PROMPT-CAM: Making Vision Transformers Interpretable for Fine-Grained Analysis

Supplementary Material

The supplementary is organized as follows.

- Appendix A: Related Work
- Appendix B: Details of Architecture Variant (cf. subsection 2.4 of the main paper)
- Appendix C: Dataset Details (cf. subsection 3.1 of the main paper)
- Appendix D: Inner Workings of Visualization (cf. subsection 2.3 of the main paper)
- Appendix E: Additional Experiment Settings (cf. subsection 3.1 of the main paper)
- Appendix F: Additional Experiment Results and Analysis (cf. subsection 3.2 of the main paper)
- Appendix G: More visualizations of different dataset (cf. Figure 4 of the main paper)

A. Related Work

Pre-trained Vision Transformer. Vision Transformers (ViT) [9], pre-trained on massive amounts of data, has become indispensable to modern AI development. For example, ViTs pre-trained with millions of image-text pairs via a contrastive objective function (*e.g.*, a CLIP-ViT model) show an unprecedented zero-shot capability, robustness to distribution shifts and serve as the encoders for various power generative models (*e.g.* Stable Diffusion [35] and LLaVA [19]). Domain-specific CLIP-based models like BioCLIP [38] and RemoteCLIP [18], trained on millions of specialized image-text pairs, outperform general-purpose CLIP models within their respective domains. Moreover, ViTs trained with self-supervised objectives on extensive sets of well-curated images, such as DINO and DINOv2 [4, 29], effectively capture fine-grained localization features that explicitly reveal object and part boundaries. We employ DINO, DINOv2, and BioCLIP as our backbone models in light of our focus on fine-grained analysis.

Prompting Vision Transformer. Traditional approaches to adapt pre-trained transformers—full fine-tuning and linear probing—face challenges: the former is computationally intensive and prone to overfitting, while the latter struggles with task-specific adaptation [22, 23]. Prompting, first popularized in natural language processing (NLP), addressed such challenges by prepending task-specific instructions to input text, enabling large language models like GPT-3 to perform zero-shot and few-shot learning effectively [3].

Recently, prompting has been introduced in vision transformers (ViTs) to enable efficient adaptation while leveraging the vast capabilities of pre-trained ViTs [12, 42, 53]. Visual Prompt Tuning (VPT) [12] introduces learnable embedding vectors, either in the first transformer layer or across layers, which serve as "prompts" while keeping the backbone frozen. This offers a lightweight and scalable alternative to full fine-tuning, achieving competitive performance on a diverse range of tasks while preserving the pre-trained features.

Explainable methods. Understanding the decision-making process of neural networks has gained significant traction, particularly in tasks where model transparency is critical. Explainable methods (XAI) focus on post-hoc analysis to provide insights into pre-trained models without altering their structure. Methods like Class Activation Mapping (CAM) [52] and Gradient-weighted CAM (Grad-CAM) [37] visualize class-specific contributions by projecting gradients onto feature maps. Subsequent improvements, such as Score-CAM [46] and Eigen-CAM [25], incorporate global feature contributions or principal component analysis to generate more detailed explanations. Despite these advancements, many XAI methods produce coarse, low-resolution heatmaps, which can be imprecise and fail to fully capture the model's decision-making process.

Interpretable methods. In contrast, interpretable methods provide a direct understanding of predictions by aligning intermediate representations with human-interpretable concepts. Early approaches such as ProtoPNet [6] utilized "learnable prototypes" to represent class-specific features, enabling visual comparison between input features and prototypical examples. Extensions like ProtoConcepts [21], ProtoPFormer [51], and TesNet [47] have refined this approach, integrating prototypes into transformer-based architectures to achieve higher accuracy and interoperability. More recent advancements leverage transformer architectures to enable interpretable decision-making. For example, Concept Transformers utilize query-based encoder-decoder designs to discover meaningful concepts [34], while methods like INTR [31] employ competing query mechanisms to elucidate how the model arrives at specific predictions. While these approaches offer fine-grained interpretability, they require substantial modifications to the backbone, leading to increased training complexity and longer computational times for new datasets.

PROMPT-CAM aims to overcome the shortcomings of both approaches. The special prediction mechanism encourages explainable, class-specific attention that is aligned well with model predictions. Simultaneously, we leverage pre-trained ViTs by simply modifying the usage of task-

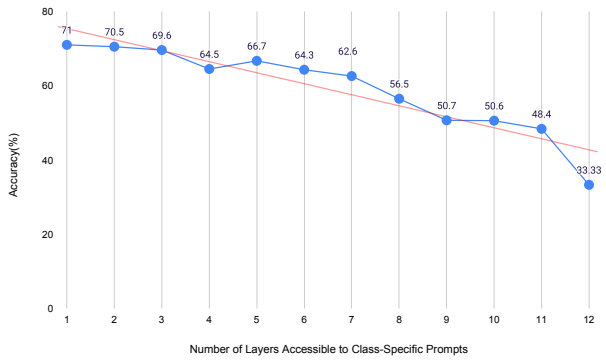

Figure 10. **Accuracy versus the number of layers (from last layer to first) attended by class-specific prompts.** As the number of attended layers increases in class-specific prompts, accuracy decreases, highlighting the importance of class-agnostic prompts. The more class-agnostic prompts a model has, the better trait localization and higher accuracy are achieved.

specific prompts without altering the backbone architecture.

B. Details of Architecture Variant

In this section, we explore variations of PROMPT-CAM by experimenting with the placement of class-specific prompts within the vision transformer (ViT) architecture. While PROMPT-CAM-SHALLOW introduces class-specific prompts in the first layer and PROMPT-CAM-DEEP applies them in the final layer, we also investigate injecting these prompts at various intermediate layers. Specifically, we control the layer depth at which class-specific prompts are added and analyze their impact on feature interpolation.

In PROMPT-CAM-SHALLOW, class-specific prompts are introduced at the first layer ($i = 1$), allowing them to interact with patch features across all transformer layers (*i.e.*, $\boldsymbol{E}_i$, $i = 0, \cdots, N-1$) without using class-agnostic prompts. As we increase the layer index i where class-specific prompts are added, the number of layers class-specific prompts interact decreases. At the same time, the number of preceding class-agnostic prompts increases, which interacts with the preceding $(i-1)$ layers (mentioned in subsection 2.2).

In Figure 11, we demonstrate the relationship between the number of layers accessible to class-specific prompts and their ability to localize fine-grained traits effectively. The visualization provides a clear pattern: as the prompts attend only to the last layer (first row) (same as PROMPT-CAM-DEEP), their focus is highly localized on discriminative traits, such as the red patch on the wings of the "Red-Winged Blackbird." This precise focus enables the model to excel in fine-grained trait analysis.

As we move downward through the rows, class-specific prompts attending to increasingly more layers (from top to bottom), the attention maps become progressively more diffused. For instance, in the middle rows (e.g., rows 6–8), the attention begins to cover broader regions of the object rather than the trait of interest. This diffusion correlates with a drop in accuracy, as seen in the accuracy plot, Figure 10.

In the bottom rows (e.g., rows 10–11), the attention becomes scattered and unfocused, covering irrelevant regions. This fails to correctly classify the object. The accuracy plot confirms this trend: as the class-specific prompts attend to more layers, accuracy steadily decreases.

C. Dataset Details

Table 3. Dataset statistics (Animals).

	Animals							
	Bird	CUB	Dog	Pet	Insects	Fish	Moth	RareS.
# Train Images	84,635	5,994	12,000	3,680	52,603	35,328	5,000	9,584
# Test Images	2,625	5,795	8,580	3,669	22,619	7,556	1,000	2,399
# Labels	525	200	120	37	102	414	100	400

Table 4. Dataset statistics (Plants & Fungi and Objects).

