Supplementary material for "HIVE: Evaluating the Human Interpretability of Visual Explanations"

In this document, we provide additional details on some sections of the main paper. Code is available at `https://princetonvisualai.github.io/HIVE`.

Section A: We provide more information on the evaluation tasks.
Section B: We provide more information on the four evaluated methods and the modifications we made to their original explanation form.
Section C:: We provide more information about our human studies.
Section D: We report additional results and analyses.

- **Section D.1:** We supplement Section 5.2 of the main paper and discuss the *agreement* study results with vs. without examples from the predicted class.
- **Section D.2:** We supplement Section 5.3 of the main paper and provide more details on our analysis with automatic evaluation metrics.
- **Section D.3:** We supplement Section 5.4 of the main paper and discuss the participants' similarity ratings and decisions. We also provide a plot of the participant vs. ProtoPNet prototype similarity ratings.
- **Section D.4:** We supplement Section 5.5 of the main paper and provide the full results of our subjective evaluation.
- **Section D.5:** We supplement Section 5.6 of the main paper and provide the full results of our interpretability-accuracy tradeoff study.

Section E: We show the simple decision tree model for fruit classification we used to introduce ProtoTree.
Section F: We show snapshots of our full user interface.

A Details on the evaluation tasks

Agreement task. For each image, we show one model prediction-explanation pair and ask the participants how confident they are in the model's prediction. We show 10 images in total (5 correct, 5 incorrect predictions in random order). Participants rate their confidence in the given prediction on a 4-point scale (1: confident prediction is incorrect, 2: somewhat confident prediction is incorrect, 3: somewhat confident prediction is correct, 4: confident prediction is correct).
Distinction task. For each image, we show four model prediction-explanation pairs for it (in random order) and ask the participants to identify the correct prediction based on the explanations. For GradCAM [10] and BagNet [1], participants are tasked with 10 sample images (5 correct and 5 incorrect predictions), each of which is shown with four heatmaps. On correctly predicted samples, the four heatmaps correspond to the top-4 predicted classes. On incorrectly predicted ones, we show heatmaps for the top-3 predicted classes and the heatmap of the ground-truth class. For ProtoPNet [2], we show four correctly predicted samples in total. Each sample is presented with four explanations corresponding to the top-4 predicted classes. We reduce the total number of samples and focus on correctly predicted samples due to the complexity of the ProtoPNet

explanations; even with this change, the ProtoPNet study duration is twice as long as that of GradCAM and BagNet. For ProtoTree [7], we show 10 correctly predicted samples in total and ask participants to select the correct decisions on the two final nodes which lead to four (2^2) different predictions.

Additionally for ProtoPNet [2] and ProtoTree [7], we ask participants to rate the similarity of prototype-region pairs in both tasks using a 4-point Likert scale (1: not similar, 2: somewhat not similar, 3: somewhat similar, 4: similar).

B Details on the evaluated interpretability methods

GradCAM [10]. For our ImageNet [9] studies, we generate GradCAM explanations for the ResNet50 [5] model in the `torchvision` library which achieves 76.1% classification accuracy. For our CUB studies, we generate GradCAM explanations for a ResNet50 [5] model we trained on the CUB [11] training set. This model achieves 81.0% accuracy on the CUB test set. We used the code by Gildenblat et al. [4] to generate GradCAM visualizations.[1] For the *agreement* task, we generate the GradCAM heatmap for the model prediction and normalize it into the [0, 1] range. For the *distinction* task, we generate four GradCAM heatmaps for each image: for correct predictions, we generate heatmaps for the top-4 predictions; for incorrect predictions, we generate heatmaps for the top-3 predictions and for the ground-truth class. We identify the local minimum and maximum of the four heatmaps, and then normalize the heatmaps into the [0, 1] range. This way, we preserve the intensity difference between heatmaps for different predictions. See Fig. A1 for an example set of GradCAM explanations.
BagNet [1]. For our ImageNet studies, we use the BagNet33 model trained by the original authors which achieves 66.7% accuracy on ImageNet classification. For our CUB studies, we train a BagNet33 model on the CUB training set. This model achieves 74.2% accuracy on the CUB test set. For the *agreement* task, we use the authors' code as is and normalize each heatmap individually by clipping the values above the 99th percentile.[2] On the other hand, for the *distinction* task, we normalize the four heatmaps together so that we preserve the intensity difference. See Fig. A2 for an example set of BagNet explanations.

[1] `https://github.com/jacobgil/pytorch-grad-cam`
[2] `https://github.com/wielandbrendel/bag-of-local-features-models`

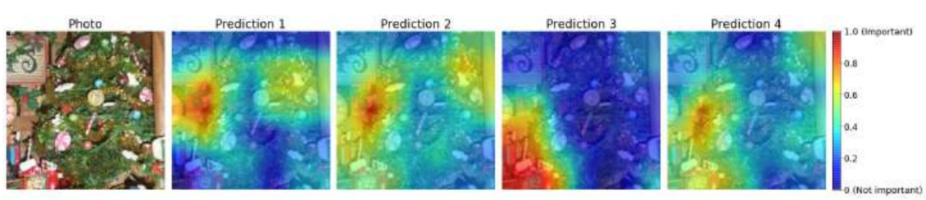

Fig. A1. GradCAM explanations shown in the *distinction* task.

ProtoPNet [2]. For ProtoPNet, we used the ResNet34-based model trained by Hoffmann et al. [6]. We pruned 331 prototypes from this model to improve interpretability. The resulting model has 1669 prototypes and achieves 79.9% accuracy on the CUB [11] test set. For generating explanations, we used the code by the original authors with some modifications which we describe below.[3] In our studies, given an explanation, participants are asked to rate the similarity of each prototype-region pair, then either rate the level of confidence in the prediction's correctness *(agreement)* or select the correct class *(distinction)*. To make ProtoPNet's explanations more suitable for these tasks, we made the following modifications to the original explanation form.

– The ProtoPNet model calculates evidence for all classes using the learned prototypes, then predicts the class with the highest evidence. However, we deemed it is unrealistic to ask users to review explanations for all 200 bird classes in CUB. Hence, we only present explanations for one (*agreement*) or four (*distinction*) classes and ask users to examine them.
– The original explanation (Fig. A3 left) shows activation maps, similarity scores, class connection weights, and the total class evidence. In our version (Fig. A3 right) we remove them as we seek to investigate what participants rate as similar and not.
– In the original explanation, prototypes are presented in the order of highest to lowest similarity. In ours, we randomly shuffle the order of prototypes because we don't want to skew the participants' region-prototype similarity ratings.

ProtoTree [7]. For ProtoTree, we used the model trained by the original authors which achieves 81.7% accuracy on the CUB [11] test set. This model is a pruned tree of depth 10 and 511 nodes. We used the authors' code to generate explanations with some modifications we describe below.[4]

– Same as what we did for ProtoPNet explanations, we removed the similarity scores as we seek to investigate what participants rate as similar and not.
– For the local explanation, we converted the horizontal explanation (Fig. A4) into a vertical one (Fig. A5). A vertical explanation is a better representation of how the model reasons, as the model starts from the root node and proceeds down the tree until it reaches one of the bottom leaves. Further, it is easier for the participants to examine the explanation by scrolling up and down.

[3] `https://github.com/cfchen-duke/ProtoPNet`
[4] `https://github.com/M-Nauta/ProtoTree`

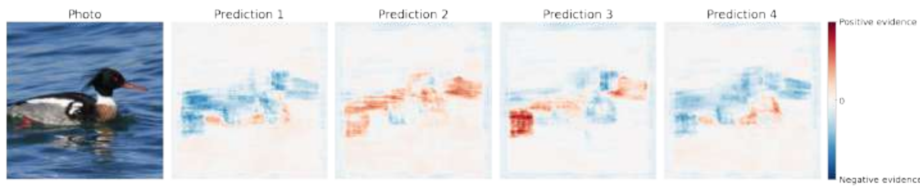

Fig. A2. BagNet explanations shown in the *distinction* task.

