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]

[1]School of Informatics, University of Edinburgh
[2]Dept. of Computer Eng., Middle East Technical University (METU) [3]Orbital
[4]DTU Compute, Technical University of Denmark [5]Dept. of Biology, University of Copenhagen
[6]LTCI, Télécom Paris, Institut Polytechnique de Paris [7]Technical University of Munich (TUM)
[8]Helmholtz Munich [9]MCML [10]MDSI [11]Robotics & AI Center (ROMER), METU

Abstract

Large-scale vision-language models such as CLIP achieve remarkable zero-shot recognition but remain opaque, while Concept Bottleneck Models offer interpretability through human-defined concepts but require concept supervision and cannot generalize to unseen classes. We introduce EZPC, which bridges these paradigms by projecting CLIP's joint image-text embeddings into a human-interpretable concept space learned from language descriptions. The projection is learned via alignment and reconstruction objectives that preserve CLIP's semantic structure while enabling faithful, concept-level explanations without additional supervision. Experiments on five benchmarks (CIFAR-100, CUB-200-2011, Places365, ImageNet-100, and ImageNet-1k) demonstrate that EZPC maintains CLIP's zero-shot accuracy while providing meaningful explanations, offering a principled step toward interpretable vision-language models. Code is available at https://github.com/oonat/ezpc.

1. Introduction

The rapid integration of machine learning into high-stakes domains has intensified the demand for models that are not only accurate but also transparent. Concept Bottleneck Models (CBMs) [13] address this by decomposing predictions into two stages: mapping inputs to human-understandable concept activations, and predicting class labels from these activations. However, classical CBMs require explicit concept-level annotations and operate under a closed-world assumption, limiting their scalability. Recent efforts leveraging vision-language models [22, 35, 37] reduce annotation costs but still require task-specific training and cannot generalize to unseen classes.

In contrast, vision-language models such as CLIP [24], ALIGN [10], and SigLIP [38] demonstrate strong open-vocabulary generalization by aligning images and text in a shared semantic space, enabling zero-shot classification without task-specific training. However, this generalization comes at the cost of interpretability. CLIP's high-dimensional embeddings offer little insight into what visual or semantic properties drive a particular decision.

We propose **EZPC** ("**E**xplaining CLIP **Z**ero-shot **P**redictions Through **C**oncepts"), a method that bridges the interpretability of CBMs with CLIP's generalization ability. Our method introduces a lightweight linear projection that decomposes CLIP's image-text embeddings into a shared concept space, enabling faithful concept-level explanations while preserving zero-shot capabilities. EZPC aligns CLIP's representations with a predefined concept basis using two complementary objectives: (i) a matching loss that enforces alignment between learned and known concept embeddings, and (ii) a reconstruction loss that preserves CLIP's similarity structure in the concept space.

Contributions. Our key contributions are as follows:

- We propose a novel method that decomposes CLIP's image-text embeddings into a shared concept space, enabling interpretable zero-shot predictions.
- We introduce two training objectives, matching and reconstruction, that jointly align concept projections with CLIP's latent space while preserving semantic fidelity.
- We demonstrate through quantitative and qualitative experiments on five benchmarks that EZPC provides human-interpretable explanations of CLIP predictions with minimal performance loss.

Related Work. Zero-shot learning (ZSL) aims to recognize unseen categories without explicit training data, progressing from attribute-based methods [6, 15, 16] and word embeddings [7, 20, 23], through embedding-based approaches [1, 8, 21, 26] and weight-learning methods [5, 11, 18, 31], to large-scale vision-language models that learn open-vocabulary recognition through contrastive learning. Subsequent VLM works such as CoOp [40], BLIP [17],

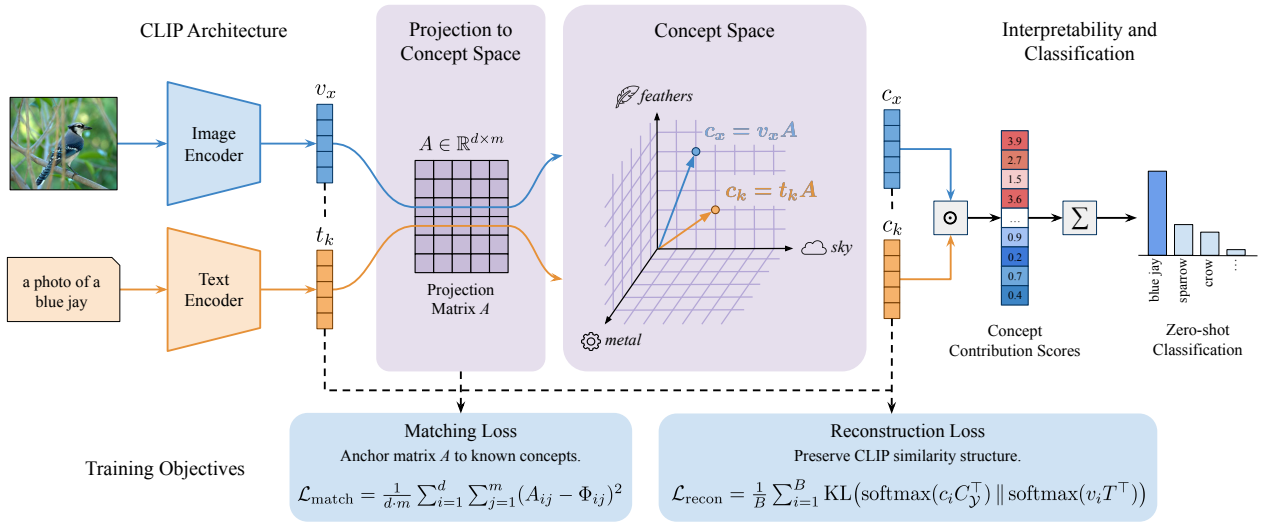


Figure 1. **Overview of EZPC.** CLIP image and text embeddings are projected into a shared concept space using a learnable matrix A. The projected representations c_x and c_k provide (i) concept-based explanations via a Hadamard product and (ii) class logits via a dot-product in concept space. Training jointly optimizes a matching loss and a reconstruction loss to preserve CLIP's predictive behavior.

and PaLI-Gemma [2] expanded capabilities through prompt tuning and cross-modal retrieval, though their internal representations remain opaque.

On the interpretability side, CBMs [13] predict human-defined attributes before classification, with subsequent works improving robustness [12, 28, 34, 37], discovering concepts automatically [25], and integrating textual guidance [22, 29, 35, 36], though most remain limited to closed-world settings. Recent efforts to merge CLIP with concept-based interpretability [19, 33] include Gandelsman et al. [9] decomposing CLIP's encoder across patches and attention heads, SpLiCE [3] decomposing embeddings into sparse concept combinations with expensive per-image optimization, and Z-CBM [32] extending CBMs to zero-shot settings through concept bank regression. In contrast, EZPC learns a single, unified projection that jointly preserves semantic alignment and interpretability, yielding efficient, faithful explanations of zero-shot predictions.

2. Method

2.1. Concept Decomposition

Let $\mathcal{Y} = \{y_1, \ldots, y_K\}$ denote a set of K candidate class labels. We define a learnable projection matrix $A \in \mathbb{R}^{d\times m}$ that should map CLIP's d-dimensional embedding space into a concept space with m interpretable dimensions. For this, we want each of the m columns of A to correspond to a distinct, human-interpretable concept direction in CLIP's space (e.g., *feathers, metal, sky*). We describe how this is achieved through initialization and training in Section 2.2.

Now, let f_{img} and f_{text} denote CLIP's image and text encoders, respectively, and define the normalized embeddings

$$v_x = \frac{f_{\text{img}}(x)}{\|f_{\text{img}}(x)\|}, \qquad t_k = \frac{f_{\text{text}}(y_k)}{\|f_{\text{text}}(y_k)\|}, \tag{1}$$

where $v_x \in \mathbb{R}^d$ is the image embedding and $t_k \in \mathbb{R}^d$ is the text embedding of the k-th class label. We stack the class embeddings into a matrix $T = [t_1; \ldots; t_K] \in \mathbb{R}^{K\times d}$.

Using A, we then compute concept activations for images and labels in the shared concept space as

$$c_x = v_x A, \qquad C_{\mathcal{Y}} = TA, \tag{2}$$

where $c_x \in \mathbb{R}^m$ and $C_{\mathcal{Y}} \in \mathbb{R}^{K\times m}$. Each row of $C_{\mathcal{Y}}$ gives the concept activations of the corresponding class label.

2.2. Training Objectives

Our goal is to learn a concept projection that preserves CLIP's similarity structure while remaining interpretable. For this, we jointly optimize two complementary objectives. **(1) Matching Loss.** We initialize A from a set of concept embeddings relevant to the target domain, such as visual attributes. For a set of m concepts, we compute and stack their CLIP text embeddings to form a matrix $\Phi \in \mathbb{R}^{d\times m}$, such that each column corresponds to a concept phrase (e.g., *has feathers, made of metal*). We initialize $A = \Phi$ and use a mean-squared matching loss to keep A close to this interpretable basis throughout training:

$$\mathcal{L}_{\text{match}} = \frac{1}{d\cdot m}\sum_{i=1}^{d}\sum_{j=1}^{m}(A_{ij}-\Phi_{ij})^2. \tag{3}$$

This anchoring ensures that the columns of A remain aligned with known concept directions, preserving interpretability.

(2) Reconstruction Loss. To ensure that the decomposition preserves CLIP's zero-shot similarity structure, we introduce a reconstruction loss based on the KL divergence between the original CLIP similarity distribution and the concept-based distribution. Given a batch of B image embeddings $\{v_i\}_{i=1}^{B}$, we define:

$$\mathcal{L}_{\text{recon}} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{KL}\big(\mathrm{softmax}(c_i\, C_{\mathcal{Y}}^{\top}) \,\|\, \mathrm{softmax}(v_i T^{\top})\big)\,, \tag{4}$$

where $c_i = v_i A$ and $C_{\mathcal{Y}} = TA$ are the concept activations defined in Section 2.1. This enforces that the concept-space similarity distribution remains consistent with CLIP's original predictions, ensuring semantic faithfulness.

