Explaining CLIP Zero-shot Predictions Through Concepts

CVPR 2026

Onat Ozdemir[1,2]

Anders Christensen[3,4,5]

Stephan Alaniz[6]

Zeynep Akata[7,8,9,10]

Emre Akbas[2,8,11]

1 The University of Edinburgh 2 ODTÜ METU 3 Orbital 4 DTU 5 University of Copenhagen 6 Télécom Paris 7 TUM 8 Helmholtz Munich 9 MCML 10 MDSI 11 ROMER

Contrastive Language-Image Pre-Training

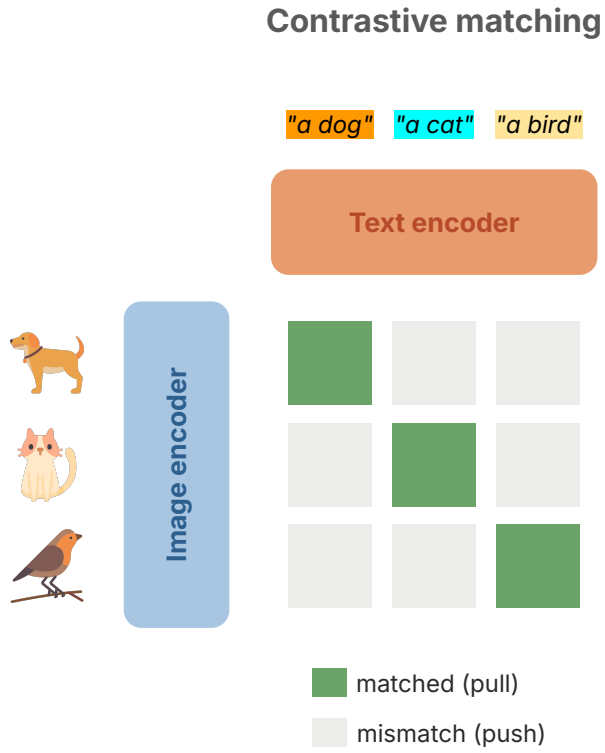
Shared embedding space for images and text

Trained on 400M image-text pairs with a contrastive objective:

- Pull matched image-text pairs together
- Push mismatched pairs apart

After training, classification becomes a similarity comparison, no task-specific labels needed.

Open-vocabulary zero-shot recognition



Concept Bottleneck Models

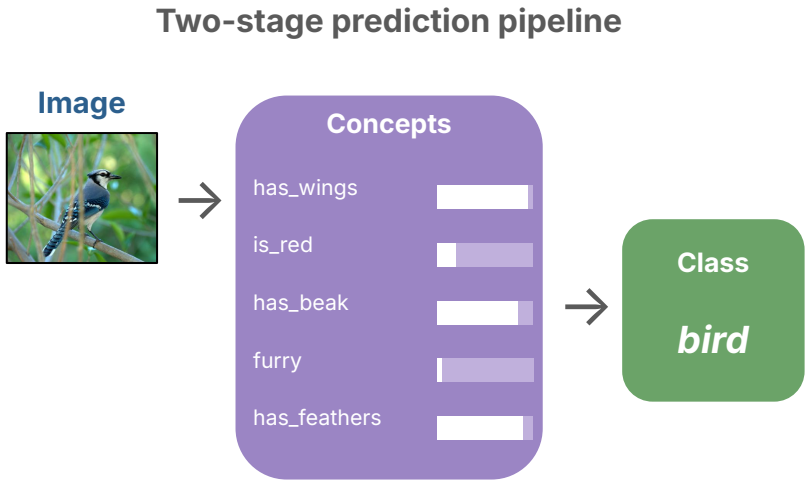
Predictions routed through human concepts

Two-stage prediction:

