Concepts from Representations: Post-hoc Concept Bottleneck Models via Sparse Decomposition of Visual Representations

Shizhan Gong[1], **Xiaofan Zhang**[2], **Qi Dou**[1]

[1]The Chinese University of Hong Kong, Hong Kong, China
[2]Shanghai Jiao Tong University, Shanghai, China
szgong22@cse.cuhk.edu.hk, xiaofan.zhang@sjtu.edu.cn, qidou@cuhk.edu.hk

Abstract

Deep learning has achieved remarkable success in image recognition, yet their inherent opacity poses challenges for deployment in critical domains. Concept-based interpretations aim to address this by explaining model reasoning through human-understandable concepts. However, existing post-hoc methods and ante-hoc concept bottleneck models (CBMs), suffer from limitations such as unreliable concept relevance, non-visual or labor-intensive concept definitions, and model/data-agnostic assumptions. This paper introduces **P**ost-hoc **C**oncept **B**ottleneck **M**odel via **Re**presentation **D**ecomposition (**PCBM-ReD**), a novel pipeline that retrofits interpretability onto pretrained opaque models. PCBM-ReD automatically extracts visual concepts from a pre-trained encoder, employs multimodal large language models (MLLMs) to label and filter concepts based on visual identifiability and task relevance, and selects an independent subset via reconstruction-guided optimization. Leveraging CLIP's visual-text alignment, it decomposes image representations into linear combination of concept embeddings to fit into the CBMs abstraction. Extensive experiments across 11 image classification tasks show PCBM-ReD achieves state-of-the-art accuracy, narrows the performance gap with end-to-end models, and exhibits better interpretability.

Code — https://github.com/peterant330/PCBM_ReD

1 Introduction

Deep learning has made significant strides in various image recognition tasks. However, the complexity of these models, which is often necessary to achieve high accuracy, leads to an opaque behavior. This limits the broader application in critical fields such as medical imaging analysis (Gong et al. 2025b) and autonomous driving (Omeiza et al. 2021). Concept-based interpretation is a subfield of explainable artificial intelligence that aims to use human understandable concepts to explain the model behaviors. Concept-based methods usually represent the semantic features learned by the network with concept labels, so that human can understand which features are responsible for the final predictions.

Concept-based methods can be divided into post-hoc methods and ante-hoc approaches (Fig. 1). Post-hoc methods associate latent space representations to human-understandable concepts, either supervised with user-defined concepts (Bau et al. 2017; Fong and Vedaldi 2018) or by unsupervised pattern recognition (Ge et al. 2021; Vielhaben, Bluecher, and Strodthoff 2023). However, these methods face some major limitations. Firstly, there is no guarantee that the extracted concepts faithfully reflect the reasoning process of the network. Additionally, there may be no intuitive and human-understandable causal relationship between the extracted concepts and targets, making it difficult for users to understand the mechanism of the network. Moreover, automated concepts mining process can generate a great number of concepts with low semantic value, which are difficult to be labeled by humans (Kim et al. 2018).

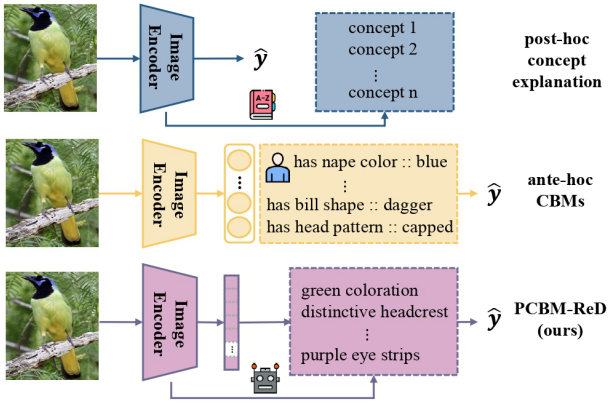


Figure 1: PCBM-ReD extracts concepts from the pre-trained image encoder and reconstruct the visual representation with the concepts, which gives faithful interpretation and can take the best advantage of the encoder's representation power.

Concept Bottleneck Models (CBMs) (Koh et al. 2020) are ante-hoc approaches that integrate concept associations into neural networks through modifications in model design or training. They predict a set of interpretable concepts and use a linear function to generate final predictions based on these concepts. This structure allows users to trace the links between concepts and class labels, facilitating error correction. CBMs rely on user-defined concepts, which can be either manually crafted by experts (Koh et al. 2020; Yun et al.

2022) or generated by large language models (LLMs) (Yang et al. 2023; Oikarinen et al. 2023). However, the current CBMs exhibit several weaknesses: human-crafted concepts can be labor-intensive and may lack comprehensive coverage (Koh et al. 2020), while LLMs-generated concepts often include non-visual features (e.g., the taste of food or bird behavior, which are not visually inferred) (Yang et al. 2023). Additionally, these pre-defined concepts are data-independent and model-agnostic, which can hinder their effectiveness, especially if a dataset is biased towards certain subpopulation or if the image encoder cannot capture certain features (e.g., color or spatial relationships (Kamath, Hessel, and Chang 2023)). Furthermore, current methods do not guarantee the independence of generated concepts, which is essential for effective interventions in CBMs. These limitations hinder CBMs from achieving complete feature space interpretability (Kim et al. 2018).

