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 have achieved remarkable success in zero-shot image recognition, yet their predictions remain largely opaque to human understanding. In contrast, Concept Bottleneck Models provide interpretable intermediate representations by reasoning through human-defined concepts, but they rely on concept supervision and lack the ability to generalize to unseen classes. We introduce EZPC that bridges these two paradigms by explaining CLIP's zero-shot predictions through human-understandable concepts. Our method projects CLIP's joint image-text embeddings into a concept space learned from language descriptions, enabling faithful and transparent explanations without additional supervision. The model learns this projection via a combination of alignment and reconstruction objectives, ensuring that concept activations preserve CLIP's semantic structure while remaining interpretable. Extensive experiments on five benchmark datasets, CIFAR-100, CUB-200-2011, Places365, ImageNet-100, and ImageNet-1k, demonstrate that our approach maintains CLIP's strong zero-shot classification accuracy while providing meaningful concept-level explanations. By grounding open-vocabulary predictions in explicit semantic concepts, our method offers a principled step toward interpretable and trustworthy vision-language models. Code is available at https://github.com/oonat/ezpc.

1. Introduction

The rapid integration of machine learning into real-world systems has intensified the demand for models that are not only accurate but also transparent and trustworthy. In high-stakes domains such as medical imaging, autonomous driving, and scientific discovery, understanding why a model makes a particular prediction is as critical as the prediction itself. Despite their impressive capabilities, modern deep networks remain largely black boxes, making it difficult to interpret their internal reasoning or diagnose their failures.

Concept Bottleneck Models (CBMs) [13] address this issue by introducing an intermediate layer composed of human-understandable concepts. These models decompose the prediction process into two stages: (1) mapping inputs to concept activations, and (2) predicting the final class label based on these activations. This structure provides an interpretable interface between perception and decision-making, allowing users to inspect, validate, or even modify concept activations to understand or correct model behavior.

However, classical CBMs are constrained by two major limitations. First, they require dense concept supervision, which is often expensive or infeasible to collect. Second, they operate under a closed-world assumption: CBMs are trained and evaluated on a fixed set of classes and thus fail to generalize to unseen categories or new domains. These constraints limit their scalability and applicability to open-vocabulary recognition problems. Recent efforts to mitigate these issues by 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 (VLMs) such as CLIP [24], ALIGN [10], and SigLIP [38] demonstrate strong open-vocabulary generalization by aligning images and text in a shared semantic space. CLIP, for instance, learns to associate visual and textual information at scale, enabling zero-shot classification by comparing an image embedding with textual embeddings of candidate labels. Without task-specific training, CLIP can accurately recognize objects from natural-language descriptions, which is a significant leap toward flexible, general-purpose perception.

Yet, this generalization comes at the cost of interpretability. CLIP's embeddings are high-dimensional and entangled, offering little insight into what visual or semantic properties drive a particular decision. As a result, users cannot eas-