(3) Total Loss. The overall objective combines both terms with a balancing coefficient λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{match}} + \lambda\, \mathcal{L}_{\text{recon}}. \tag{5}$$

2.3. Concept-based Inference

At inference, we perform zero-shot classification directly in the concept space:

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\arg\max}\ \langle c_x, c_k\rangle, \tag{6}$$

where c_k denotes the k-th row of $C_{\mathcal{Y}}$. We define a vector of concept-wise interaction scores between image x and class y_k as

$$s_{x,k} := c_x \odot c_k, \tag{7}$$

where the j-th entry of $s_{x,k}$ is large when both the image and the class strongly activate the j-th concept. The overall concept-space similarity decomposes as

$$\langle c_x, c_k\rangle = \sum_{j=1}^{m} s_{x,k}^{(j)}. \tag{8}$$

Each dimension of $s_{x,k}$ quantifies how strongly a specific concept contributes to the image-text alignment. Since these scores compose the prediction logit directly, the explanations are faithful to the model's decision process by construction.

3. Experiments

Datasets. We evaluate our approach on five benchmarks covering diverse visual domains: CIFAR-100 [14], CUB-200-2011 (CUB) [30], ImageNet-100 [27], ImageNet-1k [27], and Places365 [39]. Each dataset is partitioned into seen and unseen classes following an $80/20$ split, enabling evaluation under both zero-shot and generalized zero-shot settings.

Concept space. We adopt concept sets from LF-CBM [22], originally generated using GPT-3 [4]. Concept vocabulary sizes are: CIFAR-100 (*892*), CUB (*370*), ImageNet-1k (*4,751*), and Places365 (*2,544*). For datasets with fewer concepts, we merge their vocabularies with ImageNet-1k's larger pool to obtain a richer set of interpretable attributes.

Evaluation Metrics. In the standard zero-shot setting, prediction is made by selecting the class with the highest cosine similarity to the image embedding. In the generalized zero-shot (GZS) setting, the label space includes both seen and unseen classes. We report accuracies on seen (Acc_S) and unseen (Acc_U) classes, and their harmonic mean (H) as a balanced measure of generalization.

Baselines. We compare against CLIP [24], Z-CBM [32], and SpLiCE [3], all evaluated with identical backbones using official implementations. We use the CLIP RN50 backbone for all experiments except the backbone ablation. Ablations are conducted on ImageNet-100 unless otherwise noted.

3.1. Quantitative Analysis

Results. Table 1 reports generalized zero-shot performance. On CIFAR-100, ImageNet-100, and CUB, EZPC remains within roughly 1% harmonic mean of CLIP, while on ImageNet-1k the gap is larger ($\sim$5%), reflecting the difficulty of this large-scale setting. Across all datasets, EZPC substantially outperforms Z-CBM and SpLiCE, often by 10–15% in harmonic mean, demonstrating that meaningful interpretability can be achieved while keeping performance competitive with CLIP.

Backbone sensitivity. As summarized in Table 2, larger and more expressive backbones improve performance for EZPC. This indicates that the concept-based decomposition scales naturally with model capacity, maintaining interpretability across different architectures.

Impact of Concept Vocabulary Size We analyze how the number of concepts affects both predictive performance and explanation fidelity. We randomly subsample the concept vocabulary and train EZPC with $m \in \{250, 500, 1000, 2000, 3000, 4751\}$ concepts. For each size, we repeat training with three random seeds. Results are shown in Table 3. Performance improves monotonically with the number of concepts, with diminishing returns beyond approximately 3000 concepts. This shows that EZPC is robust to moderate reductions in vocabulary size, while larger vocabularies mainly improve prediction fidelity.

3.2. Cross-Dataset Experiments

To assess the generalization ability of our model beyond its training distribution, we perform cross-dataset experiments where the concept bottleneck is trained on a source dataset (ImageNet-100) and evaluated on distinct target datasets (CIFAR-100 and CUB) without any fine-tuning.

Setup. We train using only source-dataset images, but use both source and target class names and concepts to build the shared semantic space. No images from the target dataset are used during training. During evaluation, we treat classes from the source dataset (e.g., ImageNet-100) as *seen* and

Table 1. **Generalized zero-shot performance across five datasets.** Each dataset is partitioned into seen (80%) and unseen (20%) classes. All models use the CLIP RN50 backbone. EZPC retains strong performance compared to CLIP while introducing interpretability.

	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 [24]	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 [32]	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

Table 2. **Effect of backbone architecture on zero-shot and generalized zero-shot performance.** Larger backbones consistently improve both zero-shot and generalized zero-shot performance.

Backbone	Variant	Zero-shot		Generalized		
		Seen	Unseen	Seen	Unseen	H
CLIP RN50	Base	0.706	0.855	0.680	0.707	0.693
	EZPC	0.699	0.851	0.675	0.690	0.682
CLIP ViT-B/32	Base	0.729	0.887	0.703	0.715	0.709
	EZPC	0.724	0.879	0.694	0.716	0.705
CLIP ViT-L/14	Base	0.839	0.925	0.821	0.836	0.828
	EZPC	0.832	0.924	0.812	0.831	0.822
SigLIP ViT-SO400M/14	Base	0.882	0.972	0.871	0.889	0.880
	EZPC	0.880	0.972	0.870	0.886	0.878

Table 3. **Effect of concept vocabulary size m on performance, tested on ImageNet-100 under the generalized ZSL setting using the RN50 backbone.**

m	Top-1 Agree. (%)	Seen Acc.	Unseen Acc.	H
250	69.05 ± 0.34	0.5527 ± 0.0027	0.5637 ± 0.0105	0.5581 ± 0.0054
500	80.95 ± 1.35	0.6252 ± 0.0064	0.6423 ± 0.0090	0.6336 ± 0.0072
1000	89.09 ± 0.33	0.6622 ± 0.0028	0.6827 ± 0.0032	0.6723 ± 0.0030
2000	91.73 ± 0.17	0.6702 ± 0.0017	0.6890 ± 0.0026	0.6795 ± 0.0004
3000	92.58 ± 0.26	0.6742 ± 0.0010	0.6900 ± 0.0026	0.6820 ± 0.0010
4751 (full)	92.92	0.6745	0.6900	0.6821

classes from the target dataset (e.g., CUB) as *unseen*. For the generalized zero-shot setting, we merge all categories from both datasets and jointly predict across this combined label space, a substantially more challenging scenario than standard zero-shot transfer. We report *Seen* and *Unseen* accuracy under both standard zero-shot and generalized zero-shot settings.

Results. Table 4 reports cross-dataset transfer performance when the projection is trained on ImageNet-100 and evaluated on CIFAR-100 and CUB. For CIFAR-100, EZPC produces zero-shot and generalized zero-shot accuracies that are close to CLIP. The seen accuracies differ by less than 0.5%, and the unseen accuracies are within roughly 2-3%. In the generalized setting, EZPC achieves a harmonic mean about 3% higher than CLIP. For CUB, the differences are similarly small: EZPC is within about 1-2% of CLIP on both seen and unseen zero-shot accuracies, and the harmonic mean differs by roughly 1%. These results indicate that the concept projection learned from ImageNet-100 transfers reasonably well to both object-centric and fine-grained domains, maintaining performance close to CLIP without any fine-tuning.

Table 4. **Cross-dataset transfer results for EZPC trained on ImageNet-100 compared to CLIP.** We evaluate on two target datasets: CIFAR-100 and CUB.

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

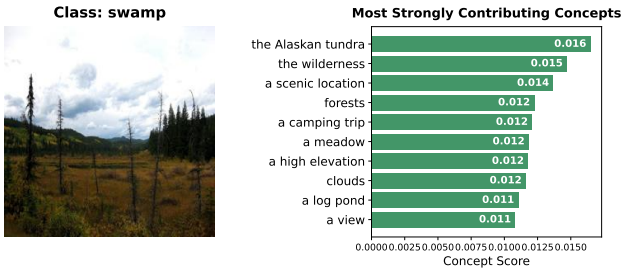


Figure 2. **Places365 image-level explanations.** For the given image, EZPC displays the top-10 activated concepts that contribute most to the zero-shot prediction.

3.3. Qualitative Analysis

3.3.1. Image-level Explanations

To compute image-level explanations, we project both the image and predicted class label embeddings into concept space, compute their element-wise product, and select the top-10 activated concepts. Figure 2 shows a Places365 example, where scene-level concepts such as *the Alaskan tundra* and *the wilderness* emerge for the class *swamp*. While most activated concepts are semantically relevant, some reflect class-level associations rather than the visual content of the specific image.

3.3.2. Class-level Explanations

We obtain class-level explanations by averaging concept activations across nine randomly sampled images from a target class. Figure 3 presents an example from CUB. The *Cardinal* class highlights *a small, sparrow-like bird*, *a red crest on the head*, and *a red head*. Some activated concepts are not meaningful (e.g., *a checkered or solid red sauce*, *heavy build*, and *a pack of dholes*), likely stemming from noise in CLIP's embedding space or limitations of the concept vocabulary.

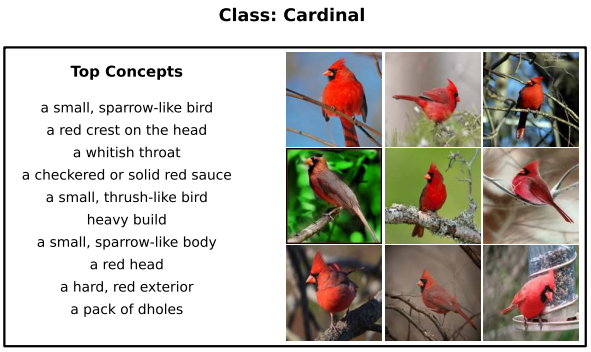


Figure 3. **CUB class-level concept explanations.** For the given class, we average concept activations over nine sampled images.