In this work, we introduce a novel pipeline that retrofits interpretability onto pretrained opaque foundation models, called **P**ost-hoc **C**oncept **B**ottleneck **M**odel via **Re**presentation **D**ecomposition (**PCBM-ReD**). Given a pretrained image encoder, we first apply automatic concept extraction (Fel et al. 2023) to mine the generalizable features encoded within the foundation models. To further ensure they are visually-identifiable, human-understandable, and task-related, we use multimodal large language models (MLLMs) to label concepts by summarizing descriptions of top-activated images for each concept and scoring them based on prior knowledge of the task. We then propose a reconstruction-guided concept selection algorithm that selects a subset of concepts, whose embedding spans are independent and completely define the visual representation space. For image-concept association, we utilize the visual-text alignment property of CLIP (Radford et al. 2021), which allows us to decompose the visual representation into a weighted sum of the concepts' text embeddings. By eliminating residual terms and fitting a linear function to the reconstructed representation, we can develop a model that adheres to the abstractions of CBMs while preserving the representational power of the original opaque visual encoder.

We conduct comprehensive experiments to demonstrate the efficacy of our proposed method. We evaluate the model performance on 11 image classification tasks, including common object recognition, fine-grained types, texture, and action classification, as well as domain-specific tasks such as medical diagnosis and satellite image object recognition. Our main finding is that the model achieves state-of-the-art classification accuracy, with a reduced gap compared to end-to-end models, compared to existing CBMs. Furthermore, as we aim to mimic the behavior of the end-to-end model using CBM, our approach exhibits similar properties, including zero-shot and few-shot capabilities. Additionally, we conduct human evaluation to demonstrate the improved interpretability of PCBM-ReD. Our main contributions include:

- We propose a data-driven scheme for creating and selecting concepts that align with the data distribution and the image encoder's representation capabilities.
- We propose constructing CBMs by leveraging CLIP's visual-text alignment to sparsely decompose the visual representation into concept embeddings.
- We demonstrate that PCBM-ReD achieves SOTA classification accuracy across various tasks, along with better interpretability and robust zero/few-shot capabilities.

2 Related Work

Explanations in Computer Vision. Explainable computer vision primarily falls into two categories: post-hoc explanations and interpretable models by design. Post-hoc explanation methods generate saliency maps to identify which input features influence the decisions of neural networks (Simonyan, Vedaldi, and Zisserman 2013; Selvaraju et al. 2017; Gong, Dou, and Farnia 2024; Gong et al. 2025a). However, these methods do not guarantee that the explanations faithfully reflect the model's reasoning process (Rudin 2019). In contrast, interpretable models by design ensure that explanations align with the models' reasoning (Chen et al. 2019; Nauta, Van Bree, and Seifert 2021; Ma et al. 2024; Tan, Zhou, and Chen 2024b). Our work builds upon CBMs (Koh et al. 2020), a type of interpretable models.

Concept Bottleneck Models. CBMs (Koh et al. 2020; Zarlenga et al. 2022; Kim et al. 2023; Gong et al. 2025b) predict outcomes by linearly combining an intermediate layer of human-understandable attributes. The original CBM relies on handcrafted and manually annotated attributes. CompDL (Yun et al. 2022) replaces manual annotations with CLIP scores, but still depends on concepts designed by human. LaBo (Yang et al. 2023) and label-free CBM (Oikarinen et al. 2023) further automate concept generation using LLMs. Early CBMs often underperformed compared to end-to-end models. To improve it, Post-hoc CBM (Yuksekgonul, Wang, and Zou 2022) introduces a residual connection from image features to predictions, while Res-CBM (Shang et al. 2024) approximates this connection by incrementally adding new concepts. OpenCBM (Tan, Zhou, and Chen 2024a) detects missing concepts from an open vocabulary. Our method reduces residuals from the beginning by mining concepts that can well reconstruct the image representations.

CLIP Interpretation. CLIP (Radford et al. 2021) utilizes contrastive learning on a large dataset of paired text and images to link images with their textual descriptions. This approach effectively groups similar concepts together and shows strong performance in downstream tasks, such as zero-shot classifications (Saha, Van Horn, and Maji 2024). Several studies focusing on the interpretation of CLIP have shown that its homogeneous feature space allows the visual embeddings to be decomposed into multiple concepts represented by concepts' text embeddings (Moayeri et al. 2023; Chen et al. 2023; Gandelsman, Efros, and Steinhardt 2023, 2024). This provides the theoretical foundation for our method, which constructs CBMs by directly decomposing visual representations into concepts representation, thus better preserving predictive power.

3 Method

Consider a training set of image-label pairs $\mathcal{D} = \{(\mathbf{x}, y) \in \mathbb{R}^{d_x} \times \mathcal{Y}\}$, and a bottleneck C made of N_C concepts,