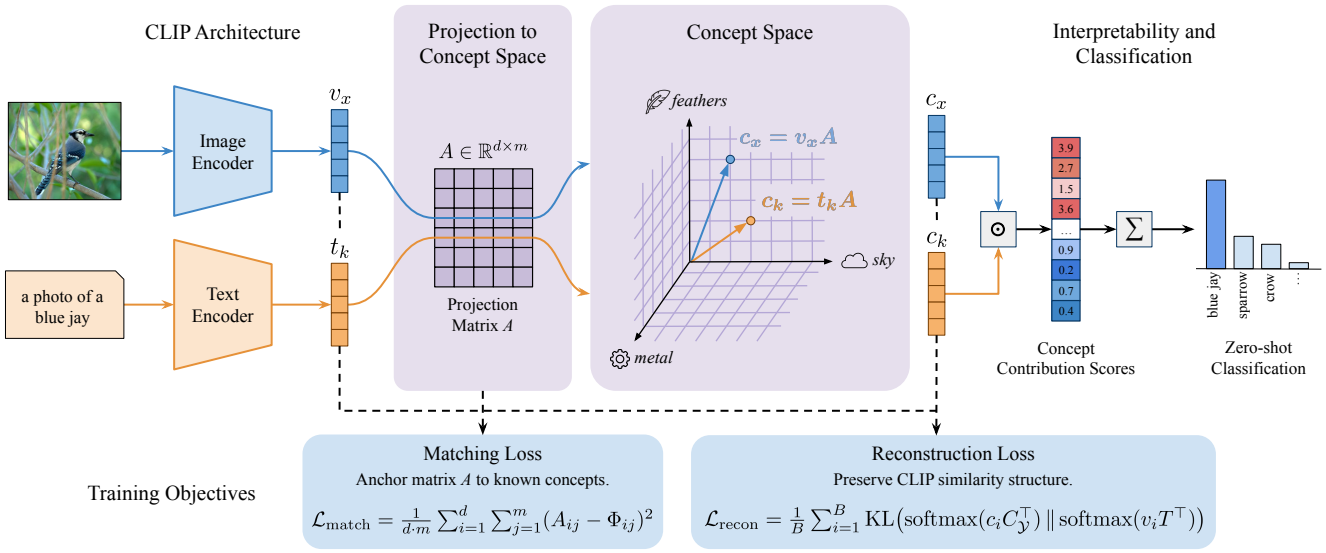


Figure 1. **Overview of EZPC.** CLIP image and text embeddings are projected into a shared concept space using a learnable matrix A. The projected representations c_x and c_k provide (i) concept-based explanations via a Hadamard product and (ii) class logits via a dot-product in concept space. Training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

ily understand why CLIP associates an image with a given label, nor can they trace these predictions back to human-interpretable reasoning.

In this paper, we aim to bridge the gap between the interpretability of CBMs and the generalization ability of CLIP. We propose a method that explains CLIP's zero-shot predictions through human-understandable concepts. Instead of retraining CLIP or relying on annotated concepts, our method introduces a lightweight decomposition that projects CLIP's image-text embeddings into a shared concept space. This enables faithful, concept-level explanations of CLIP's predictions while maintaining its zero-shot capabilities.

Our method, "Explaining CLIP Zero-shot Predictions Through Concepts" (**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. The resulting model not only interprets CLIP's predictions through meaningful concepts but also retains high zero-shot accuracy across diverse datasets.

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.

2. Related Work

Our work lies at the intersection of zero-shot learning, vision-language modeling, and concept-based interpretability. Below, we review the most relevant research in each area and discuss how our approach relates to prior work.

Zero-shot Learning. Zero-shot learning (ZSL) aims to recognize unseen categories without explicit training data for those classes. Early work achieved this by leveraging human-defined attributes as semantic intermediaries [6, 15, 16], enabling recognition of novel classes through shared attributes. Later approaches replaced attributes with distributed word embeddings such as Word2Vec, GloVe, and BERT to establish semantic correspondences between visual and linguistic spaces [7, 20, 23].

Embedding-based ZSL methods [1, 8, 21, 26] projected both images and labels into a shared latent space, where similarity was computed for classification. Other approaches [5, 11, 18, 31] instead aim to learn weights for zero-shot classification. More recently, large-scale vision-language models such as CLIP [24], ALIGN [10], and SigLIP [38] have shown remarkable zero-shot generalization by aligning visual and textual modalities through contrastive learning on massive image-text datasets. These models removed the need for explicit semantic attribute design and established a new paradigm for open-vocabulary recognition.

Vision-Language Models. Vision-language models jointly learn image and text representations in a shared embedding space. CLIP's contrastive learning objective allows image embeddings to align with corresponding text embeddings, enabling zero-shot recognition of arbitrary categories described in natural language. Subsequent works such as CoOp [40],

BLIP [17], and PaLI-Gemma [2] expanded the scale and adaptability of VLMs, incorporating prompt tuning, caption generation, and cross-modal retrieval.

