Supplemental Materials

A1. The significance t-test for comparing two groups

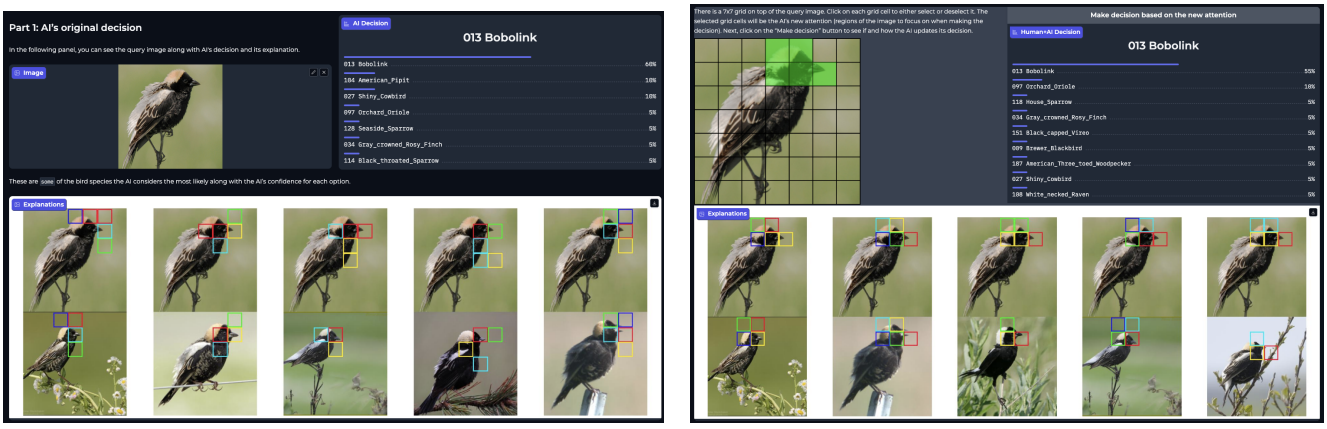
With data collected from the human study, we compute the per-user average accuracy over our controlled, balanced image sets (i.e., the ratio of samples where AI correctly classified and misclassified is approx. 50/50) and report in Table 1. Yet, we find that dynamic explanations (CHM-Corr++ users score 73.57%), which one might naturally assume to be superior, do not show a significant improvement over static ones (CHM-Corr users score 72.68%) in helping users making more accurate decisions.

The findings, with a t-statistic of -0.321 and a p-value of 0.749, indicate that the average accuracy levels for users exposed to both types of explanations are not significantly different. This challenges the common belief that dynamic, interactive content inherently boosts user comprehension or performance [30, 55].

A2. The shortcomings of CHM-Corr classifier

For some samples, AI fails to classify the input image correctly regardless of how the attention is directed towards the image (see Fig. A4). This indicates that improving attention alone does not suffice for classifiers to accurately classify these samples, suggesting the need for insights into developing new models that focus on more than just improving attention mechanisms. Moreover, the underlying nature of the classifier contributes to this issue. Given that the classifier employs a k-Nearest Neighbors (kNN) algorithm to retrieve a set of candidate samples, there is a possibility that the ground-truth class may not appear within the candidate pool. Consequently, no matter how the CHM-Corr model re-ranks these candidates, it may never correctly identify the top-1 class.

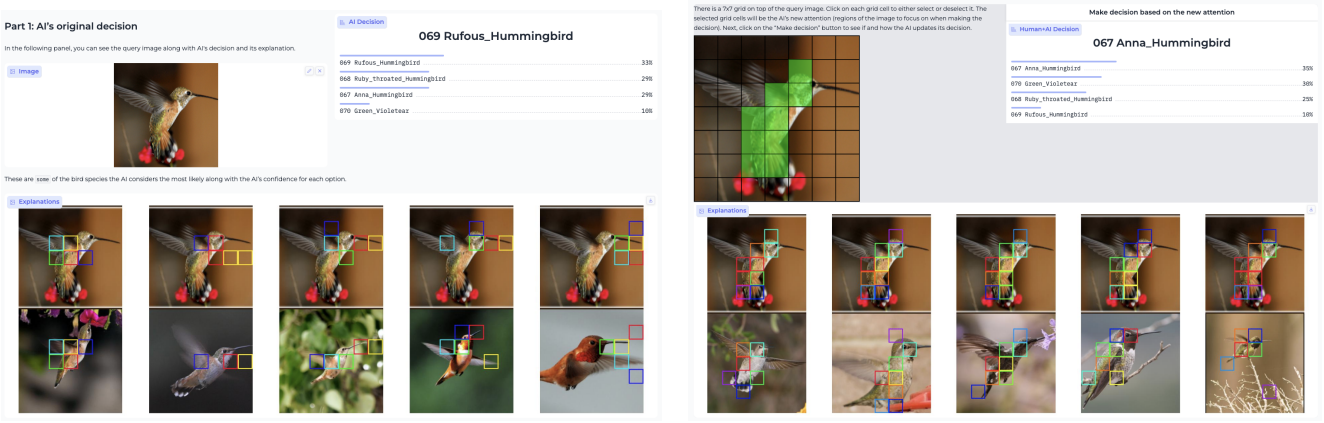
A3. How users interact with explanations



(a) Static explanation for Bobolink.

