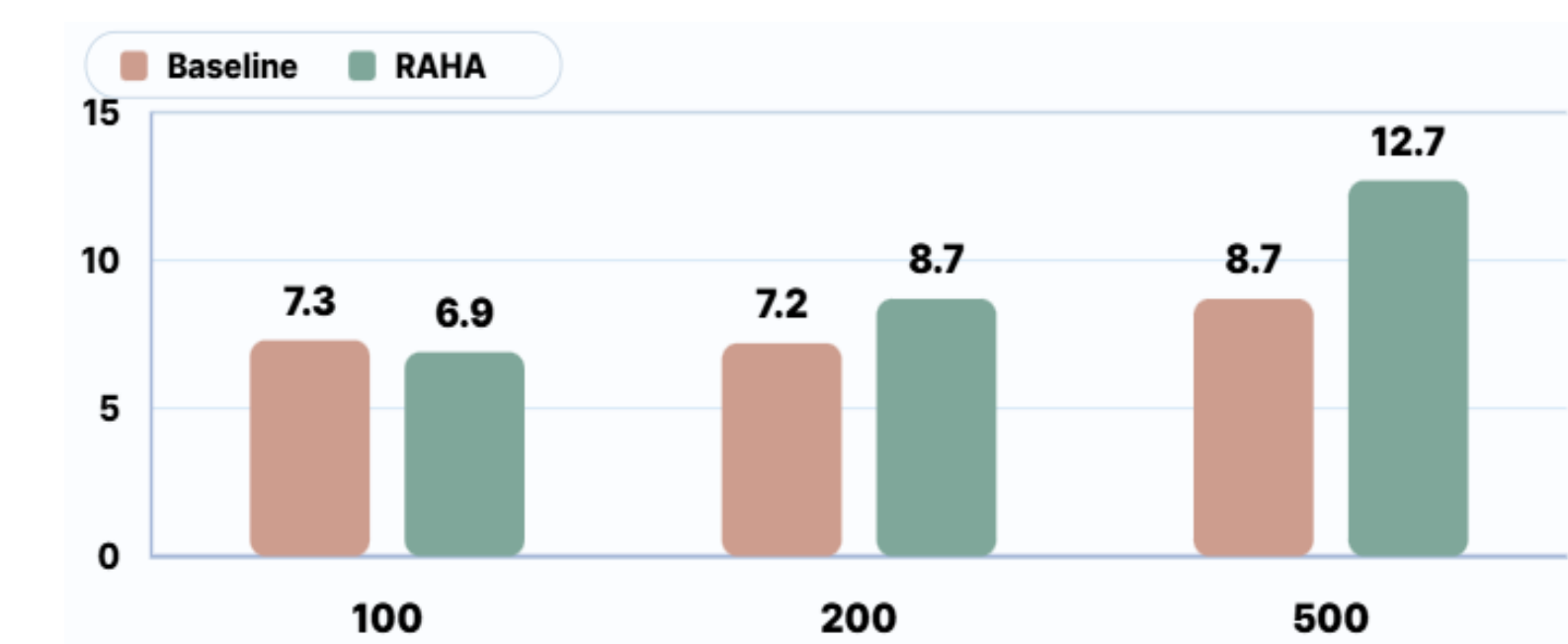
Qualitative Comparison



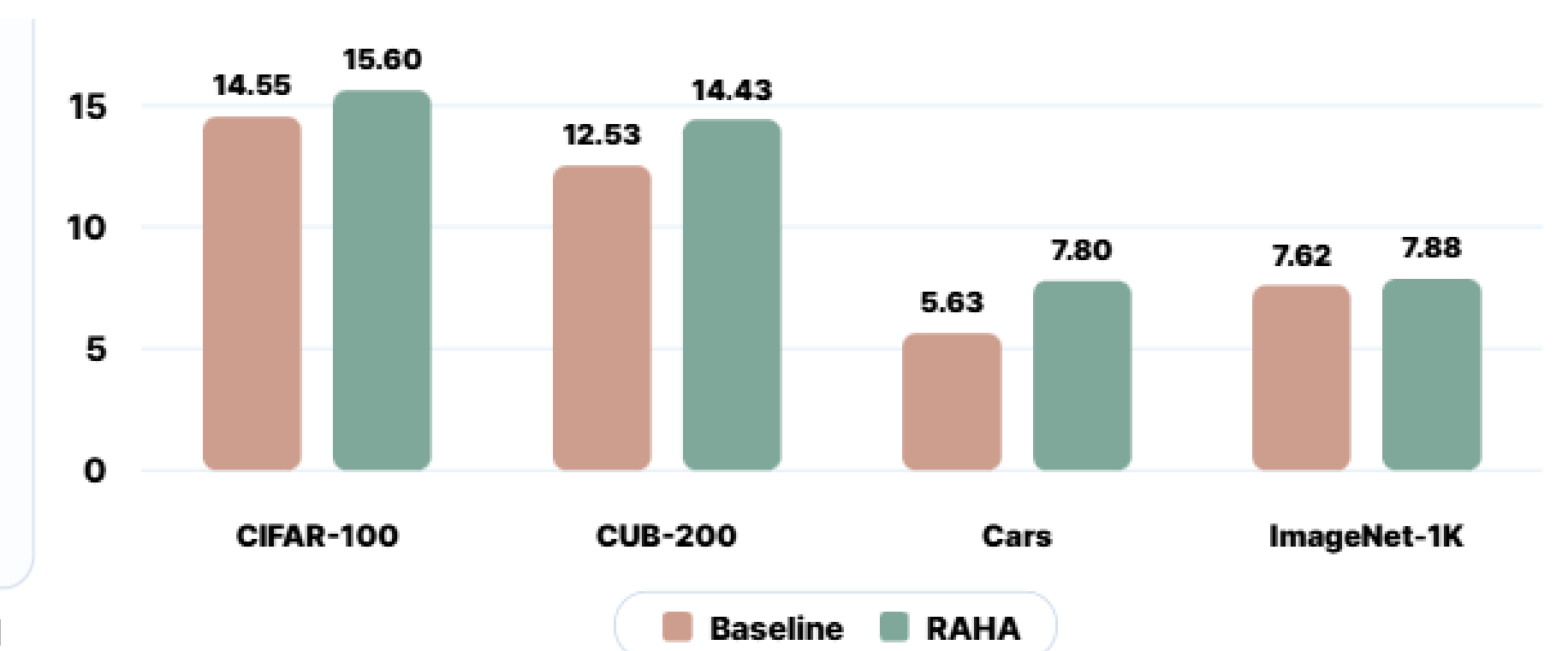
Experiments



Cross-Architecture Transfer