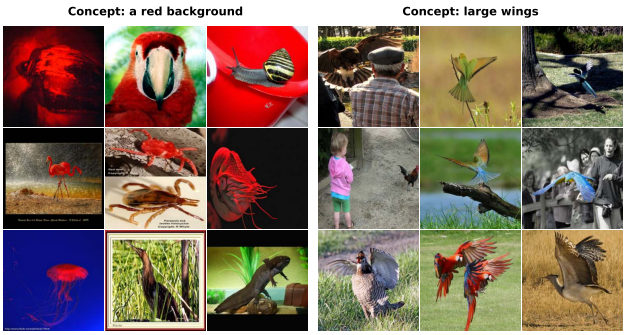


Figure 4. **ImageNet-100 concept clustering results.** For each concept, we retrieve the nine images with the highest activations.

3.3.3. Concept Clustering

We analyze the learned concept space by performing concept-based image retrieval: for a given concept, we retrieve the nine images with the highest activation. Figure 4 shows clusters from ImageNet-100. The concept *a red background* retrieves images with prominent red tones, and *large wings* retrieves bird photos, confirming that concept dimensions capture visually grounded semantics.

3.3.4. Concept-Region Alignment

To evaluate whether the learned concept space produces spatially meaningful explanations, we extract patch-level features from the CLIP RN50 backbone, project them into concept space via A, and visualize the resulting spatial activation maps. Figure 5 shows an example of the *Indigo Bunting* class from CUB: a positive concept (*a blue-gray body*) produces high activations localized on the bird, while a negative concept (*a red face*) shows near-zero activation, confirming that concept dimensions capture spatially grounded semantics. We further quantify localization quality using ground-truth segmentation masks, reporting *Pointing Accuracy*, *Inside Activation Ratio*, *IoU@10%*, and *IoU@20%*. As shown in Table 5, the positive concept achieves 96.7% pointing accuracy and substantially higher IoU scores than the negative concept.

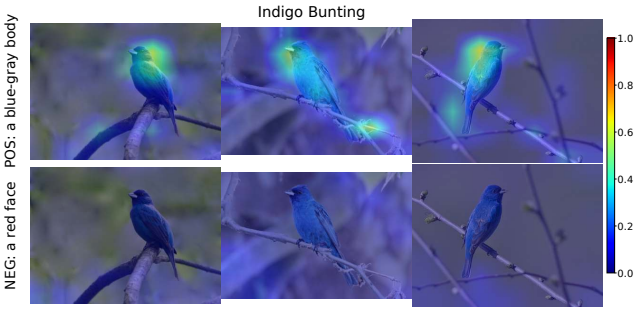


Figure 5. **Region-level concept alignment for Indigo Bunting.** Top row: positive concept (*a blue-gray body*). Bottom row: negative concept (*a red face*).

Table 5. **Quantitative evaluation of concept-region alignment using CUB ground-truth segmentation masks.** Positive concept consistently localizes on the object, while unrelated (negative) concept shows near-zero alignment.

Metric	Positive Concept	Negative Concept
Pointing Accuracy ↑	0.967 ± 0.180	0.017 ± 0.128
Inside Activation Ratio ↑	0.507 ± 0.191	0.031 ± 0.054
IoU@10% ↑	0.423 ± 0.159	0.019 ± 0.067
IoU@20% ↑	0.408 ± 0.148	0.044 ± 0.087

Table 6. **Inference time comparison** on ImageNet-100 (NVIDIA H100). We report median per-image latency (ms) with 95% confidence intervals.

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

3.4. Time Analysis

A key advantage of EZPC is its computational efficiency. Unlike SpLiCE [3] and Z-CBM [32], which require solving an optimization problem per image at inference time, EZPC performs a single matrix multiplication ($v_x A$) on top of CLIP's forward pass. As shown in Table 6, EZPC adds only $\sim$0.1 ms per image over CLIP (5.90 vs. 5.77 ms), whereas Z-CBM and SpLiCE incur $94\times$ and $59\times$ overhead respectively, making EZPC well-suited for large-scale deployment and interactive analysis.

4. Conclusion

We introduced EZPC, a method that explains CLIP's zero-shot predictions through human-interpretable concepts by learning a shared concept projection that preserves CLIP's semantic structure via matching and reconstruction objectives. Across five benchmarks, EZPC maintains strong zero-shot accuracy comparable to CLIP while providing meaningful explanations at both the image and class levels, demonstrating that open-vocabulary recognition and interpretability need not be mutually exclusive.

Acknowledgments

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.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[3] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). In *Advances in Neural Information Processing Systems*, 2024.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[5] Anders Christensen, Massimiliano Mancini, A. Sophia Koepke, Ole Winther, and Zeynep Akata. Image-free classifier injection for zero-shot classification. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[6] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013.

[9] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations*, 2024.

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[11] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Injae Kim, Jongha Kim, Joonmyung Choi, and Hyunwoo J Kim. Concept bottleneck with visual concept filtering for explainable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[13] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020.

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[15] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[16] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[18] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[19] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations*, 2023.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[21] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

[22] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, 2014.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[25] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, 2024.

[26] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 2015.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[28] Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[29] Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. In *Advances in Neural Information Processing Systems*, 2024.

[30] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[31] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models. *arXiv preprint arXiv:2502.09018*, 2025.

[33] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023.

[34] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023.

[35] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[36] Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, and Changsheng Xu. Language guided concept bottleneck models for interpretable continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[37] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023.

[38] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Explaining CLIP Zero-shot Predictions Through Concepts

Supplementary Material

A. Implementation Details

This section provides additional details on concept construction, training procedures, and evaluation settings used for EZPC. We omit definitions and equations already introduced in the main paper and focus on implementation-specific clarifications.

A.1. Backbone and Embedding Extraction

We primarily use the CLIP RN50 backbone, with additional experiments using ViT-B/32, ViT-L/14 (via OpenAI CLIP), and SigLIP ViT-SO400M/14 (via OpenCLIP). For every dataset, we precompute:

- ℓ_2-normalized image embeddings for all train/validation images,
- text embeddings for all dataset class names,
- text embeddings for the corresponding concept vocabulary.

All embeddings remain frozen during the training of EZPC.

A.2. Concept Vocabulary per Dataset

We follow the LF-CBM [22] and use their GPT-3-generated concept sets.

- **ImageNet-1k & ImageNet-100:** We use the 4,751 ImageNet-derived GPT-3 concepts from LF-CBM. No additional concepts are merged.
- **CIFAR-100, CUB, Places365:** The original LF-CBM concept sets for these datasets are limited. Therefore, we merge the dataset's own LF-CBM concepts with the larger ImageNet concept set to obtain a sufficiently expressive concept space. After merging, duplicate concepts are removed.

All concepts are encoded once using the corresponding model's text encoder.

A.3. Training the Concept Projection Matrix A

The learnable matrix $A \in \mathbb{R}^{d\times m}$ maps CLIP/SigLIP embeddings into the m-dimensional concept space. The matrix is initialized as

$$A^{(0)} = \Phi, \tag{9}$$

where Φ is the CLIP concept embedding matrix.

The training objective consists of a matching term and a reconstruction term, as defined in the main paper. The scalar λ (typically $\lambda = 1$, and $\lambda = 5$ for CUB and Places365) controls the relative weight of the reconstruction loss. No orthogonality or sparsity regularizers are used.

We optimize only A (all backbone parameters remain frozen) using Adam with learning rate 10^{-2} for 10,000 iterations. After every epoch we renormalize all concept vectors

$$A_{:,j} \leftarrow \frac{A_{:,j}}{\|A_{:,j}\|_2}, \tag{10}$$

which stabilizes the concept geometry and prevents drift during training.

A.4. Zero-Shot and Generalized Zero-Shot Evaluation

Let $\mathcal{Y}_S$ and $\mathcal{Y}_U$ denote the seen and unseen class sets.

Zero-shot learning (ZSL). Evaluation is performed over unseen classes only. For each test image x, the predicted label is

$$\hat{y} = \underset{k:\, y_k\in\mathcal{Y}_U}{\arg\max} \langle c_x, c_k\rangle, \tag{11}$$

and accuracy is computed over the unseen test set $\mathcal{D}_U$ as

$$\mathrm{Acc}_U = \frac{1}{|\mathcal{D}_U|} \sum_{(x,y_k)\in\mathcal{D}_U} \mathbb{1}[\hat{y} = y_k]. \tag{12}$$

Generalized zero-shot learning (GZSL). Predictions are made over the combined label set $\mathcal{Y}_G = \mathcal{Y}_S \cup \mathcal{Y}_U$. We report accuracies on seen and unseen classes separately (Acc_S, Acc_U), along with the harmonic mean

$$H = \frac{2 \times \mathrm{Acc}_S \times \mathrm{Acc}_U}{\mathrm{Acc}_S + \mathrm{Acc}_U}. \tag{13}$$

A.5. Time Analysis Protocol

We measure inference latency on the ImageNet-100 validation set using a single NVIDIA H100 GPU. For each method, we report two quantities: (i) the *embedding time*, which measures only the concept decomposition step (excluding the shared CLIP forward pass), and (ii) the *full pipeline time*, which includes CLIP encoding plus the method-specific decomposition.

To obtain stable estimates, we first run warm-up iterations, then time each method over the full validation set and record per-image latencies. We report the median latency together with 95% confidence intervals.

For EZPC, the decomposition reduces to a single matrix multiplication $c_x = v_x A$, which adds negligible overhead to CLIP's forward pass. To verify this statistically, we apply a Wilcoxon signed-rank test comparing per-image latencies of CLIP and EZPC, yielding $p = 0.31$ (no significant difference). In contrast, SpLiCE requires iterative sparse coding

per image, and Z-CBM performs a retrieval-based regression over the concept bank, both of which involve iterative optimization and result in substantially higher latency.

A.6. Concept-Region Alignment

Generating spatial heatmaps. To produce concept-level spatial activation maps, we extract patch-level representations from the CLIP RN50 backbone. We register a forward hook on `layer4` of the ResNet visual encoder to capture the spatial feature map before attention pooling, yielding $N_p = 7 \times 7 = 49$ patch embeddings. Each patch is passed through CLIP's attention pooling projections (v_proj, c_proj) and ℓ_2-normalized to obtain patch embeddings $\{p_i\}_{i=1}^{N_p}$ in $\mathbb{R}^d$.

