Attribute-formed Class-specific Concept Space: Endowing Language Bottleneck Model with Better Interpretability and Scalability

Jianyang Zhang[1*], Qianli Luo[2*], Guowu Yang[1,3], Wenjing Yang[4],
Weide Liu[5], Guosheng Lin[6], Fengmao Lv[2,7,†]

[1]University of Electronic Science and Technology of China [2]Southwest Jiaotong University
[3]Institute of Electronics and Information Industry Technology of Kash [4]University of Minnesota
[5]Harvard University [6]Nanyang Technological University
[7]Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education

zhang_jy.1@qq.com, qianlil@my.swjtu.edu.cn, guowu@uestc.edu.cn, wjyang2987@gmail.com,
weide001@e.ntu.edu.sg, gslin@ntu.edu.sg, fengmaolv@126.com

CVPR 2025

Background

传统的CBM/LBM通常会存在三个问题

- **共享概念池容易产生伪相关推理**

dark brown

long wings

wide sea of blue

forest background

served on a plate

...

- **难以扩展到新类别**

原来：[black body, long wings, red beak]　　新类别概念：curved crest,...

- **CLIP 的全局 [CLS] 特征不够细粒度**

sim(f_cls, color concept)

sim(f_cls, wing concept)　　f_cls 是整体图像表示，不一定能准确表达每个细粒度属性。

sim(f_cls, texture concept)

