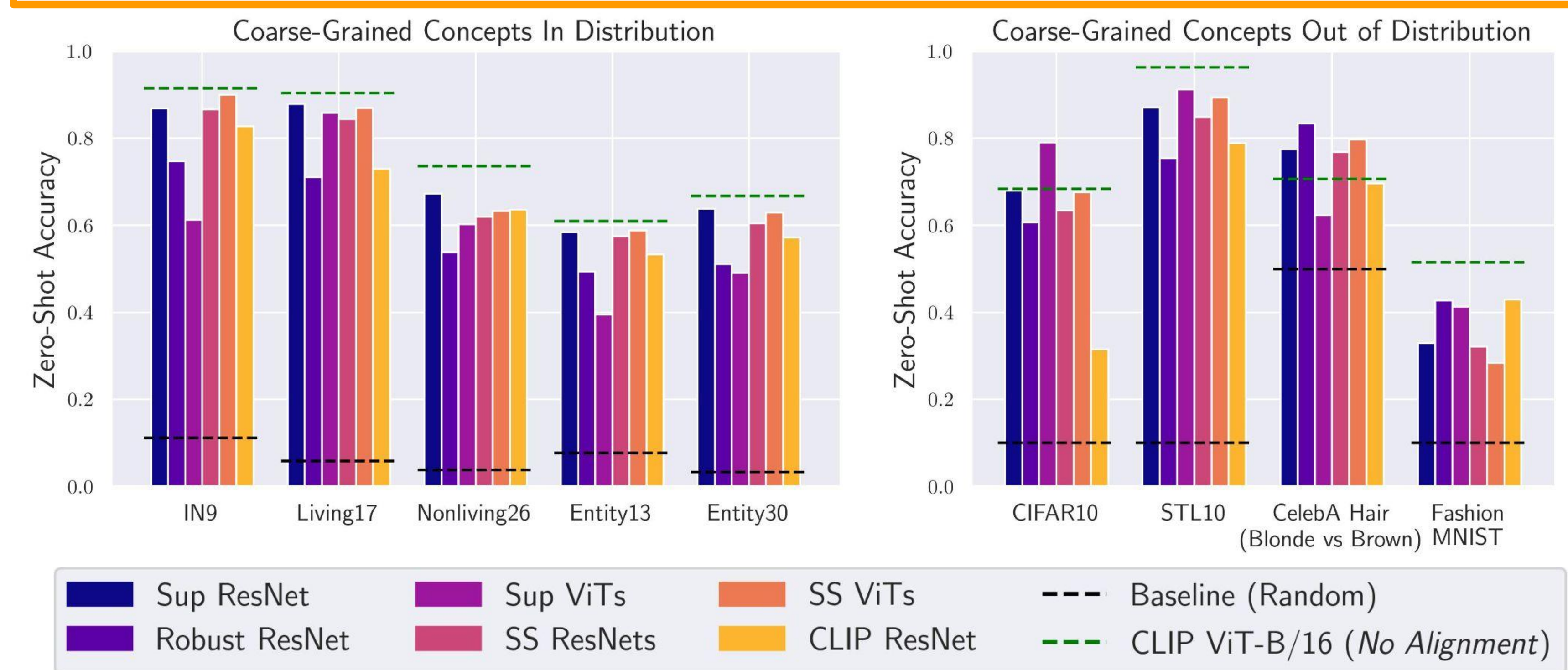
Text-to-Concept (and back) via Cross-Model Alignment