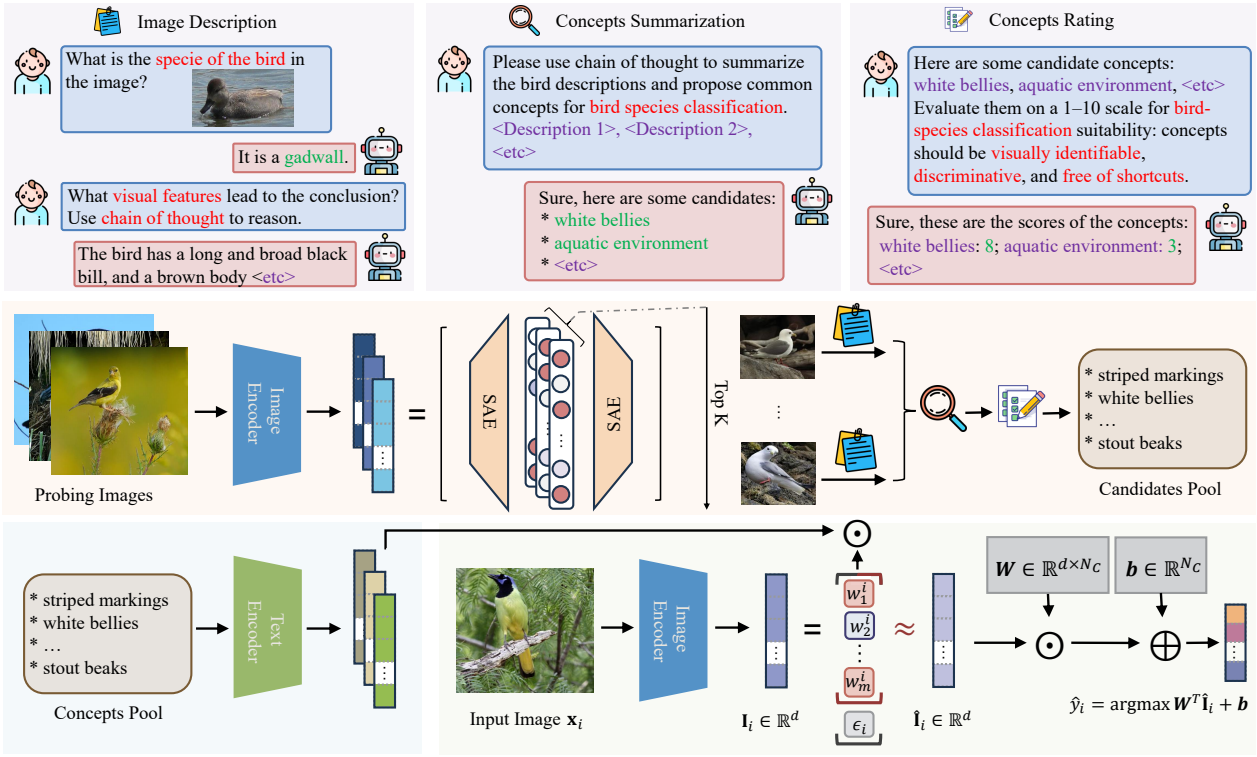


Figure 2: We present an overview of **P**ost-hoc **C**oncept **B**ottleneck **M**odel via **Re**presentation **D**ecomposition (**PCBM-ReD**). First, we extract concepts from the learned representation, and use MLLMs to summarize and score the concepts. Second, we apply a concept selection algorithm to choose concepts and construct the bottleneck. Third, we perform sparse decomposition and reconstruct the image embedding by concepts. A linear layer is trained to predict the targets with the fitted embedding.

$C := \{c_1, c_2, \cdots, c_{N_C}\}$. CBM make prediction by composing two functions, $\hat{y} = f(g(\mathbf{x}))$, where $g : \mathbb{R}^{d_x} \to \mathbb{R}^{N_C}$ maps $\mathbf{x}$ to a score for each element of the bottleneck, and $f : \mathbb{R}^{N_C} \to \mathcal{Y}$ makes the final prediction on the label space given the concept scores. Fig. 2 presents an overview.

3.1 Data-driven Concept Discovery

Most existing methods for constructing bottlenecks rely on either handcrafted concepts (Koh et al. 2020) or concepts generated by LLMs (Yang et al. 2023). However, these concept creation processes are typically independent of the training data or the image encoder, which can result in sub-optimal concepts that fail to capture variations in specific data distributions or are challenging for the visual encoder to distinguish. To overcome these limitations, we propose an approach that automatically extracts concepts from a pre-trained image encoder. We leverage the pre-trained multi-modal alignment model, CLIP (Radford et al. 2021), which consists of an image encoder $\mathcal{I}$ and a text encoder $\mathcal{T}$. The image encoder transforms an image $\mathbf{x}_i$ into a representation $\mathbf{I}_i = \mathcal{I}(\mathbf{x}_i) \in \mathbb{R}^d$. If the encoder effectively captures structured patterns, its latent space should be disentangled into subspaces representing distinct concepts (Bau et al. 2017; Oikarinen and Weng 2022). To do so, we apply sparse autoencoder (SAE) (Bricken et al. 2023), which enables us to represent the visual embedding $\mathbf{I}_i$ with the concept bank atoms $\mathbf{u}_i \in \mathbb{R}^k$, such that $\mathbf{I}_i \approx \mathbf{V}\mathbf{u}_i$. The SAE represents $\mathbf{u}_i$ by a neural network ψ (i.e., $\mathbf{u}_i = \psi(\mathbf{I}_i)$) and enforces sparsity on $\mathbf{u}_i$ (Fel et al. 2023), known to effectively address the challenge of polysemanticity (Bricken et al. 2023). Other dictionary learning paradigms are also applicable.

Ideally, each column of $\mathbf{V}$ represents a concept and the corresponding value of $\mathbf{u}_i$ reflects the significance of the concept in $\mathbf{I}_i$. To transform the concepts into human-understandable form, we propose a MLLM-based pipeline for labeling and scoring the concepts. We sample a reasonable number of images from the training data as probing images, and obtain their corresponding $\mathbf{u}_i$. Then for each concept, we select the top K images with largest concept scores. We then prompt the MLLMs with chain-of-thought to summarize the concept. As shown in Fig. 2, for each image, we prompt MLLMs to describe the visual features that can help identify the category of the image. After collecting image descriptions for the top K images, we ask LLMs to summarize these descriptions and generate candidate concepts that is useful for a specific classification problem. This process results in a large number of candidate concepts, which may or may not be good concepts for the classification tasks. We then ask LLMs to score these concepts. We prompt the LLMs to only assign high scores to concepts that are visually identifiable, discriminatory, and free of shortcuts (e.g., concepts describing the background). We filtered out candi-