Background

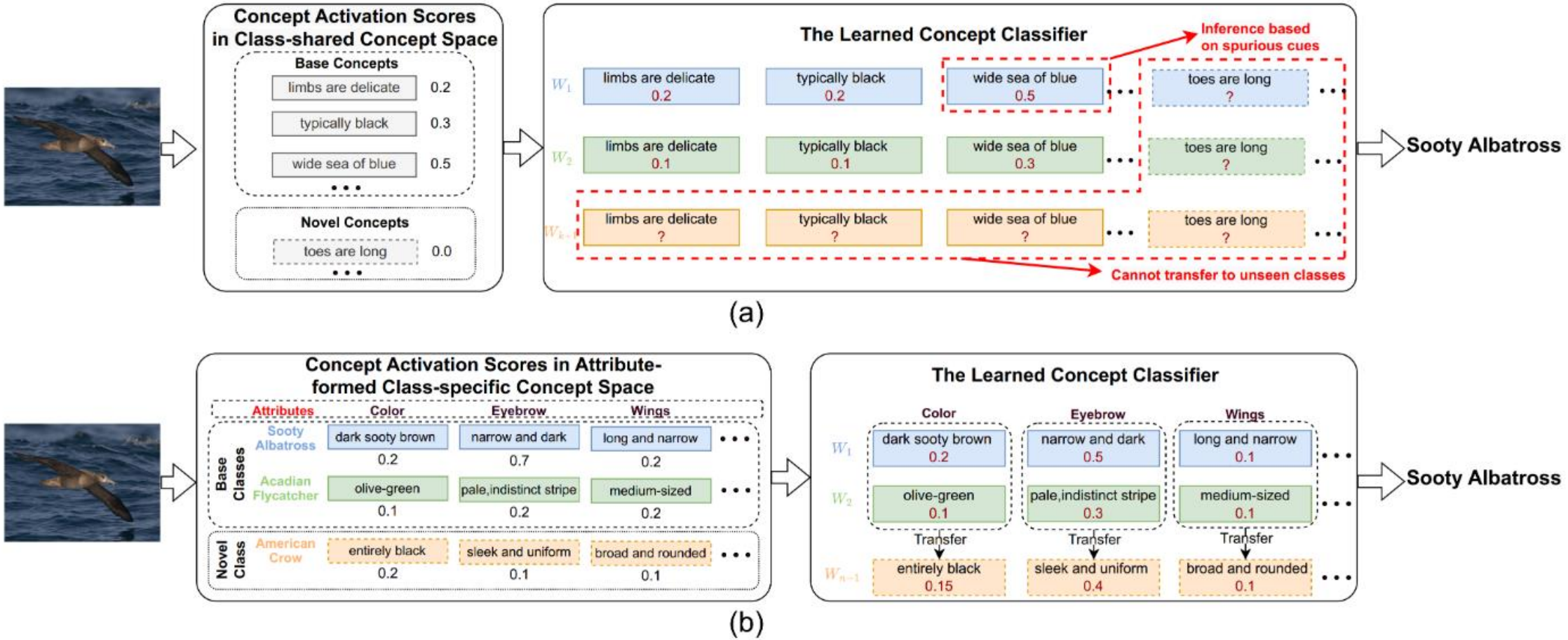


Figure 1. Illustration of the scenario for concept classification. (a) Existing Language Bottleneck Models [32, 33] (b) Our Attribute-formed Language Bottleneck Model. Existing LBMs suffer spurious cue inference as they may make decisions based on non-essential or background concepts. Additionally, their cross-class scalability is also limited, as expanding the concept space may be necessary for unseen classes. On the contrary, our approach predicts classes solely based on their corresponding concepts to avoid the spurious cue problem, and also ensures the cross-category consistent concept space by sharing the unified attribute set, allowing transfer to unseen classes.