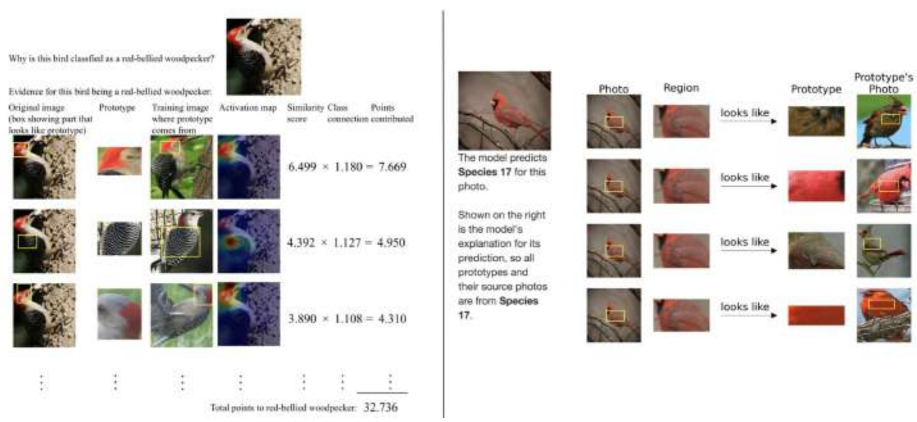

Fig. A3. ProtoPNet original and modified explanations. The original explanation (left) taken from the original paper [2] contains details such as activation maps, similarity scores, and class connection weights. In our version (right), we remove these to abstract away the complexities and have the participants focus on examining the similarity between prototypes and their matched image regions.

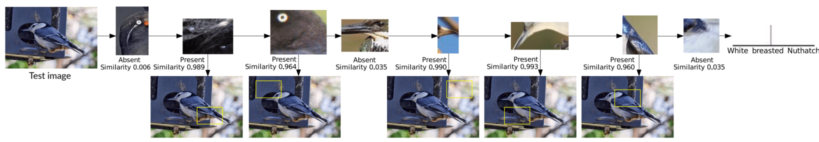

Fig. A4. ProtoTree original explanation. We show the original explanation displayed in Fig. 9 of the original paper [7]. See Fig. A5 for our modified explanation.

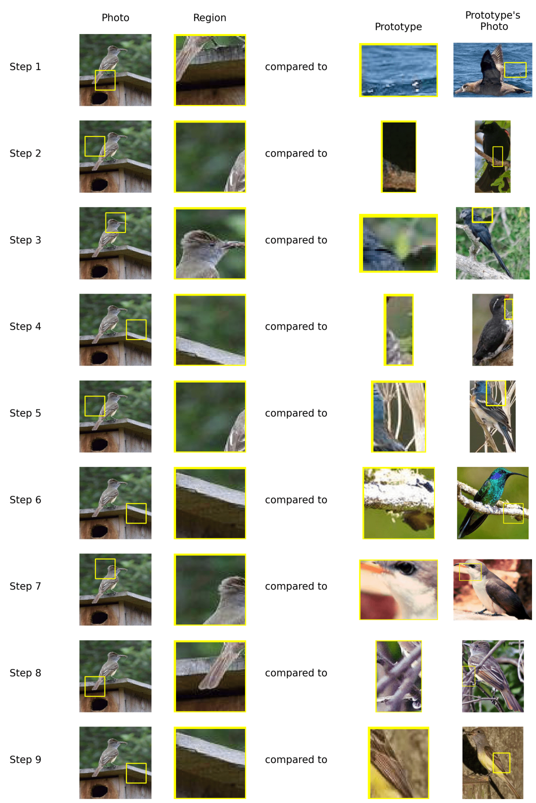

Fig. A5. ProtoTree modified explanation. See Fig. A4 for the original explanation.

C Details on the human studies

We ran our study through Human Intelligence Tasks (HITs) deployed on Amazon Mechanical Turk (AMT). We recruited participants who are US-based, have done over 1000 HITs, and have prior approval rate of at least 98%. For each study, we deployed 10 HITs, each with a different set of input images and explanations. To reduce the variance with respect to the input, we had 5 participants complete each HIT, so each study had 50 participants. Participants were compensated based on the state-level minimum wage of $12/hr.

The demographic distribution was: man 60%, woman 38%, non-binary 1%, no gender reported 1%; White 74%, Black/African American 9%, Asian 7%, no race/ethnicity reported 7%, Hispanic/Latino/Spanish Origin of any race 2%, American Indian/Alaska Native 1%, Native Hawaiian/Other Pacific Islander 0%. The self-reported machine learning experience was 2.5 ± 1.0, between "2: have heard about..." and "3: know the basics..." The average study duration was 6.9 ± 3.5 minutes for GradCAM, 6.6 ± 3.5 for BagNet, 13.6 ± 6.2 for ProtoPNet, and 10.4 ± 3.1 for ProtoTree.

D Additional results and analyses

D.1 Agreement study results with vs. without example images

In Section 5.2 of the main paper, we described the results of our *agreement* study. Here we provide additional results.

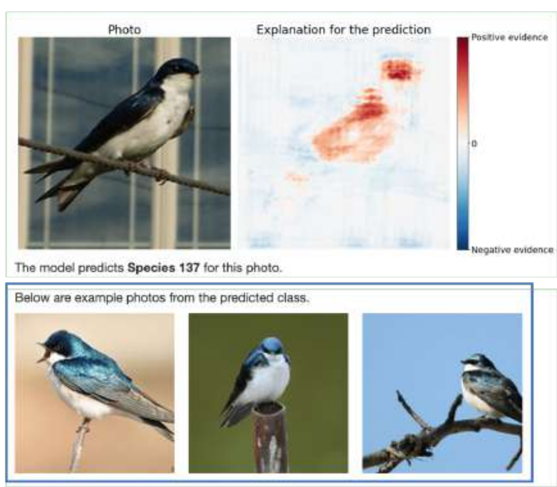

Fig. A6. BagNet *agreement* study input with example images. For the study version with example images, we additionally show three example images from the predicted class (highlighted in the blue box).

For GradCAM and BagNet, we run another version of the *agreement* study where we show three example images from the predicted class, in addition to the test image, prediction, and heatmap (see Fig. A6). Since ProtoPNet and ProtoTree explanations consist of source images of the learned prototypes, we take this measure to provide similar supplementary information for GradCAM and BagNet. As expected, participants improve on the task when they see example images from the predicted class (5.3% overall improvement for GradCAM, 7.1% for BagNet). However, even with the help of example images, participants tend to believe in incorrect predictions, which suggests that incorrect top-1 predictions from high-performance models such as ResNet50 and BagNet are oftentimes convincing. Between CUB and ImageNet, task accuracy is overall higher on CUB, but both yield similar insights. See Tab. A1 for full results.

Table A1. Agreement study results with vs. without examples. For each study, we show mean accuracy, standard deviation of the participants' performance, and mean confidence rating in parentheses. *Italics* denotes methods with accuracy not statistically significantly different from 50% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. **In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct.** For example, for GradCAM on CUB, participants thought 72.4% of correct predictions were correct and $100 - 32.8 = 67.2\%$ of incorrect predictions were correct. These results reveal an issue of *confirmation bias.* **Comparing results with vs. without example images from the predicted class, participants improve on the task when they see examples, but still tend to believe in incorrect predictions.** See Appendix D.1 for a discussion.

CUB	GradCAM [10]	w/ examples	BagNet [1]	w/ examples
Correct	72.4% ± 21.5 (2.9)	83.2% ± 15.7 (3.3)	75.6% ± 23.4 (3.0)	**83.6% ± 17.3 (3.3)**
Incorrect	32.8% ± 24.3 (2.8)	36.8% ± 22.8 (2.8)	*42.4% ± 28.7 (2.7)*	***44.4% ± 30.5 (2.6)***
ImageNet	GradCAM [10]	w/ examples	BagNet [1]	w/ examples
Correct	70.8% ± 26.6 (2.9)	**78.4% ± 25.6 (3.2)**	66.0% ± 27.2 (2.8)	77.2% ± 23.3 (3.2)
Incorrect	***44.8% ± 31.6 (2.7)***	*43.6% ± 32.4 (2.6)*	35.6% ± 26.9 (2.7)	*42.8% ± 32.7 (2.6)*

D.2 Analysis with automatic evaluation metrics

In Section 5.3 of the main paper, we briefly summarized our analysis with automatic evaluation metrics. Here we discuss the results in more detail.