Each patch embedding is then projected into the concept space via A to obtain $z_i = p_i A \in \mathbb{R}^m$. We normalize each z_i by its maximum absolute value and mean-center across concepts to ensure comparable activation scales. For a given concept j, the spatial heatmap is formed by extracting the j-th coordinate of each normalized patch, applying ReLU, and reshaping to the 7×7 grid. The result is bilinearly upsampled to the original image resolution for visualization.

Quantitative evaluation metrics. We evaluate spatial alignment on CUB-200-2011 using ground-truth segmentation masks. For each class, we manually specify a *positive* concept (e.g., *a blue-gray body* for Indigo Bunting) and a *negative* concept (e.g., *a red face*), compute heatmaps for both across all images of that class, and compare against the binary segmentation mask M (resized to the patch grid). We report:

- **Pointing Accuracy**: fraction of images where the maximally activated patch falls inside M.
- **Inside Activation Ratio**: proportion of total activation mass falling inside M.
- **IoU@τ%**: intersection-over-union between M and the binary mask obtained by thresholding the heatmap at the $(100-\tau)$-th percentile. We report IoU@10% and IoU@20%.

For each metric, we report the mean and standard deviation across all images in the class. As shown in Table 5 of the main paper, positive concepts consistently localize on the object region, while negative concepts produce near-zero alignment scores.

B. Faithfulness & Causal Validation

A core requirement for concept-based interpretability is faithfulness: the degree to which the concepts identified as important are causally responsible for the model's predictions. In this section, we evaluate whether the learned concept space A discovered by EZPC produces faithful explanations of CLIP's zero-shot classifier.

All experiments are conducted on **ImageNet-100** using the **CLIP RN50** backbone. Because ImageNet-100 is large, we evaluate faithfulness on a randomly sampled subset of validation images, ensuring stable estimates while keeping computation tractable.

B.1. Concept Removal for Causal Testing

As defined in the main paper, the concept-wise interaction score between image x and class y_k is $s_{x,k} = c_x \odot c_k$, where the j-th entry $s_{x,k}^{(j)}$ measures how strongly concept j contributes to the prediction. Let $\mathcal{J}_n(x, y_k)$ denote the indices of the top-n concepts ranked by $s_{x,k}^{(j)}$. To test causal influence, we ablate the top-n influential concepts by removing their contribution from the scoring function

$$f'(x, y_k) = f(x, y_k) - \sum_{j \in \mathcal{J}_n(x, y_k)} s_{x,k}^{(j)}, \tag{14}$$

where $f(x, y_k) = \langle c_x, c_k \rangle$ is the concept-space logit reconstructed via EZPC.

Faithfulness is quantified as the expected logit drop

$$\Delta_n = \mathbb{E}_{(x, y_k)} \big[f(x, y_k) - f'(x, y_k) \big]. \tag{15}$$

Higher Δ_n values imply stronger causal reliance on the discovered concepts.

B.2. Faithfulness Results

Mean Logit Drops. Table 7 reports the mean logit drop and prediction flip rate after ablating the top-n most influential concepts, averaged across the sampled ImageNet-100 validation images.

Table 7. **Faithfulness results on ImageNet-100 (CLIP RN50).** Mean logit drop and prediction flip rate after ablating the top-n most influential concepts.

Top-n	Logit Drop	Flip Rate
1	0.0306	0.059
3	0.0816	0.099
5	0.1263	0.132
10	0.2256	0.169

Both metrics increase monotonically with n: ablating more high-ranked concepts results in larger logit drops and higher flip rates, confirming that the learned concept directions are causally involved in reconstructing CLIP's similarity structure.

Distributional Effects. Figure 6 shows the full empirical distributions of logit drops for $n = \{1, 3, 5, 10\}$.

As n increases, the distributions shift consistently to the right, indicating that the effect is not limited to a few outlier images but holds broadly across the dataset.

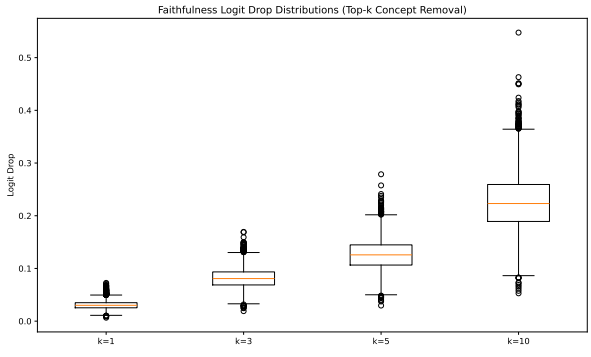


Figure 6. **Faithfulness distributions for n=1, 3, 5, 10.** Removing more highly ranked concepts yields consistently larger drops in model confidence.

B.3. Causal Intervention: Top-10 vs. Random-10

To distinguish causal influence from mere correlation, we compare removing the top-10 influential concepts with removing 10 random concepts

$$\Delta_{\text{top-10}} \quad \text{vs.} \quad \Delta_{\text{rand-10}}. \tag{16}$$

Table 8 reports prediction flip statistics when removing the top-10 most influential concepts compared to removing 10 random concepts. Removing the top-10 concepts changes the predicted class for 16.9% of the evaluated samples, whereas removing 10 random concepts causes flips in only 1.4%. The large gap between these two settings indicates that the discovered concepts correspond to directions that are causally used by CLIP for decision making rather than reflecting spurious correlations.

Table 8. **Top-10 vs Random-10 concept removal.** Prediction flip counts and flip rates on ImageNet-100 (5000 samples).

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

Figure 7 shows that top-10 removal induces a much larger logit decrease than random removal, whose distribution remains tightly centered near zero. This separation is a strong indicator of true causal involvement: if concept scores merely reflected correlations or dataset priors, random removal would produce similar changes, which it does not.

B.4. Discussion and Takeaways

Across all faithfulness and causal tests, the learned concept space A demonstrates:

- **Causal responsibility:** Removing top-ranked concepts reliably degrades classifier confidence.
- **Stable, monotonic attribution:** Logit drops increase smoothly with n, consistent with the additive structure of the concept-space logit.

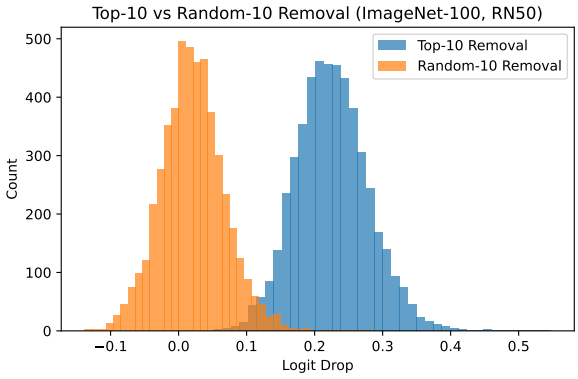


Figure 7. **Causal intervention analysis.** Removing the top-10 influential concepts produces a substantially larger drop in the predicted class logit than removing 10 random concepts.

- **Robustness to dataset variability:** Distributions shift consistently, indicating that effects are widespread and not image-specific.
- **No need for sparsity:** Despite dense activations, explanations remain faithful. Concept rankings are sufficient for causal attribution.

These results confirm that the discovered concept axes form a faithful and causally meaningful basis for explaining CLIP's zero-shot decisions. The concept directions are not merely descriptive; they directly mediate model predictions in a measurable and controllable manner.

C. Concept Space Structure Analysis

This section provides an extended analysis of the geometry and activation behavior of the learned concept space A compared to the original CLIP concept space Φ. All analyses are conducted on **ImageNet-100** using the **CLIP RN50** backbone, matching the quantitative setup of the main paper. Our goal is to understand how optimization alters the concept space, whether semantic identity is preserved, and whether the learned space is geometrically stable, coherent, and suitable for interpretation.

C.1. How the Concept Spaces Are Obtained

The original CLIP concept matrix Φ is created by encoding each concept name j using the CLIP text encoder:

$$\Phi_j = \text{normalize}\big(f_{\text{text}}(\text{“a photo of } j\text{”})\big). \tag{17}$$

The learned concept space A is obtained from our optimization objective, using the same concept vocabulary.

For an image x and class label y_k, we compute the image embedding $v_x \in \mathbb{R}^d$ and text embedding $t_k \in \mathbb{R}^d$

$$v_x = \text{normalize}(f_{\text{img}}(x)), \qquad t_k = \text{normalize}(f_{\text{text}}(y_k)). \tag{18}$$

Following the explanation model of the main paper, the image-side, label-side, and joint concept activations are

$$c_x = v_x A, \quad c_k = t_k A, \quad s_{x,k} = c_x \odot c_k. \tag{19}$$

These activations form the basis for all structure measurements in this section, ensuring full consistency with the interpretability mechanism.

C.2. Summary of Quantitative Results

Table 9 summarizes the key geometric properties of A and Φ. We provide detailed explanations in the subsections below.

Table 9. **Summary of concept space statistics.** All measurements are computed on ImageNet-100 using CLIP RN50. The learned space A preserves semantic identity while becoming more compact and uniformly correlated.

Metric	CLIP (Φ)	Learned (A)
Alignment Mean / Median	-	0.651 / 0.648
Alignment Std	-	0.036
Alignment Min / Max	-	0.559 / 0.822
Total PCA Variance	0.1396	0.1047
Top-10 Variance Fraction	0.4586	0.4364
Off-diagonal Mean ($\Phi^\top \Phi$ vs. $A^\top A$)	0.6920	0.7461
Off-diagonal Std	0.0705	0.0593

These results support the main claim that our learned concept space remains semantically meaningful and structurally well-behaved.