Methods

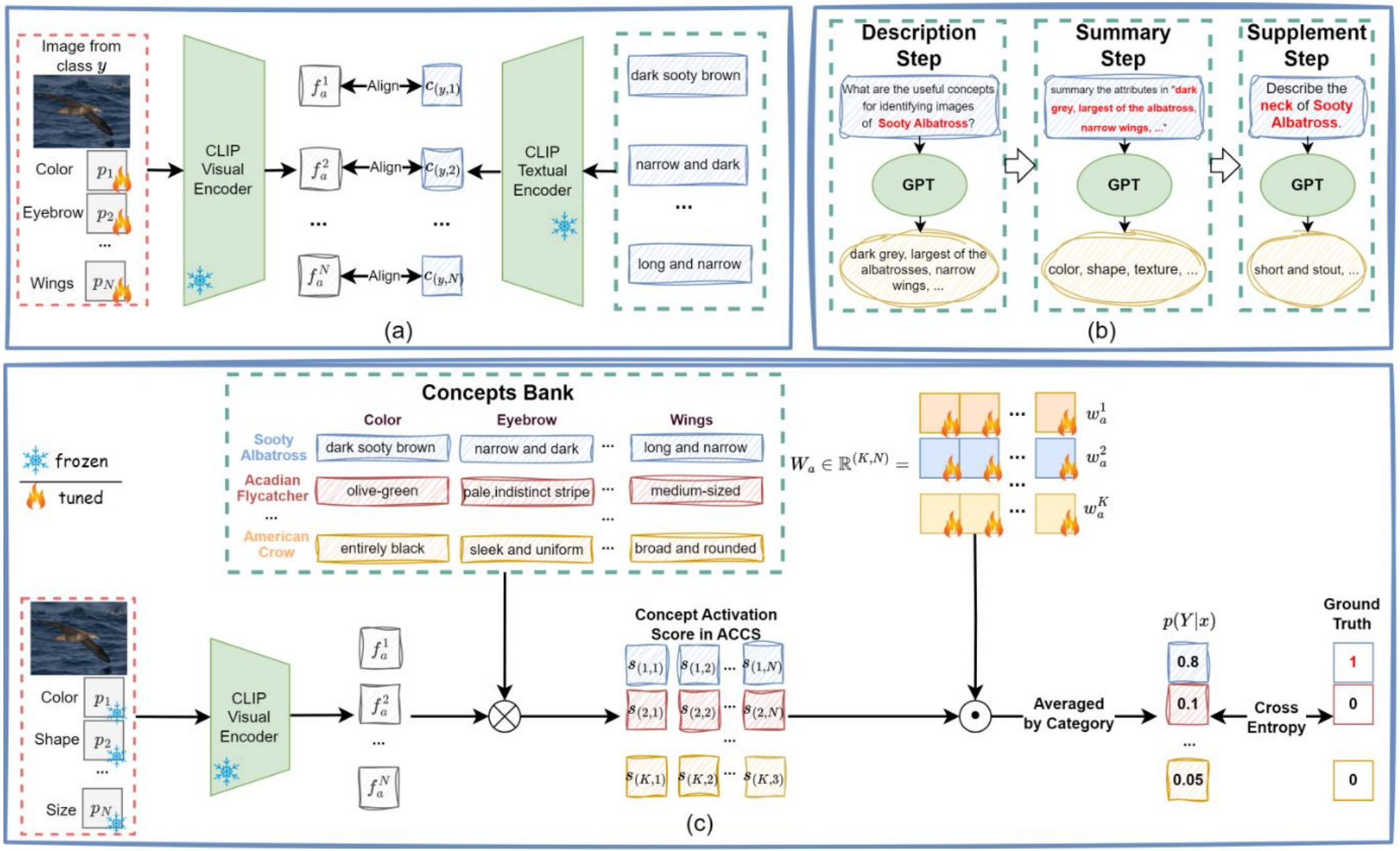


Image from class y
Color p_1
Eyebrow p_2
Wings p_N
CLIP Visual Encoder
Align
CLIP Textual Encoder
dark sooty brown
narrow and dark
long and narrow
(a)
Description Step
What are the useful concepts for identifying images of Sooty Albatross?
GPT
dark grey, largest of the albatrosses, narrow wings, ...
Summary Step
summary the attributes in "dark grey, largest of the albatross, narrow wings, ..."
GPT
color, shape, texture, ...
Supplement Step
Describe the neck of Sooty Albatross.
GPT
short and stout, ...
(b)
frozen
tuned
Concepts Bank
Color
Eyebrow
Wings
Sooty Albatross
dark sooty brown
narrow and dark
long and narrow
Acadian Flycatcher
olive-green
pale,indistinct stripe
medium-sized
American Crow
entirely black
sleek and uniform
broad and rounded
$W_a \in \mathbb{R}^{(K,N)} =$
Color p_1
Shape p_2
Size p_N
CLIP Visual Encoder
Concept Activation Score in ACCS
Averaged by Category
$p(Y|x)$
0.8
0.1
0.05
Cross Entropy
Ground Truth
1
0
0
(c)

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Attribute-formed language bottleneck model(ALBM)

类别	Color	Wings	Beak
Sooty Albatross	dark sooty brown	long and narrow	large and hooked
American Crow	entirely black	broad and rounded	thick and black
Acadian Flycatcher	olive-green	medium-sized	short and pointed