	Plants & Fungi			Objects	
	Flower	MedLeaf	Fungi	Car	Food
# Train Images	2,040	1,455	12,250	8,144	75,750
# Test Images	6,149	380	2,450	8,041	25,250
# Labels	102	30	245	196	101

We comprehensively evaluate the performance of PROMPT-CAM on a diverse set of benchmark datasets curated for fine-grained image classification across multiple domains. The evaluation includes animal-based datasets such as CUB-200-2011 (**CUB**) [45], Birds-525 (**Bird**) [33], Stanford Dogs (**Dog**) [15], Oxford Pet (**Pet**) [30], iNaturalist-2021-Moths (**Moth**) [43], Fish Vista (**Fish**) [24], Rare Species (**RareS.**) [41] and Insects-2 (**Insects**) [49]. Additionally, we assess performance on plant and fungi-based datasets, including iNaturalist-2021-Fungi (**Fungi**) [43], Oxford Flowers (**Flower**) [28] and Medicinal Leaf (**MedLeaf**) [36]. Finally, object-based datasets, such as Stanford Cars (**Car**) [16] and Food 101 (**Food**) [2], are also included to ensure comprehensive coverage across various fine-grained classification tasks. For the Moth and Fungi dataset, we extract species belonging to Noctuidae Family from taxonomic class *Animalia Arthropoda Insecta Lepidoptera Noctuidae* and species belonging to Agaricomycetes Class from taxonomic path *Fungi→ Basidiomycota*, respectively, from the iNaturalist-2021 dataset. For hierarchical classification and trait localization, we use taxonomical information from the Fish and iNaturalist-2021 dataset. We provide dataset statistics in Table 3 and Table 4.

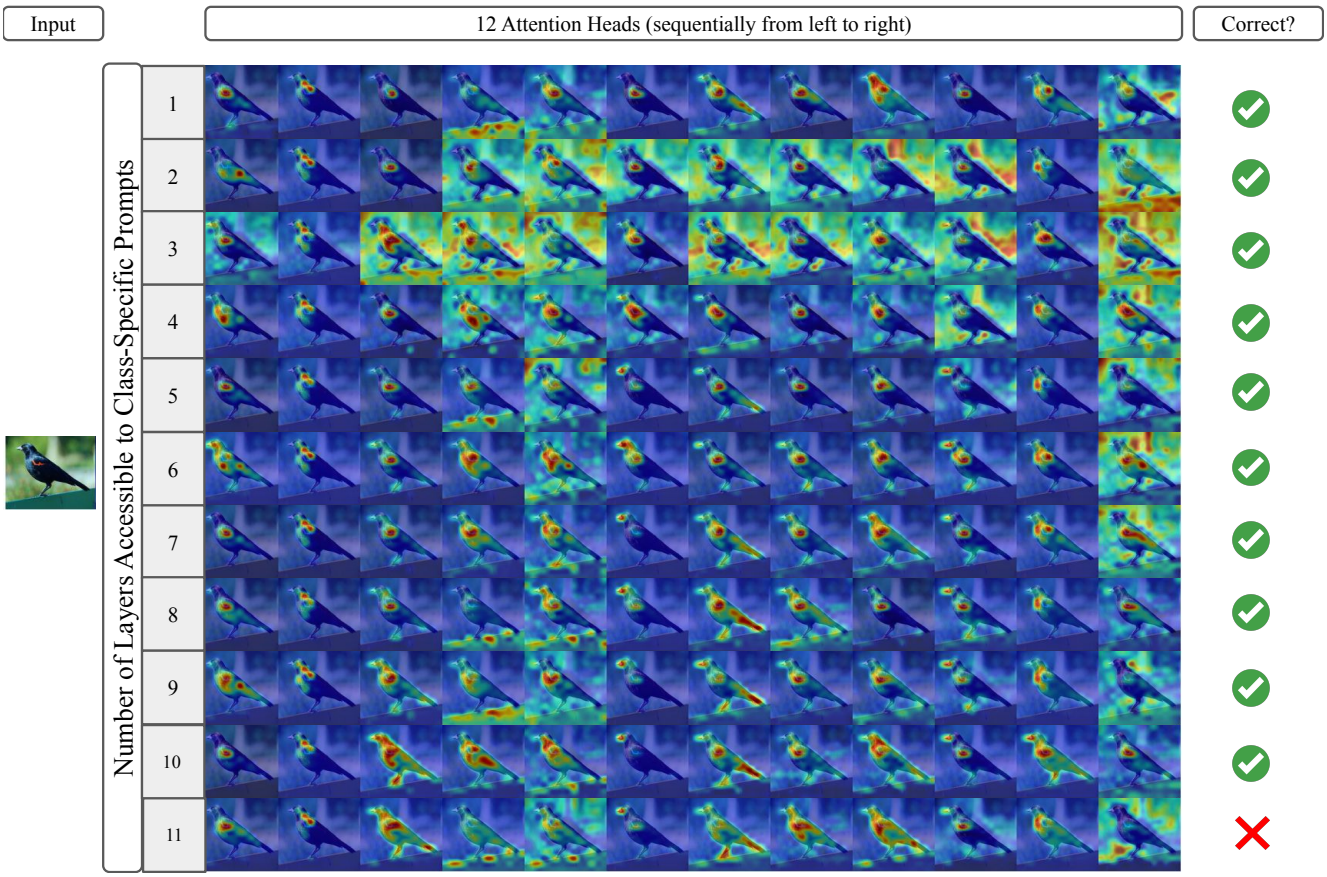

Figure 11. **Visualization of attention maps for different configurations of PROMPT-CAM.** For a random image of the "Red-Winged Blackbird" species, twelve attention heads of the last layer of PROMPT-CAM on the DINO backbone are shown for the ground truth class prompt. The first row shows class-specific prompts attending to only the last layer (as PROMPT-CAM-DEEP), resulting in highly localized attention on fine-grained traits, such as the red patch on the wings of the "Red-Winged Blackbird." As these prompts attend to increasingly more layers (progressing down the rows), the attention becomes more diffuse, covering broader regions of the object and eventually leading to a loss of focus on relevant traits.

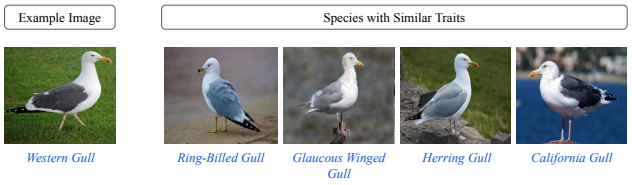

Figure 12. **Example Image of a "Western Gull" and its closest bird species, highlighting overlapping traits.** Correctly classifying the "Western Gull" requires attention to multiple subtle traits, as it shares many traits with similar species. This highlights the need to examine a broader range of attributes for accurate classification.

D. Inner Workings of Visualization

Which traits are more discriminative? As discussed in subsection 2.3, certain categories within the CUB dataset exhibit distinctive traits that are highly discriminative. For instance, in the case of the "Red-winged Blackbird," the defining features are its red-spotted black wings. Similarly, the "Ruby-throated Hummingbird" is characterized by its ruby-colored throat and sharp, long beak. However, some species require consideration of multiple traits to distinguish them from others. For example, correctly classifying a "Western Gull" demands attention to several subtle traits (Figure 12), as it shares many features with other species. This observation raises a key question: can we automatically identify and rank the most important traits for a given image of a species?

To address this, we propose a greedy algorithm that progressively "blurs" traits in a correctly classified image until its decision changes. This process reveals the traits that are both necessary and sufficient for the correct prediction.