C.3. Alignment Between A and Φ

We evaluate how much each learned concept direction A_j deviates from its original CLIP counterpart Φ_j using cosine alignment

$$\text{align}(A_j, \Phi_j) = \left\langle \frac{A_j}{\|A_j\|}, \frac{\Phi_j}{\|\Phi_j\|} \right\rangle. \tag{20}$$

Figure 8 shows that alignment scores are tightly clustered around 0.65, with a standard deviation of only 0.036 and minimum/maximum values of 0.559 and 0.822. This distribution demonstrates two important facts:

- The learned concept directions preserve **semantic grounding**.
- Optimization introduces **controlled refinements** rather than large distortions.

Thus, each learned concept remains strongly related to its original semantic meaning, ensuring stable and interpretable explanations.

C.4. PCA Geometry

To assess global geometry, we perform PCA on A and Φ, treating concepts as data points in the embedding space.

The PCA results show:

- **Lower total variance**: A has 0.1047 vs. 0.1396 in Φ.
- **Comparable top-10 variance fraction**: 0.4364 vs. 0.4586.

This means that A is more compact, but not low-rank or collapsed. The PCA scatter in Figure 9 shows that concept clusters and their relative positions are preserved, indicating that the refined space maintains CLIP's semantic organization while smoothing its geometry.

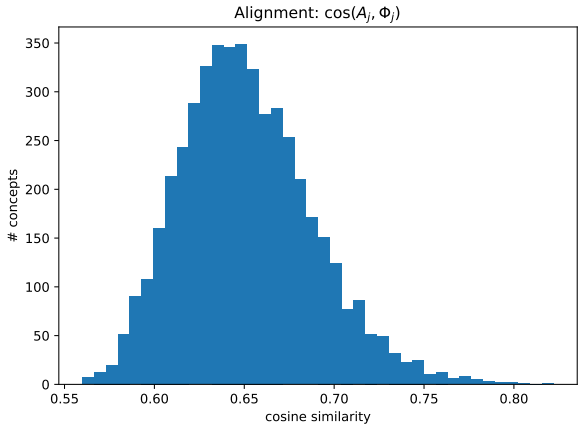


Figure 8. **Alignment distribution** between learned concept directions A_j and original CLIP directions Φ_j. High alignment values indicate semantic consistency between the learned space and CLIP.

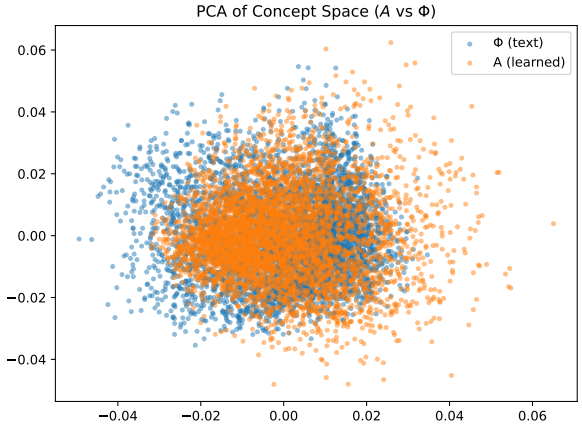


Figure 9. **PCA comparison** between CLIP concept space (Φ) and learned space (A). The learned space is more compact yet preserves the structural layout of concept groups.

C.5. Activation Density and Why Sparsity Is Not Required

We examine the density of concept activations using the joint interaction vector $s_{x,k} = c_x \odot c_k$. For each image, a concept j is counted as active if $s_{x,k}^{(j)} > \tau$, where $\tau = 0.01$. Figures 10 and 11 show the distribution of active concepts per image and the number of images activating each concept.

Across ImageNet-100:

- Each image activates **490 concepts on average** (median 190).
- Each concept is activated for **516 images on average** (median 494).

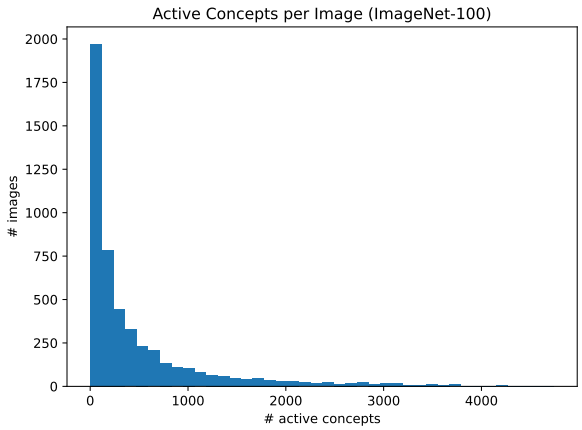


Figure 10. **Activation density**: number of active concepts per image.

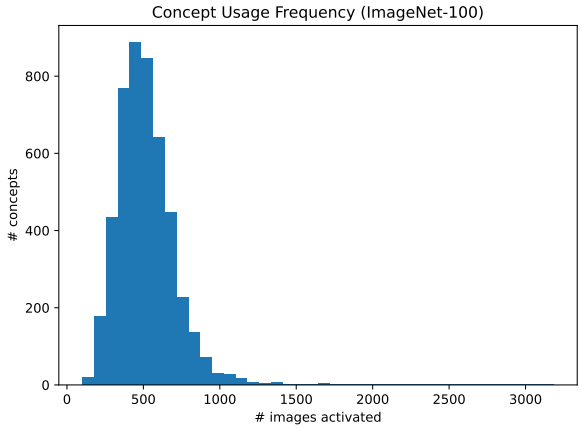


Figure 11. **Activation coverage**: number of images activating each concept.

These dense activation patterns are expected because CLIP embeddings are intrinsically distributed. Importantly, **sparsity is not required** for concept-based interpretability in our method. Explanations rely on identifying and ranking the most influential concept directions (top-n), not on enforcing a few active dimensions. Dense signals still yield clear, semantically aligned top concepts, as demonstrated in the qualitative results of the main paper.

D. CLIP-EZPC Fidelity

We evaluate how closely EZPC reproduces the predictions of the original CLIP model. This measures the faithfulness of the learned concept space to the teacher model. We report the following metrics: *Top-1 agreement*, *Spearman rank correlation*, *Kendall rank correlation*, and *KL divergence*. Results are shown in Table 10.

Fidelity remains high across datasets, indicating that EZPC preserves the ranking structure of CLIP predictions while enabling concept-based explanations. Lower agreement on larger datasets, such as ImageNet-1k, reflects the increased semantic ambiguity rather than the failure of the concept model.

Table 10. **Prediction-level fidelity between CLIP and EZPC (RN50 backbone) with label consistency, ranking preservation, and KL.** Top-1 agreement measures label consistency. Spearman and Kendall correlations quantify ranking preservation. Kendall correlation is computed over the top-50 CLIP-ranked classes.

Dataset	Top-1 Agree. (%)	Spearman	Kendall	KL
CIFAR-100	80.26	0.959	0.733	6.79×10^{-6}
IN-100	92.92	0.994	0.904	6.10×10^{-6}
CUB	84.03	0.994	0.876	6.83×10^{-6}
ImageNet-1k	72.37	0.924	0.536	7.98×10^{-5}
Places365	79.92	0.972	0.715	1.46×10^{-5}

Table 11. **Effect of the reconstruction weighting parameter λ.** Larger λ values improve both zero-shot and generalized performance, with moderate to high settings giving the best results.

	Zero-shot		Generalized Zero-shot		
λ	Seen	Unseen	Seen	Unseen	H
0.01	0.377	0.508	0.347	0.371	0.358
0.1	0.654	0.820	0.626	0.633	0.630
1	0.699	0.851	0.675	0.690	0.682
10	0.707	0.859	0.681	0.709	0.695
100	0.706	0.857	0.680	0.704	0.692
1000	0.707	0.857	0.680	0.708	0.694

E. Additional Ablation Studies

E.1. Effect of λ.

Table 11 shows that larger λ values improve quantitative performance by emphasizing the reconstruction loss, which better preserves CLIP's similarity structure. However, Figure 12 reveals the opposite trend qualitatively: $\lambda = 1$ produces image-relevant concepts, whereas higher values (e.g., $\lambda = 100$) introduce unrelated activations. This trade-off is expected; smaller λ strengthens the matching loss, keeping learned concept directions closer to CLIP's concept embeddings and yielding more interpretable explanations.

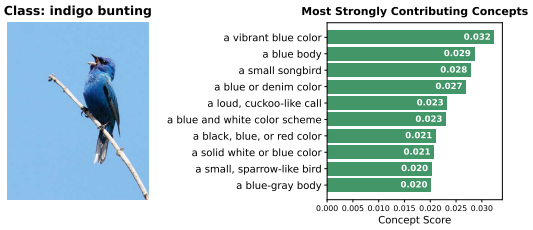
E.2. Impact of Training Objectives

We analyze the contribution of each component of the training objective used to learn the projection matrix A. Our method optimizes a combination of a matching loss $\mathcal{L}_{\text{match}}$ and a reconstruction loss $\mathcal{L}_{\text{recon}}$.
We compare the following settings:

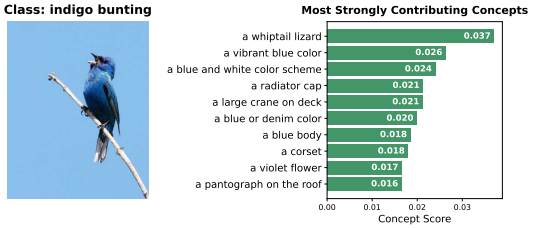
- No training ($A = \Phi$)
- Matching loss only
- Reconstruction loss only
- Full objective ($\mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{recon}}$)

Results calculated on ImageNet-100 using the generalized zero-shot setting are shown in Table 12.

Without any training ($A = \Phi$) or with the matching loss alone, the model achieves near-zero accuracy. This is ex-



(a) Image-level analysis for $\lambda = 1$.



(b) Image-level analysis for $\lambda = 100$.

Figure 12. **Qualitative comparison of image-level explanations for different λ values.** For $\lambda = 1$, EZPC produces semantically consistent concept activations. For larger values (e.g., $\lambda = 100$), unrelated concepts appear among the top activations.