假设有 K 个类别，N_a 个属性。

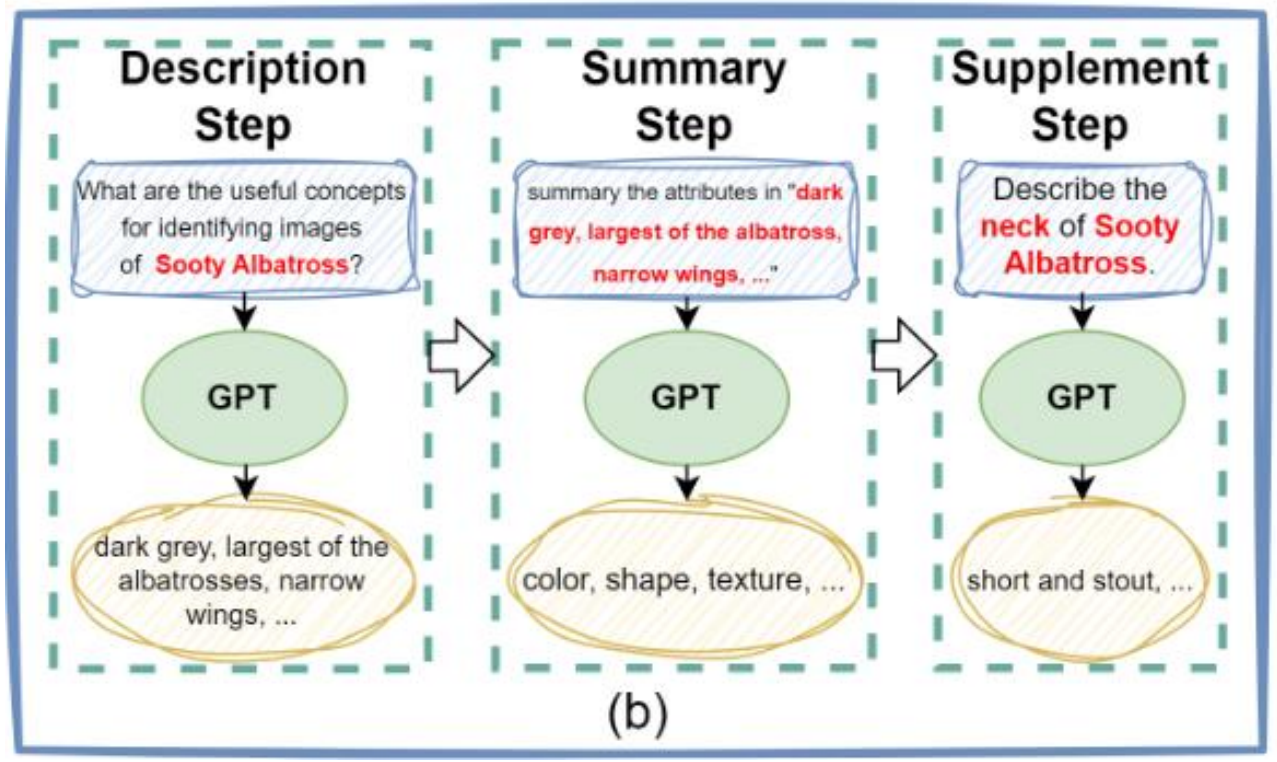
每个类别每个属性一个概念，所以总概念数是：

$$K \times N_a$$

输入图像后，模型计算一个概念激活矩阵：

$$S \in \mathbb{R}^{K \times N_a}$$

Description, summary, and supplement strategy(DSS)



$$\boldsymbol{c}_i = \mathrm{LLM}\left(g_i, q_{des}\right), \tag{9}$$

$$\boldsymbol{A} = \mathrm{LLM}\left(\{\boldsymbol{c}_i\}_{i=1}^{K}, q_{sum}\right). \tag{10}$$

$$\boldsymbol{c}_i^j = \mathrm{LLM}\left(g_i, a_j, q_{vis}\right), \tag{11}$$

Description, summary, and supplement strategy(DSS)

	Aircraft	CUB	DTD	Flowers102	Food101	OxfordPets	CIFAR-10	CIFAR-100	ImageNet
CLIP-GPT	22	7	8	18	16	7	-	-	17
ALBM	23	37	33	26	29	12	11	21	55