We further analyzed GradCAM heatmaps set using three automatic evaluation metrics: pointing game [13], energy-based pointing game (energy game) [12], and intersection-over-union (IoU) [14]. Pointing game considers a heatmap correct when its highest-intensity point lies inside the segmentation/bounding-box annotation. Energy game calculates how much energy in a heatmap falls inside the segmentation/bounding-box annotation. IoU captures the amount of overlap between a binarized heatmap (according to some threshold) and the segmentation/bounding-box annotation. For all three metrics, higher values indicate better localization quality.

We evaluate up to three GradCAM explanations per image, all using the same segmentation/bounding-box annotation for the ground-truth class: heatmaps for the ground-truth class, predicted class, and class with the second-highest score. Results are summarized in Tab. A2. For CUB heatmaps, we calculate the three metrics on the entire test set (top table). For ImageNet heatmaps, we calculate the metrics on 5,000 randomly sampled validation images. Since ImageNet images sometimes have multiple bounding box annotations, we report results evaluated with one bounding box that yields the best result (middle table) and results evaluated with the union of bounding boxes (bottom table). We find that all three metrics are highest on the ground-truth/predicted class heatmaps for correctly predicted samples. However, we find that these metrics are also high for other heatmaps, even when they are for wrong classes.

Next, we calculate these metrics on images/heatmaps we showed the participants and analyze our human study results. In the *agreement* study, we find near-zero correlation between participants' confidence in the model prediction and localization quality of heatmaps. In the *distinction* study, we also do not see meaningful relationships between participants' choices and these automatic metrics, possibly because all four heatmaps have similar localization quality. These observations are consistent with the findings of [8,3], i.e., automatic metrics poorly correlate with human performance in post-hoc attribution heatmap evaluation. Overall, our analysis reveals a limitation of automatic metrics.

D.3 Similarity judgment of humans vs. prototype-based models

In Section 5.4 of the main paper, we quantified the gap between prototype-based models and human users' notion of similarity. Here we show a plot of participant vs. ProtoPNet prototype similarity rating (Fig. A7). There is no significant negative correlation between the two. This result suggests a gap between ProtoPNet and human judgments of similarity.

Nonetheless, we find that participants are consistent in their similarity ratings and decisions. When examining ProtoPNet and ProtoTree explanations, on average participants assign higher similarity ratings to prototypes of the class they select to be correct (2.9 out of 4 for both ProtoPNet *agreement* and *distinction* tasks, 2.4 for ProtoTree *agreement*) and lower similarity ratings to prototypes of the class they select to be incorrect (2.0 and 2.1 for ProtoPNet *agreement* and *distinction*, 2.0 for ProtoTree *agreement*). The similarity ratings between the two groups are statistically significantly different in all studies. This suggests that participants understand how the model reasons (i.e., they predict the bird class whose prototypes appear most similar to the given photo).

Table A2. Evaluation of GradCAM heatmaps using automatic metrics. We report the mean and standard deviation of three automatic evaluation metrics calculated on heatmaps for the ground-truth class, predicted class, and class with the second-highest score. **All three metrics are highest on the ground-truth/predicted class heatmaps for correctly predicted samples. However, these metrics are also high for other heatmaps, even when they are for wrong classes.**

CUB [11] heatmaps evaluated with the segmentation mask				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	0.92 ± 0.27	0.12 ± 0.07	0.38 ± 0.15
	Second-highest	0.74 ± 0.44	0.09 ± 0.06	0.24 ± 0.15
Incorrect	GT	0.73 ± 0.45	0.08 ± 0.06	0.23 ± 0.16
	Predicted	0.83 ± 0.37	0.09 ± 0.06	0.29 ± 0.15
	Second-highest	0.80 ± 0.40	0.09 ± 0.06	0.26 ± 0.15
ImageNet [9] heatmaps evaluated with the bounding box that yields the best result				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	0.95 ± 0.22	0.27 ± 0.13	0.60 ± 0.28
	Second-highest	0.93 ± 0.26	0.26 ± 0.13	0.60 ± 0.27
Incorrect	GT	0.91 ± 0.29	0.23 ± 0.14	0.52 ± 0.31
	Predicted	0.82 ± 0.38	0.22 ± 0.15	0.52 ± 0.33
	Second-highest	0.84 ± 0.37	0.23 ± 0.15	0.52 ± 0.33
ImageNet [9] heatmaps evaluated with the union of the bounding boxes				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	0.95 ± 0.22	0.29 ± 0.13	0.65 ± 0.26
	Second-highest	0.93 ± 0.26	0.28 ± 0.13	0.64 ± 0.26
Incorrect	GT	0.91 ± 0.29	0.24 ± 0.14	0.56 ± 0.30
	Predicted	0.82 ± 0.38	0.24 ± 0.15	0.56 ± 0.32
	Second-highest	0.84 ± 0.37	0.24 ± 0.15	0.56 ± 0.32

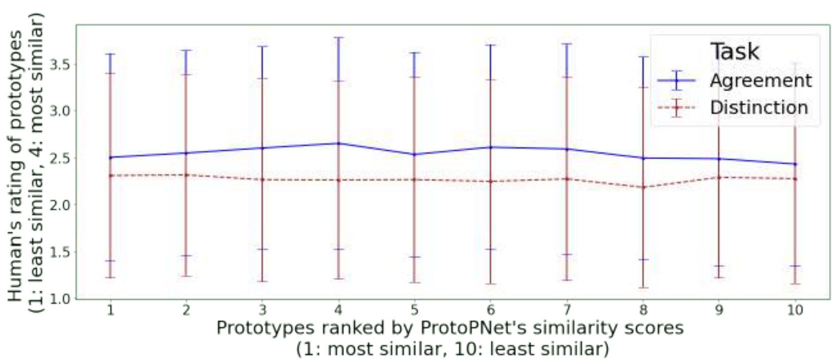

Fig. A7. Participant vs. ProtoPNet prototype similarity rating. There exists a gap between ProtoPNet's similarity scores and human judgments of similarity (Spearman's $\rho = -0.25$, $p = 0.49$ for *distinction*; $\rho = -0.52$, $p = 0.12$ for *agreement*).

D.4 Subjective evaluation results

In Section 5.5 of the main paper, we summarized our subjective evaluation results. Here we provide the full results.

In Tab. A3, we report the participants' self-rated level of understanding of the given model's reasoning process. Overall, the participants rated their level of understanding between 3 (fair) and 4 (good). Interestingly, we find that the rating tends to decrease after the participants see their task performance. Several participants indicated that their performance was lower than what they expected: "I thought I would do a bit better!", "my score wasn't as high as I would have liked", "I was surprised that my score was not very much higher than random guessing. I thought I had a good idea of the model, especially making judgements about the amount of positive and negative evidence, but it seems I did not." No one suggested the opposite. This trend suggests that participants might have been disappointment in their task performance, which in turn led them to lower their self-rated level of method understanding.

Table A3. Participants' self-rated level of method understanding. We report the mean and standard deviation of the participants' self-rating of their method understanding. Participants provide ratings three times: after reading about the method (post-intro), after completing the task (post-task), and after learning about their task performance (post-results). The rating tends to *decrease* after the participants see their task performance ($p < 0.05$).