Beyond classification, VLMs have become a powerful foundation for interpretability studies. Their multimodal embeddings naturally encode high-level semantics that can be probed with linguistic queries or concept prompts [22, 25]. However, these models lack explicit interpretability mechanisms; while they capture semantics implicitly, the internal representations remain opaque and difficult to translate into human-understandable concepts.

Concept Bottleneck Models. Concept Bottleneck Models [13] improve interpretability by decomposing predictions into human-defined concepts. A CBM first predicts interpretable attributes (e.g., *has wings, is red*) and then uses these attributes for downstream classification. This design provides transparency and controllability: users can inspect or modify the concept activations to understand and alter model decisions.

Subsequent works have enhanced CBMs by improving their robustness [12, 28, 34, 37], discovering latent concepts automatically [25], or integrating textual guidance from large language models [22, 29, 35, 36]. Despite these advances, most CBMs remain limited to closed-world settings, requiring concept supervision and predefined label sets.

CLIP Interpretability and Zero-shot CBMs. Recent works [19, 33] have sought to merge CLIP's open-vocabulary recognition with the interpretability of concept-based reasoning. Gandelsman et al. [9] decomposes CLIP's image encoder across patches, layers, and attention heads, using CLIP's text encoder to interpret component-wise contributions. SpLiCE [3] decomposes CLIP's embeddings into sparse linear combinations of concept vectors, producing per-instance concept explanations but at the cost of expensive test-time optimization for each image. Z-CBM [32] extends the CBM paradigm to the zero-shot setting by reconstructing embeddings from concept banks, yet it depends on large concept repositories and costly regressions.

In contrast, EZPC introduces a unified, trainable decomposition that aligns CLIP's vision-language space with a shared, interpretable concept basis. Unlike prior instance-specific or retrieval-based methods, EZPC learns a single projection that jointly preserves semantic alignment and interpretability, yielding efficient, faithful explanations of zero-shot predictions.

3. Method

We propose a method that explains CLIP's zero-shot predictions through a human-interpretable concept space. Our method learns a linear decomposition of CLIP's joint image-text embeddings into concept activations. This decomposition preserves the semantic relationships inherent in CLIP's representation while making them interpretable.

3.1. 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 3.2.

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.

3.2. 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:

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

where $c_i = v_i A$ and $C_{\mathcal{Y}} = TA$ are the concept activations defined in Section 3.1. 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:

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

3.3. 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 the concept scores compose the prediction logit directly, the explanations provided by EZPC are faithful to the model's decision process by construction. Thus, the model not only classifies unseen classes in a zero-shot manner but also provides fine-grained, faithful explanations identifying which concepts drive its decisions.

4. Experiments

Datasets. We evaluate our approach on five benchmark datasets covering diverse visual domains and levels of granularity: 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, ensuring that unseen categories are completely excluded during training. This setup enables evaluation under both the zero-shot (unseen-only) and generalized zero-shot (joint seen-unseen) settings.

The chosen datasets span a wide range of visual abstraction levels: CIFAR-100 and ImageNet-100 capture compact, object-centric imagery; CUB offers fine-grained bird species classification; ImageNet-1k provides large-scale diversity; and Places365 focuses on scene-level understanding. This diversity allows us to assess the interpretability and generalization of EZPC across varying domains.

Concept space. We adopt the human-interpretable concept sets introduced by LF-CBM [22], where concepts were originally generated using GPT-3 [4] based on the class names in each dataset. Specifically, the concept vocabulary sizes are as follows: CIFAR-100 (*892 concepts*), CUB (*370 concepts*), ImageNet-1k (*4,751 concepts*), and Places365 (*2,544 concepts*). Since CIFAR-100, CUB, and Places365 contain relatively few concepts, we merge their original concept vocabularies with ImageNet-1k's larger concept pool to obtain a richer and more transferable set of interpretable attributes. This unified concept space ensures consistent coverage of visual semantics across datasets and provides a fair basis for cross-domain comparison in our zero-shot evaluation.