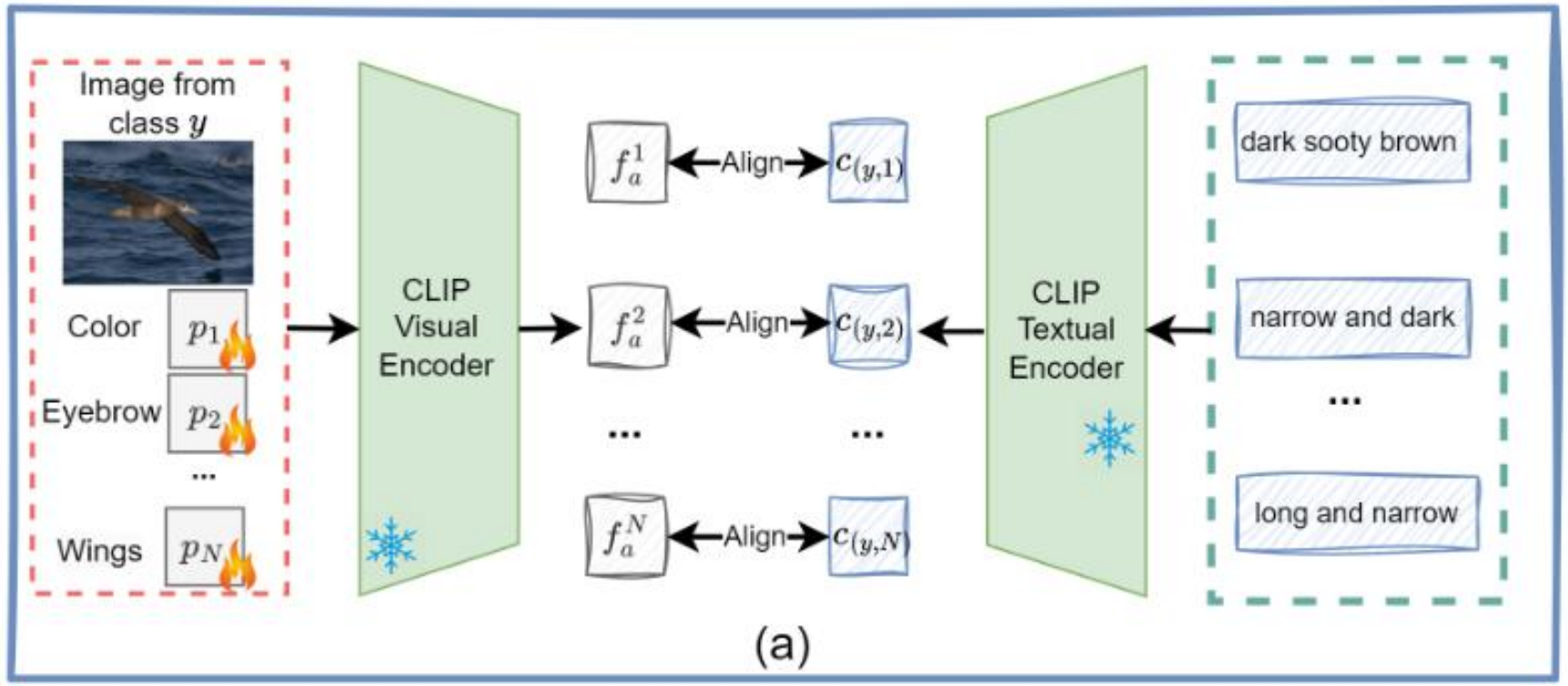
Table 1. The size of the attribute set on each dataset.

CLIP-GPT	ALBM
size, length, fur texture, fur color, eye color, ear shape, distinctive features	fur, size, breed, appearance, body, color, snout, head, legs, tail, eyes, ears

Table 2. Comparison of the collected attributes on OxfordPets.

Methods

Visual attribute prompt learning(VAPL)




```
A = {Color, Shape, Texture, Wings, Beak, Tail, ...}  →  p_color, p_shape, p_texture, p_wings, p_beak, p_tail, ...
```

$$\left[\boldsymbol{f}_a^1, \boldsymbol{f}_a^2, ..., \boldsymbol{f}_a^N\right] = \mathcal{V}\left(\left[x, \boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_N\right]\right), \quad (6)$$

$$s_{(i,j)} = \mathrm{sim}\left(\boldsymbol{f}_a^j, \boldsymbol{c}_{(i,j)}\right), \quad (7)$$

$$\mathcal{L}_p = \frac{1}{N_a} \sum_{j=1}^{N_a} -\log\left(\frac{\exp\left(s_{(y,j)}\right)}{\sum_{i=1}^{K} \exp\left(s_{(i,j)}\right)}\right), \quad (8)$$

Methods

Visual attribute prompt learning(VAPL)

where f_a^i is the feature that represents information of x on the i-th attribute. Notably, to prevent these visual attribute prompts from interfering with the feature extraction process, we masked the attention between these prompts and the attention from the image tokens to these prompts. Based on $[f_a^1, f_a^2, ..., f_a^N]$, the concept activation score S can be

[image patch tokens, p_color, p_texture, p_shape, ...]