Table 12. **Ablation study on the training objective of projection matrix A using RN50 on IN-100 (generalized zero-shot setting).**

Training Setting	Seen Acc.	Unseen Acc.	H
$A = \Phi$ (0-step, no training)	0.013	0.0	0.0
Matching loss only ($\mathcal{L}_{\text{match}}$)	0.013	0.0	0.0
Reconstruction loss only ($\mathcal{L}_{\text{recon}}$)	0.680	0.708	0.693
Full objective ($\mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{recon}}$)	0.674	0.690	0.682

pected as the raw CLIP concept embeddings do not form a basis that can reconstruct the original similarity structure, and the matching loss by itself only regularizes A toward Φ without learning to preserve predictive information. The reconstruction loss alone achieves the highest harmonic mean ($H = 0.693$), indicating that it is the primary driver of classification performance. Adding the matching loss in the full objective slightly reduces the harmonic mean to 0.682, but as shown in Figure 12, the matching loss plays a critical role in maintaining concept interpretability. Without it, learned concept directions drift from their original semantic meaning, producing less interpretable explanations despite higher quantitative scores. The full objective, therefore, represents a trade-off between predictive fidelity and explanation quality.

F. Additional Qualitative Visualizations

This section presents additional qualitative results that complement the analysis in the main paper, covering image-level, class-level, concept-clustering, and region-level alignment visualizations:

- **Image-level explanations (Figures 13 to 16)** Top activated concepts for individual predictions.
- **Class-level explanations (Figures 17 to 20)** Average concept activations across a random subset of images per class.
- **Concept clustering (Figures 21a to 21d)** Clusters of images that strongly activate the given concept.
- **Region-level concept alignment (Figures 22a to 22c)** Spatial localization of positive and negative concepts within individual images.

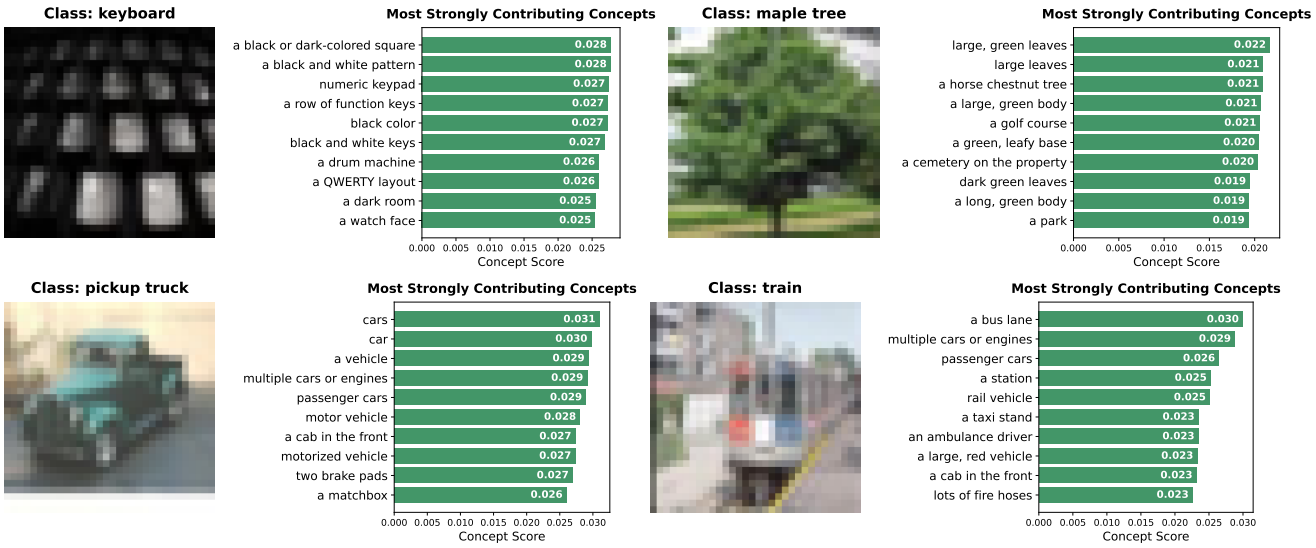


Figure 13. CIFAR-100 image-level explanations.

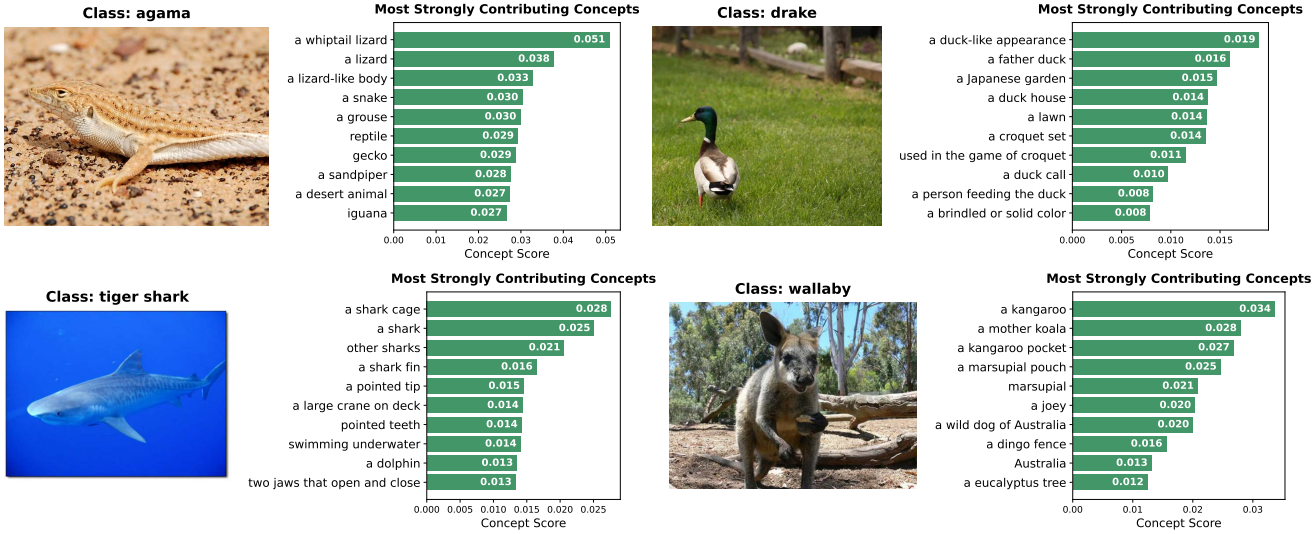


Figure 14. ImageNet-100 image-level explanations.

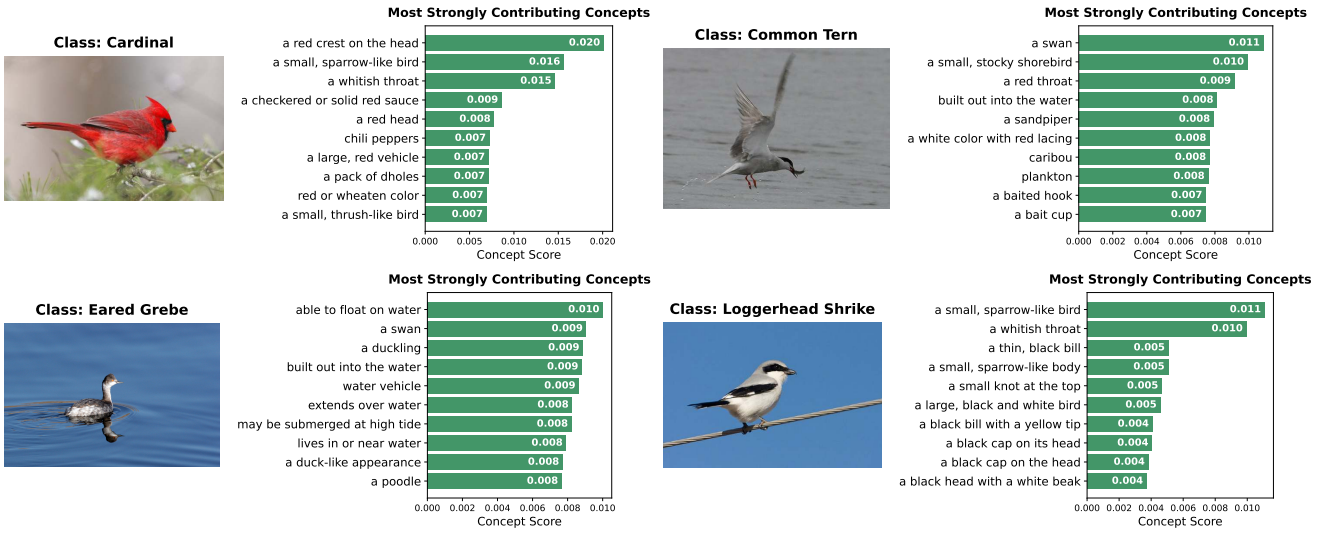


Figure 15. CUB image-level explanations.

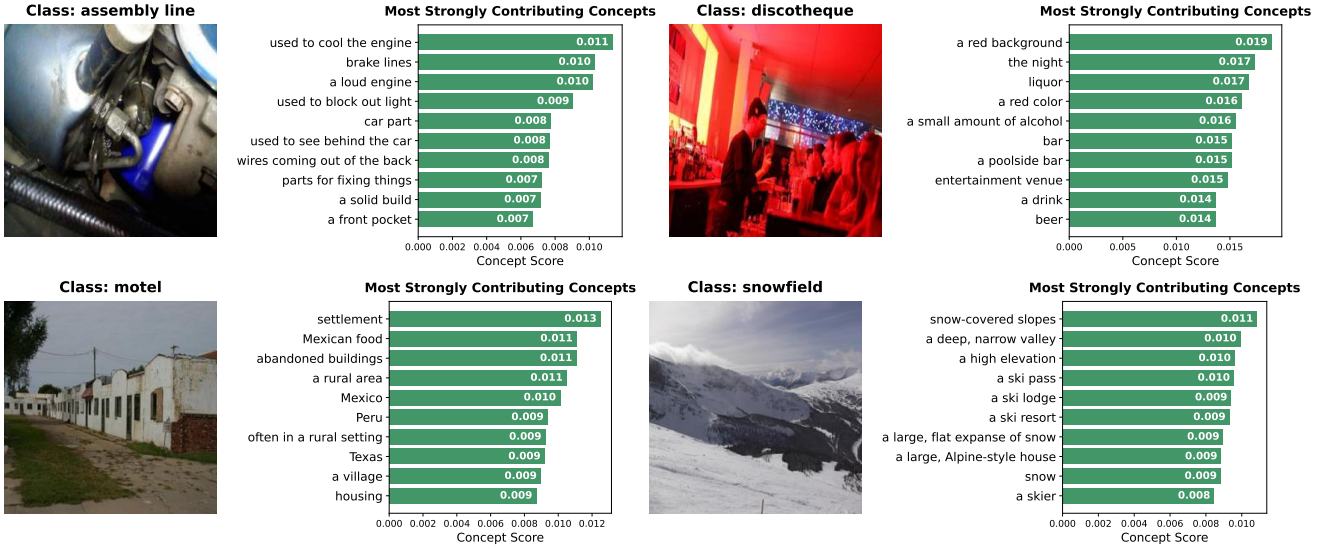


Figure 16. Places365 image-level explanations.