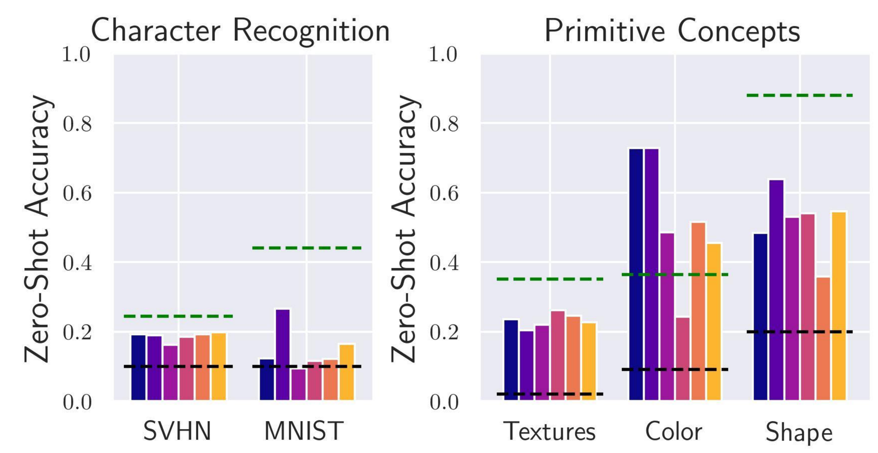
Concept Activation Vectors ✨*Directly from Text*✨

Mazda Moayeri*, Keivan Rezaei*, Maziar Sanjabi, Soheil Feizi

JUNE 18-22, 2023 CVPR XAI4CV Workshop VANCOUVER, CANADA
ICML International Conference On Machine Learning 40 Years

Application 1: **Converting off-the-shelf Vision Encoders into Zero-shot Classifiers**

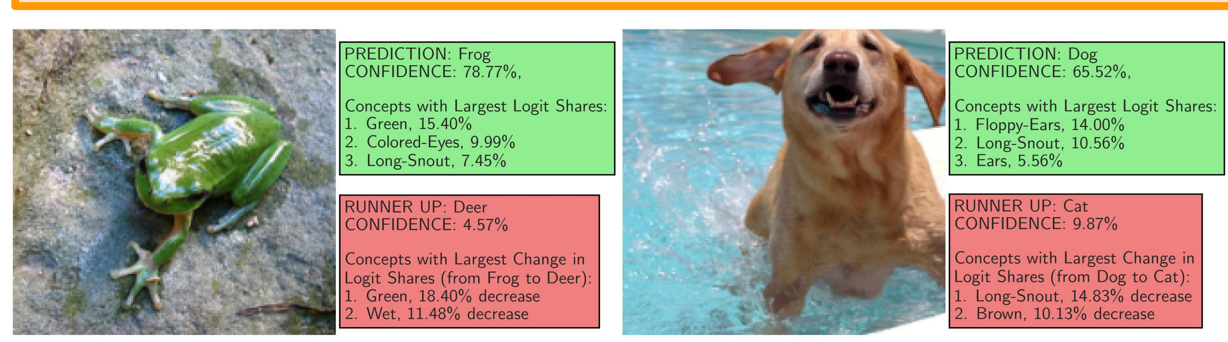




Competitive, and *at times* better, zero-shot accuracy than CLIP, for various [much simpler] encoders.

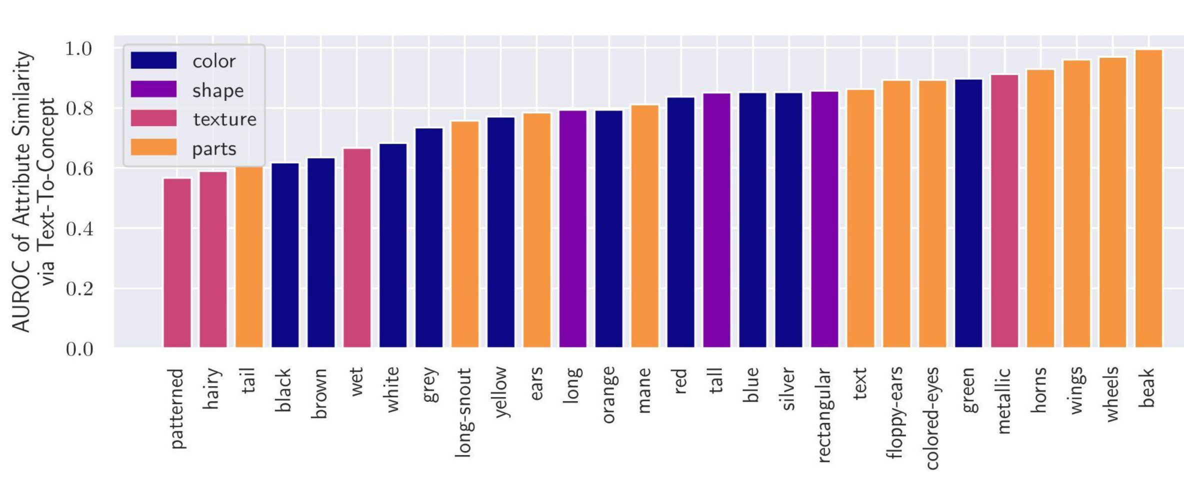
Notable gains over CLIP occur for tasks that involve color recognition.

