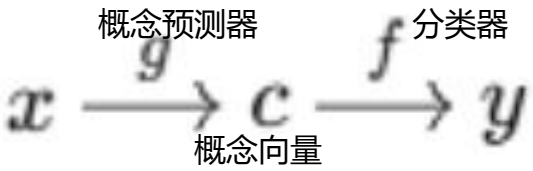
南京航空航天大學

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

Shizhan Gong1, Xiaofan Zhang2, Qi Dou1 1The Chinese University of Hong Kong, Hong Kong, China 2Shanghai Jiao Tong University, Shanghai, China szgong22@cse.cuhk.edu.hk, xiaofan.zhang@sjtu.edu.cn, qidou@cuhk.edu.hk

AAAI 2026

Introduction



Ante-hoc:
可解释性强
可干预性好
因果链路清晰
训练目标明确

Post-hoc
概念自动挖掘
分类效果好

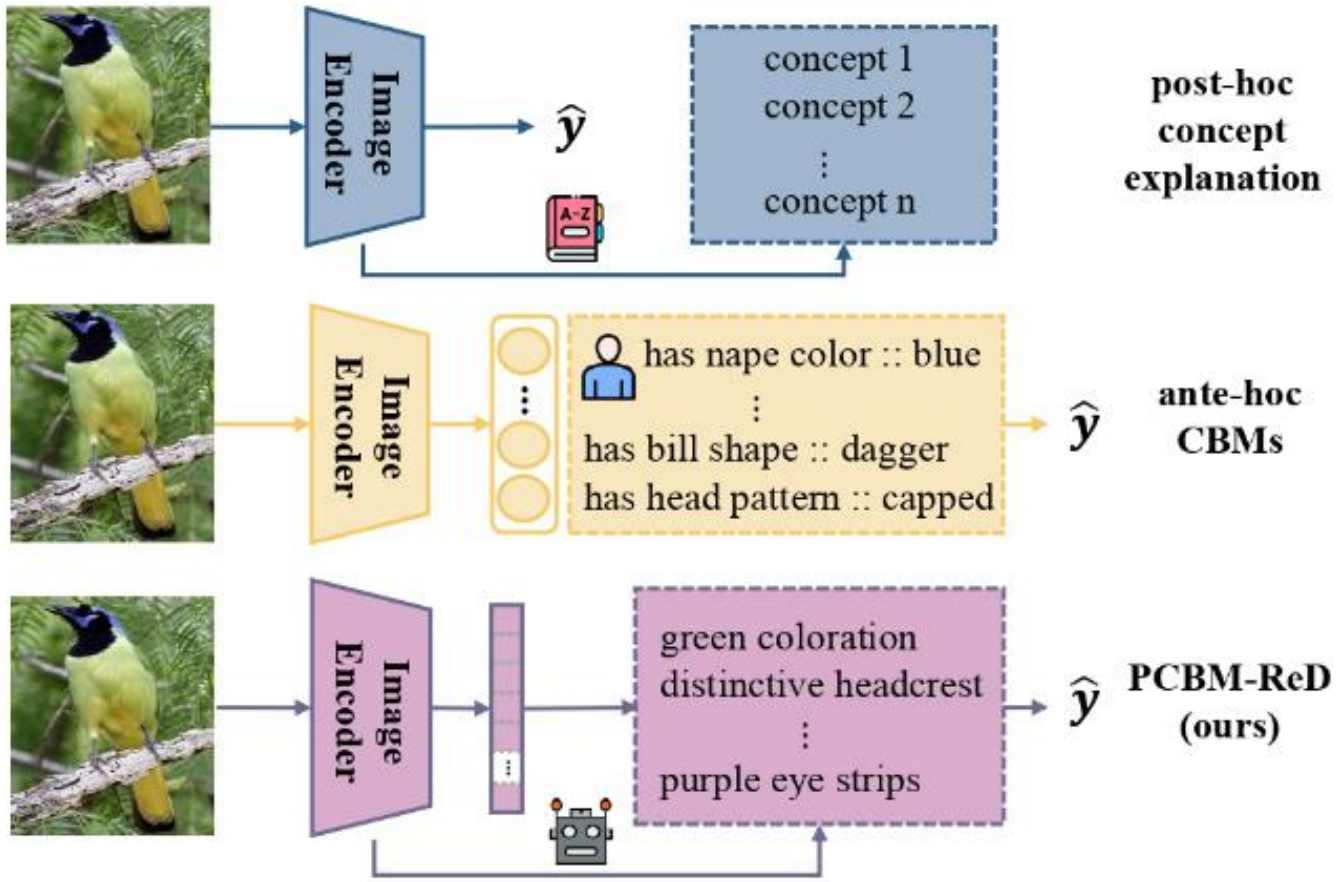


Figure 1: PCBM-ReD extracts concepts from the pre-trained image encoder and reconstruct the visual representation with the concepts, which gives faithful interpretation and can take the best advantage of the encoder's representation power.

Method

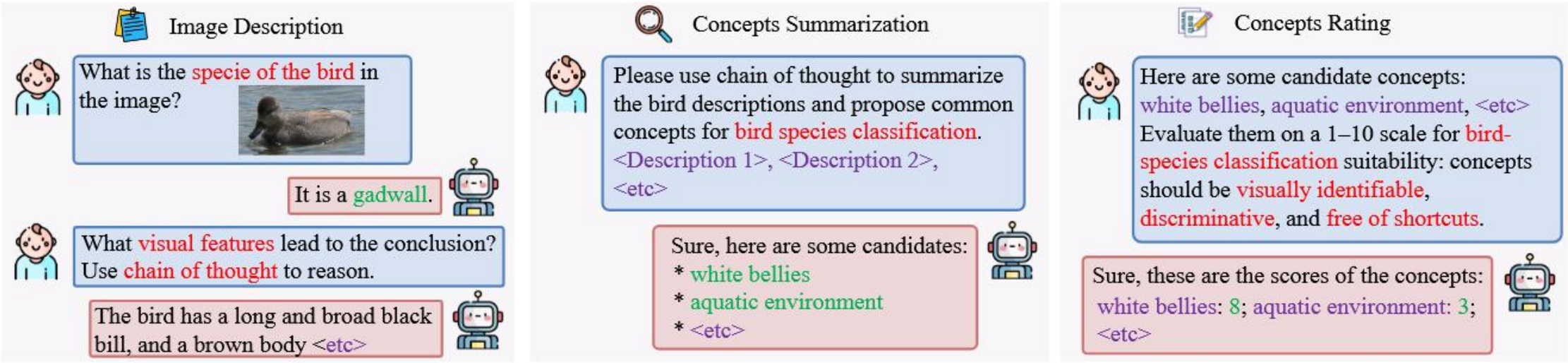
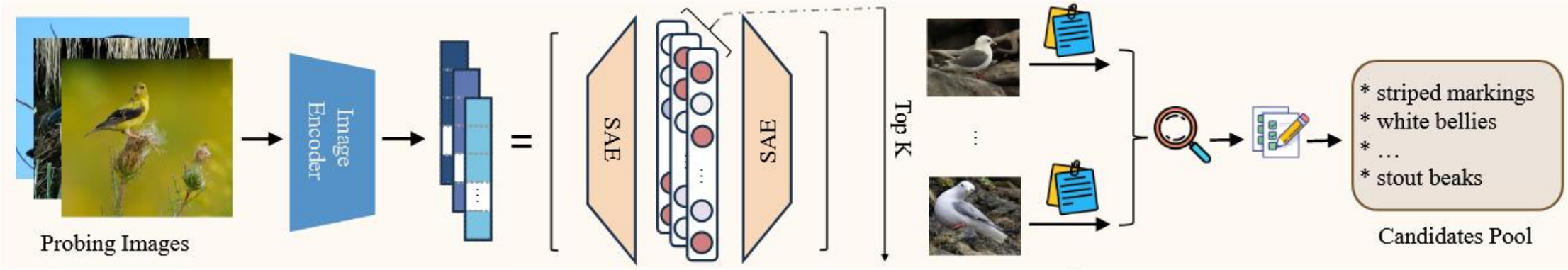
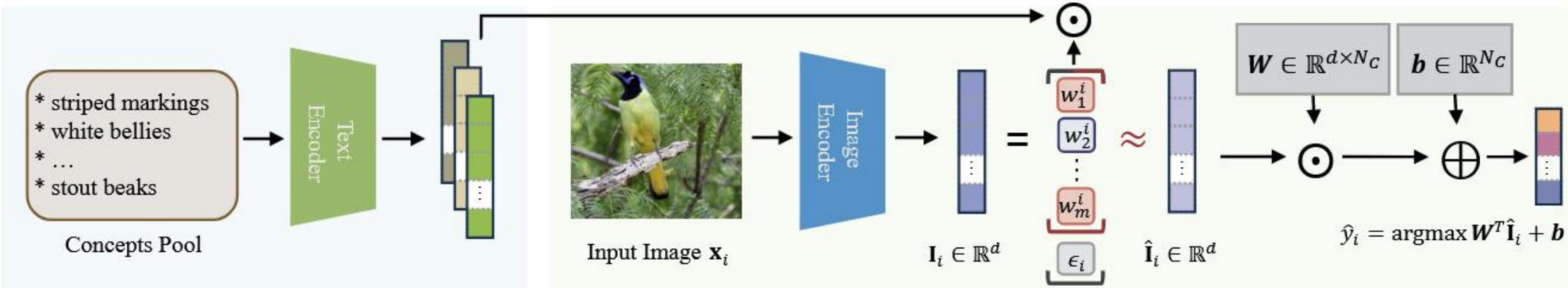