Greedy approach for identifying discriminative traits: Suppose class c is the true class and the image is correctly classified. In the first greedy step, for each attention head,

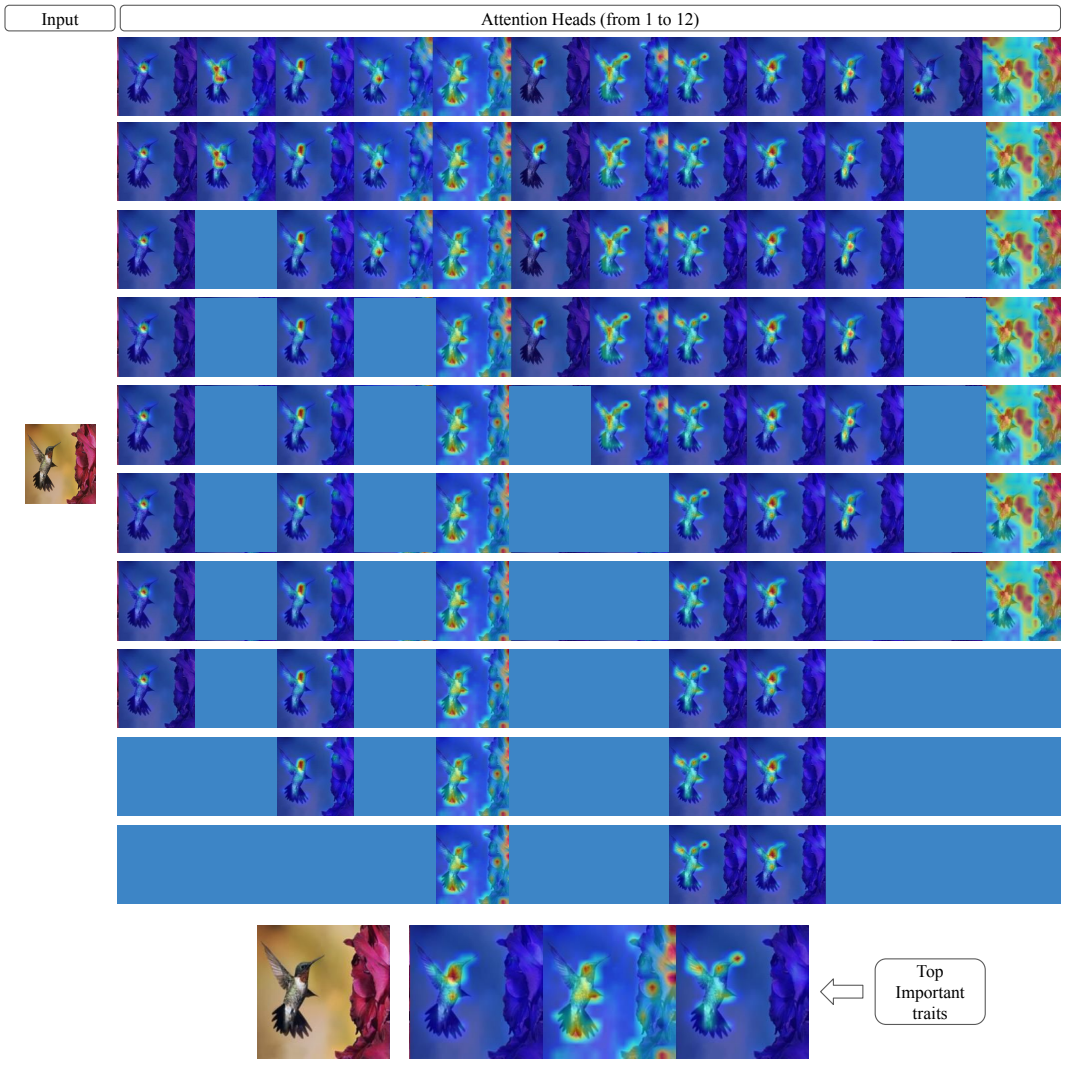

Figure 13. **Greedy approach to identify and rank important traits for species classification.** For the species "Ruby Throated Hummingbird", we progressively blur attention heads (from top to bottom), retaining only the traits necessary for correct classification, using the PROMPT-CAM on the DINO backbone. The blurred attention heads are shown in solid blue color.

$r = 1, \cdots, R$ (R attention heads), we iteratively replace the attention vector $\boldsymbol{\alpha}_{N-1}^{c,r}$, with a uniform distribution:

$$\boldsymbol{\alpha}_{N-1}^{c,r} \leftarrow \frac{1}{M}\mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^M$ is a vector of all ones, and M is the number of patches. This replacement effectively assigns equal importance to all patches in the attention weights, thereby "blurring" the r-th head's contribution to class c. After this modification, we recalculate the score $s[c]$ in Equation 1.

For each iteration, we select the attention head r^* that, when blurred, results in the highest probability for the correct class c. This head r^* is then added to B_a (set of blurred attention heads), as the *blurred* head with the *highest* $s[c]$ is the *least* important and contributes the least discriminative information for class c. We repeat this process, iteratively blurring additional heads and updating B_a, until blurring any remaining head not in B_a changes the model's prediction. In Figure 13, for an image of "Ruby Throated Hummingbird" we show this greedy approach, by progressively blurring out the attention heads in each step, retaining only necessary traits.

Attention head vs species. In addition to image-level analysis, we conduct a species-level investigation to determine whether certain attention heads consistently focus on important traits across all images of a species. Using the greedy approach discussed in the above paragraph, we analyze each correctly classified image of a species c to iteratively select the attention head r^* that minimally impacts the probability of the correct class c. We then examine how the probability $s[c]$ changes as attention heads are progressively blurred

