Beyond Accuracy: Metrics that Uncover What Makes a 'Good' Visual Descriptor

Ethan Lin[1*], Linxi Zhao[1], Atharva Sehgal[2], Jennifer J. Sun[1],
[1]Cornell University, [2]University of Texas at Austin

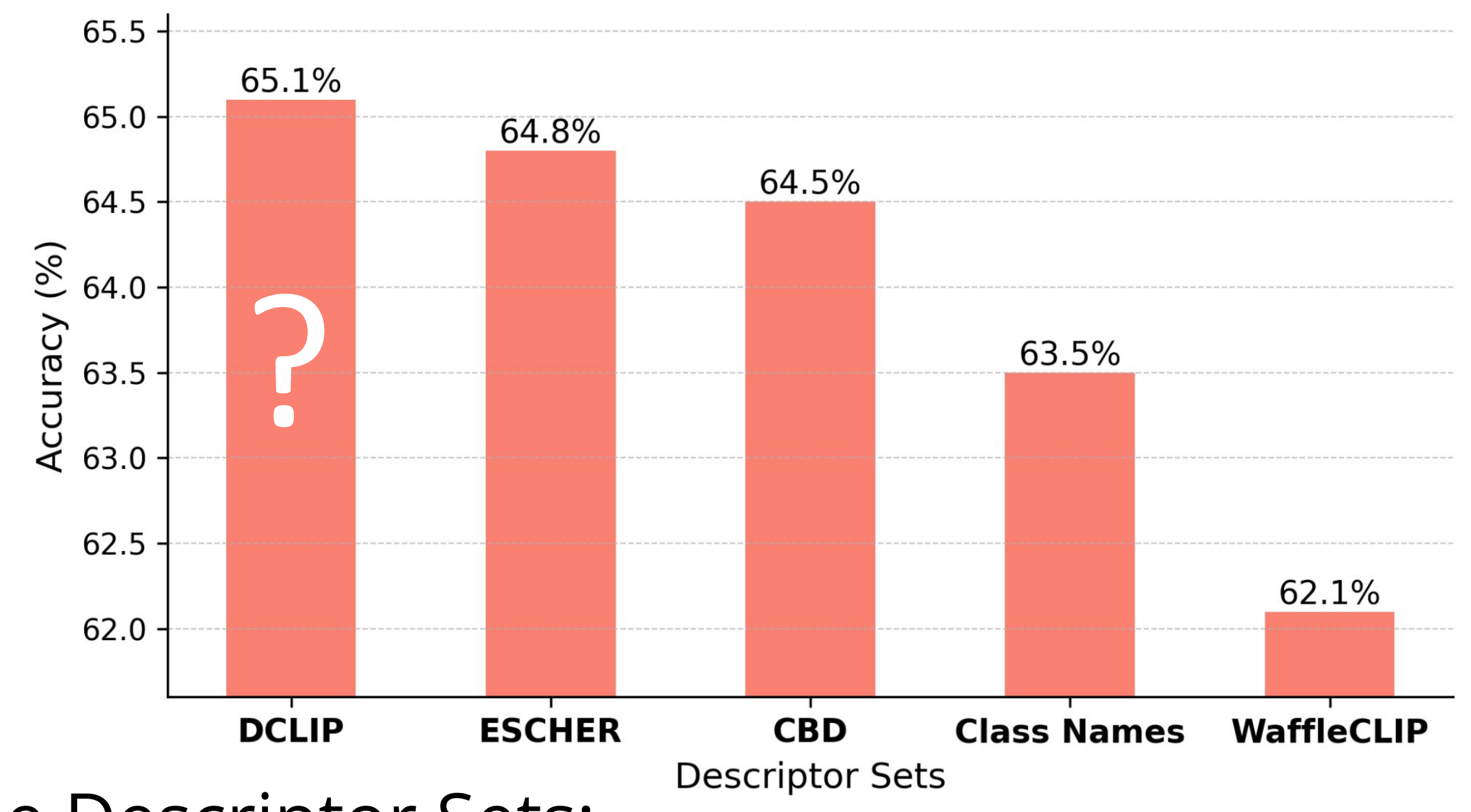
CVPR Nashville JUNE 11-15, 2025

High accuracy doesn't always mean high-quality descriptors. We introduce **Global Alignment** and **CLIP Similarity**—two alignment-based metrics that evaluate the relationship between text-based visual descriptors and the underlying vision-language models beyond accuracy.

MOTIVATION

Text-based visual descriptors are widely used with vision-language models for concept discovery and classification. While often evaluated by classification accuracy, this metric alone offers limited insight into descriptor quality or interpretability.