Dataset	Method	Study	Post-intro	Post-task	Post-results
CUB [11]	GradCAM [10]	Agreement	3.7 ± 0.9	3.8 ± 0.9	3.3 ± 1.1
		Agreement w/ examples	3.7 ± 1.0	3.9 ± 0.7	3.4 ± 1.0
		Distinction	3.4 ± 1.0	3.5 ± 1.2	3.6 ± 0.8
	BagNet [1]	Agreement	3.5 ± 1.0	3.7 ± 0.8	3.3 ± 1.1
		Agreement w/ examples	3.7 ± 0.8	3.9 ± 0.8	3.6 ± 1.0
		Distinction	3.8 ± 0.7	4.0 ± 0.8	3.9 ± 0.8
	ProtoPNet [2]	Agreement	3.9 ± 0.8	4.0 ± 0.8	3.7 ± 0.8
		Distinction	4.1 ± 0.8	3.9 ± 0.8	3.7 ± 1.1
	ProtoTree [7]	Agreement	3.7 ± 0.8	3.7 ± 1.0	3.4 ± 0.8
		Agreement (tree)	3.7 ± 0.7	3.5 ± 0.9	3.2 ± 1.1
		Distinction	3.4 ± 1.0	3.6 ± 1.1	3.3 ± 1.2
ImageNet [9]	GradCAM [10]	Agreement	3.7 ± 0.9	3.9 ± 0.9	3.0 ± 1.0
		Agreement w/ examples	3.4 ± 0.8	3.7 ± 0.8	3.5 ± 0.9
		Distinction	3.9 ± 0.9	3.7 ± 1.0	3.7 ± 1.0
		Distinction w/ labels	3.9 ± 0.9	3.8 ± 1.0	3.8 ± 0.9
	BagNet [1]	Agreement	3.7 ± 0.8	3.9 ± 0.7	3.4 ± 1.0
		Agreement w/ examples	3.8 ± 0.9	3.9 ± 0.9	3.3 ± 1.0
		Distinction	3.9 ± 0.8	3.9 ± 0.8	3.8 ± 1.0
		Distinction w/ labels	3.8 ± 0.9	4.0 ± 0.8	3.8 ± 0.8
Mean across all studies			3.7 ± 0.9	3.8 ± 0.9	3.5 ± 1.0

D.5 Interpretability-accuracy tradeoff results

In Section 5.6 of the main paper, we summarized our interpretability-accuracy tradeoff study results. Here we provide more details.

In Tab. A4 and Fig. A8, we show the full results of our interpretability-accuracy tradeoff study. We report the accuracy of the evaluated interpretable model and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. Across all studies, we find that participants require the baseline model to have higher accuracy than the evaluated interpretable model, and input a higher accuracy requirement for higher-risk settings. On average, participants require the baseline model to have +6.2% higher accuracy for low-risk (e.g., bird species recognition for scientific or educational purposes), +8.2% for medium-risk (e.g., object recognition for automatic grocery checkout), and +10.9% for high-risk (e.g., scene understanding for autonomous driving) settings.

We observe this trend in the participants' written responses as well. Most participants write that they would use the baseline model only when it has higher accuracy than the evaluated interpretable model: "I would need the black box model to give me a nice boost in accuracy, or I would just stick to the bagnet model, since it is pretty accurate." On the contrary, participants exhibit different levels of desire for interpretability. Some deem interpretability as important: "Understanding how a prediction works is important. For me to accept a model with no explanations, the level of accuracy needs to be higher", "I prefer to understand how models work, so the black box model has to be significantly better than the other model for me to use it. As the stakes become higher, I want its accuracy to be higher because there's no way for me to question or check its progress if it's wrong." Other participants willingly tradeoff interpretability for accuracy: "I don't need to know how it works. So, as long as it's marginally better, it should be used", "I don't care about not having an explanation, so if the accuracy of a different model has just a 1% improvement in performance then I would choose the better performing model."

Nonetheless most participants express a need for higher-accuracy models in higher-risk settings: "The higher the risk, the more accurate I need it to be in order to feel confident using it", "If I were to choose to use a model that did not provide reasoning for me to utilize in evaluating how the decision was made I would need to know that the model would give me significantly better results, especially in a high-risk scenario as described above, but even in the medium risk setting, being able to asses the reasoning of the model is an invaluable tool and I would only be willing to give it up for significant increases in accuracy."

Table A4. Interpretability-accuracy tradeoff results. We report the mean and standard deviation of the additional accuracy participants require for the baseline model, to use it over the model with explanations. For example in the GradCAM *agreement* study with CUB, participants require the baseline model to have +5.6% accuracy beyond the model that comes with GradCAM explanations and achieves 81.0% accuracy, in the low-risk setting. See Fig. A8 for a visualization of the results.

Dataset	Method	Study	Orig	Low-risk	Med-risk	High-risk
CUB [11]	GradCAM [10]	Agreement	81.0	+5.6 (±6.9)	+6.2 (±5.7)	+7.7 (±7.0)
		Agreement w/ examples		+4.2 (±6.1)	+5.7 (±5.7)	+7.7 (±7.5)
		Distinction		+2.9 (±6.9)	+4.5 (±5.2)	+8.1 (±6.9)
	BagNet [1]	Agreement	74.2	+6.8 (±7.9)	+7.8 (±8.1)	+12.3 (±9.2)
		Agreement w/ examples		+6.1 (±7.1)	+8.1 (±6.3)	+10.7 (±9.2)
		Distinction		+7.0 (±8.1)	+8.8 (±7.4)	+8.4 (±8.4)
	ProtoPNet [2]	Agreement	79.9	+5.8 (±6.6)	+7.8 (±4.9)	+9.4 (±6.6)
		Distinction		+4.1 (±7.9)	+6.1 (±6.4)	+9.7 (±7.1)
	ProtoTree [7]	Agreement	81.7	+3.8 (±6.5)	+4.2 (±6.3)	+5.1 (±6.5)
		Agreement (tree)		+3.7 (±5.5)	+5.8 (±5.1)	+6.7 (±6.6)
		Distinction		+5.1 (±5.7)	+6.4 (±5.8)	+9.2 (±6.2)
ImageNet [9]	GradCAM [10]	Agreement	76.1	+6.3 (±7.6)	+8.1 (±8.6)	+11.8 (±10.7)
		Agreement w/ examples		+4.8 (±6.8)	+8.6 (±7.6)	+11.4 (±10.8)
		Distinction		+5.3 (±7.3)	+9.8 (±6.7)	+12.4 (±8.6)
		Distinction w/ labels		+7.6 (±7.7)	+9.3 (±7.9)	+13.2 (±9.0)
	BagNet [1]	Agreement	66.7	+9.9 (±7.5)	+14.1 (±9.5)	+17.5 (±11.1)
		Agreement w/ examples		+9.7 (±8.5)	+13.2 (±10.4)	+17.6 (±13.0)
		Distinction		+7.9 (±9.3)	+9.6 (±9.2)	+11.2 (±11.2)
		Distinction w/ labels		+11.4 (±9.2)	+12.4 (±10.4)	+16.6 (±11.6)
Mean across all studies				+6.2 (±7.7)	+8.2 (±7.9)	+10.9 (±9.7)

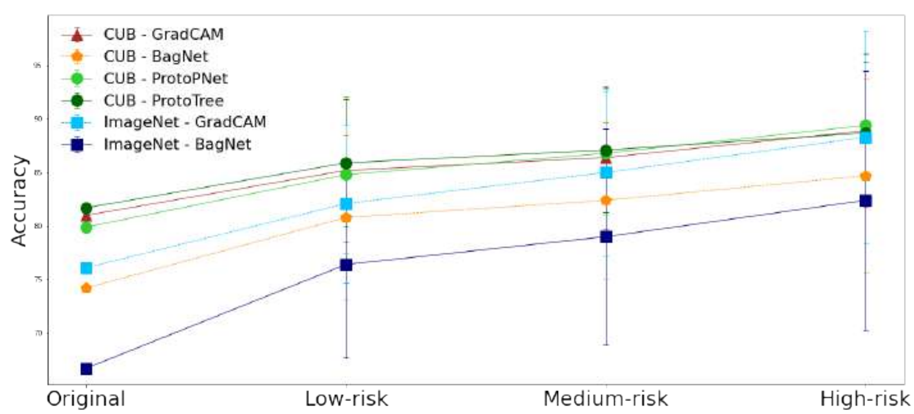

Fig. A8. Visualization of the interpretability-accuracy tradeoff results. This plot shows that participants desire higher accuracies for the baseline model, especially in higher-risk settings. See Tab. A4 for the full results.