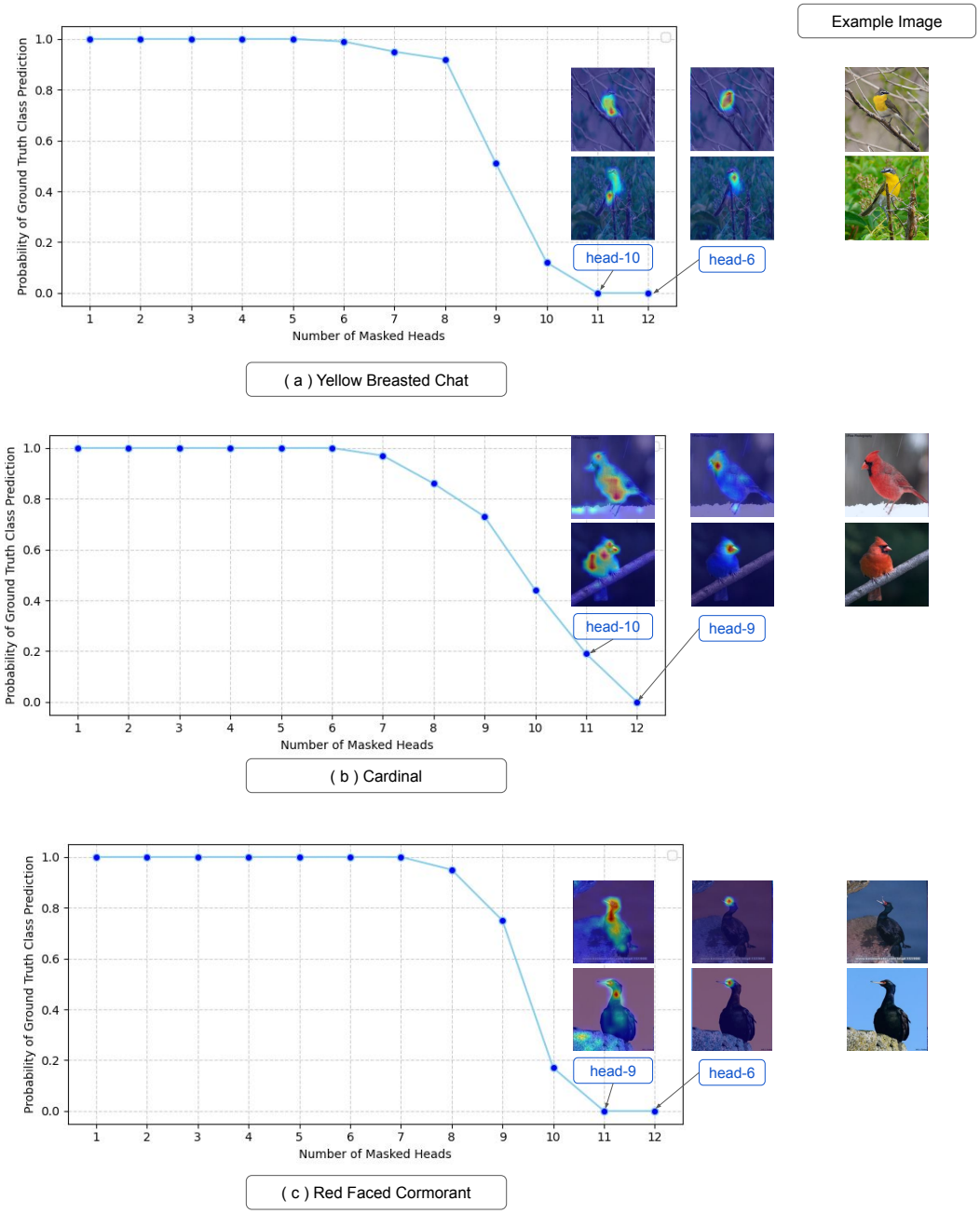

Figure 14. **Visualization of ground truth class probability vs. the number of masked heads at the species level in PROMPT-CAM.** The left plots show how the probability of the ground truth class changes for all correctly classified images in a species, as heads are progressively masked in the greedy approach discussed in Appendix D. For class (a) "Yellow Breasted Chat," the probability drops significantly after masking eight heads, indicating that the last four heads are critical. The top two heads, head-6 and head-10, focus on the yellow breast and lower belly. For class (b) "Cardinal," the top 2 heads, head-9 and head-10, attend to the black pattern on the face and the red belly. In class (c) "Red Faced Cormorant," the critical heads, head-6 and head-9, emphasize the red head and the neck's shape. These results highlight the interpretability of PROMPT-CAM in identifying essential traits for each species.

or masked for all images of a species. This analysis, visualized in Figure 14, demonstrates that for most species in the CUB dataset, approximately four attention heads capture traits critical for class prediction. In the Figure 14, we highlight the top-2 attention heads for example images from various species. The results reveal that these heads

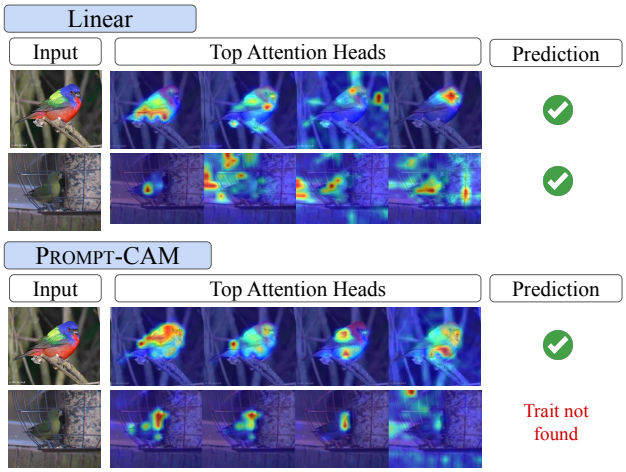

Figure 15. **Comparison of top attention heads for PROMPT-CAM and Linear probing on two images of the species "Painted Bunting."** For the correctly classified image by both, PROMPT-CAM focuses on meaningful traits such as the blue head, wings, tail, and red lower belly, while Linear probing produces noisy and less diverse heatmaps. For the other image, Linear probing relies on global memorized attributes for correct classification, whereas PROMPT-CAM attempts to identify object-specific traits, resulting in an interpretable misclassification due to poor visibility of key features.

consistently focus on important, distinctive traits for their respective species. For instance, in the case of the "Cardinal", head-9 focuses on the black stripe near the beak, while head-10 attends to the red breast color—traits essential for identifying the species. Similarly, for "Yellow-breasted Chat" and "Red-faced Cormorant", attention heads consistently highlight relevant features across their respective species. These findings emphasize the robustness of our approach in identifying class-specific discriminative traits and the flexibility of choosing any number of ranked important traits per species.

E. Additional Experiment Settings

E.1. Implementation Details

Dataset-specific settings. For DINO backbone, the learning rate varied across datasets within the set $\{0.01, 0.1, 0.125\}$, selected based on dataset-specific characteristics. For Bird and MedLeaf, training was conducted for 30 epochs. For all other datasets, training was conducted for 100 epochs. For DINOv2 backbone, the learning rate varied across datasets within the set of $\{0.005, 0.01\}$, selected based on dataset-specific characteristics. For Insect, CUB, and Bird, training was conducted for 130 epochs. For all other datasets, training was conducted for 100 epochs. For DINOv2 backbone, the learning rate varied across datasets within the set of $\{0.05, 0.01\}$, selected based on dataset-specific characteristics. For all datasets, training was conducted for 100 epochs. A batch size of 64 was used for all datasets and all backbones.

Optimization settings. Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. Weight Decay 0.0 was used for all datasets for DINO, 0.001 for the rest. A cosine learning rate scheduler was applied, with a warmup period of 10 epochs and cross-entropy loss was used.

E.2. Baseline Methods

We used **XAI methods** Grad-CAM, Score-CAM, and Eigen-CAM to compare PROMPT-CAM performance with them on a quantitative scale. For qualitative comparison, we compare with a variety of **interpretable methods**, ProtoPFormer, TesNet, INTR, and ProtoPConcepts shown in Figure 6.

F. Additional Experiment Results

Model performance analysis. As discussed in subsection 2.3, we analyze misclassified examples by PROMPT-CAM, illustrated in Figure 5. We attribute the slight decline in accuracy of PROMPT-CAM to its approach of forcing prompts to focus on the object itself and its traits, rather than relying on surrounding context for classification. In Figure 15, we compare the heatmaps of two images of the species "Painted Bunting". The first image, I_c, is correctly classified by both PROMPT-CAM and Linear probing, while the second image, I_m, is correctly classified by Linear probing but misclassified by PROMPT-CAM. The image I_m presents additional challenges: it is poorly lit, further from the camera, and depicts a less common gender of the species in the CUB dataset.

For I_c, the top heatmaps from Linear probing appear noisy and less diverse compared to PROMPT-CAM. In contrast, PROMPT-CAM exhibits a more meaningful focus, with its top attention heads targeting the blue head, part of the wings, the tail, and the red lower belly—traits characteristic of the species.

In the case of I_m, although Linear probing predicts the image correctly, its top attention heads fail to focus on consistent traits. Instead, they appear to rely on global features memorized from the training dataset, resulting in a lack of meaningful interpretation. On the other hand, PROMPT-CAM, despite misclassifying I_m, focuses its attention on traits within the object itself. The heatmaps reveal that PROMPT-CAM attempts to identify relevant features, but the lack of visible traits in the image leads to an interpretable misclassification.