Accuracy for different descriptor sets on the CUB Dataset



Example Descriptor Sets:
CBD: Zero-shot LLM-generated descriptors.
ESCHER: Iteratively refined CBD descriptors.
DCLIP & WaffleCLIP: [class names] + [random attack string]

Example

This is a **Great-Crested Flycatcher.**

DCLIP prediction: **Great-Crested Flycatcher** ✓

CBD prediction: **Acadian Flycatcher** ✗

DCLIP descriptors are qualitatively worse than CBD but have better accuracy.

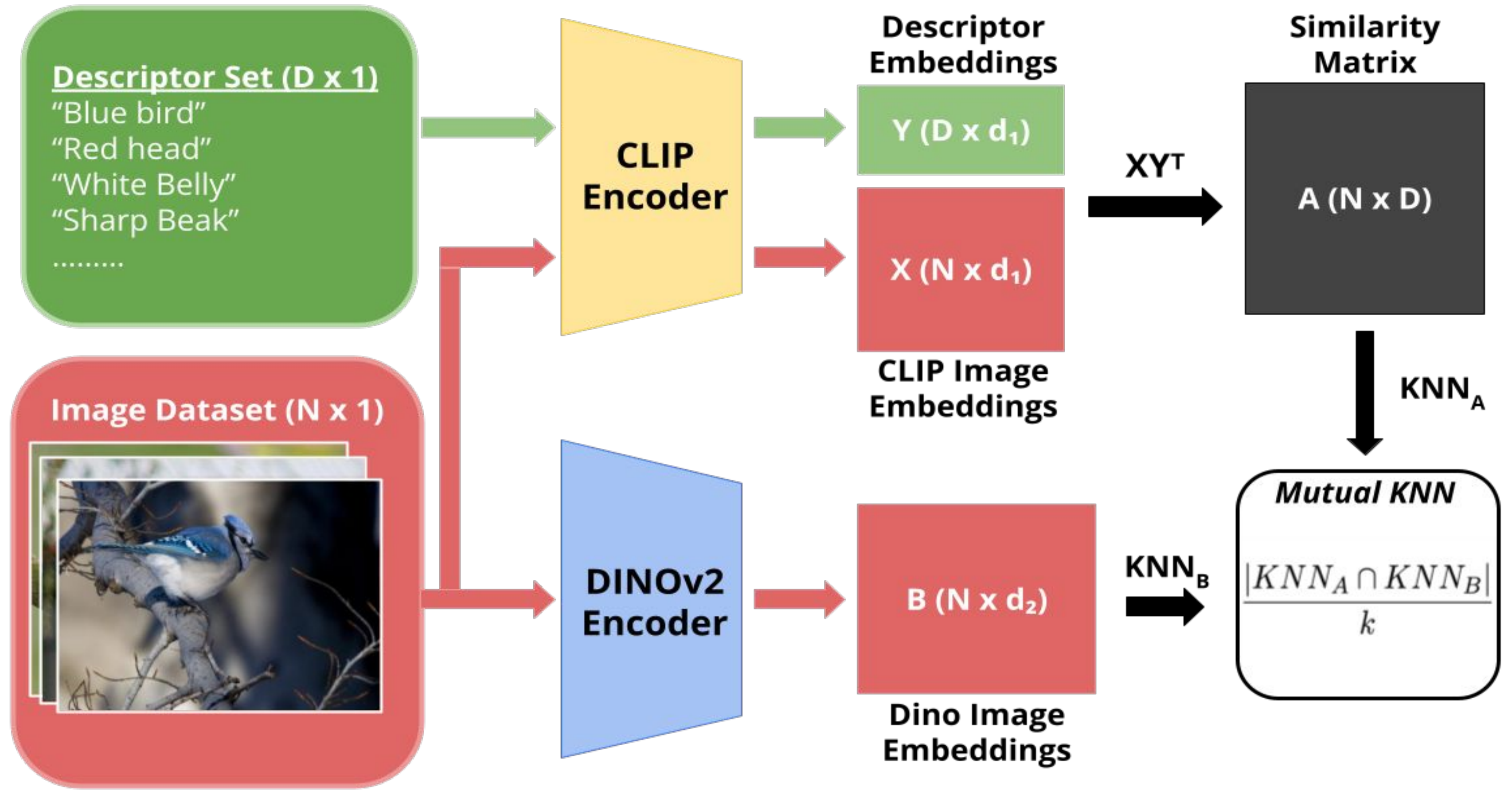
DCLIP Descriptors for **Great-Crested Flycatcher**
- [Class], which is **a blackbird with red and yellow wing markings.**
- [Class], which is **a black cap and nape, with a white face and throat.**
- [Class], which is a **short forked tail.**
- [Class], which is a **yellow throat**
- [Class], which is a **long, thin neck.**

CBD Descriptors for **Great-Crested Flycatcher**
- [Class], which is a **small to medium-sized songbird.**
- [Class], which is a **white breast with black streaks.**
- [Class], which has **grey, brown, or black.**
- [Class], which has **yellow eyes.**
- [Class], which is a **black bill.**

Global Alignment

Research Question: Can we evaluate the quality of the representation space induced by a visual descriptor set?