Evaluation Metrics. In the standard zero-shot setting, given an image x and a set of class names $\mathcal{Y} = \{y_1, y_2, \dots, y_K\}$, the prediction is made by selecting the class whose text embedding has the highest cosine similarity with the image embedding

$$\hat{y} = \underset{y_k \in \mathcal{Y}}{\arg\max} \langle v_x, t_k \rangle, \tag{9}$$

where v_x and t_k denote the normalized CLIP embeddings for the image and text, respectively.

In the generalized zero-shot (GZS) setting, the label space includes both seen and unseen classes, $\mathcal{Y}_{\text{GZS}} = \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$. We report accuracies on seen (Acc_S) and unseen (Acc_U) classes, and use their harmonic mean (H) as a balanced measure which penalizes models that overfit to either seen or unseen categories and thus reflects overall generalization capability.

Baselines. We compare EZPC against three baselines: (1) the original CLIP [24], (2) the zero-shot concept bottleneck variant Z-CBM [32], and (3) the sparse linear decomposition approach SpLiCE [3]. All models are evaluated with identical backbones to ensure comparability using their official implementations with default hyperparameters.

For all experiments, except the backbone ablation, we use the CLIP RN50 backbone, as it is widely adopted in prior work. Unless otherwise noted, ablations are conducted on ImageNet-100, which provides a convenient testbed for evaluating different settings. We use $\lambda = 1$ for all datasets, except CUB and Places365, where $\lambda = 5$ gives better quantitative and qualitative results.

4.1. Quantitative Analysis

Results. Table 1 reports generalized zero-shot performance across the five datasets. On CIFAR-100, ImageNet-100, and CUB, EZPC remains within roughly 1% harmonic mean of CLIP, showing that most of CLIP's accuracy is retained despite the addition of an interpretable concept layer. On ImageNet-1k, the gap is larger (around 5%), reflecting the increased difficulty of this large-scale setting. For Places365, EZPC is about 1% below CLIP and close to Z-CBM. Across the remaining datasets, EZPC provides clear improvements

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

over Z-CBM and SpLiCE, often exceeding them by 10-15% in harmonic mean. Overall, these results demonstrate that EZPC offers meaningful interpretability while keeping performance competitive with CLIP and substantially stronger than prior concept-based baselines.

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.

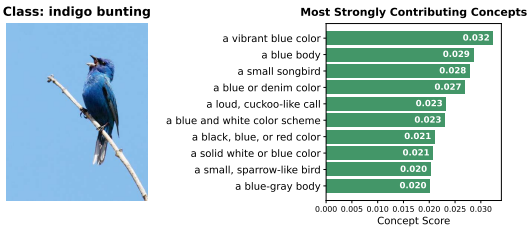
Effect of λ. Table 3 shows that larger λ values improve quantitative performance by emphasizing the reconstruction loss, which better preserves CLIP's similarity structure. However, Figure 2 reveals the opposite trend qualitatively: $\lambda = 1$ produces image-relevant concepts, whereas higher values (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.

4.2. Cross-Dataset Experiments

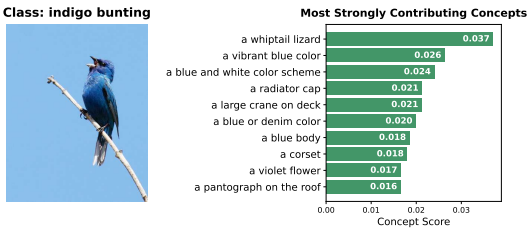
To rigorously assess the generalization ability of our model beyond its training distribution, we perform cross-dataset experiments where the concept bottleneck is trained on a source dataset (ImageNet-100) and evaluated on distinct target datasets (CIFAR-100 and CUB) without any fine-tuning. This setup examines whether the learned projection matrix A captures transferable semantics that remain valid across domains with different visual granularity and label taxonomies.

Table 3. **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



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



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

Figure 2. **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.