Method	Interpret.	ImageNet	CIFAR10	CIFAR100	FOOD	Aircraft	Flower	CUB	UCF	DTD	HAM	RESISC	Average
Fully-supervised Setting													
Linear Probe	✗	83.90	98.10	87.48	93.17	64.03	99.45	84.54	90.67	81.68	83.18	94.98	87.38
LaBo	✓	83.97	97.75	86.04	92.45	61.42	99.35	81.90	90.11	77.30	**81.39**	91.22	85.72
Res-CBM	✓	82.98	97.77	83.01	90.17	54.67	97.85	79.27	88.37	75.77	75.72	91.67	83.39
V2C-CBM	✓	84.15	98.03	86.41	92.84	60.71	98.94	83.12	-	78.49	81.12	92.86	-
PCBM-ReD (ours)	✓	**84.48**	**98.05**	**87.27**	**93.16**	**62.95**	**99.39**	**84.80**	**90.38**	**81.44**	**81.39**	**93.31**	**86.97**
Zero-shot Classification Setting													
CLIP	✗	72.90	95.57	78.28	90.91	31.77	79.46	62.19	75.31	57.09	58.21	64.87	69.69
PCBM-ReD (ours)	✓	72.89	95.56	78.24	90.91	32.01	79.58	62.03	75.36	57.09	58.51	64.87	69.73
CuPL	✗	73.45	95.82	78.59	91.23	35.43	80.80	64.43	76.00	62.65	57.91	71.88	71.65
PCBM-ReD + CuPL (ours)	✓	73.43	95.80	78.59	91.22	35.52	80.67	64.62	75.91	62.77	58.11	71.88	71.68

Table 1: Test accuracy (%) of PCBM-ReD on 11 image classification benchmarks. We report performance of both fully-supervised setting and zero-shot setting. For fully-supervised setting, we compare PCBM-ReD with linear probe, LaBo and Res-CBM. For zero-shot setting, we utilize two strategies, i.e., vanilla CLIP and CuPL. CLIP ViT-L/14 is used as the backbone.

dates with low scores, leaving only high-quality concepts.

3.2 Reconstruction-guided Concept Selection

The LLM-based scoring ensures that the resulting concepts are both human-understandable and relevant to specific tasks. However, the use of a limited sample size for summarizing these concepts can result in noisy outputs with low coverage of the overall dataset. Moreover, the candidate pool may include overlapping or repetitive concepts. To address these issues, it is crucial to select a subset of important and independent concepts that can still comprehensively cover the entire representation space.

We take a set of N probing images sampled from the training data, with $\mathbf{I}_1, \cdots, \mathbf{I}_N$ denoting their image embeddings in the joint text-image space. Given a collection of concepts $\mathcal{C}$, we can reconstruct the image representations by linear combination of the text representations of the concepts within the concept set (denoted as $\mathbf{R}(\mathcal{C}) \in \mathbb{R}^{Mc \times d}$). We further define the reconstruction error to be the Frobenius norm of the difference between the original and the reconstructed image embedding. Our goal is to identify the optimal subset that minimize the reconstruction error:

$$\min_{\mathcal{C}} \sum_{i=1}^{N} \min_{\beta_i(\mathcal{C})} \|\mathbf{I}_i - \mathbf{R}(\mathcal{C})^T \beta_i(\mathcal{C})\|_F^2, \tag{1}$$

where $\beta_i(\mathcal{C})$ is the coefficients of linear combination for reconstructing $\mathbf{I}_i$. While the subset selection is a discrete optimization problem without close-form solution, we propose a greedy algorithm that select concepts step-wisely. Additionally, although the inner optimization of coefficients β_i has analytical solution, we need to solve the optimization problem for each concept within the set $\mathcal{C}$, which can be computational-intensive when the size of $\mathcal{C}$ is large. We propose an algorithm for efficient computation, as illustrated in Alg. 1. Detailed explanation of the algorithm can be found in Appendix. The algorithm can incrementally select new concepts, minimizing reconstruction error to the greatest extent, while ensuring that the newly added concepts are linearly independent from the existing ones. The algorithm stops when the selected concepts reach a pre-defined value, or when all new concepts are linearly dependent on the existing concepts. It is important to note that, unlike the selection algorithms presented in previous work (Chen et al. 2019; Yang et al. 2023), this selection scheme is entirely unsupervised, making it suitable for zero/few-shot applications.

3.3 Post-hoc Class-concept Association

Concept Scores Assignment. Multimodal learning enables the alignment of representations from different modalities into a joint space. Several studies (Gandelsman, Efros, and Steinhardt 2023, 2024) have demonstrated that the image embeddings of CLIP can be represented as a weighted sum of text representations. Therefore, rather than relying on manual annotations or text-image similarity scores to generate concept score supervision, we propose to directly decompose the image representation into concept-related directions within the joint representation space. To ensure high interpretability, we aim for this decomposition to be sparse, utilizing only a few key concepts to explain the image representation. Mathematically, we express this as:

$$\mathbf{I}_i = \hat{\mathbf{I}}_i + \epsilon_i = \sum_{j=1}^{m} w_j^i \mathbf{c}_j + \epsilon_i, \tag{2}$$

where $\mathbf{c}_j$ is the embedding of concept c_j and ϵ_i is residue. We apply a sparse coding algorithm (e.g., orthogonal matching pursuit (Pati, Rezaiifar, and Krishnaprasad 1993)) to approximate $\mathbf{I}_i$ as the sum above, where only n of the w_j^i are non-zero, for some $n < m$.