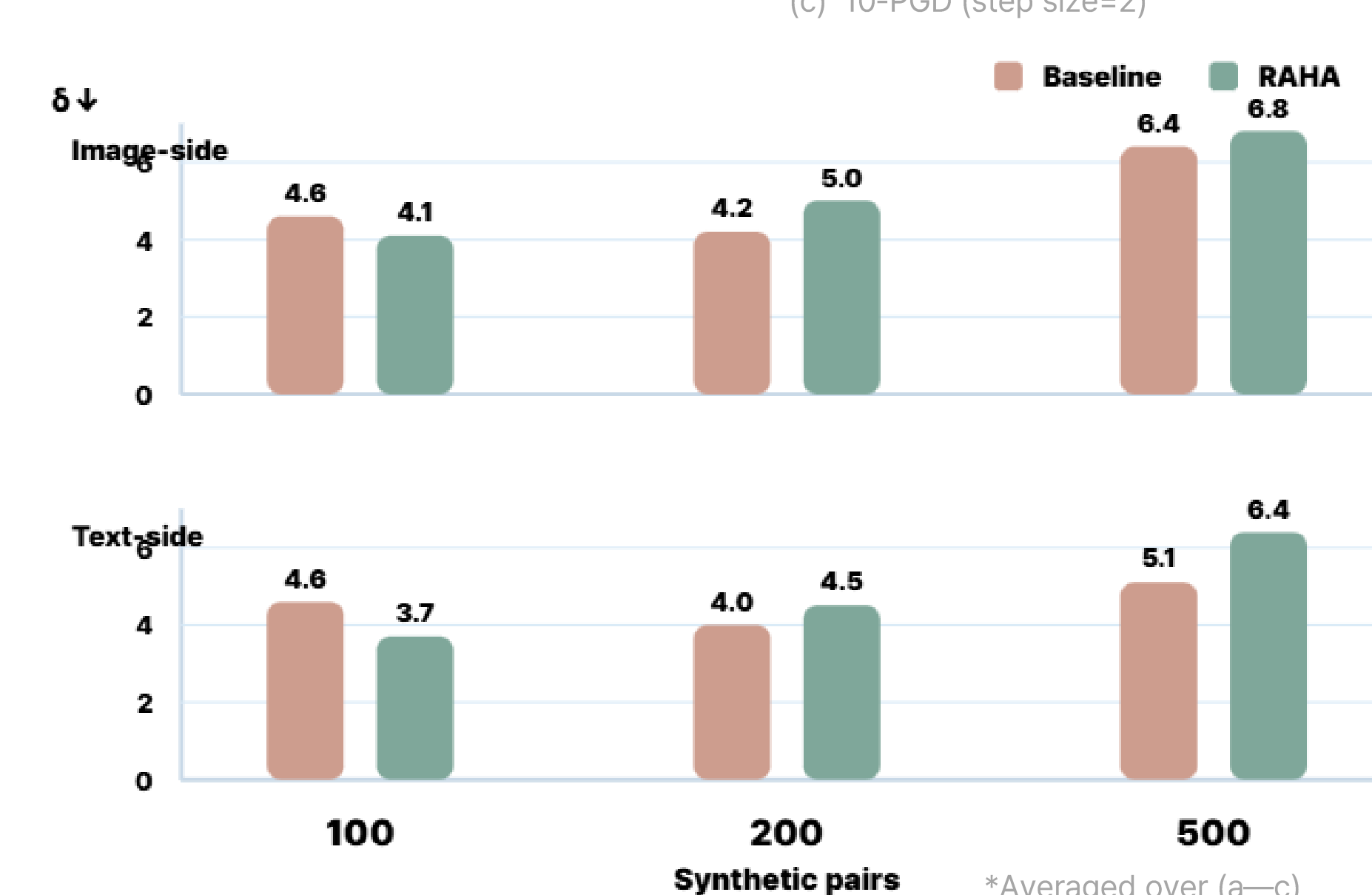


Prompted Image Classification



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

Robustness to Noise



Component Analysis