1. Image → concept activations (e.g. has_wings, is_red, has_beak)
2. Concept activations → class label

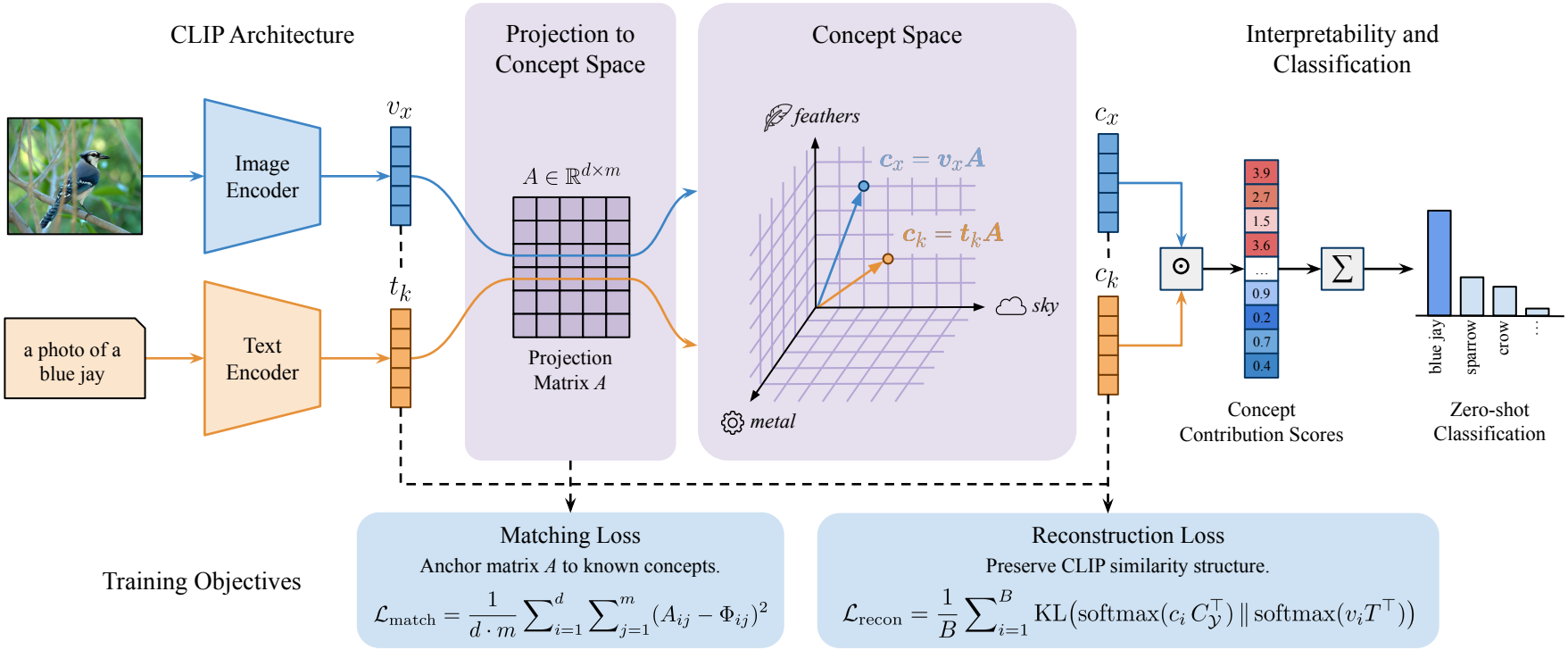
Users can inspect, validate, or even edit concept activations to understand or correct model behavior.

Interpretable by design



Decisions are traceable through human-readable concepts

Method


CLIP Architecture
Image Encoder
Text Encoder
a photo of a blue jay
v_x
t_k
Projection to Concept Space
$A \in \mathbb{R}^{d \times m}$
Projection Matrix A
Concept Space
feathers
$c_x = v_x A$
$c_k = t_k A$
sky
metal
Interpretability and Classification
c_x
c_k
3.9
2.7
1.5
3.6
...
0.9
0.2
0.7
0.4
Concept Contribution Scores
blue jay
sparrow
crow
...
Zero-shot Classification
Training Objectives
Matching Loss
Anchor matrix A to known concepts.
$\mathcal{L}_{\text{match}} = \frac{1}{d \cdot m} \sum_{i=1}^{d} \sum_{j=1}^{m} (A_{ij} - \Phi_{ij})^2$
Reconstruction Loss
Preserve CLIP similarity structure.
$\mathcal{L}_{\text{recon}} = \frac{1}{B} \sum_{i=1}^{B} \text{KL}\left(\text{softmax}(c_i C_{\mathcal{Y}}^{\top}) \,\|\, \text{softmax}(v_i T^{\top})\right)$