E Simple decision tree used for explaining ProtoTree

One additional challenge of evaluating the ProtoTree model is that participants may not be familiar with decision trees. To mitigate this challenge, we introduce a simple decision tree model for fruit classification before introducing ProtoTree. This simple decision tree model takes in an input image and makes an output classification (Class A, B, C, D, E) based on three decision nodes. We first walk through the participants through an example. We then present two warm-up exercises so that the participants can become more familiar with decision trees. When the participants submit their answers, we also provide the correct answer and the reason for it. Participants achieved 86.5% performance on this task, implying that the low task accuracy for ProtoTree is not due to a lack of comprehension of decision trees. See Fig. A9 for the UI.

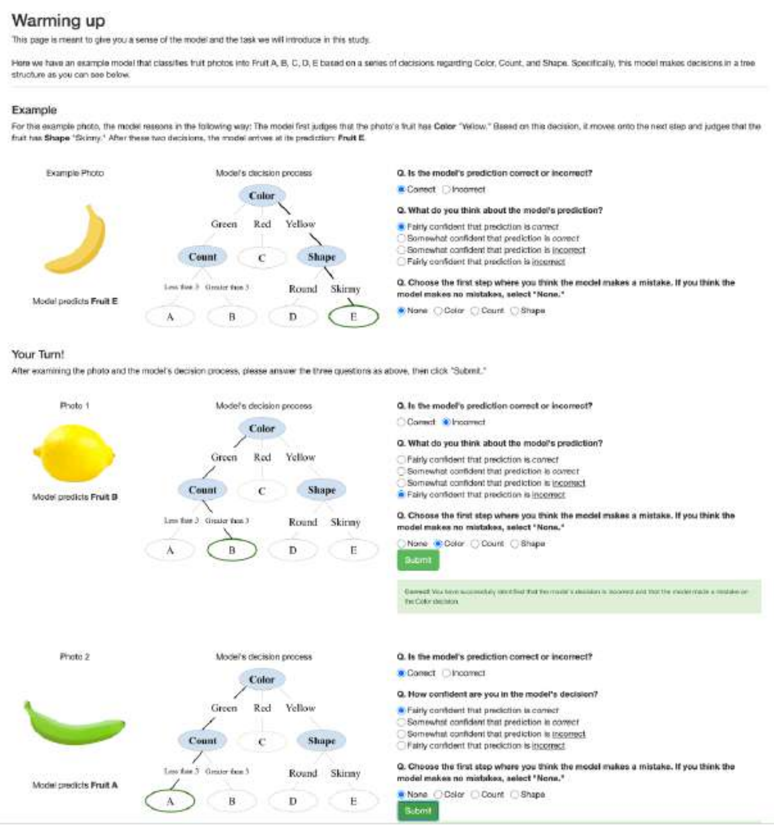

Fig. A9. A simple decision example. We use this model to introduce participants to decision trees before explaining the more complex ProtoTree. See Appendix E for details.

F UI snapshots

In Section 4 of the main paper, we outlined our study design. Here we provide snapshots of our study UIs in the following order.

1. Study introduction. For each participant, we first briefly introduce the study and receive their informed consent. The consent form was approved by the IRB and acknowledges that participation is voluntary, refusal to participate will involve no penalty or loss of benefits, etc. See Fig. A10.

2. Demographics and background. To help future researchers calibrate our results and do proper comparison, we request optional demographic data regarding gender identity, race and ethnicity. We also ask the participant's experience with machine learning. See Fig. A11.

3. Method introduction. We introduce each interpretability method/model in simple terms. See Fig. A12.

4. Task preview and first subjective evaluation. To encourage participants to carefully read the method explanation, we show a preview of the task they will complete along with a correct and incorrect prediction. Participants then answer their first subjective evaluation question. In Fig. A13 we shown an example from the ProtoPNet *agreement* study.

5. Task. Participants then proceed onto the main task. We show the UI for the following 8 studies:

- GradCAM *distinction* (Fig. A14)
- GradCAM *agreement* (Fig. A15)
- Bagnet *distinction* (Fig. A16)
- Bagnet *agreement* (Fig. A17)
- ProtoPNet *distinction* (Fig. A18)
- ProtoPNet *agreement* (Fig. A19)
- ProtoTree *distinction* (Fig. A20)
- ProtoTree *agreement* (Fig. A21)

6. Second and third subjective evaluation. After the task, participants complete their second subjective evaluation question. We then disclose their task performance and ask the third subjective evaluation question. These questions allow us to investigate if the participants' self-rated level of method understanding undergoes any changes throughout the study. See Fig. A22.

7. Interpretability-accuracy tradeoff. Finally, we investigate the tradeoff participants are willing to make when comparing the evaluated interpretable model against a baseline model that doesn't come with any explanation. We present three scenarios to the participants: low-risk (e.g., scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for self-driving cars). We then ask them to input the minimum accuracy of a baseline model that would convince them to use the baseline model over the model that comes with explanations and briefly describe their reasoning. See Fig. A23.

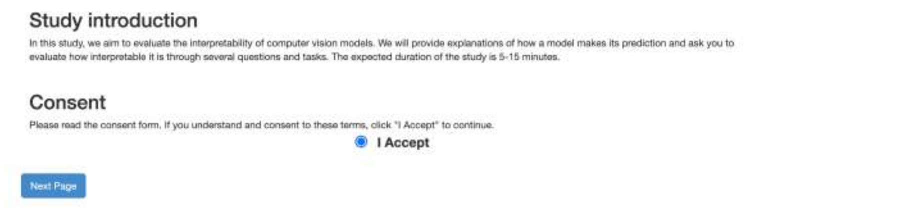

Fig. A10. 1. Study introduction.

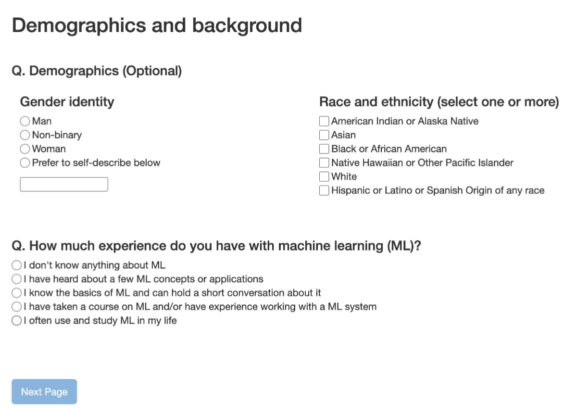

Fig. A11. 2. Demographics and background.