1. prompt 之间不要互相 attention
2. image tokens 不要 attention 到 prompts
3. prompts 可以 attention 到 image tokens

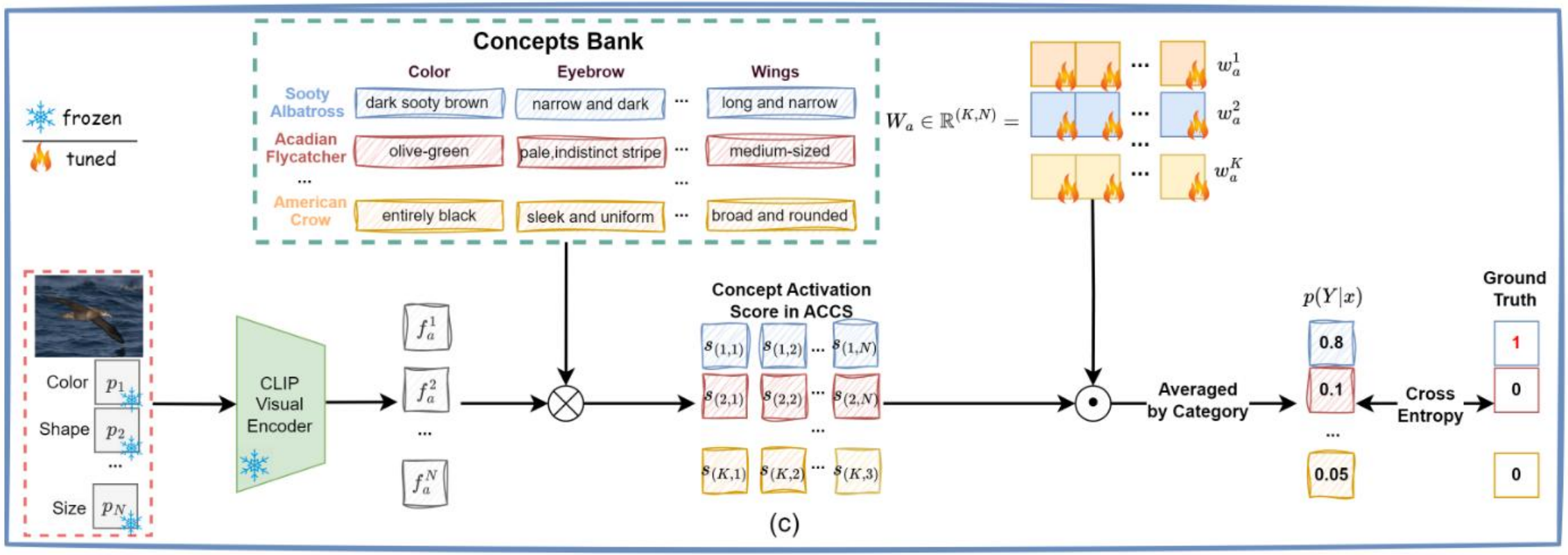
假设一张训练图像真实类别是 Sooty Albatross。

它的概念是：

属性	概念
Color	dark sooty brown
Wings	long and narrow
Beak	large and hooked

$$\mathcal{L}_p = \frac{1}{N_a} \sum_{j=1}^{N_a} -\log \left(\frac{\exp\left(s_{(y,j)}\right)}{\sum_{i=1}^{K} \exp\left(s_{(i,j)}\right)} \right), \qquad (8)$$