Setup. During evaluation, we treat classes from the source dataset (e.g., ImageNet-100) as *seen* and 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 evalu-

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. *Seen* classes correspond to ImageNet-100 categories, while *unseen* classes correspond to the target dataset.

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

ated 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.

4.3. Qualitative Analysis

4.3.1. Image-level Explanations

For image-level explanations, EZPC follows a straightforward procedure: it projects both the image embedding and the predicted class label embedding into the concept space, computes their element-wise product, and selects the top-10 activated concepts that drive the zero-shot prediction. Figure 3a shows examples from CIFAR-100: for the class *sea*, top concepts include *a beach*, *reefs*, and *a hammock*, while *wardrobe* activates *a wooden or plastic body* and *wood*. Figure 3b shows Places365 examples, 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 visual content of the specific image: for *sea*, concepts like *a hammock* and *a paddleboard* are related to the class but not visible in the image.

4.3.2. Class-level Explanations

To obtain class-level explanations, we randomly sample nine images from a target class and compute the mean concept activations across these images. This averaged concept profile highlights the concepts that are most representative of the class as a whole, rather than a single instance. Figure 4 presents such examples from CUB and ImageNet-100. For instance, the CUB class *Cardinal* activates concepts such as *a red crest on the head* and *a red head*, while the ImageNet-100 class *Lorikeet* highlights *parrot*, *red, blue, and yellow feathers*, and *brightly colored feathers*. However, not all activated concepts are meaningful: for Cardinal, concepts such as *heavy build*, *a hard, red exterior*, and *a pack of dholes* appear irrelevant. Such cases may stem from noise in CLIP's embedding space, limitations of the concept vocabulary, etc.

4.3.3. Concept Clustering

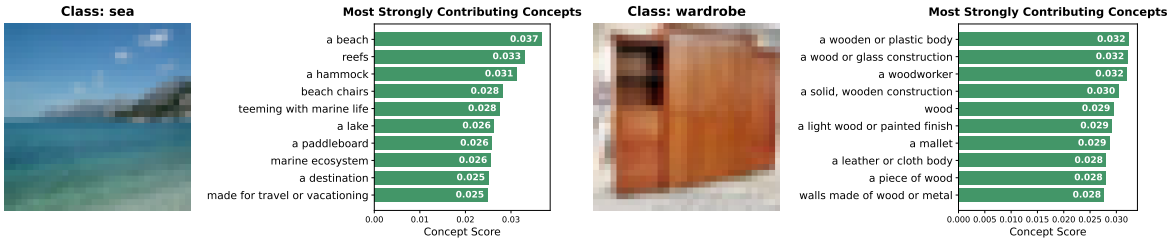
We further analyze the semantic structure of the learned concept space by performing concept-based image retrieval. For a given concept, we compute the image-side concept activation for every image in the dataset and retrieve the nine images with the highest activation. This procedure reveals whether individual concept dimensions correspond to visually coherent directions in the embedding space. Figure 5 shows example clusters from ImageNet-100 and Places365. For the concept *a red background*, the retrieved images contain prominent red tones; *large wings* retrieves bird images; and *large buildings* retrieves architectural scenes. The strong thematic consistency across retrieved images confirms that individual concept dimensions capture interpretable and visually grounded semantics, rather than encoding arbitrary or entangled directions in CLIP's latent space.

4.3.4. Concept-Region Alignment

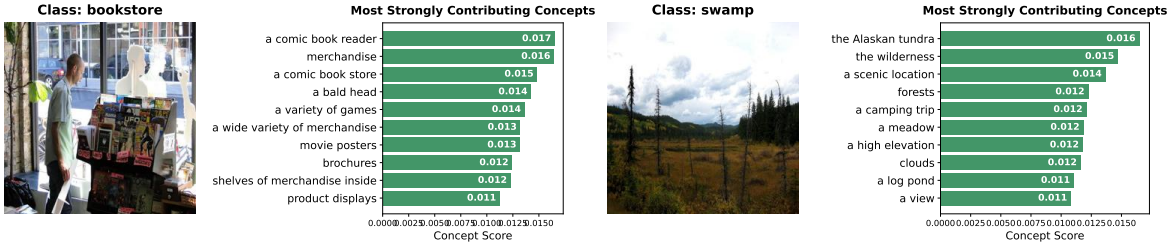
To evaluate whether the learned concept space produces spatially meaningful explanations, we analyze region-level alignment between concept activations and object locations. We extract patch-level features from the CLIP RN50 backbone, project them into the concept space via A, and visualize the resulting spatial activation maps. Figure 6 shows an example on CUB, where 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.