In Figure 16, the comparison between Linear Probing and PROMPT-CAM in the attention heatmaps reveals a fundamental difference in their classification and trait identification approach. As shown in the heatmaps, Linear Probing

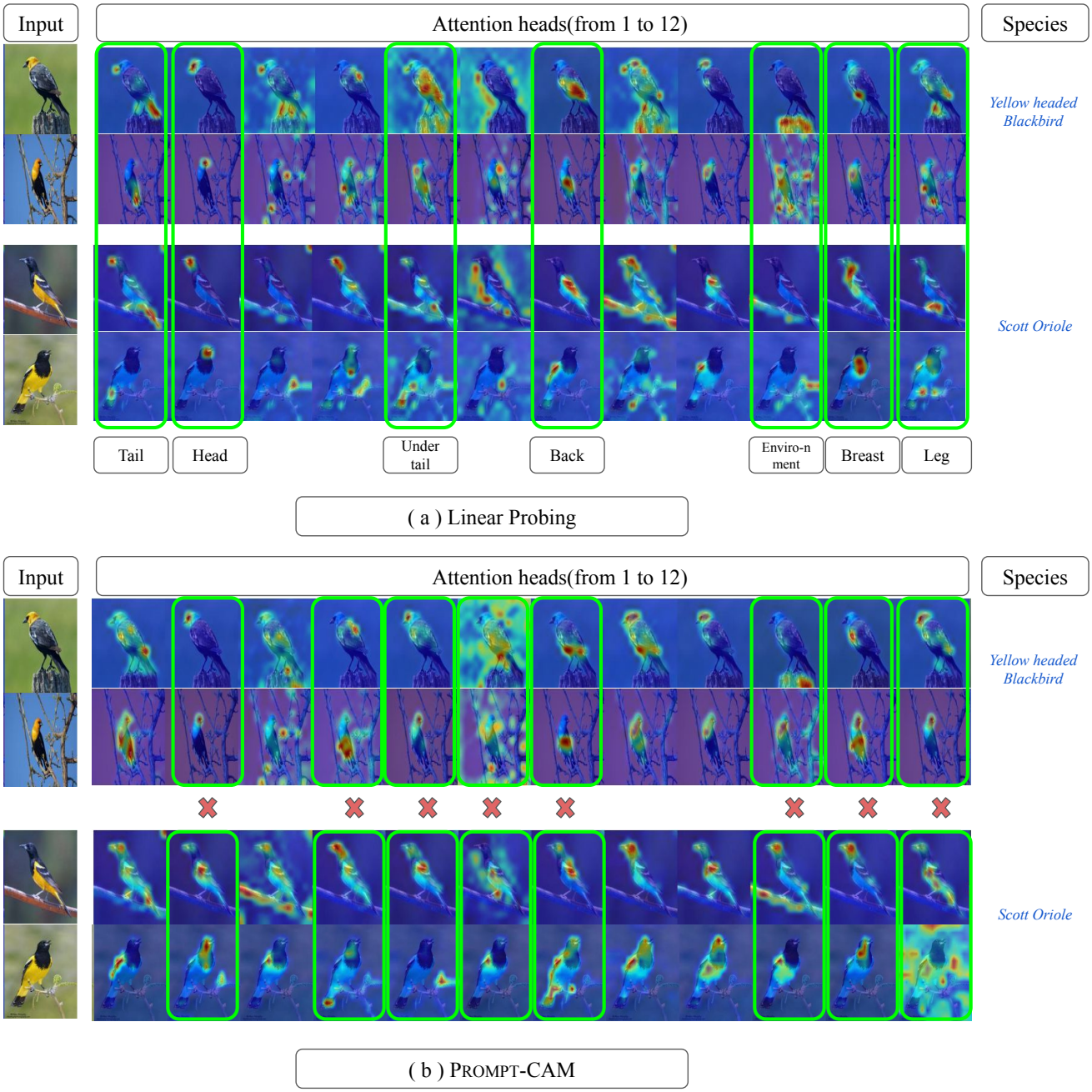

Figure 16. **Comparison of attention heatmaps for Linear Probing and PROMPT-CAM.** On random images of "Yellow Headed Blackbird" and "Scott Oriole" from the CUB dataset, in (a), Linear Probing consistently focuses on similar body parts (e.g., tail, head, under-tail, wings) across all species, showing limited adaptability to traits specific to each class. In contrast, (b) PROMPT-CAM (using pretrained DINO) dynamically adapts its attention to focus on distinct and meaningful traits required for class-specific identification. For instance, PROMPT-CAM highlights traits such as the yellow head and breast for "Yellow Headed Blackbird" and the wing pattern for "Scott Oriole".

uniformly distributes its attention across similar body parts, such as the tail, head, and wings, irrespective of the species being analyzed. This behavior indicates that Linear Probing relies on global patterns that may not be specific to any particular class. In contrast, for each species, PROMPT-CAM focuses on specific traits important for differentiating one class from another. For example, in the case of the "Yellow Headed Blackbird," PROMPT-CAM emphasizes the yellow head and breast, traits unique to the species. Similarly, for the "Scott Oriole," the yellow breast and wing patterns are

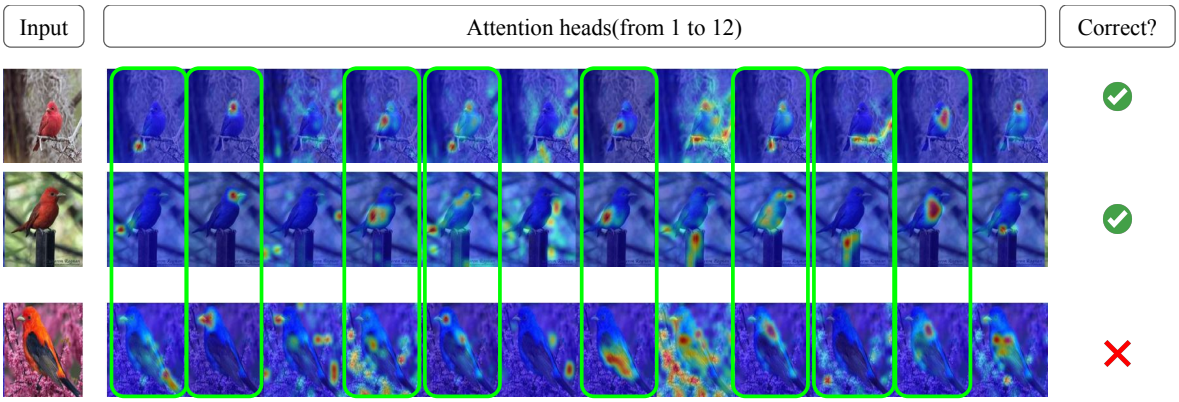

Figure 17. **Attention heatmaps of cls-token for Linear Probing on misclassified images.** For some random images of "Scarlet Tanager" from the CUB dataset, Linear Probing highlights the same body parts across images, failing to provide meaningful insights into misclassifications.

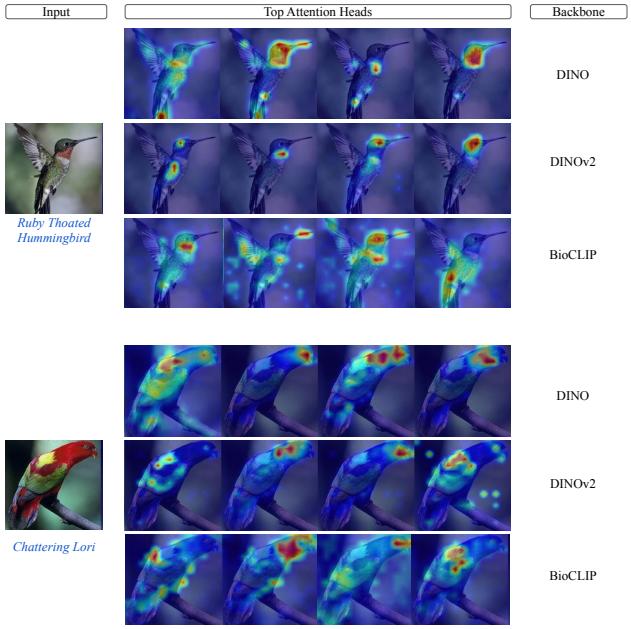

Figure 18. **Visualization of top attention heads of PROMPT-CAM for DINO, DINOv2 and BioCLIP backbones.** For random correctly classified images from "Ruby Throated Hummingbird" and "Chattering Lori" species from Bird Dataset, top-4 attention heads (from left to right) are shown. PROMPT-CAM can identify and locate meaningful important traits for species regardless of pre-trained visual backbone used.

prominently highlighted. By prioritizing traits essential for species identification, PROMPT-CAM provides a more robust and meaningful framework for understanding model decisions.