Algorithm 1: Concepts Selection Algorithm

Input: Image embedding for N images stacked as rows in a matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, a pool of M concepts $\{c_i\}_{i=1}^{M}$, their corresponding text representations $\{\mathcal{T}(c_i)\}_{i=1}^{M}$, selected concepts size m, identity matrix $\mathbf{E}$

Output: A set of selected candidates $\mathcal{C}$

Initialization: $\mathcal{C} \leftarrow \phi, \mathcal{C}_0 \leftarrow \{c_i\}_{i=1}^{M}$

$c^* \leftarrow \arg\max_{c \in \mathcal{C}_0} \left\| \mathbf{X} \cdot \left(\mathbf{E} - \frac{\mathcal{T}(c)\mathcal{T}(c)^T}{\mathcal{T}(c)^T \mathcal{T}(c)} \right) \right\|_F^2$

$\mathcal{C} \leftarrow \mathcal{C} \cup \{c^*\}, \mathcal{C}_0 \leftarrow \mathcal{C}_0 \setminus \{c^*\}$

for *i in* $[2, ..., m]$ **do**

  $\mathbf{P} \leftarrow \mathbf{R}(\mathcal{C})(\mathbf{R}(\mathcal{C})^T \mathbf{R}(\mathcal{C}))^{-1} \mathbf{R}(\mathcal{C})^T$

  for *c in* $\mathcal{C}_0$ **do**

    $z \leftarrow \mathcal{T}(c)^T (\mathbf{E} - \mathbf{P}) \mathcal{T}(c)$

    if $z = 0$ **then**

      $\mathcal{C}_0 \leftarrow \mathcal{C}_0 - \{c\}$

    else

      $\mathbf{Q} \leftarrow \mathcal{T}(c)\mathcal{T}(c)^T / z$

      $\mathbf{L}(c) \leftarrow \mathbf{PQP} - \mathbf{QP} - \mathbf{PQ} + \mathbf{P} + \mathbf{Q}$

  if $\mathcal{C}_0 = \phi$ **then**

    break

  $c^* \leftarrow \arg\min_{c \in \mathcal{C}_0} \| \mathbf{X} (\mathbf{E} - \mathbf{L}(c)) \|_F^2$

  $\mathcal{C} \leftarrow \mathcal{C} \cup \{c^*\}, \mathcal{C}_0 \leftarrow \mathcal{C}_0 \setminus \{c^*\}$

Label Predictor. After decompose each image embedding $\mathbf{I}_i$ into the sum of concept embedding, we discard the residue and retain only the fitted representation $\hat{\mathbf{I}}_i$. We then fit a linear layer to predict the class label from the fitted representation, expressed as $\hat{y}_i = \arg\max(\mathbf{W}^T \hat{\mathbf{I}}_i + \mathbf{b})$. This function can be reformulated as $\hat{y}_i = \arg\max(\sum_{j=1}^{m} w_j^i \mathbf{W}^T \mathbf{c}_j + \mathbf{b})$. As a result, it satisfies the CBM abstraction, with the coefficients $[w_1^i, \cdots, w_m^i]$ representing the concept scores, and $\mathbf{W}^T[\mathbf{c}_1, \cdots, \mathbf{c}_m]$ serving as the class-concept weight matrix. Unlike traditional CBMs that train a sparse class-concept weight matrix to improve the model interpretability, our method enforces sparsity on the concept score side by applying sparse decomposition to the visual representation.

Weight Matrix Initialization. Since $\hat{\mathbf{I}}_i$ is an approximation of $\mathbf{I}_i$, it shares similar properties with $\mathbf{I}_i$. One advantage of $\mathbf{I}_i$ is its zero-shot capability, stemming from the alignment of image and text representations. Ideally, $\hat{\mathbf{I}}_i$ would exhibit similar zero-shot ability. To better leverage this zero-shot prior, we propose to initialize $\mathbf{W}$ with the text embeddings of the prompt "`This is a photo of [cls]`".

4 Experiments

4.1 Dataset and Baselines

Following the benchmark proposed by (Yang et al. 2023), we use 11 image classification datasets spanning a diverse set of domains, including (1) Common objects: ImageNet (Deng et al. 2009), CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009); (2) Fine-grained objects: Food-101 (Bossard, Guillaumin, and Van Gool 2014), FGVC-Aircraft (Maji et al. 2013), Flower-102 (Nilsback and Zisserman 2008), CUB-200-2011 (Wah et al. 2011); (3) Actions: UCF-101 (Soomro 2012); (4) Textures: DTD (Cimpoi et al. 2014); (5) Skin tumors: HAM10000 (Tschandl, Rosendahl, and Kittler 2018), and (6) Satellite images: RESISC45 (Cheng, Han, and Lu 2017). We use the same train/dev/test splits as (Yang et al. 2023) for all datasets. For all experiments, we train on the training set and report the test accuracy. We compare our model, PCBM-ReD, with end-to-end linear probing, as implemented from CLIP (Radford et al. 2021), as well as several CBMs, including original CBM (Koh et al. 2020), PCBM (Yuksekgonul, Wang, and Zou 2022), CompDL (Yun et al. 2022), label-free CBM (Oikarinen et al. 2023), LaBo (Yang et al. 2023), Res-CBM (Shang et al. 2024), CDM (Panousis, Ienco, and Marcos 2023), DN-CBM (Rao et al. 2024), V2C-CBM (He et al. 2025), and VLG-CBM (Srivastava, Yan, and Weng 2024).

Method	Interpret.	CIFAR10	CIFAR100	CUB	Average
Linear Probe	✗	88.80	70.10	72.14	77.01
Original CBM	✓	-	-	65.13	-
CompDL	✓	-	-	54.19	-
PCBM	✓	84.50	56.00	63.63	68.04
Label-free CBM	✓	86.40	65.13	62.40	71.31
CDM	✓	86.50	67.60	**72.26**	75.45
DN-CBM	✓	87.60	67.50	68.38	74.49
VLG-CBM	✓	**88.63**	66.48	66.03	73.71
PCBM-ReD (ours)	✓	88.61	**70.03**	72.01	**76.88**

Table 2: Test accuracy (%) comparison between PCBM-ReD and baselines for fully-supervised setting. We use CLIP RN50 as the backbone for all methods.