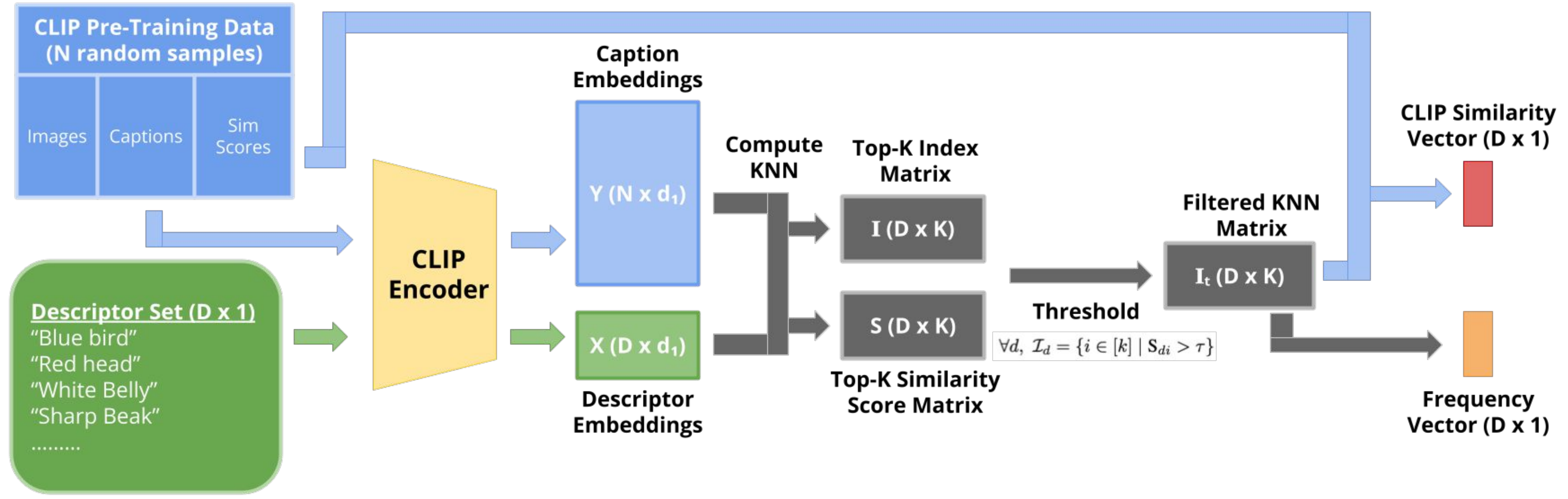
Key Idea: Use representational alignment metrics to compare the visual descriptor space to a reference image embedding space.



CLIP Similarity Metric

Research Question: Can we evaluate how well a descriptor aligns with CLIP's pre-training data?

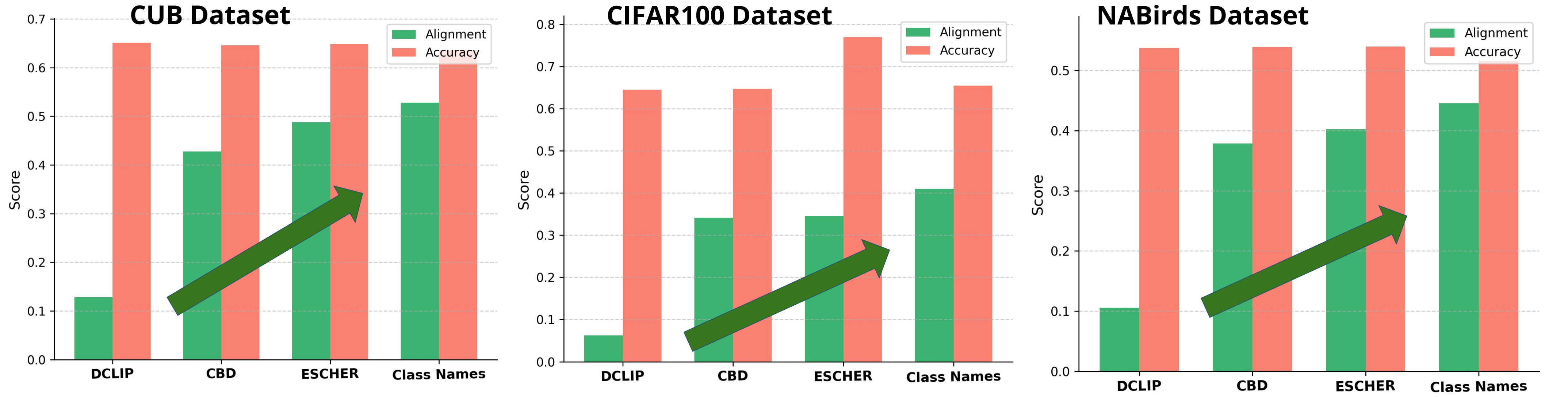
Key Idea: Find CLIP pre-training captions that match a descriptor using KNN in CLIP embedding space and a threshold hyperparameter τ.