Methods



concept activation score $\boldsymbol{S} = \mathbf{C} \cdot \boldsymbol{f}^T$

$$p\left(Y=j|x\right) = \frac{\exp\left(\boldsymbol{w}_a^j \cdot \boldsymbol{s}_y^T\right)}{\sum_{i=1}^{K} \exp\left(\boldsymbol{w}_a^i \cdot \boldsymbol{s}_i^T\right)}, \tag{3}$$

$$\mathcal{L}_w = \frac{1}{|\mathfrak{D}_s|} \sum_{x \in \mathfrak{D}_s} -\log\left(p\left(Y=y|x\right)\right). \tag{4}$$

$$\boldsymbol{w}_a^{K+j} = \sum_{i=1}^{K} \left(\frac{\exp\left(\boldsymbol{n}_i \cdot \boldsymbol{n}_{K+j}^T\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{n}_l \cdot \boldsymbol{n}_{K+j}^T\right)} \cdot \boldsymbol{w}_a^i \right), \tag{5}$$

Experiments

	Approach	Aircraft	CUB	DTD	Flowers102	Food101	OxfordPets	CIFAR-10	CIFAR-100	ImageNet
Unexplainable	ZS-CLIP [25]	32.6	63.4	53.2	79.3	91.0	93.6	86.0	55.6	71.4
Training-free Language Bottleneck	VDCLIP [20]	-	3.9	18.6	-	15.2	11.9	-	-	23.4
	CuPL [24]	**19.9**	-	37.2	-	66.3	33.9	75.2	40.6	59.2
	CLIP-GPT [19]	13.4	11.4	40.0	11.7	48.4	31.9	-	-	44.3
	LaBo* [33]	15.7	16.2	37.9	34.2	52.2	-	64.0	31.1	37.8
	ALBM* (ours)	18.0	**25.0**	**48.5**	**54.9**	**75.4**	**35.9**	**83.1**	**43.1**	**64.6**

Table 3. Comparison with zero-shot CLIP and training-free language bottlenecks in the zero-shot setting. * denote zero-shot predictions based on their collected concept sets, while "-" indicates that the original approaches didn't collect a concept set for the dataset.

	Approach	Aircraft		CUB		DTD		Flowers102		Food101		OxfordPets		CIFAR-10		CIFAR-100		ImageNet	
		Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel
Unexplainable	ZS-CLIP [25]	37.2	44.5	69.9	60.1	61.2	71.4	83.2	82.7	93.7	94.9	95.1	98.2	91.1	93.7	66.9	60.2	77.2	72.3
	LP-CLIP [25]	50.9	-	86.4	-	80.7	-	98.6	-	91.6	-	93.3	-	91.1	-	71.3	-	78.5	-
Training-free Language Bottleneck	VD-CLIP [20]	-	-	7.0	5.5	31.0	21.4	-	-	19.9	18.0	14.8	22.5	-	-	-	-	20.7	34.0
	CuPL [24]	22.7	30.5	-	-	51.2	43.8	-	-	71.6	78.3	42.8	49.4	88.5	92.8	49.4	49.7	59.8	66.8
	CLIP-GPT [19]	14.2	18.4	18.6	15.6	52.0	49.6	11.0	16.5	60.8	57.5	46.0	46.1	-	-	-	-	70.0	22.1
Language Bottleneck Model	LaBo [33]	**42.9**	-	76.9	-	77.0	-	87.6	-	**90.8**	-	-	-	89.6	-	55.6	-	71.7	-
	CLBM [32]	-	-	67.4	-	-	-	52.0	-	-	-	60.0	-	-	-	51.4	-	-	-
Base-to-Novel	ALBM (ours)	38.7	**33.0**	**91.9**	**27.8**	**78.6**	**60.5**	**91.7**	**32.4**	88.5	**86.8**	**79.2**	**61.1**	**90.8**	**93.6**	**59.3**	**55.1**	**75.0**	**73.9**

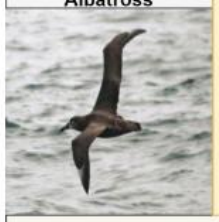
Table 4. Comparison with unexplainable CLIP, Training-free Language Bottlenecks, and Language Bottleneck Models on the base-to-novel setting, where Training-free Language Bottlenecks are zero-shot learning approaches, Language Bottleneck Models are trained on base classes, and "-" indicates that the original approaches didn't collect the concept set for the dataset or unavailable for novel classes.