Image Description
What is the specie of the bird in the image?
It is a gadwall.
What visual features lead to the conclusion? Use chain of thought to reason.
The bird has a long and broad black bill, and a brown body <etc>
Concepts Summarization
Please use chain of thought to summarize the bird descriptions and propose common concepts for bird species classification. <Description 1>, <Description 2>, <etc>
Sure, here are some candidates:
* white bellies
* aquatic environment
* <etc>
Concepts Rating
Here are some candidate concepts: white bellies, aquatic environment, <etc> Evaluate them on a 1–10 scale for bird-species classification suitability: concepts should be visually identifiable, discriminative, and free of shortcuts.
Sure, these are the scores of the concepts: white bellies: 8; aquatic environment: 3; <etc>


Probing Images
Image Encoder
=
SAE
SAE
Top K
* striped markings
* white bellies
* …
* stout beaks
Candidates Pool

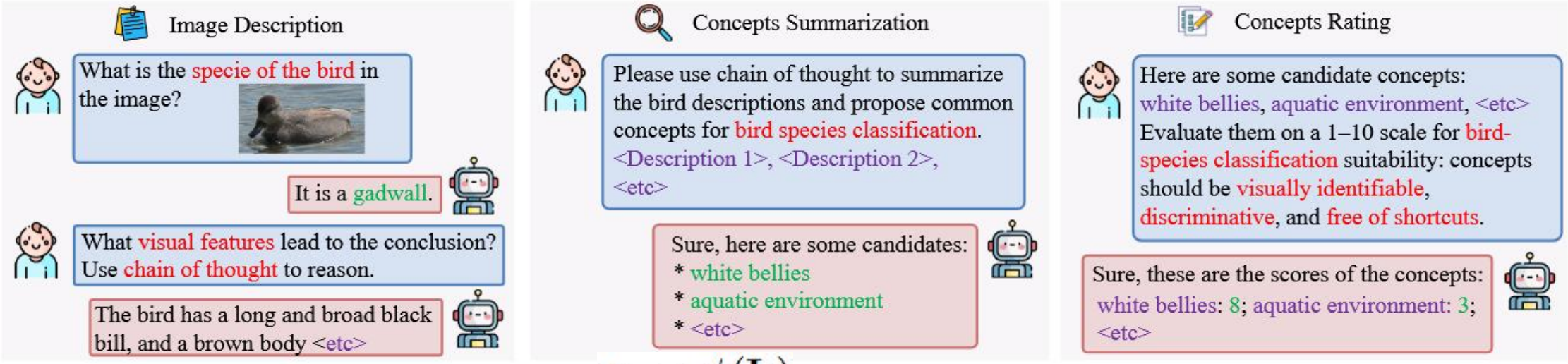

* striped markings
* white bellies
* …
* stout beaks
Concepts Pool
Text Encoder
Input Image $\mathbf{x}_i$
Image Encoder
$\mathbf{I}_i \in \mathbb{R}^d$
=
w_1^i
w_2^i
w_m^i
ϵ_i
≈
$\hat{\mathbf{I}}_i \in \mathbb{R}^d$
$\boldsymbol{W} \in \mathbb{R}^{d \times N_C}$
$\boldsymbol{b} \in \mathbb{R}^{N_C}$
$\hat{y}_i = \text{argmax}\, \boldsymbol{W}^T \hat{\mathbf{I}}_i + \boldsymbol{b}$