Quantitative Results

within 1%
of CLIP's harmonic mean

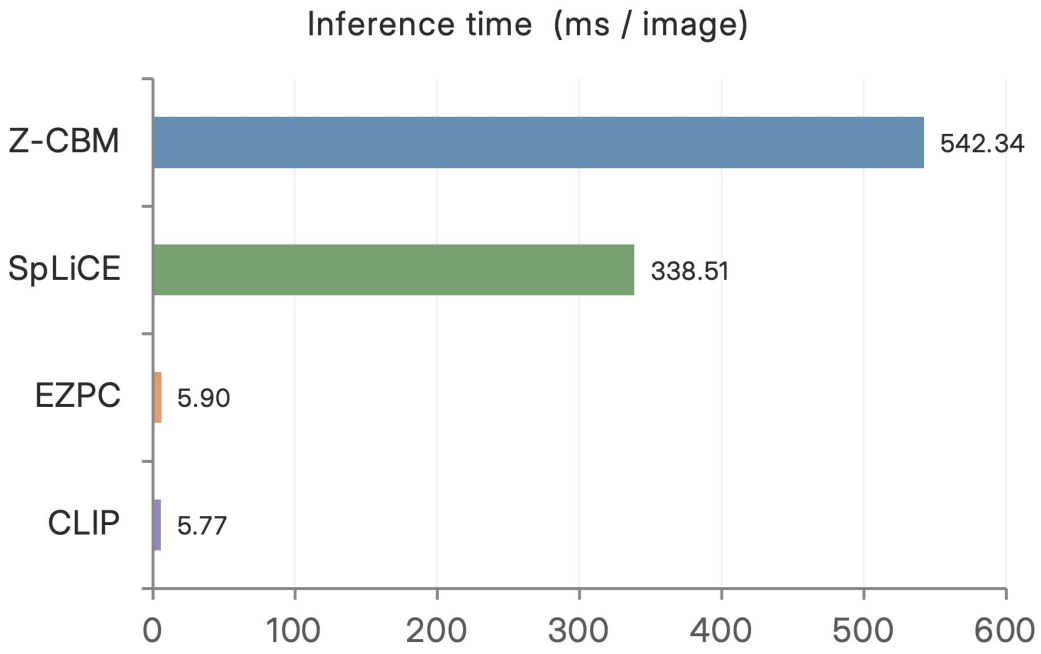
+10-15%
vs. Z-CBM & SpLiCE

5 benchmarks
object · fine-grained · scene

	CIFAR-100			ImageNet-100			CUB			ImageNet-1k			Places365		
Model	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
CLIP	0.370	0.454	0.408	0.680	0.707	0.693	0.468	0.481	0.474	0.513	0.548	0.530	0.350	0.375	0.362
Z-CBM	0.319	0.425	0.365	0.592	0.579	0.585	0.183	0.195	0.189	0.439	0.486	0.462	**0.349**	0.365	**0.357**
SpLiCE	0.248	0.298	0.270	0.371	0.409	0.389	0.100	0.053	0.070	0.275	0.331	0.300	0.276	0.288	0.282
EZPC	**0.365**	**0.449**	**0.403**	**0.675**	**0.690**	**0.682**	**0.457**	**0.473**	**0.465**	**0.468**	**0.494**	**0.481**	0.339	**0.366**	0.352

Generalized zero-shot setting. All models use CLIP RN50. Best non-CLIP score per column in bold.

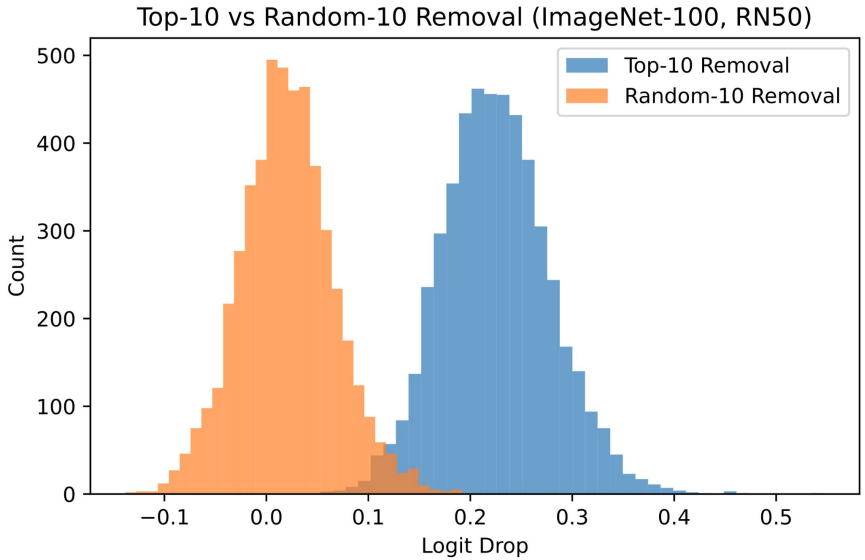
Inference Time



No added latency!