Experiments

Food101

Apple Pie		Attributes	Concepts	Score		Concepts	Score	Baby Back Ribs		Attributes	Concepts	Score		Concepts	Score
	ALBM (ours)	Center	typically filled with a mixture of sliced apples, sugar, ...	3.28	LaBo	bright, sunny brown	1.02		ALBM (ours)	Feature	tender, flavorful pork ribs that are typically slow-cooked and ...	2.83	LaBo	black pepper on the ribs	1.02
		State	typically referring to its freshness, such as freshly baked, room...	3.26		served with a dollop of whipped cream	1.02			Size	typically around 10-13 ribs per rack, each rib averaging about ...	2.24		12 inches long and 6 inches wide	1.01
		Cutting	typically sliced into wedge-shaped pieces, ensuring each slice ...	2.50		ends of the dough are crimped together	1.01			Surroundings	typically garnished with sides like coleslaw, baked beans ...	2.32		11 pairs	1.01

CUB

Black Footed Albatross		Attributes	Concepts	Score		Concepts	Score	Laysan Albatross		Attributes	Concepts	Score		Concepts	Score
	ALBM (ours)	Size	medium to large seabird with a wingspan ranging ...	4.33	LaBo	lays 4-5 eggs per clutch	1.09		ALBM (ours)	Size	medium to large seabird with a wingspan ranging ...	8.74	LaBo	white head with a black cap	1.08
		Head	dark sooty-brown with a robust and streamlined shape, featuring ...	4.28		gray with a black eye mask	1.07			Beak	large, yellowish with a hooked tip, well-suited for catching fish ...	2.97		diet consists mostly of fish, squid, and crustans	1.07
		Back and belly	generally dark sooty brown in color, with the back slightly ...	4.25		grayish-olive with two white bars	1.05			Appearance	large seabird with a wingspan of up to 7 feet, predominantly ...	2.63		most numerous albatross species	1.07

Aircraft

B707-320		Attributes	Concepts	Score		Concepts	Score	B727-200		Attributes	Concepts	Score		Concepts	Score
	ALBM (ours)	Recognition	known for its distinctive four-engine configuration and ...	3.56	LaBo	tristar-shaped tail	1.18		ALBM (ours)	Model	a variant of the Boeing 727 series, specifically designed for ...	3.84	LaBo	delivered to federal express in 1986	1.13
		Variant	includes different sub-models like the 707-320B and 707-320C ...	3.39		very versatile	1.12			Variant	represents a specific model or configuration within the ...	3.26		each with a thrust of 17,000 pounds-force (76 kn)	1.11
		Type	a commercial jet airliner designed for medium to long-haul ...	3.08		entered service with pan american world airways in 1958	1.09			Color	typically features a combination of white with additional colors ...	3.20		introduced into service with eastern air lines in 1964	1.09

Flowers102

Carnation		Attributes	Concepts	Score		Concepts	Score	Pink Primrose		Attributes	Concepts	Score		Concepts	Score
	ALBM (ours)	Color	typically pink, red, white, or yellow petals with variations ...	1.75	LaBo	national flower of Spain and Monaco	1.04		ALBM (ours)	Color	a delicate pink hue with subtle variations ranging from pale ...	3.92	LaBo	5 petals that are all symmetrical and evenly spaced	1.02
		Structure	has a layered arrangement of ruffled petals, ...	1.37		classic symbol of mother's day	1.01			Center	a bright yellow core surrounded by delicate, overlapping pink ...	2.88		pastel pink color	1.01
		Shape	typically a round, ruffled bloom with multiple layers of frilled ...	1.30		realistic pink color	1.01			Shape	typically a five-petaled, symmetrical form with each petal slightly ...	2.83		known as the common primrose, english primrose, or flower primrose	1.01

Figure 3. Case study of bottlenecks constructed by ALBM and LaBo, where red texts indicate spurious cues, scores indicate concept activations. The top three highest-weighted concepts for each category are shown. Categories and datasets are selected randomly.

Thanks