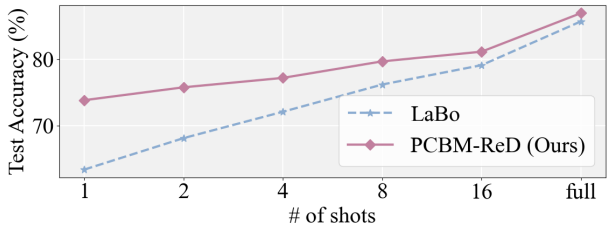


Figure 3: Few-shot test accuracy (%) comparison. Average test accuracy on 11 datasets is reported. Shot means the number of labeled images for each class.

4.2 Implementation Details

We use Llama-3.2-11B-Vision-Instruct to generate image descriptions and use DeepSeek-V3 to summarize and score the concepts. We use CLIP models from OpenCLIP (Cherti et al. 2023) with ViT-L/14 as the default backbone. To train the linear head, we use Adam optimizer with the batch size of 64 and the learning rate of 5×10^{-5}. All experiments were conducted on NVIDIA GeForce RTX 4090 GPUs. More details are provided in the Appendix.

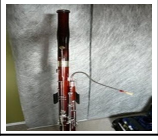
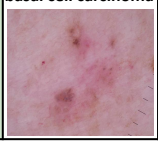
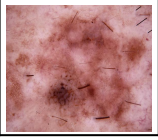
	Class Name	Top Concepts	Class Name	Top Concepts	Class Name	Top Concepts	Class Name	Top Concepts
ImageNet	cock 	1. an animal with a black-tipped tail 2. a bird with intricate plumage and orange beak 3. bird with ruffled feathers	Tibetan terrier 	1. a dark brown creature with black and white stripes 2. an animal with a black-tipped tail 3. a black dog with a short snout	basson 	1. a long-handled tool 2. a cylindrical tube with a shiny finish 3. a polished gold or silver musical instrument	notebook 	1. rectangular device with round corners and screen 2. a QWERTY keyboard layout 3. black and white device with buttons
CUB	horned puffin 	1. distinctive beak shapes and curves 2. black and white plumage 3. yellow beak with colored tip	frigatebird 	1. glossy black bird with long beaks 2. white bellied bird with black markings 3. black and vivid color combinations	nighthawk 	1. brown and gray color patterns 2. distinctive white stripes above eyes 3. prominent eye patterns	forsters tern 	1. long, distinctive tails 2. blue-gray plumage with white belly 3. orange legs and feet
Flower	passion flower 	1. flower with a central stamen 2. purple-colored flowers 3. unique corona structures of Passiflora genus	purple coneflower 	1. thistle-like blooms 2. red and yellow flowers 3. long, pointed purple petals	peruvian lily 	1. dark purple centres with yellow stamens 2. thin, long stamens 3. flowers with a pouch-like structure	gazania 	1. vibrant petals in a spiral pattern 2. large, bright yellow flowers with dark centers 3. yellow petals with red veining
HAM10000	basal cell carcinoma 	1. mix of brown, black, gray colors 2. brown area surrounded by pinkish hue 3. darker brown dots and globules	melanocytic nevi 	1. mix of brown, black, gray colors 2. central darker area and lighter periphery 3. notched and irregular shape	melanoma 	1. brownish-red patch and white surrounding area 2. darker brown center 3. multicomponent pigmented lesion	vascular lesions 	1. brown dots and globules 2. combination of brown and pink colors 3. a lighter periphery

Figure 4: Several example explanations generated by PCBM-ReD. The examples are sampled from the test set of 41 datasets, which have correct predictions. We also show their corresponding top concepts that contribute the most to the logits.

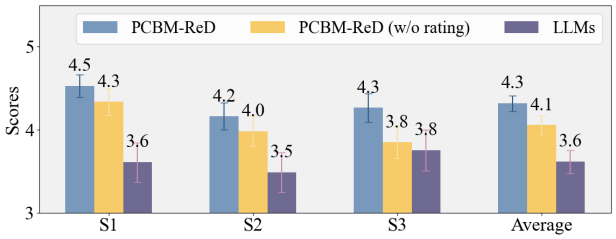


Figure 5: Human evaluation. Volunteers rate the explanation on a scale of 1 to 5 (5 = very agree). **S1**: The explanations are visually identifiable features. **S2**: The explanations faithfully describe the image. **S3**: There is a causal relationship between the explanation and the prediction.

4.3 Main Results

Comparison with End-to-end Model. Table 1 presents the test accuracy of both PCBM-ReD and the linear probe. Our findings indicate that the performance difference between PCBM-ReD and the linear probe is minimal, with an average test accuracy that is just 0.41% lower, despite improved interpretability. Notably, for several datasets, PCBM-ReD even outperforms the end-to-end model. In general, the performance of PCBM-ReD is affected by the concepts, which in turn depend on the quality of the description. For datasets like HAM, general MLLMs may struggle to accurately describe the skin tumors using specific terminology. Domain-specific MLLMs may further improve the performance.

Comparison with other CBMs. We also compare PCBM-ReD's performance with other CBMs in Table 1 and 2. PCBM-ReD outperforms other language-guided CBMs such as LaBo and label-free CBM. It shows an average test accuracy that is 1.25% higher than LaBo and 5.57% higher than label-free CBM. It also achieves greater accuracy than CompDL and the original CBM, even though it does not depend on manually constructed concepts and annotations.