Two Metrics

Frequency: number of captions above the similarity threshold τ.
CLIP Similarity: average similarity between each matched image-caption pair.

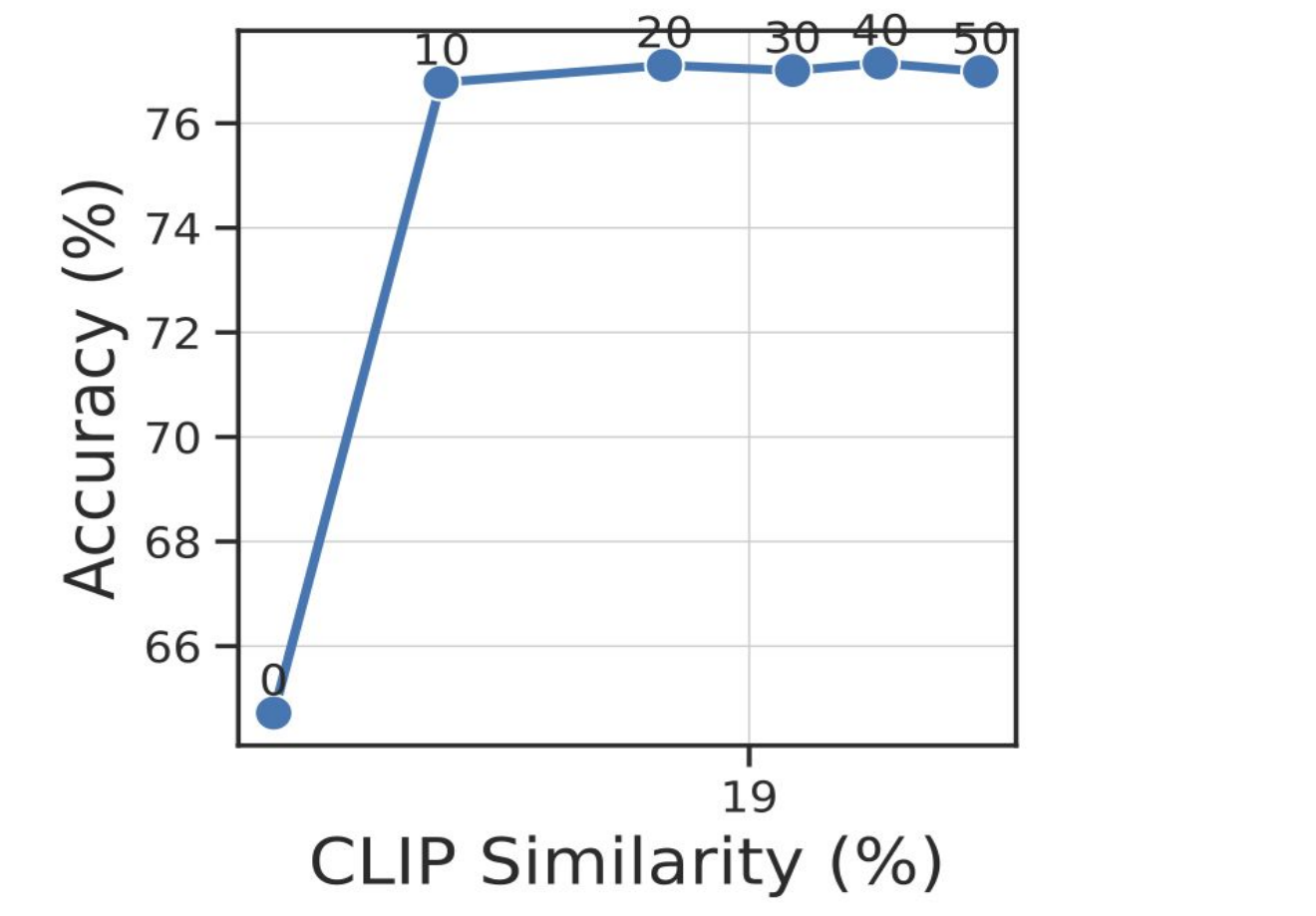
Global Alignment Results



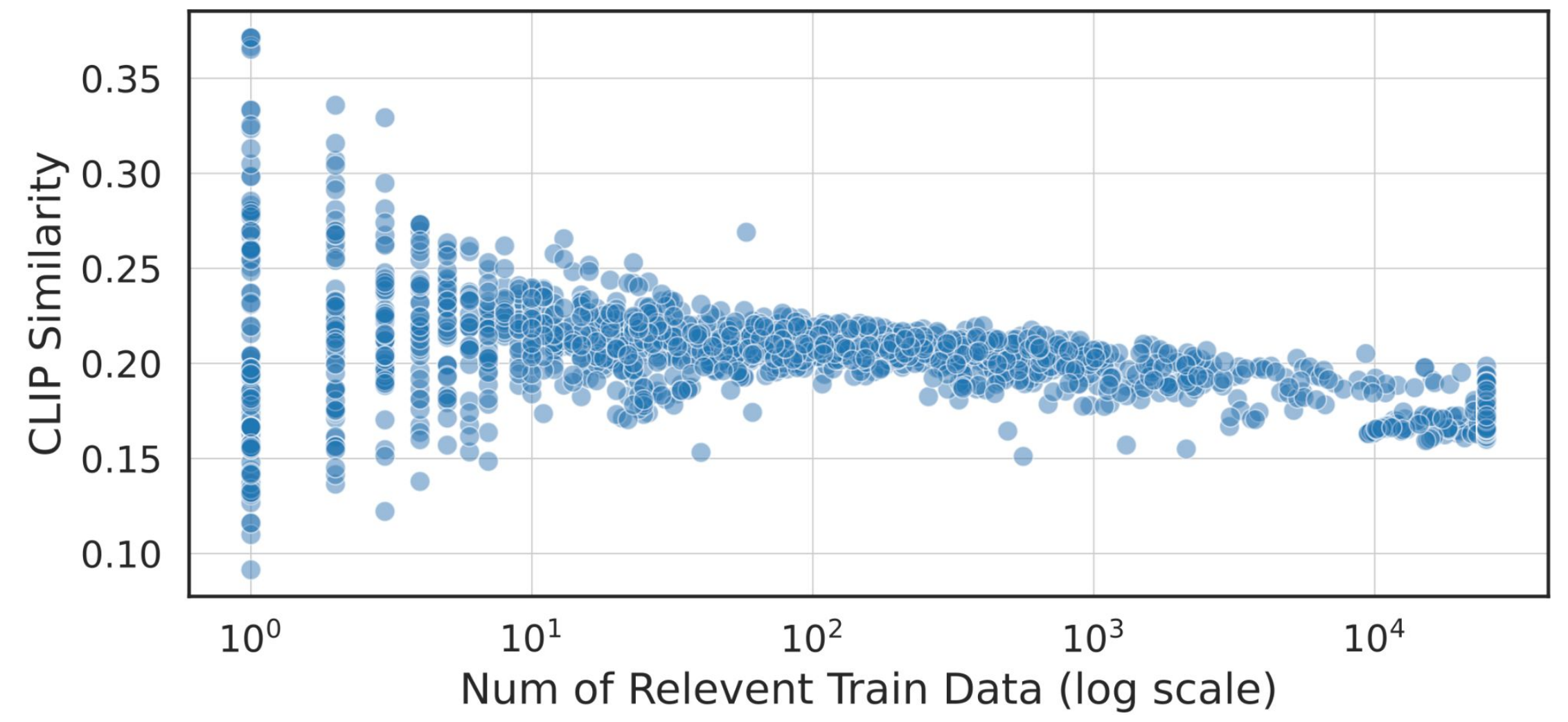
1) Different descriptor generation methods produce different quality representation spaces.
2) Global Alignment can captures trends in descriptor quality that accuracy cannot!

CLIP Similarity Results

CLIP Sim Across ESCHER Iterations on CIFAR100



CLIP Sim vs. Frequency on NABirds Descriptors



KEY TAKEAWAYS

1) As ESCHER iteratively refines descriptors, they become better aligned with CLIP's pre-training data.
2) CLIP tends to favor less frequent, more visually grounded descriptors over generic high-frequency ones.