Furthermore, in Figure 17, we present attention heatmaps for random images of the "Scarlet Tanager" species from the CUB dataset, generated using Linear Probing. Linear Probing consistently assigns attention to the same body parts (e.g., wings, head) across images, without providing meaningful insights into the reasons for misclassification. In contrast, PROMPT-CAM (as shown in Figure 8 and Figure 15) provides a more interpretable explanation for misclassifications. When PROMPT-CAM misclassifies an image, it is evident that the misclassification occurs due to the absence of the necessary trait in the image, demonstrating its focus on biologically relevant and class-specific traits.

This analysis underscores PROMPT-CAM prioritizes interpretability, ensuring that its classifications are based on meaningful and consistent traits, even at the cost of a slight accuracy decline.

Human assessment of trait identification settings. In subsection 3.2, we discussed how we measured robustness of PROMPT-CAM with assessment from human observers. To evaluate the effectiveness of trait identification, in the human assessment, we compared PROMPT-CAM, TesNet [47], and ProtoConcepts [21]. A total of 35 participants with no prior knowledge of the models participated in the study. Participants were presented with a set of top attention heatmaps (PROMPT-CAM and INTR) or prototypes generated by each method and image-specific class attributes found in CUB dataset. Then they were asked to identify and check the traits they perceived as being highlighted in the heatmaps. The traits were taken from the CUB dataset, where image-specific traits are present. We used four random correctly classified images by every method, from four species "Cardinal", "Painted Bunting", "Rose Breasted Grosbeak" and "Red faced Cormorant" to generate attention heatmaps/prototypes.

The assessment revealed that participants recognized 60.49% of the traits highlighted by PROMPT-CAM, significantly outperforming TesNet and ProtoConcepts, which achieved recognition rates of 39.14% and 30.39%, respec-

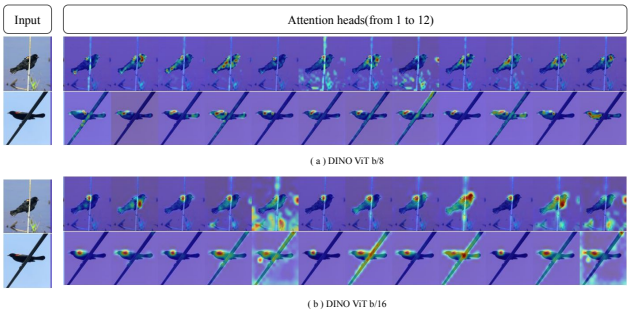

Figure 19. **Visualization of attention heads for pre-trained DINO backbone variants.** For correctly classified images of "Red winged blackbird", with PROMPT-CAM, both DINO ViT b/16 and DINO ViT b/8 backbones can capture traits for classification.

tively. These findings demonstrate PROMPT-CAM's superior ability to emphasize and communicate relevant traits effectively to human observers.

PROMPT-CAM on different backbones. We implement PROMPT-CAM on multiple pre-trained vision transformers, including DINO, DINOv2, and Bioclip. In Table 5, we present the accuracy of PROMPT-CAM across various datasets using different backbones: DINO (ViT-Base/16), DINOv2 (ViT-Base/14), and Bioclip (ViT-Base/16). For each model, we visualize the top-4 attention heads on the Bird Dataset in Figure 18. Notably, Bioclip achieves higher accuracy on biology-specific datasets, which we attribute to its pre-training on an extensive biology-focused dataset, enabling it to develop a highly specialized feature space for these species. Additionally, we also evaluate PROMPT-CAM on other DINO variations, ViT-Base/8 (accuracy: 73.9%) and ViT-Small/8 (accuracy: 68.3%) on the CUB dataset, achieving comparable performance and interpretability to DINO ViT-Base/16 (accuracy: 71.9%) (shown in Figure 19). This demonstrates PROMPT-CAM's robustness, flexibility, and ease of implementation across various pre-trained vision transformer backbones and datasets.

Table 5. **Accuracy of PROMPT-CAM on different backbones.** To show the flexibility and robustness, the accuracy of PROMPT-CAM on multiple datasets is shown implemented on pre-trained vision transformers, DINO, DINOv2 and BioCLIP.

		Bird	CUB	Dog	Pet	Insects-2	Flowers	Med.Leaf	Rare Species
Ours	DINO	98.2	73.2	81.1	91.3	64.7	86.4	99.1	60.8
	DINOv2	98.2	74.1	81.3	92.7	70.6	91.9	99.6	62.2
	BioCLIP	98.6	84.0	73.1	87.2	71.8	95.7	99.6	67.1

Taxonomical hierarchy trait discovery settings. In hierarchical taxonomic classification in biology, each level in the taxonomy leverages specific traits for classification. As we move down the taxonomic hierarchy, the traits become increasingly fine-grained. Motivated by this observation, we trained and visualized traits in a hierarchical taxonomic manner using the Fish Vista dataset.

We first constructed a taxonomic tree spanning from *Kingdom* to *Species*. For the *Family* level, we aggregated all images belonging to the diverse species under their respective *Family* and performed classification to assign images to the appropriate *Family*. As shown in Figure 9, even coarse traits, such as the tail and pelvic fin, were sufficient to classify an image of the species "Amphiprion Melanopus" to its' correct *Family* (attribute information found in Fish Dataset).

At the *Genus* level, we create a new dataset for each *Family* by grouping all images from the children nodes of each *Family* and dividing them into classes by their respective *Genus*. For instance, within the "Pomacentridae" Family, finer traits like stripe patterns, pelvic fins, and tails became necessary to classify its' *Genus* accurately for the same example image. Finally, at the *Species* level, all images from the children nodes of each *Genus* were used to create a new dataset and were divided into classes. For the example image in Figure 9, distinguishing between these two species now requires looking at subtle differences such as the pelvic fin structure and the number of white stripes on the body for the same image from the "Amphiprion Melanopus" species. This hierarchical approach offers an exciting framework to discover traits in a manner that is both evolutionary and biologically meaningful, enabling a deeper understanding of trait importance across taxonomic levels.