Application 2: **Concept-Bottleneck Models** *with no concept supervision*



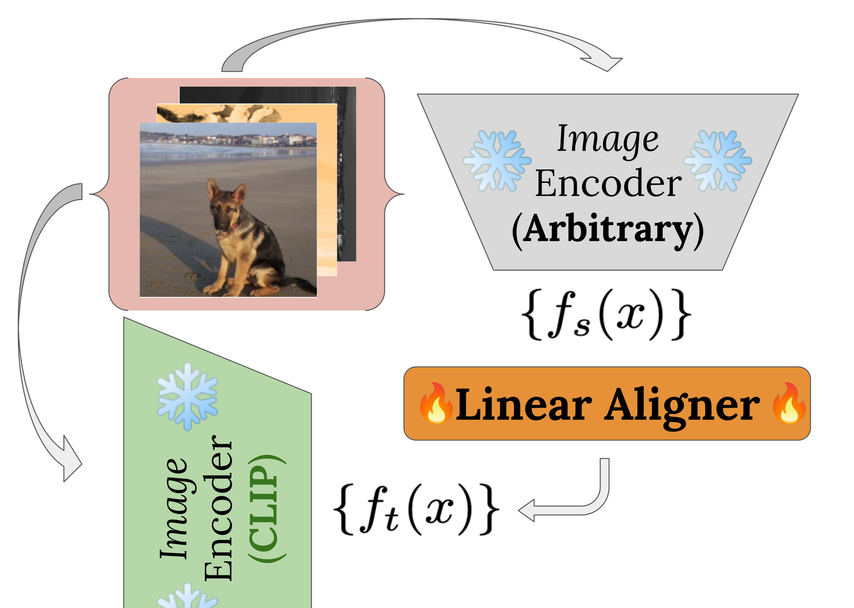
CBMs are white box models w.r.t. concept predictions, but they **require costly extra concept supervision.**

Text-to-Concept yields concept vectors for free → via training only two linear layers (aligner + final head), you can easily convert an encoder into a CBM.



Shown: interpretable inferences using RN50 fit on RIVAL10 data. (bottom) RIVAL10 Attributes are predicted reasonably well by similarity of aligned features to Text-to-Concept vectors.

😲 Key Insight: **A linear layer can map across representation spaces of diverse vision models** 😲



Cross-Model *Linear* Alignment

For the source and target image encoders,

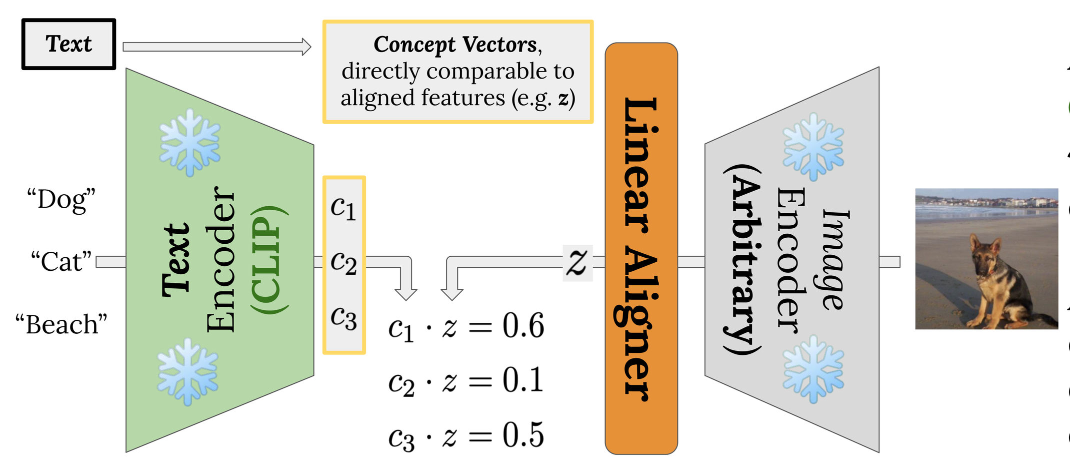
1. Take a **dataset of unlabeled images**
2. Cache their representations
3. Solve regression [***very efficient***] to obtain **linear aligner** parameters.