Figure 6. **Region-level concept alignment for Indigo Bunting.** Top row: positive concept (*a blue-gray body*). Bottom row: negative concept (*a red face*).

We further quantify localization quality on the Indigo Bunting class from CUB using ground-truth segmentation masks. We evaluate a positive concept (*a blue-gray body*) and a negative concept (*a red face*) across all images of this class, computing *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.



(a) CIFAR-100 image-level explanations.



(b) Places365 image-level explanations.

Figure 3. **Image-level Explanations.** For each image, EZPC displays the top-10 activated concepts that contribute most to the zero-shot prediction. The highlighted concepts closely correspond to salient visual characteristics of the input images.

(a) CUB class-level concept explanations.

(b) ImageNet-100 class-level concept explanations.

Figure 4. **Class-level Concept Explanations.** For each class, we average concept activations over nine sampled images. EZPC produces coherent class signatures, highlighting concepts that characterize each category.

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.

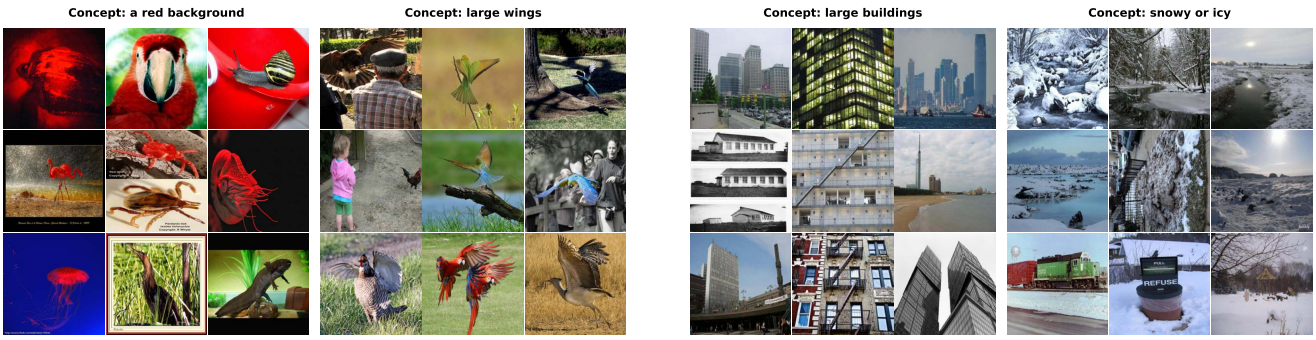
Metric	Positive Concept	Negative Concept
Pointing Accuracy ↑	0.967 ± 0.180	0.017 ± 0.128
Inside Activation Ratio ↑	0.507 ± 0.191	0.031 ± 0.054
IoU@10% ↑	0.423 ± 0.159	0.019 ± 0.067
IoU@20% ↑	0.408 ± 0.148	0.044 ± 0.087

4.4. Time Analysis

One of the key advantages of EZPC is its computational efficiency. Unlike optimization-based decomposition approaches such as SpLiCE [3] and Z-CBM [32], which require solving an optimization problem per image at inference time, our method performs a single linear projection to compute concept activations. Table 6 reports the processing time per image for four representative methods evaluated on the ImageNet-100 validation set using a single NVIDIA H100 GPU. While optimization-based methods require iterative solvers at inference time, EZPC introduces only a lightweight matrix multiplication ($v_x A$) on top of CLIP's forward pass, keeping inference latency almost identical to CLIP. 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 suitable for large-scale deployment and interactive analysis.