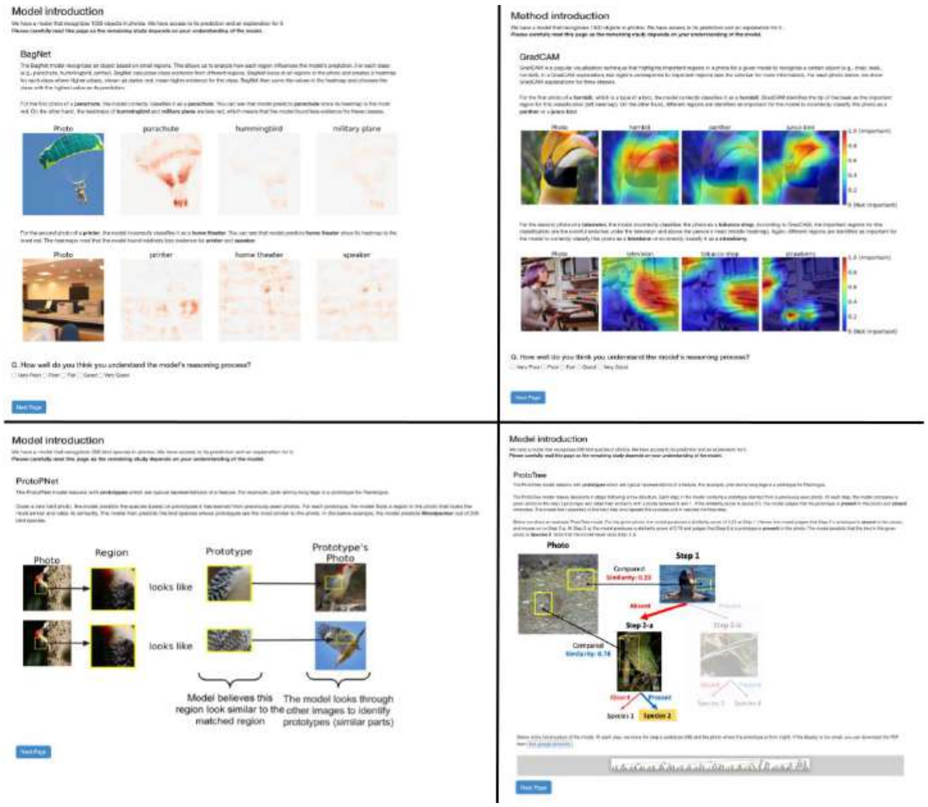

Fig. A12. 3. Method introduction. BagNet (top left), GradCAM (top right), ProtoPNet (bottom left), ProtoTree (bottom right).

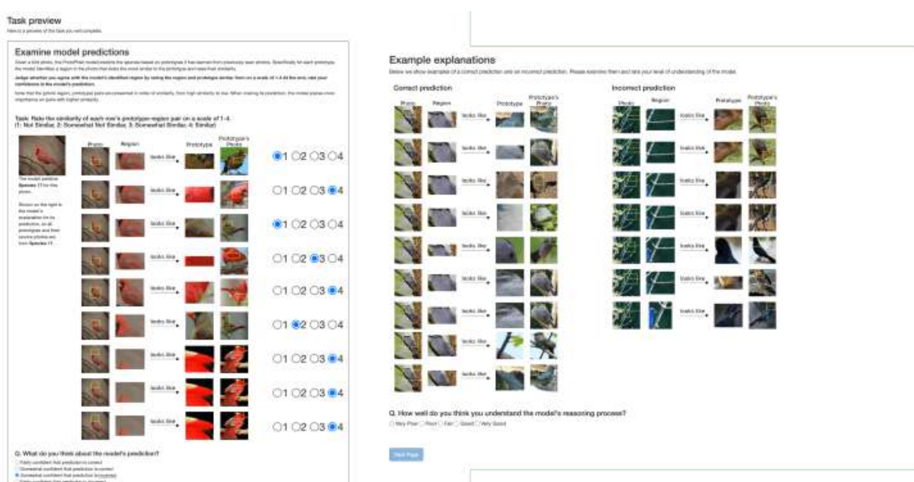

Fig. A13. 4. Task preview and first subjective evaluation.

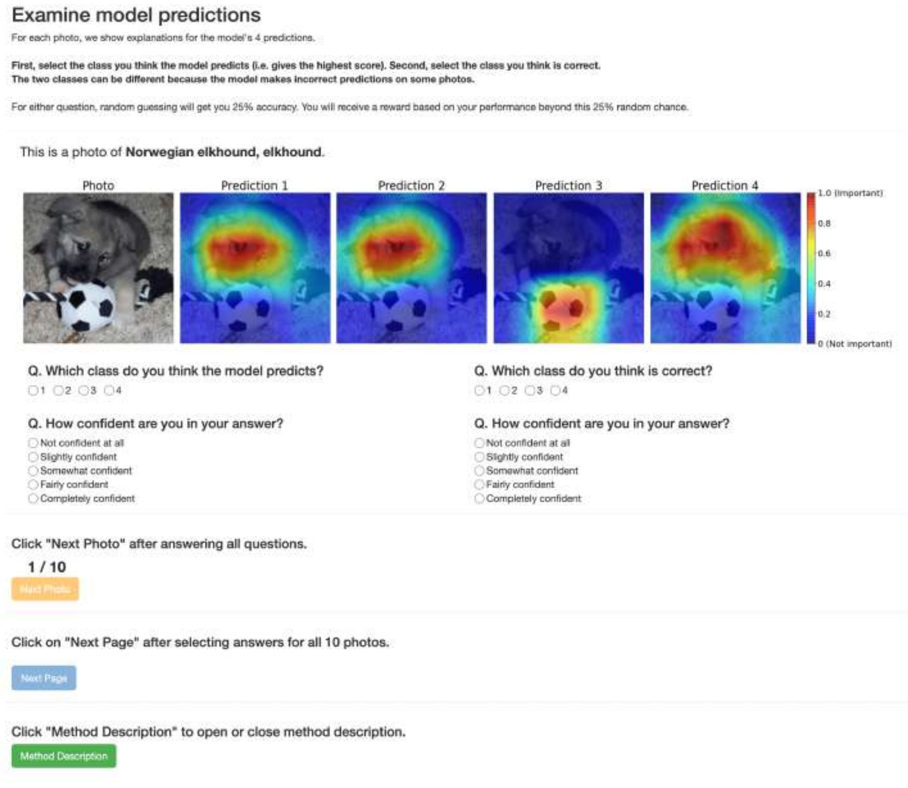

Fig. A14. 5. Task: GradCAM *distinction*.

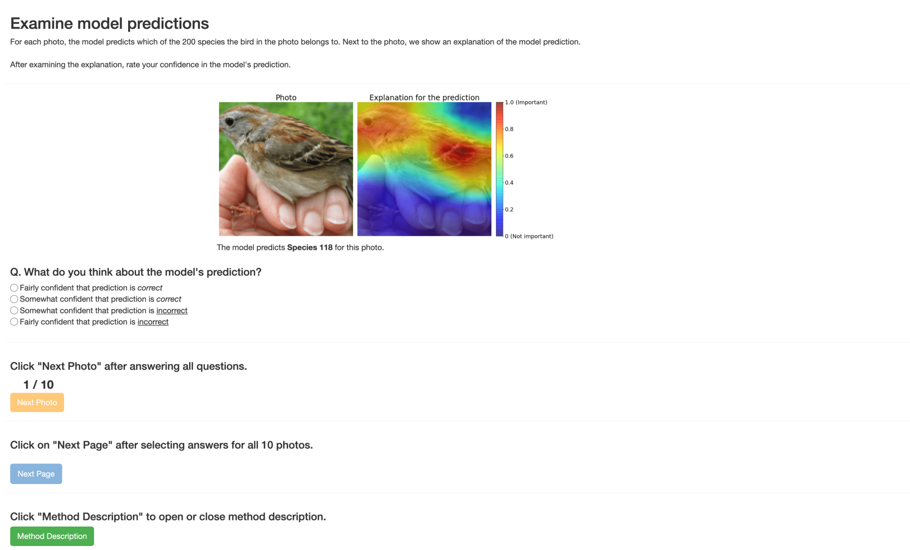

Fig. A15. 5. Task: GradCAM *agreement*.

Fig. A16. 5. Task: BagNet *distinction*.

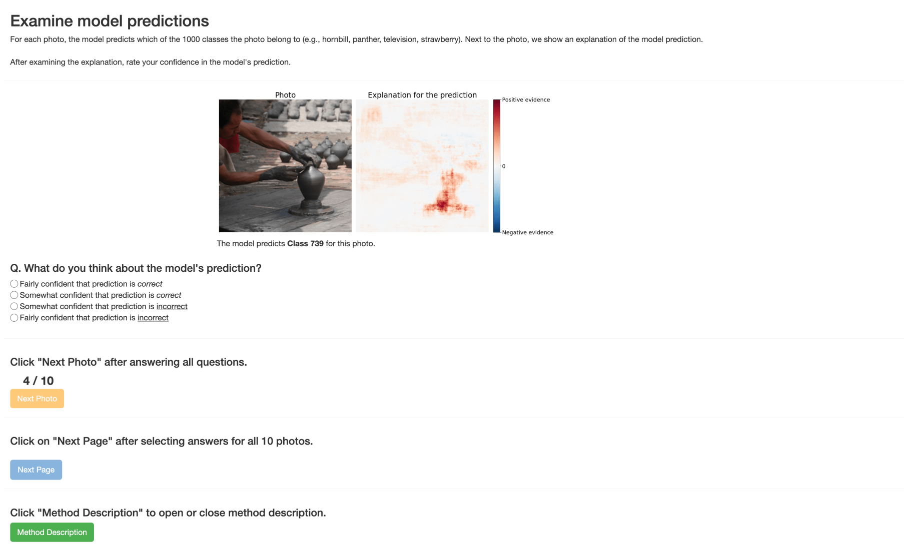

Fig. A17. 5. Task: BagNet *agreement*.