$$\min_{W,b} \sum_{x \in \mathcal{D}} \|W^T f_s(x) + b - f_t(x)\|^2$$

Despite vastly different training procedures, using a linear layer to predict representations of one encoder given those of another is surprisingly effective.

We achieve high R^2 for this regression task, and **can even use the classification head for one model on aligned features of another with minimal drop in accuracy.**

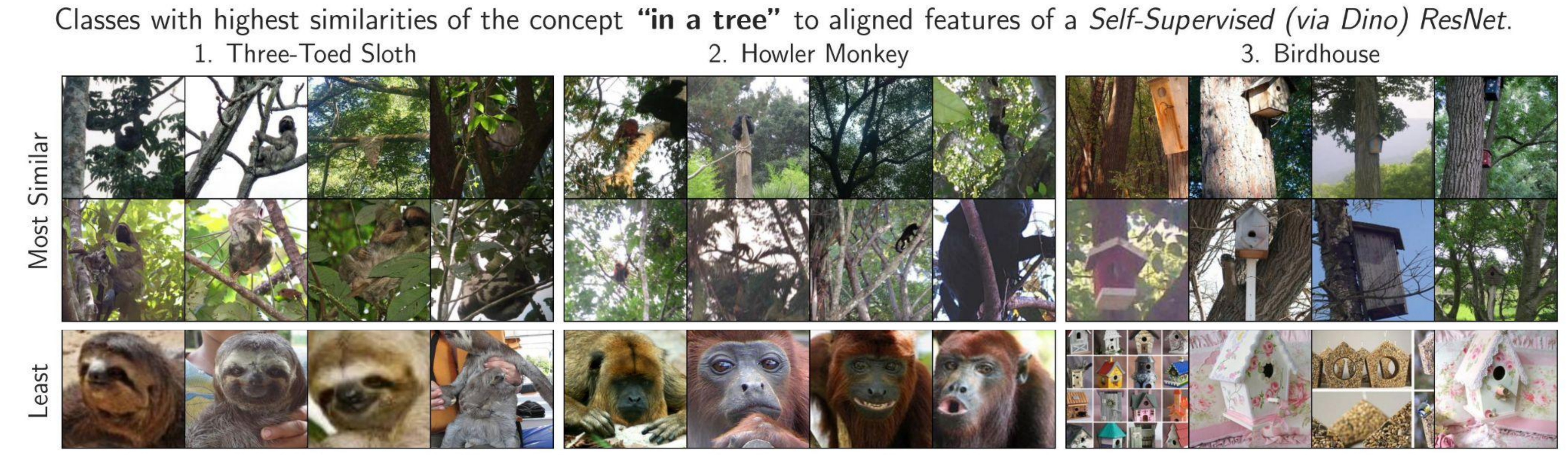
Implication: Aligning to a CLIP vision encoder → 🌟 ***Multimodal access for your Unimodal Model*** 🌟