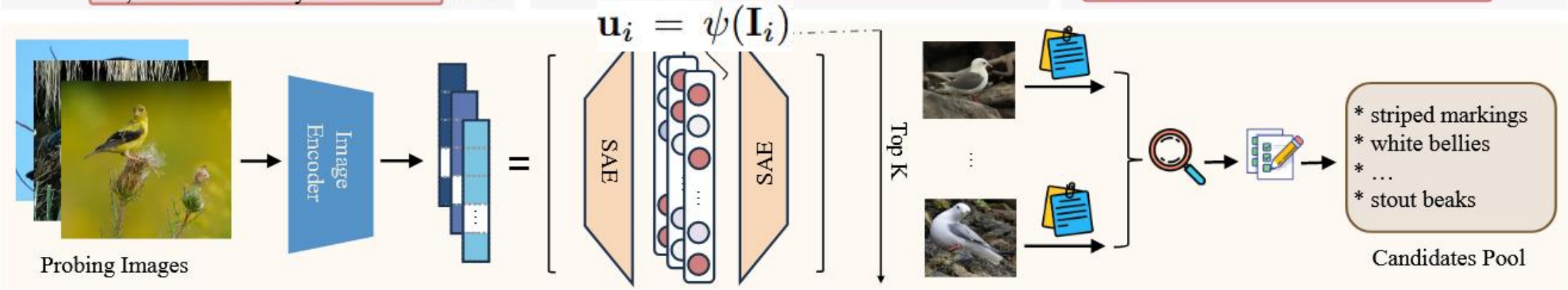
Method

Llama-3.2-11B-Vision-Instruct

DeepSeek-V3

DeepSeek-V3





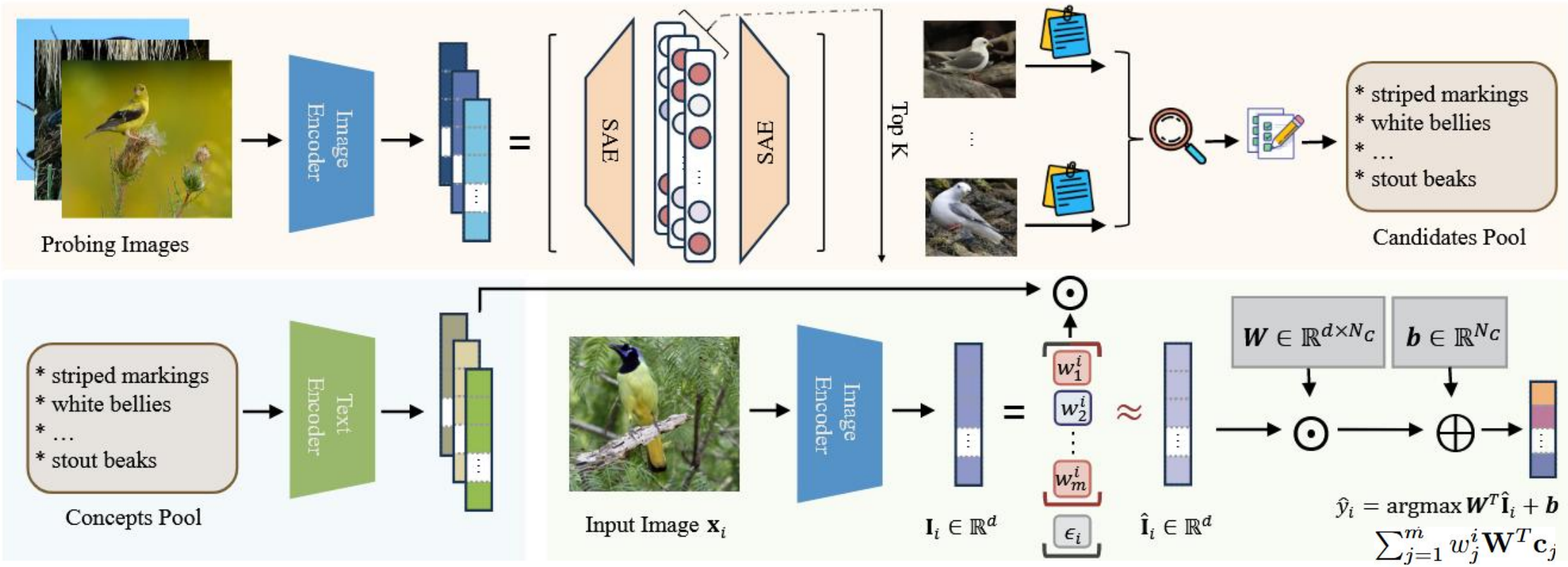
$$\mathbf{I}_i = \mathcal{I}(\mathbf{x}_i) \in \mathbb{R}^d \quad \mathbf{u}_i \in \mathbb{R}^k$$

SAE (Bricken et al. 2023)

$$\mathbf{I}_i \approx \mathbf{V}\mathbf{u}_i$$

使用稀疏自编码器能从单层transformer模型中提取大量可解释的特征

Method



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

稀疏编码算法 (Pati, Rezaiifar, and Krishnaprasad 1993)