4.4 Performance on Low-data Regime

Zero-shot Ability. Since the weighted sum of concept embeddings approximates the original image embeddings, it retains zero-shot capabilities similar to the original embeddings, something existing CBMs typically lack. In Table 1, we present zero-shot accuracy using two strategies: the vanilla approach from CLIP and CuPL (Pratt et al. 2023). For comparison, we include the performance of the original CLIP representation. The results show that our method achieves strong zero-shot performance, with average accuracy comparable to CLIP, while being interpretable.

Few-shot Performance. We also evaluate the few-shot performance of PCBM-ReD across various datasets. We follow the few-shot evaluation protocol proposed by CLIP with 1, 2, 4, 8, and 16 images randomly sampled from the training set for each class. In Fig. 3, we present the average test accuracy, alongside the accuracy of LaBo for comparison. The full results can be found in Appendix. The results show PCBM-ReD consistently outperform LaBo. On average, PCBM-ReD surpasses LaBo by 5.01%. This shows the potential application of our methods in few-shot learning.

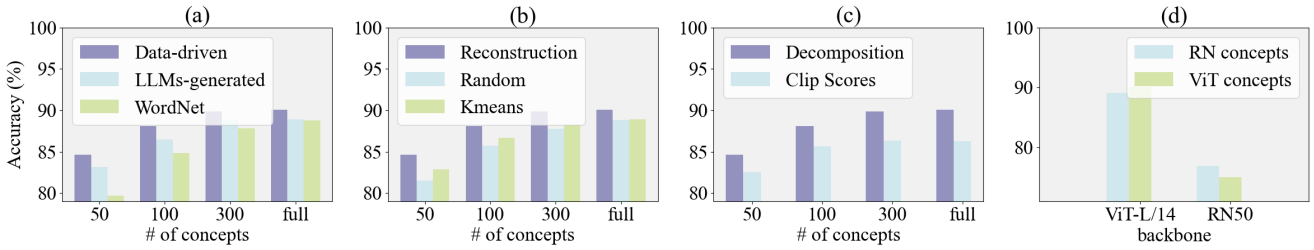


Figure 6: Ablation studies. (a) Test accuracy with different creation schemes of the concepts and the size of the bottleneck. (b) Comparison between reconstruction-guided concept selection, random sampling, and K-Means. (c) Comparison between decomposition-based and CLIP similarity-based concept score association. (d) Effects of sources of the extracted concepts.

4.5 Interpretation Evaluation

Qualitative Analysis. To evaluate PCBM-ReD's interpretability, we present examples illustrating its reasoning process. In Fig. 4, we visualize test set images alongside the top three concepts that most significantly contribute to the correct class's logit. The results show that PCBM-ReD effectively identifies key concepts in the images, which reliably predict corresponding labels. Furthermore, the concepts adapt to different data distributions. For general tasks like ImageNet classification, broad concepts (e.g., color, class) suffice to differentiate categories. In contrast, fine-grained classification requires more specific details (e.g., object parts). Additionally, the concepts are derived directly from visual features, which further enhances the interpretability.

Human Evaluation. We conduct a user study to evaluate the quality of the extracted concepts and generated explanations. Images from five datasets (Imagenet, Food-101, UCF-101, CUB-200-2011, and Flower-102) are randomly selected, and users are shown the image, prediction, and top concepts as explanations. Users rate the explanations on a scale of 1 to 5 based on: 1) whether the concepts are visually identifiable, 2) whether the concepts faithfully describe the image, and 3) whether the explanations have a clear causal link to class labels. Details of the study design are in the Appendix. For comparison, we evaluate explanations using concepts generated by prompting LLMs to describe the class based solely on prior knowledge without training image access (Yang et al. 2023), and PCBM-ReD without LLM-based concept rating. Results from 39 volunteers (Fig. 5) show that PCBM-ReD, which extracts concepts from image descriptions, produces mostly visually identifiable concepts. The concept rating step removes non-causal concepts, enabling the full PCBM-ReD to outperform all baselines.

4.6 Analytical Studies

We perform several ablation studies and report the average accuracy across CIFAR10, CIFAR100, and CUB.

Bottleneck Size. We investigate how the model's performance changes with bottleneck size. As shown in Fig. 6 (a), test accuracy improves with more concepts in the bottleneck, nearly saturating at 300. Notably, even a small bottleneck of 50 achieves reasonable accuracy. Additionally, the required bottleneck size is independent of the number of label classes. Fewer concepts also enhance interpretability.

Concept Creation Schemes. We compare our concept creation scheme with: (1) LLM-generated concepts (Yang et al. 2023) and (2) "Core" WordNet concepts (Boyd-Graber et al. 2006), using the same preprocessing and selection algorithms as PCBM-ReD. Results in Fig. 6 (a) show that our data-driven scheme outperforms both alternatives. Additionally, it avoids non-visual concepts, as confirmed by human evaluation, while also providing a performance boost.

Concept Selection Methods. We evaluate the effectiveness of the reconstruction-guided concept selection algorithm by comparing it with two baselines: random sampling and k-means clustering. As shown in Fig. 6 (b), our method consistently outperforms both baselines, with a noticeable performance gap at smaller bottlenecks. By selecting concepts that minimize reconstruction error, the reconstruction-guided approach ensures higher final accuracy.

Concept Score Association. We compare the sparsity decomposition-based concept score association and to baseline that uses CLIP similarity scores as concept scores. As shown in Fig. 6 (c), representation decomposition achieves significantly higher accuracy. By reconstructing visual representations with concepts, it ensures performance comparable to the original model, achieving higher accuracy.