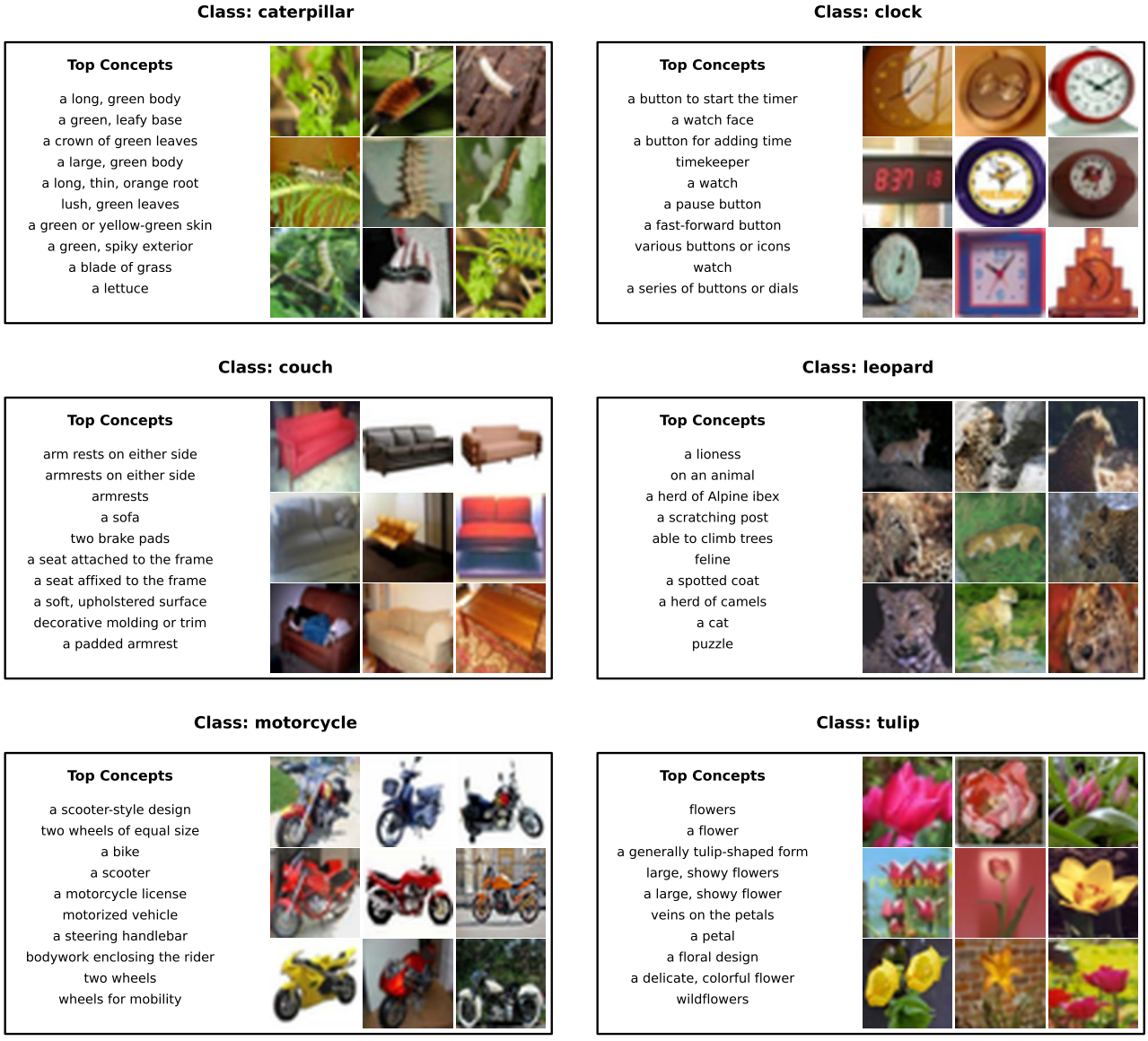


Figure 17. CIFAR-100 class-level visualizations.

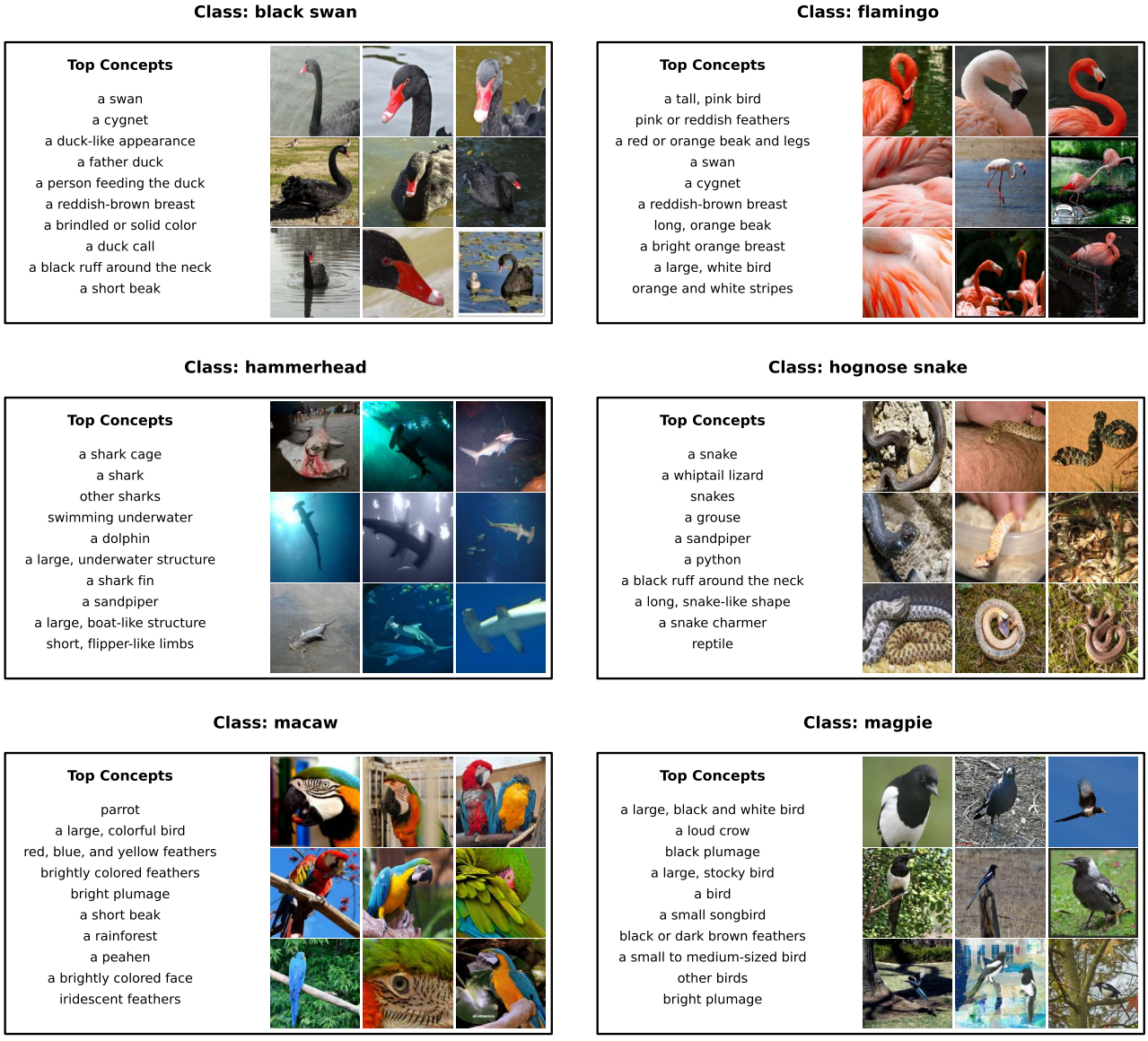


Figure 18. ImageNet-100 class-level visualizations.