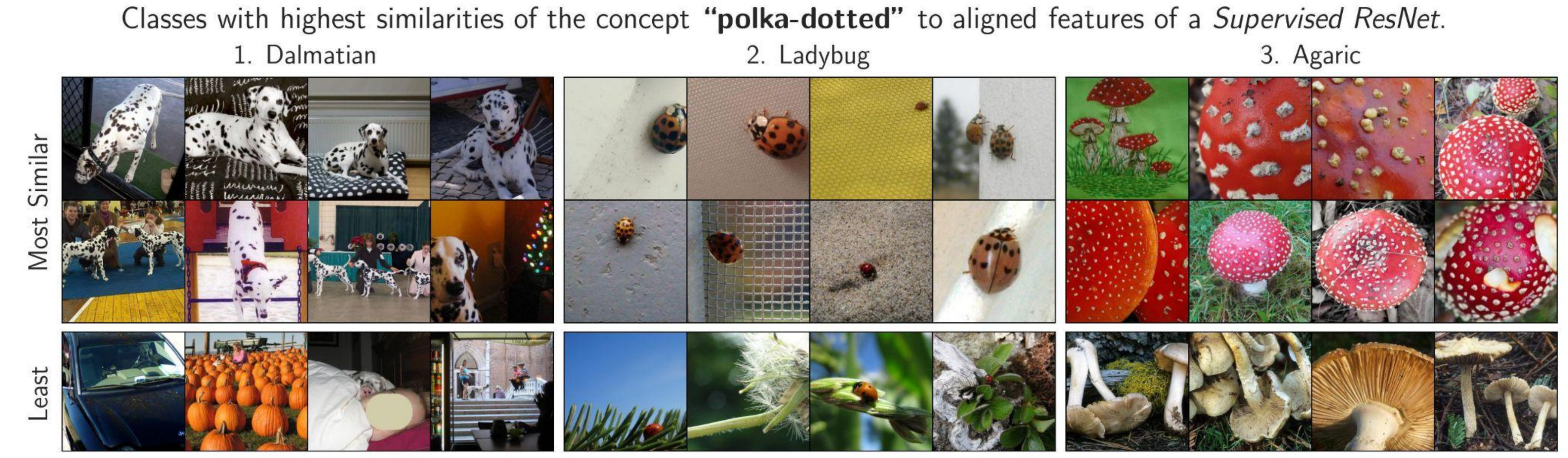
After training the aligner once, you can **use CLIP's text encoder to get *concept vectors in constant time*** w.r.t. typical cost of collecting exemplar data for every concept.

An aligned representation from the source encoder compares directly to **CLIP** text embeddings, yielding high similarity when a concept is present. See examples below.

Classes with highest similarities of the concept **"in a tree"** to aligned features of a *Self-Supervised (via Dino) ResNet.*



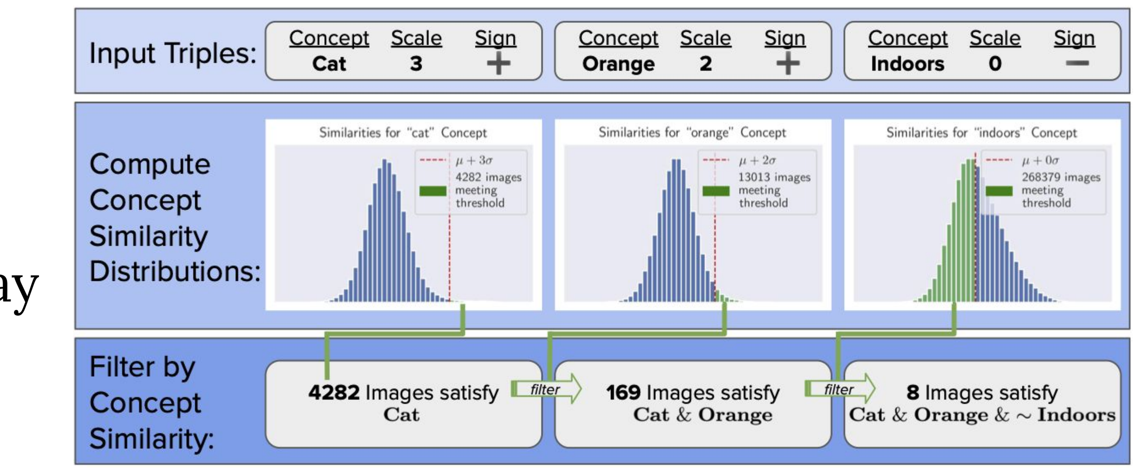
Classes with highest similarities of the concept **"polka-dotted"** to aligned features of a *Supervised ResNet.*