5. Discussion

EZPC demonstrates that interpretable concept-based reasoning can coexist with strong zero-shot recognition performance. By decomposing CLIP's vision-language embeddings into a shared concept space, EZPC reveals the underlying semantic structure that drives model predictions.



(a) ImageNet-100 concept clustering results.

(b) Places365 concept clustering results.

Figure 5. **Concept Clustering Results.** For each concept, we retrieve the nine images with the highest activation. Clusters from ImageNet-100 and Places365 show coherent semantic structure, indicating that EZPC learns interpretable concept directions.

Table 6. **Inference time comparison** on ImageNet-100 (NVIDIA H100). We report median per-image latency (ms) with 95% confidence intervals. EZPC adds negligible overhead ($p = 0.31$, Wilcoxon signed-rank test), while SpLiCE and Z-CBM are $59\times$ and $94\times$ slower due to per-image optimization and retrieval.

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	$\sim 1.0\times$

Interpretability and Faithfulness. EZPC attributes each prediction to explicit, human-understandable concepts, offering transparent reasoning paths for both seen and unseen categories. Unlike saliency-based or feature-attribution methods, which provide only localized visual cues, our concept decomposition expresses CLIP's similarity judgments in terms of global semantic concepts. This enables human inspection of the decision process, as users can identify which concepts most strongly influence each classification.

Generalization Across Domains. The method generalizes effectively across object-centric, fine-grained, and scene-centric datasets. Results on CIFAR-100 and ImageNet-100 show that the concept bottleneck retains nearly all of CLIP's discriminative power. Meanwhile, performance on CUB and Places365 demonstrates that even in challenging fine-grained or context-heavy domains, concept activations remain semantically aligned. This suggests that CLIP's internal representations inherently encode human-aligned structures that can be disentangled through our learned concept projection.

Comparison to Prior Work. In contrast to instance-specific methods such as SpLiCE [3], which require costly per-sample optimization, EZPC learns a single unified concept projection that applies to all data. Compared to retrieval-based approaches such as Z-CBM [32], it avoids concept bank search stages, reducing computational overhead. These properties make EZPC suitable for large-scale and cross-domain zero-shot applications.

Limitations and Future Directions. For completeness, we describe some limitations of EZPC. First, the linear projection assumption constrains expressive power; highly non-linear semantic relationships may not be fully captured in the concept space. Second, the interpretability depends on the quality and diversity of the concept set, biases in the concept set can affect the fidelity of explanations. Third, the current model focuses on classification; extending the approach to multimodal reasoning tasks remains an open question.

Future work could explore non-linear concept mappings, adaptive concept discovery, and integration with language models to dynamically expand the concept vocabulary. These directions could further strengthen the interpretability and faithfulness of zero-shot vision-language systems.

6. Conclusion

We introduced the EZPC, a method that explains CLIP's zero-shot predictions through human-interpretable concepts. By learning a shared concept projection that aligns image and text embeddings, EZPC enables transparent, concept-level reasoning without requiring additional supervision. Through matching and reconstruction objectives, the model preserves CLIP's semantic structure while exposing the underlying conceptual basis of its predictions.

Across five benchmarks, CIFAR-100, CUB, ImageNet-100, ImageNet-1k, and Places365, EZPC maintains strong zero-shot accuracy comparable to CLIP, while providing meaningful explanations at both the image and class levels. Qualitative analyses reveal that the learned concept vectors form coherent semantic clusters, reflecting interpretable structure within CLIP's embedding space.

Our results demonstrate that open-vocabulary recognition and interpretability need not be mutually exclusive. EZPC bridges these paradigms, offering a scalable path toward trustworthy vision-language systems. Future extensions may incorporate adaptive or hierarchical concept spaces, further deepening our understanding of the semantic organization within large-scale multimodal models.

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.