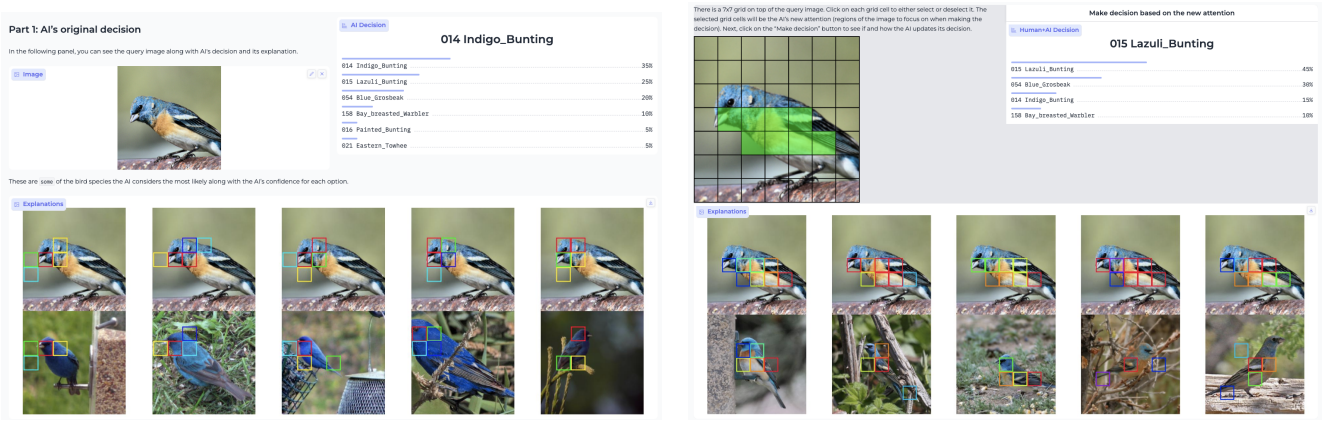
(b) Human-prompt, correspondence-based explanation for Bobolink.

Figure A1. Both dynamic and static explanations enable human users to verify that the AI is predicting the top-1 label correctly.



(a) Static explanation for Rufous Hummingbird. (b) Human-prompt, corr-based explanation for Anna Hummingbird.

Figure A2. Human intervention changes the top-1 label from Rufous Hummingbird → Anna Hummingbird that makes users more likely to reject the original, correct label.



(a) Static explanation for Indigo Bunting. (b) Human-prompt, correspondence-based explanation for Lazuli Bunting.

Figure A3. AI initially makes the wrong classification Indigo Bunting on the input image. Human intervention changes the top-1 label from Indigo Bunting → Lazuli Bunting, a more similar-looking class, encouraging users to reject the original, predicted label.

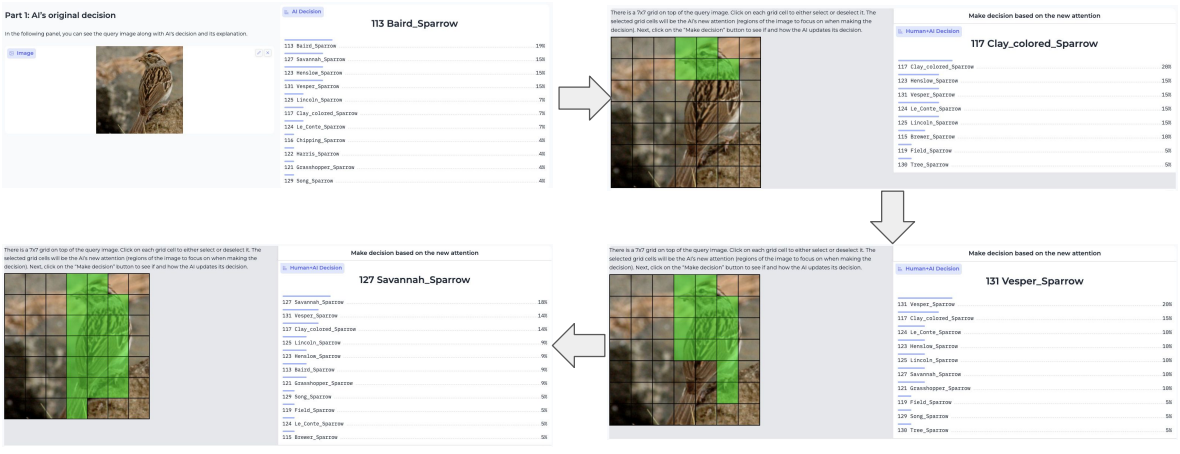


Figure A4. A sample that is unclassifiable for the classifier CHM-Corr. The ground-truth label is Chipping Sparrow. Initially, AI makes a wrong classification of Baird Sparrow. With user-guided attention, the top-1 label evolves from Baird Sparrow → Clay-colored Sparrow → Vesper Sparrow → Savannah Sparrow but none of them matches the groundtruth, making users unable to make decisions given the explanations.