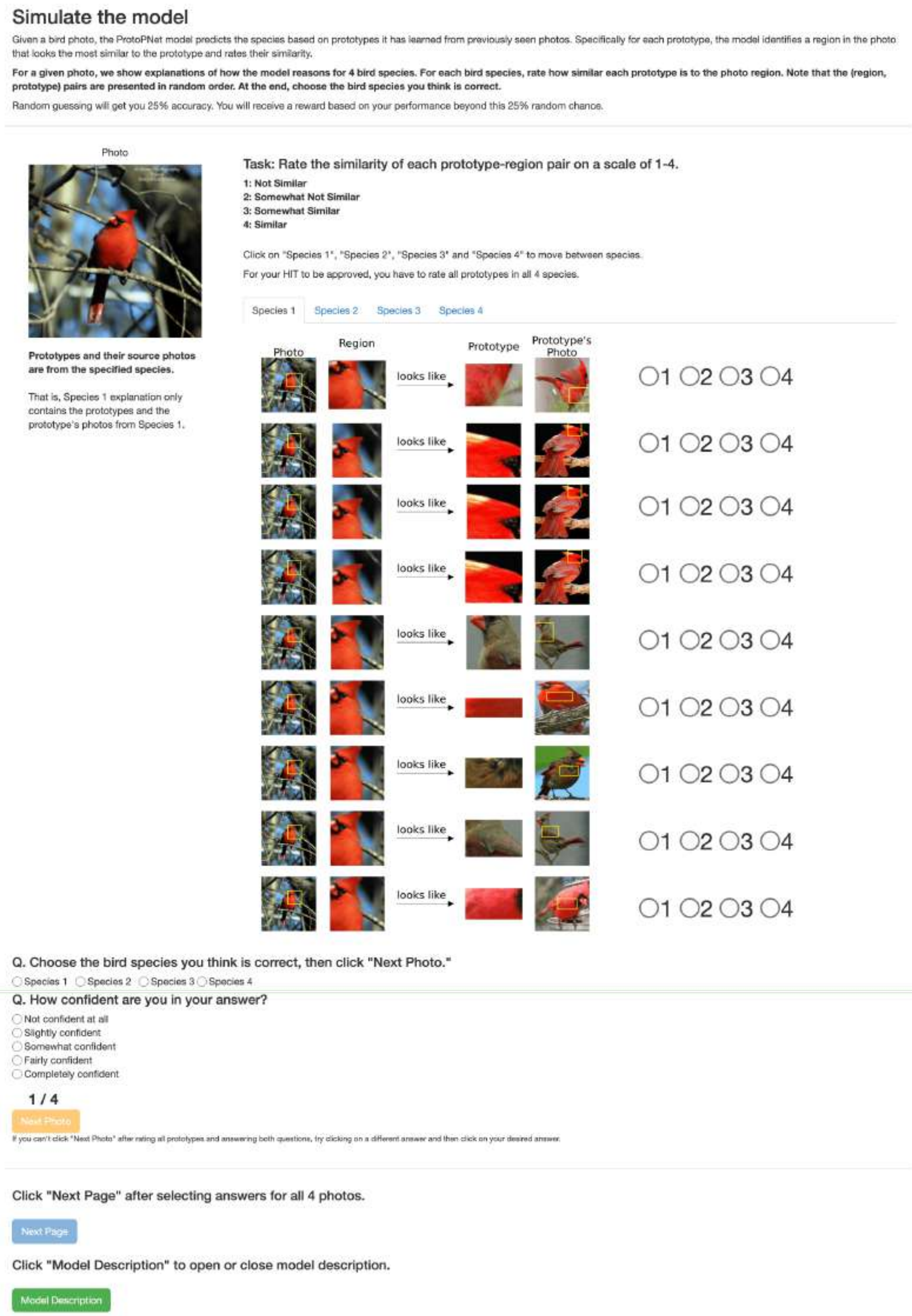

Fig. A18. 5. Task: ProtoPNet *distinction*.

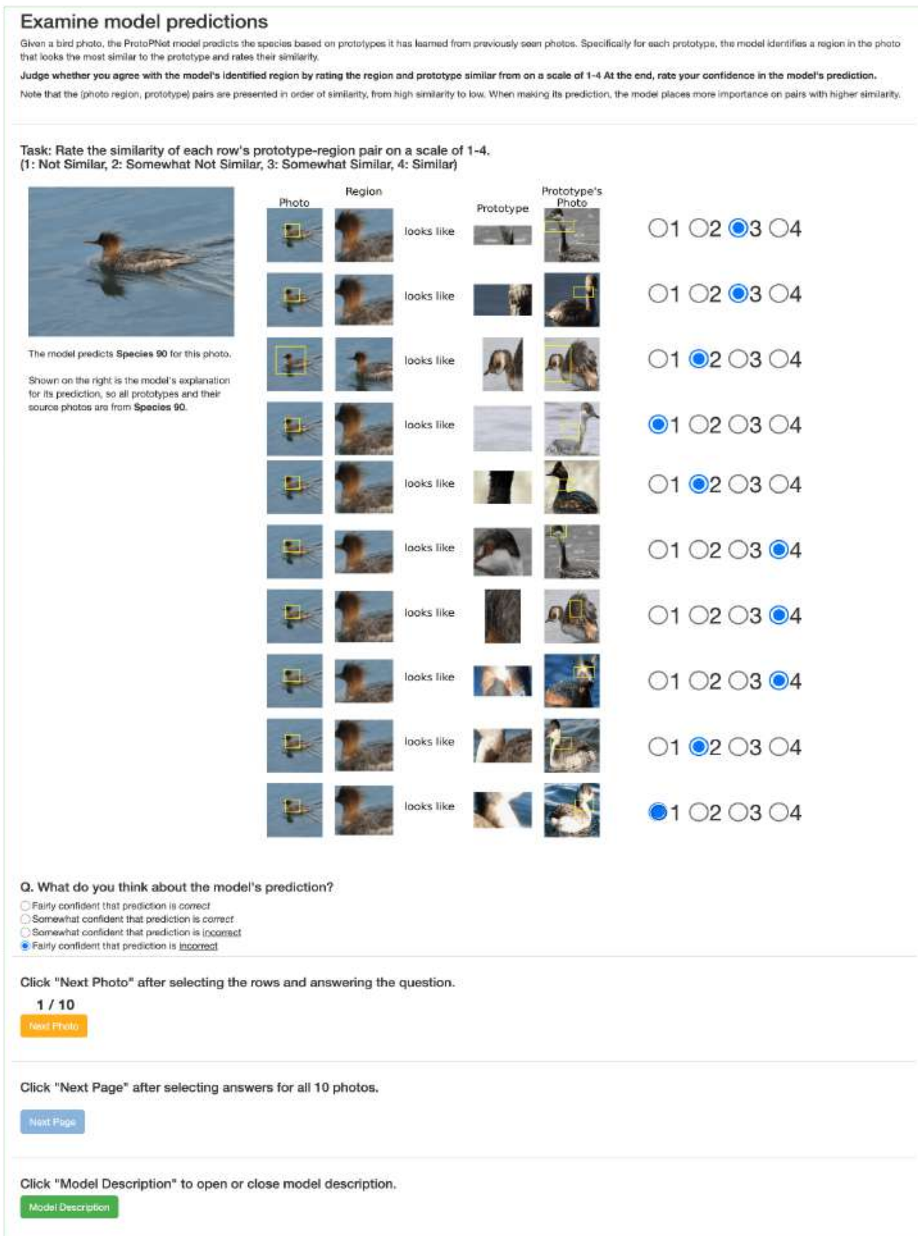

Fig. A19. 5. Task: ProtoPNet *agreement*.