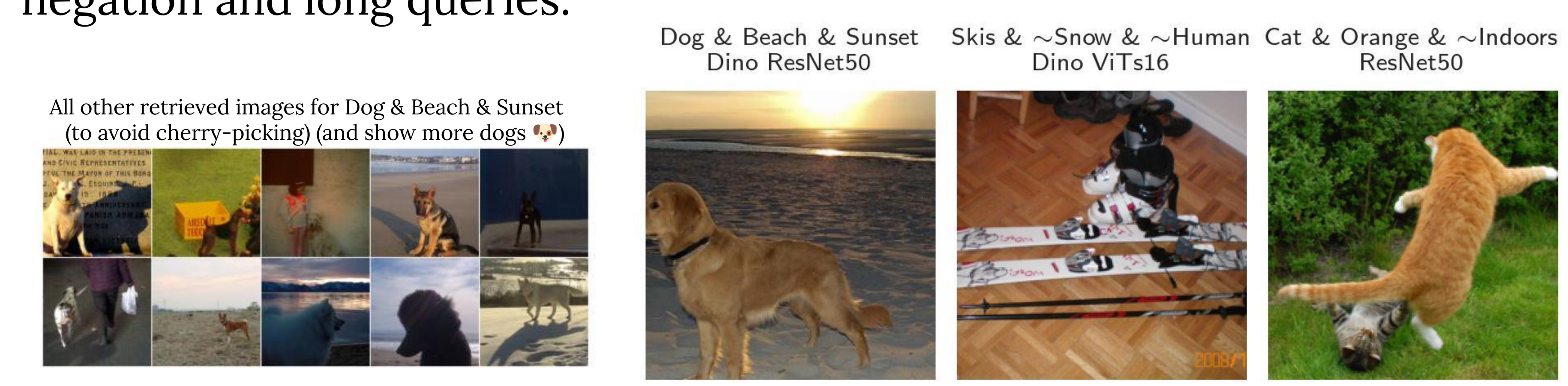


Application 3: **Image Retrieval** using the embeddings of your choice via **Concept Logic**

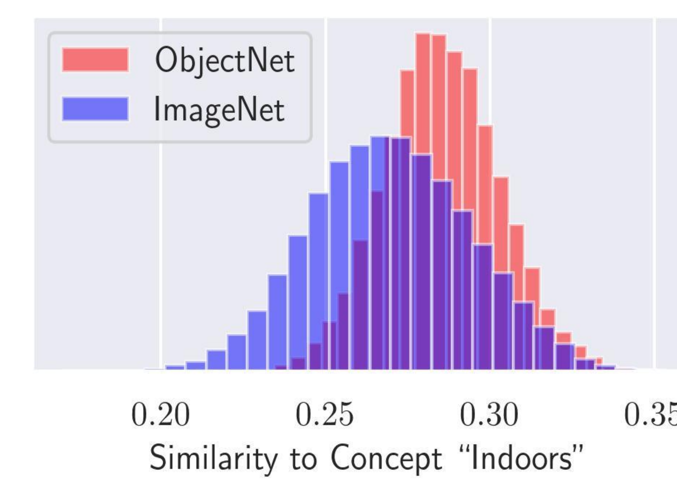
Extends CLIP's ability to retrieve images with text to arbitrary encoders.

Concept Logic is a simple way to somewhat side step text encoder limitations with negation and long queries.





Application 4: **Diagnosing Distribution Shifts**



By maintaining a bank of concept vectors (cheap via Text-to-Concept), one can track concept similarities as data (e.g. from a new deployment environment) streams in, and automatically **detect and describe data drifts** w.r.t human notions.

Example: Detecting that ObjectNet images were taken indoors, which contributes to reduced performance.

Application 5: **Decoding Latent Vectors** *to text*

Swin (S)	ResNet-50	Dino ViTs8
94.48%	95.14%	92.18%

Using CLIP decoders, we can map latent vectors for arbitrary vision encoders to text.

Shown: ImageNet classification head vectors for diversely trained vision encoders are decoded to words that adequately describe images from the class (based on human judgements).
Current challenge: Decodings are generally coarse-grain. We hope this will improve as CLIP decoders do.

Parting Thoughts

Many efforts focus on ***how*** we train models.

It might be time to shift more attention to ***what*** we train our models on.

Existing models may be capable of more than what we use them for.

How else can we allow the models we already have to work *together*?