CVPR JUNE 3-7, 2026 DENVER COLORADO

Explaining CLIP Zero-shot Predictions Through Concepts

Onat Ozdemir[1,2] Anders Christensen[3,4,5] Stephan Alaniz[6] Zeynep Akata[7,8,9,10] Emre Akbas[2,8,11]

Want to know more?

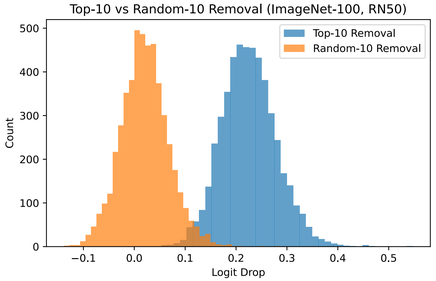
ArXiv GitHub

TL;DR

- **EZPC** explains CLIP's zero-shot predictions through human-understandable concepts.
- Learns a *single linear projection* that aligns image-text embeddings with a concept basis.

Faithful explanations, CLIP-level accuracy, no latency!

Faithfulness



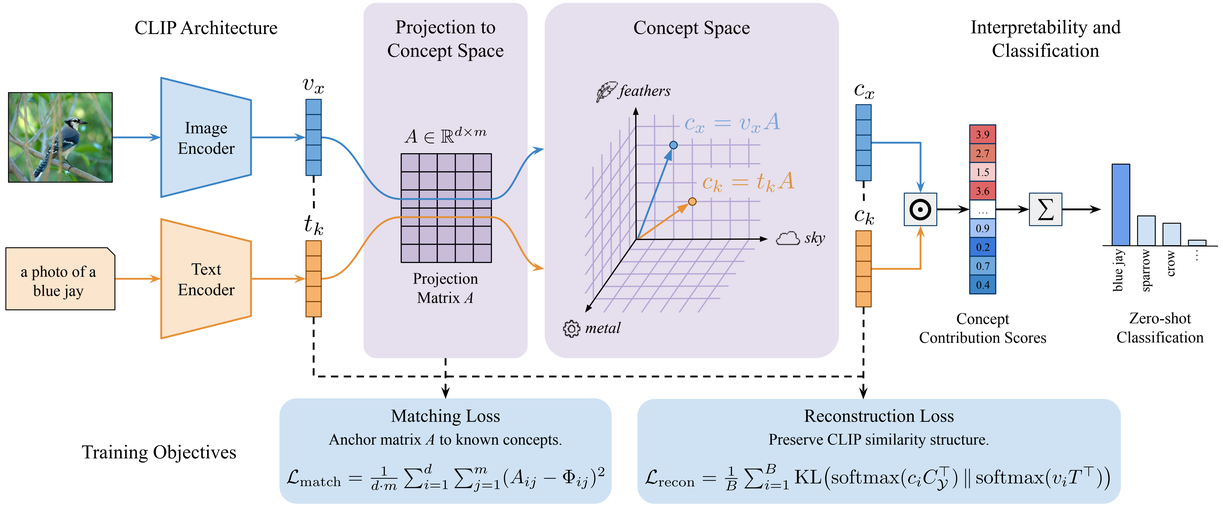
Removal type	Flip Count	Flip Rate
Top-10 concepts	845	0.169
Random-10 concepts	70	0.014

References

[1] Radford et al. *"Learning Transferable Visual Models From Natural Language Supervision."* ICML, 2021.

[2] Yamaguchi et al. *"Zero-shot Concept Bottleneck Models."* arXiv preprint, 2025.

[3] Bhalla et al. *"Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE)."* NeurIPS, 2024.



Zero-shot Quantitative Results

Model	CIFAR-100			ImageNet-100			CUB			ImageNet-1k			Places365		
	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
CLIP [1]	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 [2]	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 [3]	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

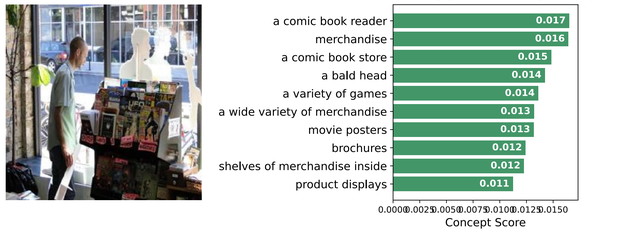
Inference Time

Method	Embedding (ms/img)	Full Pipeline (ms/img)	Overhead
CLIP	0.0001 ± 0.0000	5.77 ± 0.55	1.0×
Z-CBM	97.55 ± 1.33	542.34 ± 6.02	94.0×
SpLiCE	4.50 ± 0.54	338.51 ± 4.39	58.7×
EZPC	0.0006 ± 0.0000	5.90 ± 0.73	∼1.0×

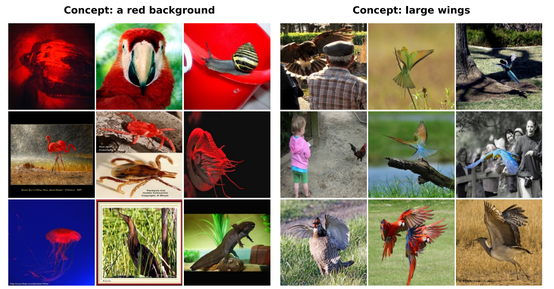
Cross-dataset Transfer

Target Dataset	Model	Zero-shot		Generalized Zero-shot		
		Seen	Unseen	Seen	Unseen	H
CIFAR-100	CLIP	0.686	0.387	0.663	0.266	0.380
	EZPC	0.684	0.363	0.659	0.296	0.409
CUB	CLIP	0.686	0.471	0.617	0.458	0.526
	EZPC	0.674	0.461	0.607	0.448	0.515

Image-level Explanations



Concept Clustering



Concept-Region Alignment

1 THE UNIVERSITY of EDINBURGH 2 ODTÜ METU 3 Orbital 4 DTU 5 UNIVERSITY OF COPENHAGEN 6 TELECOM Paris INSTITUT POLYTECHNIQUE DE PARIS 7 TUM 8 HELMHOLTZ MUNICH 9 MCML 10 MDSI 11 ROMER