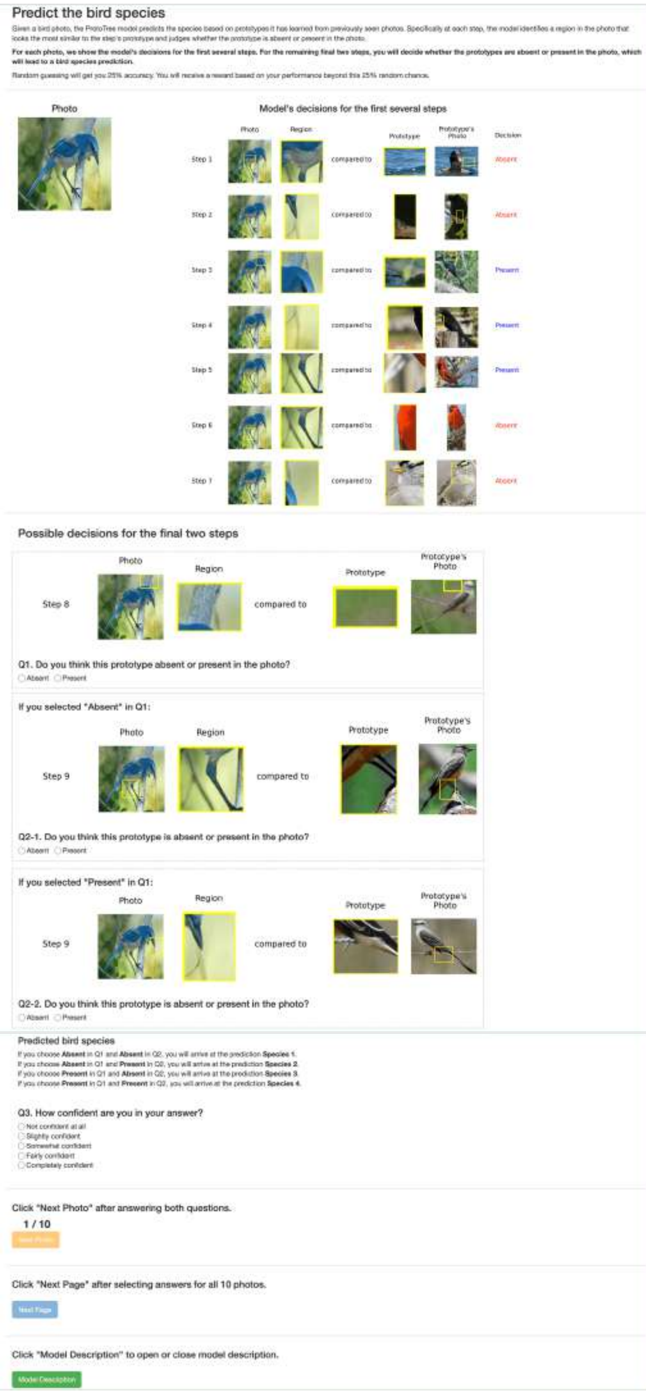

Fig. A20. 5. Task: ProtoTree *distinction*.

Fig. A21. 5. Task: ProtoTree *agreement*.

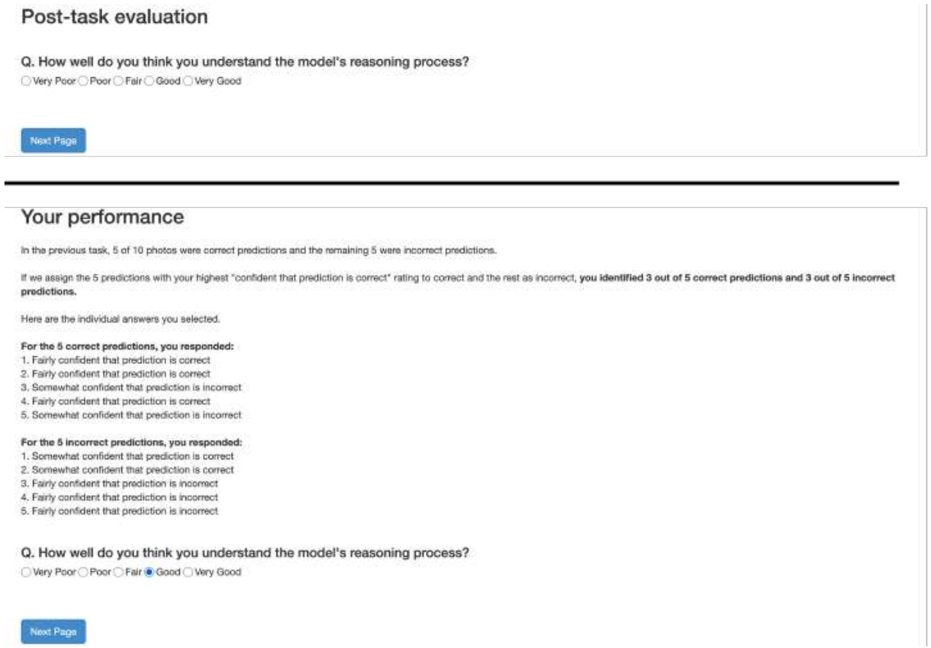

Fig. A22. 6. Second and third subjective evaluation.

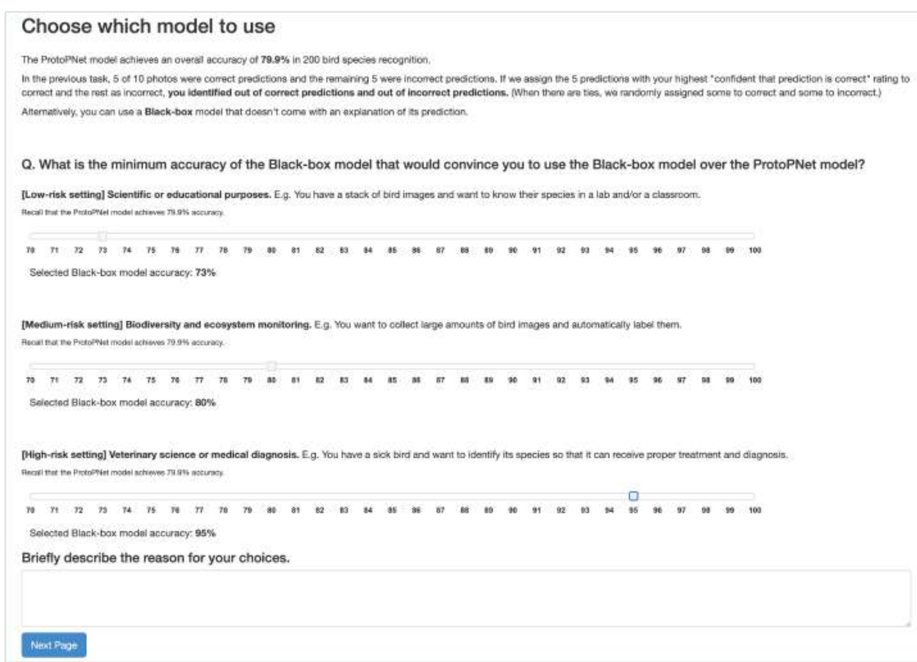

Fig. A23. 7. Interpretability-accuracy tradeoff.

References

1. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In: ICLR (2019)
2. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: NeurIPS (2019)
3. Fel, T., Colin, J., Cadène, R., Serre, T.: What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods (2021)
4. Gildenblat, J., contributors: PyTorch library for CAM methods. `https://github.com/jacobgil/pytorch-grad-cam` (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In: ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI (2021)
7. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: CVPR (2021)
8. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In: NeurIPS (2021)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
11. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
12. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: CVPR Workshops (2020)
13. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: ECCV (2016)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)