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

We propose to initialize W with the text embeddings of the prompt “This is a photo of [cls]”.

Results

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.

Results

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.

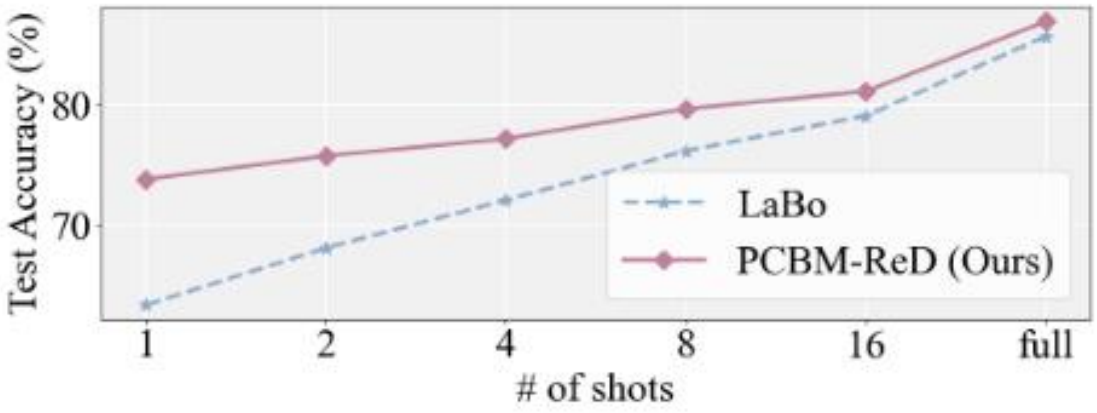


Figure 3: Few-shot test accuracy (%) comparison. Average test accuracy on 11 datasets is reported. Shot means the number of labeled images for each class.