EZPC: 5.90 ms ≈ CLIP: 5.77 ms (p = 0.31)

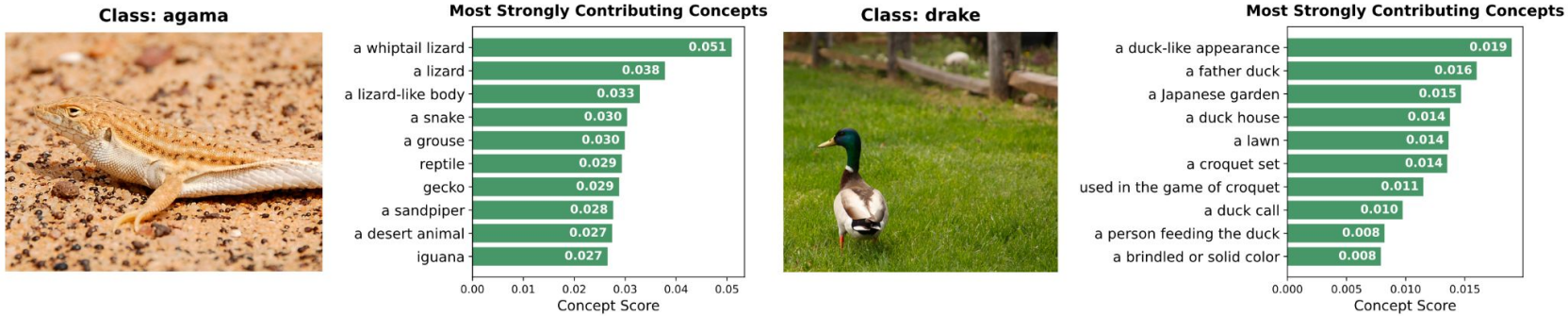
Faithfulness



Removal type	Flip Count	Flip Rate
Top-10 concepts	845	0.169
Random-10 concepts	70	0.014

Qualitative Results

Image-level Analysis



Qualitative Results

Class-level Analysis

Class: macaw

Top Concepts

parrot
a large, colorful bird
red, blue, and yellow feathers
brightly colored feathers
bright plumage
a short beak
a rainforest
a peahen
a brightly colored face
iridescent feathers

Class: magpie

Top Concepts

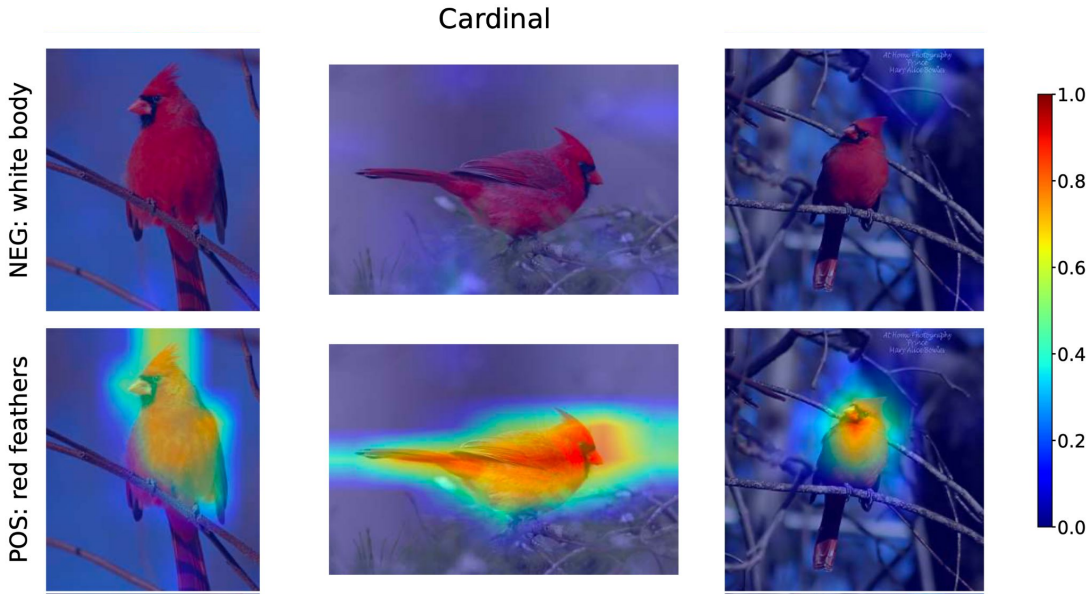
a large, black and white bird
a loud crow
black plumage
a large, stocky bird
a bird
a small songbird
black or dark brown feathers
a small to medium-sized bird
other birds
bright plumage

Qualitative Results

Concept-based Image Clustering

Qualitative Results

Concept-Region Alignment



Want to know more?

Project Page

ArXiv

GitHub

We acknowledge the computational resources provided by METU Center for Robotics and Artificial Intelligence (METU-ROMER) and TUBITAK ULAKBIM TRUBA. Dr. Alaniz is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005). Dr. Akata acknowledges partial funding by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation. Dr. Akbas gratefully acknowledges the support of TUBITAK 2219.