G. More Visualizations

In this section, we show the top-4 attention maps triggered by ground truth classes for correctly predicted classes, for some datasets mentioned Appendix C, following the same format of Figure 4. Each attention head of PROMPT-CAM for each dataset successfully identifies different and important attributes of each class of every dataset. For some datasets, if the images of a class are simple enough, we might need less than four heads to predict.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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*, pages 446–461. Springer, 2014. 2, 6

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler,

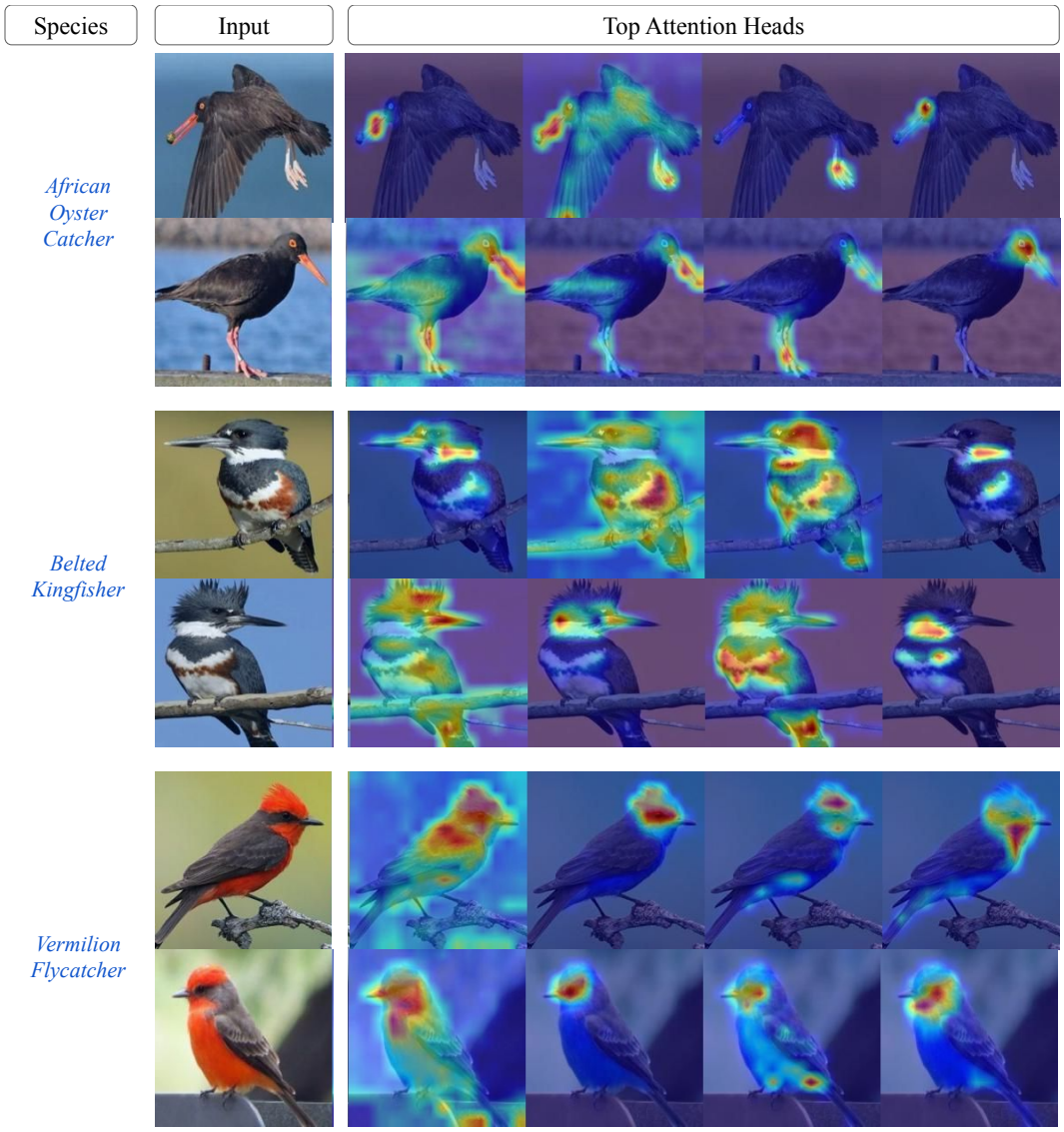

Figure 20. **Visualization of PROMPT-CAM on Bird Dataset.** We show the top four attention maps (from left to right) per correctly classified test example, triggered by the ground-truth classes. As top head indices per image may vary, traits may not align across columns.

Jeffrey Wu, Clemens Winter, Christopher 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 6, 1

[5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 2

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2, 1

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2

[8] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 36, 2024. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,

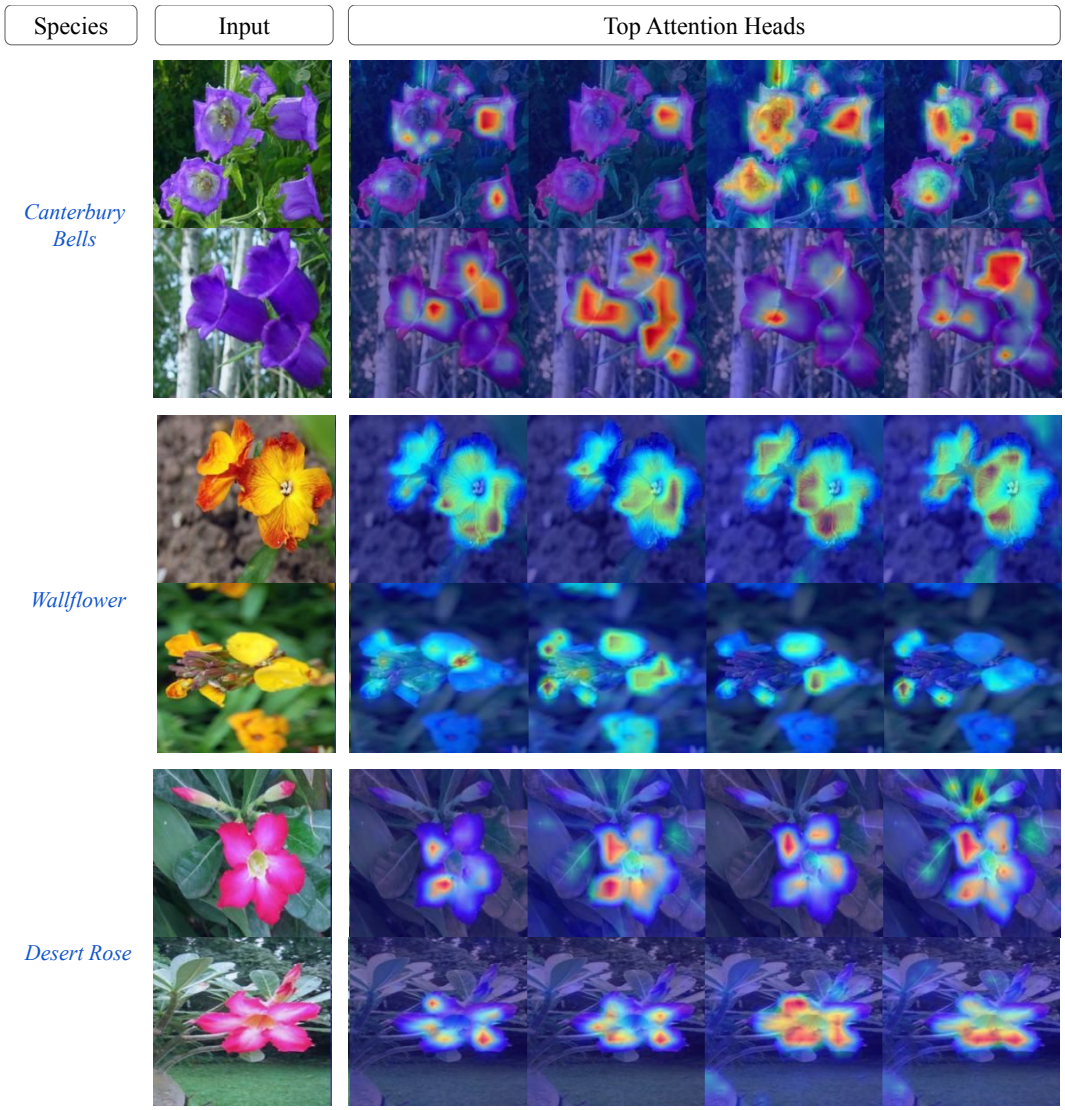

Figure 21. **Visualization of PROMPT-CAM on Flower Dataset.** We show the top four attention maps (from left to right) per correctly classified test example, triggered by the ground-truth classes. As top head indices per image may vary, traits may not align across columns.

Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3

[10] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 852–860, 2022. 2

[11] Darneisha A Jackson and Keith M Somers. The spectre of 'spurious' correlations. *Oecologia*, 86:147–151, 1991. 5

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 3, 4, 1

[13] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2, 6

[14] Rojina Kashefi, Leili Barekatain, Mohammad Sabokrou, and Fatemeh Aghaeipoor. Explainability of vision transformers: A comprehensive review and new perspectives. *arXiv preprint arXiv:2311.06786*, 2023. 2, 6

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings CVPR work-*

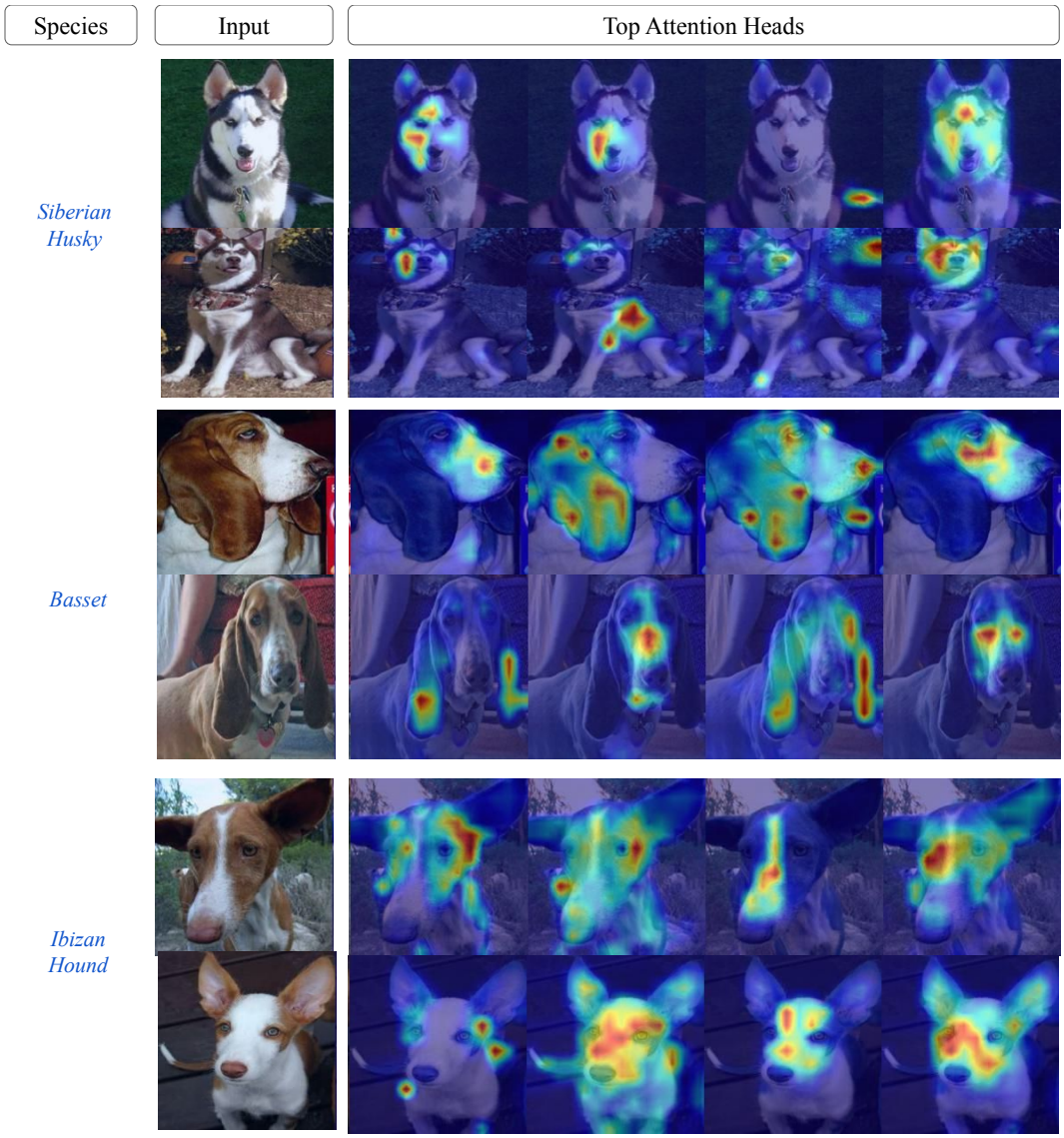

Figure 22. **Visualization of PROMPT-CAM on Dog Dataset.** We show the top four attention maps (from left to right) per correctly classified test example, triggered by the ground-truth classes. As top head indices per image may vary, traits may not align across columns.

shop on fine-grained visual categorization (FGVC), 2011. 2, 6

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2, 6

[17] Ruiwen Li, Zheda Mai, Zhibo Zhang, Jongseong Jang, and Scott Sanner. Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. *Journal of Visual Communication and Image Representation*, 92:103800, 2023. 2

[18] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, 2024. 1

[20] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 2, 3

[21] Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 1, 8

[22] Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song

Gao, Charles Stewart, Yu Su, et al. Fine-tuning is fine, if calibrated. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[23] Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Li Zhang, and Wei-Lun Chao. Lessons learned from a unifying empirical study of parameter-efficient transfer learning (petl) in visual recognition. *arXiv preprint arXiv:2409.16434*, 2024. 1

[24] Kazi Sajeed Mehrab, M Maruf, Arka Daw, Harish Babu Manogaran, Abhilash Neog, Mridul Khurana, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, Matthew J Thompson, et al. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images. *arXiv preprint arXiv:2407.08027*, 2024. 2, 6

[25] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020. 2, 6, 1

[26] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14933–14943, 2021. 2

[27] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Dreamcreature: Crafting photorealistic virtual creatures from imagination. *arXiv preprint arXiv:2311.15477*, 2023. 2

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 2, 6

[29] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 6, 1

[30] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505, 2012. 2, 6

[31] Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Carlyn, Samuel Stevens, Kaiya Provost, Anuj Karpatne, Bryan Carstens, Daniel Rubenstein, Charles Stewart, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. A simple interpretable transformer for fine-grained image classification and analysis. In *International Conference on Learning Representations*, 2024. 2, 3, 5, 6, 1

[32] V Petsiuk, A Das, and K Saenko. Rise: Randomized input sampling for explanation of black-box models. arxiv 2018. *arXiv preprint arXiv:1806.07421*, 1806. 6

[33] Gerald Piosenka. Birds 525 species - image classification. 2023. 2, 6

[34] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International conference on learning representations*, 2021. 1

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[36] Roopashree S and Anitha J. Medicinal Leaf Dataset, 2020. Mendeley Data, V1. 2, 6

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 6, 1

[38] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 6, 1

[39] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2

[40] Zhenchao Tang, Hualin Yang, and Calvin Yu-Chian Chen. Weakly supervised posture mining for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23735–23744, 2023. 2

[41] Imageomics Team. Rare Species Dataset, 2023. Dataset with 400 classes of rare species images and descriptions sourced from the Encyclopedia of Life and the IUCN Red List. 2, 6

[42] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7725–7735, 2023. 1

[43] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 2, 6

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017. 3

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 6

[46] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition workshops, pages 24–25, 2020. 1

[47] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 895–904, 2021. 6, 1, 8

[48] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-set fine-grained retrieval via prompting vision-language evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19381–19391, 2023. 2

[49] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8787–8796, 2019. 2, 6

[50] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4310–4319, 2022. 2, 3

[51] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 6, 1

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 1

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[54] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4692–4702, 2022. 2