Sources of the Concepts. Finally, we examine the impact of concept sources by using concepts extracted from different image encoders to construct CBMs. Results in Fig. 6 (d) show a performance decline when the backbones are mismatched, highlighting the importance of aligning concepts with the representational capabilities of the image encoder.

5 Limitations and Conclusion

In this paper, we propose PCBM-ReD, a novel CBM that generates concepts through post-hoc decomposition of visual representations. It achieves high accuracy, zero-shot/few-shot capabilities, and built-in interpretability. However, the approach has some limitations. First, the concept generation relies on MLLMs and LLMs, which can be suboptimal for domain-specific images that general MLLMs cannot describe precisely. Second, as a post-hoc method, the model's performance depends on the original image encoder. We will further improve the model by designing better prompts or leveraging more advanced MLLMs.

Acknowledgments

This research work was supported in part by the National Natural Science Foundation of China (Project No. 62322318 and No. 62201485), in part by the Research Grants Council of Hong Kong Special Administrative Region, China (Project No. T45-401/22-N).

References

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Boyd-Graber, J.; Fellbaum, C.; Osherson, D.; and Schapire, R. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of the third international WordNet conference*, 29–36. Citeseer.

Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; Lasenby, R.; Wu, Y.; Kravec, S.; Schiefer, N.; Maxwell, T.; Joseph, N.; Hatfield-Dodds, Z.; Tamkin, A.; Nguyen, K.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Chen, C.; Zhang, B.; Cao, L.; Shen, J.; Gunter, T.; Jose, A. M.; Toshev, A.; Shlens, J.; Pang, R.; and Yang, Y. 2023. STAIR: learning sparse text and image representation in grounded tokens. *arXiv preprint arXiv:2301.13081*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fel, T.; Boutin, V.; Béthune, L.; Cadène, R.; Moayeri, M.; Andéol, L.; Chalvidal, M.; and Serre, T. 2023. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36: 54805–54818.

Fong, R.; and Vedaldi, A. 2018. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8730–8738.

Gandelsman, Y.; Efros, A. A.; and Steinhardt, J. 2023. Interpreting CLIP's Image Representation via Text-Based Decomposition. *arXiv preprint arXiv:2310.05916*.

Gandelsman, Y.; Efros, A. A.; and Steinhardt, J. 2024. Interpreting the Second-Order Effects of Neurons in CLIP. arXiv:2406.04341.

Ge, Y.; Xiao, Y.; Xu, Z.; Zheng, M.; Karanam, S.; Chen, T.; Itti, L.; and Wu, Z. 2021. A peek into the reasoning of neural networks: interpreting with structural visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2195–2204.

Gong, S.; Dou, Q.; and Farnia, F. 2024. Structured Gradient-based Interpretations via Norm-Regularized Adversarial Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11009–11018.

Gong, S.; Haoyu, L.; Dou, Q.; and Farnia, F. 2025a. Boosting the visual interpretability of clip via adversarial fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Gong, S.; Wang, H.; Zhang, X.; and Dou, Q. 2025b. Concepts from Neurons: Building Interpretable Medical Image Diagnostic Models by Dissecting Opaque Neural Networks. In *International Conference on Information Processing in Medical Imaging*, 3–18. Springer.

He, H.; Zhu, L.; Zhang, X.; Zeng, S.; Chen, Q.; and Lu, Y. 2025. V2C-CBM: Building Concept Bottlenecks with Vision-to-Concept Tokenizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3401–3409.

Kamath, A.; Hessel, J.; and Chang, K.-W. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Kim, E.; Jung, D.; Park, S.; Kim, S.; and Yoon, S. 2023. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Ma, C.; Zhao, B.; Chen, C.; and Rudin, C. 2024. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Moayeri, M.; Rezaei, K.; Sanjabi, M.; and Feizi, S. 2023. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, 25037–25060. PMLR.

Nauta, M.; Van Bree, R.; and Seifert, C. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14933–14943.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

Oikarinen, T.; and Weng, T.-W. 2022. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*.

Omeiza, D.; Webb, H.; Jirotka, M.; and Kunze, L. 2021. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8): 10142–10162.

Panousis, K. P.; Ienco, D.; and Marcos, D. 2023. Sparse linear concept discovery models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2767–2771.

Pati, Y. C.; Rezaiifar, R.; and Krishnaprasad, P. S. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, 40–44. IEEE.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, S.; Mahajan, S.; Böhle, M.; and Schiele, B. 2024. Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery. In *European Conference on Computer Vision*.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Saha, O.; Van Horn, G.; and Maji, S. 2024. Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17542–17552.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shang, C.; Zhou, S.; Zhang, H.; Ni, X.; Yang, Y.; and Wang, Y. 2024. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11030–11040.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Soomro, K. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Srivastava, D.; Yan, G.; and Weng, L. 2024. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37: 79057–79094.

Tan, A.; Zhou, F.; and Chen, H. 2024a. Explain via Any Concept: Concept Bottleneck Model with Open Vocabulary Concepts. *arXiv preprint arXiv:2408.02265*.

Tan, A.; Zhou, F.; and Chen, H. 2024b. Post-hoc Part-prototype Networks. *arXiv preprint arXiv:2406.03421*.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Vielhaben, J.; Bluecher, S.; and Strodthoff, N. 2023. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *arXiv preprint arXiv:2301.11911*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19187–19197.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*.

Yun, T.; Bhalla, U.; Pavlick, E.; and Sun, C. 2022. Do Vision-Language Pretrained Models Learn Composable Primitive Concepts? *arXiv preprint arXiv:2203.17271*.

Zarlenga, M. E.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. *arXiv preprint arXiv:2209.09056*.