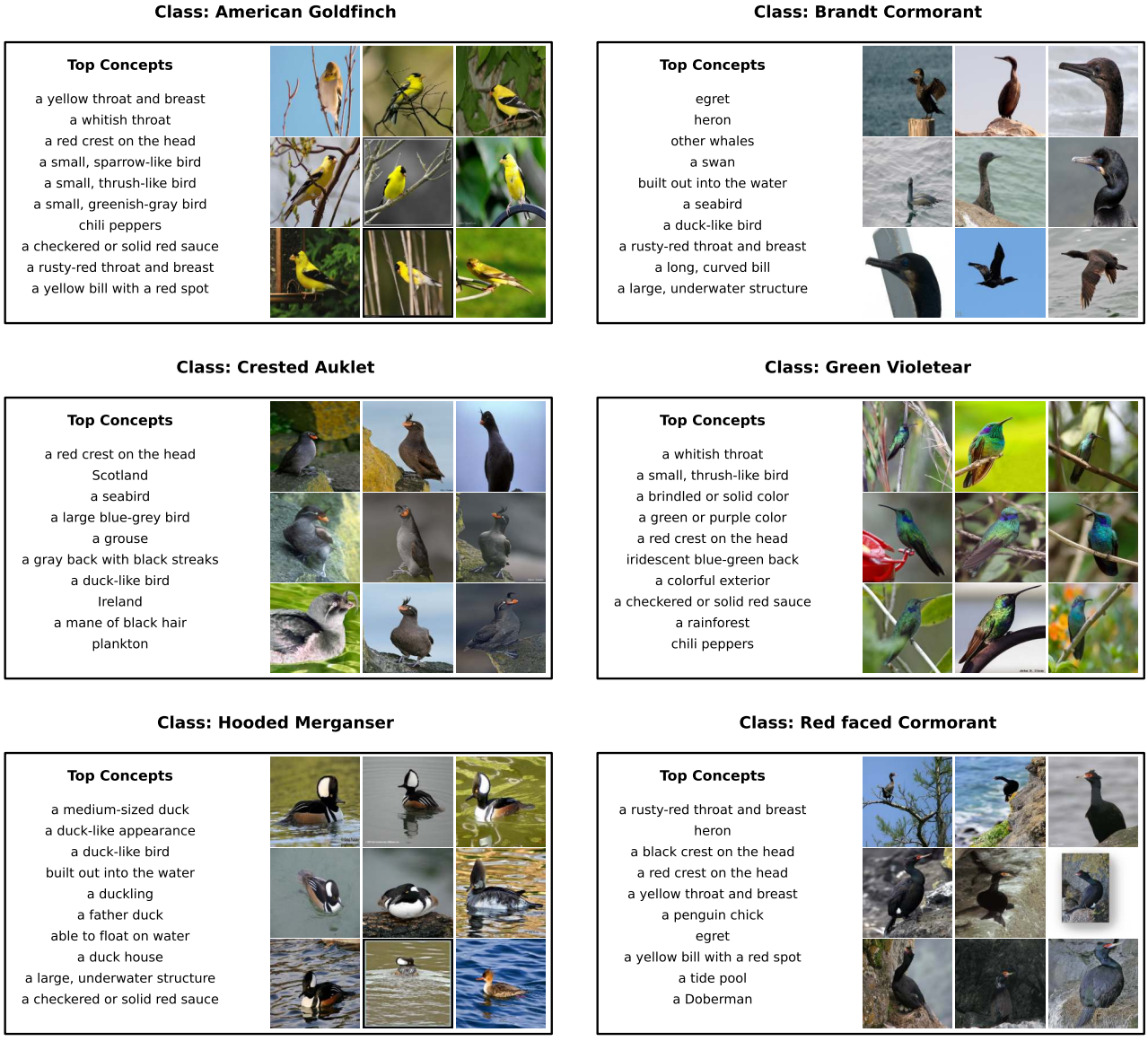


Figure 19. CUB class-level visualizations.

Class: amphitheater

Top Concepts

ancient architecture
ancient ruins
a pilgrimage site
a major pilgrimage site
a theater
Italian dish
lots of dirt and rubble
a scenic location
usually made of stone or brick
theater

Class: amusement arcade

Top Concepts

game
game center
a variety of games
a game
gambling
a place to play games
a video game console
a sports game
a slot on the top for coins
A toy

Class: bow window, indoor

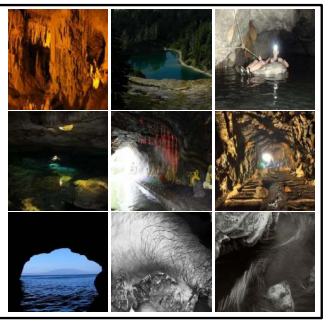
Top Concepts

a leather or fabric pouch
plants growing in the water
a room in a house or apartment
memorial plaque
large room with high ceilings
property
a flat, open area
indoor
a light, airy feel
paintings

Class: grotto

Top Concepts

a cave
a dark, hidden cave or recess
a canyon
a gorge
a deep, narrow valley
Ireland
a lightbulb at the top
a light at the top
a dark interior
under a rock



Class: skyscraper

Top Concepts

tall buildings
large buildings
a view of the cityscape
often surrounded by buildings
a city
a group of buildings
a skyline
china
high towers
place

Class: supermarket

Top Concepts

goods
stores
retailer
shopping
product displays
a brightly colored label
a checkout area
a place to store food
aisles between the shelves
a wide variety of merchandise

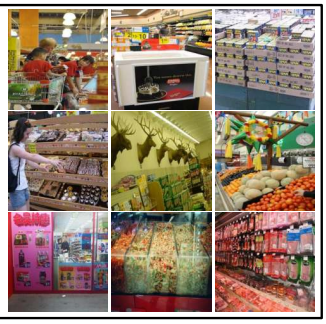
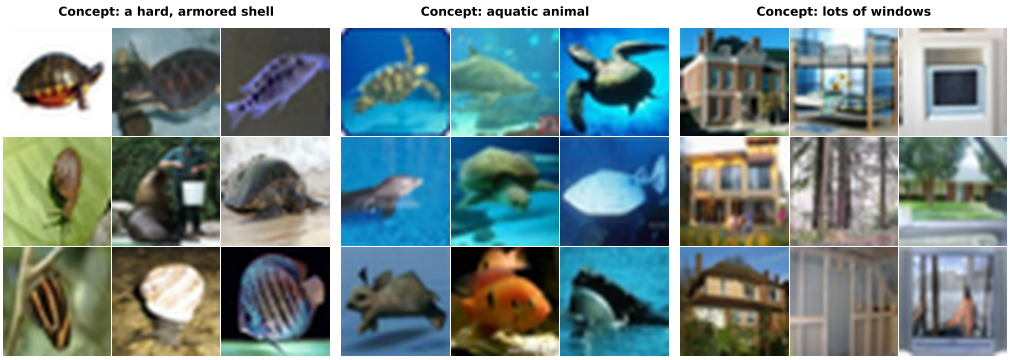


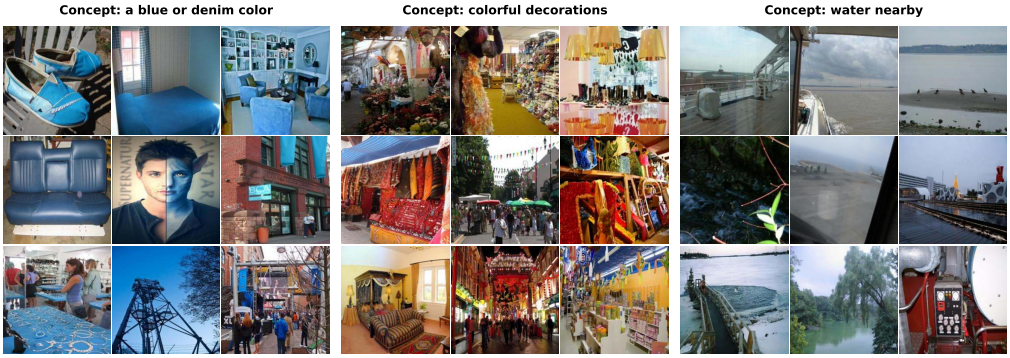
Figure 20. Places365 class-level visualizations.



(a) CIFAR-100 concept clustering examples.

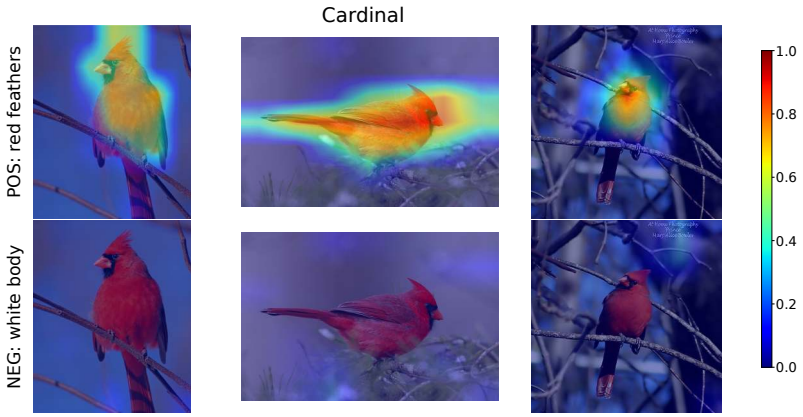
(b) ImageNet-100 concept clustering examples.

(c) CUB concept clustering examples.

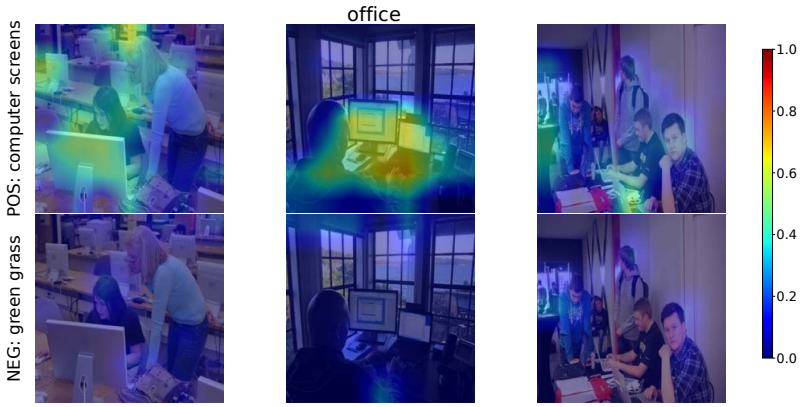


(d) Places365 concept clustering examples.

Figure 21. Concept clustering examples across datasets.



(a) **Region-level concept alignment for Cardinal class from CUB dataset.** Top row: positive concept (*red feathers*). Bottom row: negative concept (*white body*).



(b) **Region-level concept alignment for Office class from Places365 dataset.** Top row: positive concept (*computer screens*). Bottom row: negative concept (*green grass*).

(c) **Region-level concept alignment for Albatross class from ImageNet-100 dataset.** Top row: positive concept (*large wings*). Bottom row: negative concept (*yellow feathers*).

Figure 22. Region-level concept alignment examples across datasets.