Experiments

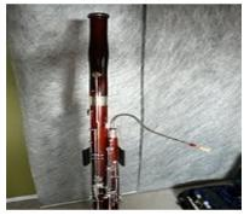
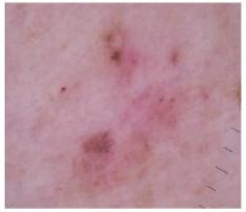
	Class Name	Top Concepts	Class Name	Top Concepts	Class Name	Top Concepts	Class Name	Top Concepts
ImageNet	cock 	1. an animal with a black-tipped tail 2. a bird with intricate plumage and orange beak 3. bird with ruffled feathers	Tibetan terrier 	1. a dark brown creature with black and white stripes 2. 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 11 datasets, which have correct predictions. We also show their corresponding top concepts that contribute the most to the logits.

Experiments

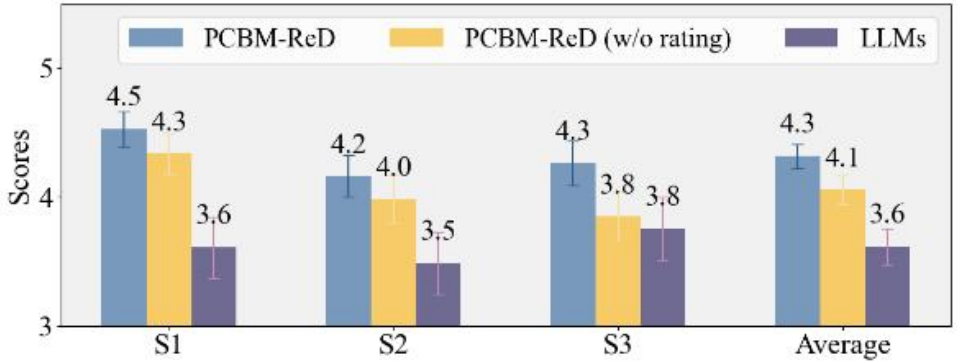


Figure 5: Human evaluation. Volunteers rate the explanation on a scale of 1 to 5 (5 = very agree). **S1**: The explanations are visually identifiable features. **S2**: The explanations faithfully describe the image. **S3**: There is a causal relationship between the explanation and the prediction.

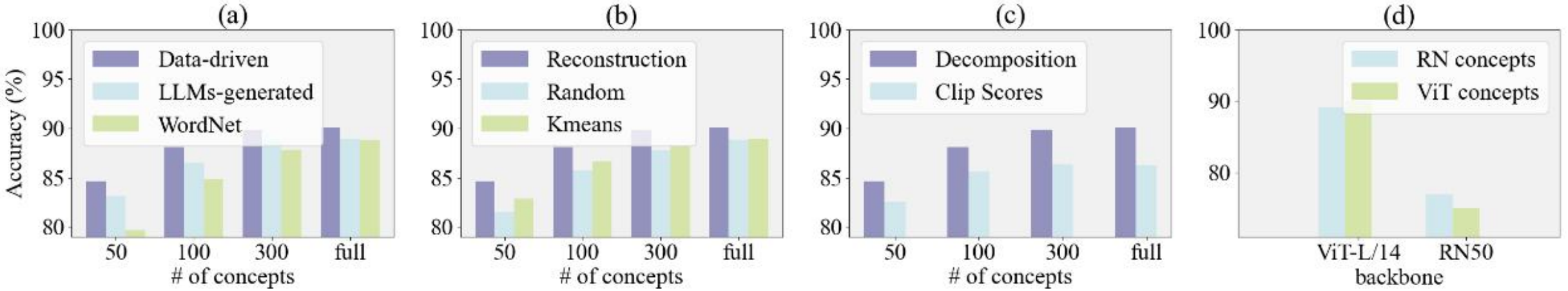


Figure 6: Ablation studies. (a) Test accuracy with different creation schemes of the concepts and the size of the bottleneck. (b) Comparison between reconstruction-guided concept selection, random sampling, and K-Means. (c) Comparison between decomposition-based and CLIP similarity-based concept score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

Thanks