Allowing humans to interactively guide machines where to look does not always improve human-AI team's classification accuracy

Giang Nguyen* Mohammad Reza Taesiri* Sunnie S. Y. Kim Anh (Totti) Nguyen

Auburn University, University of Alberta, Princeton University

* Equal contribution

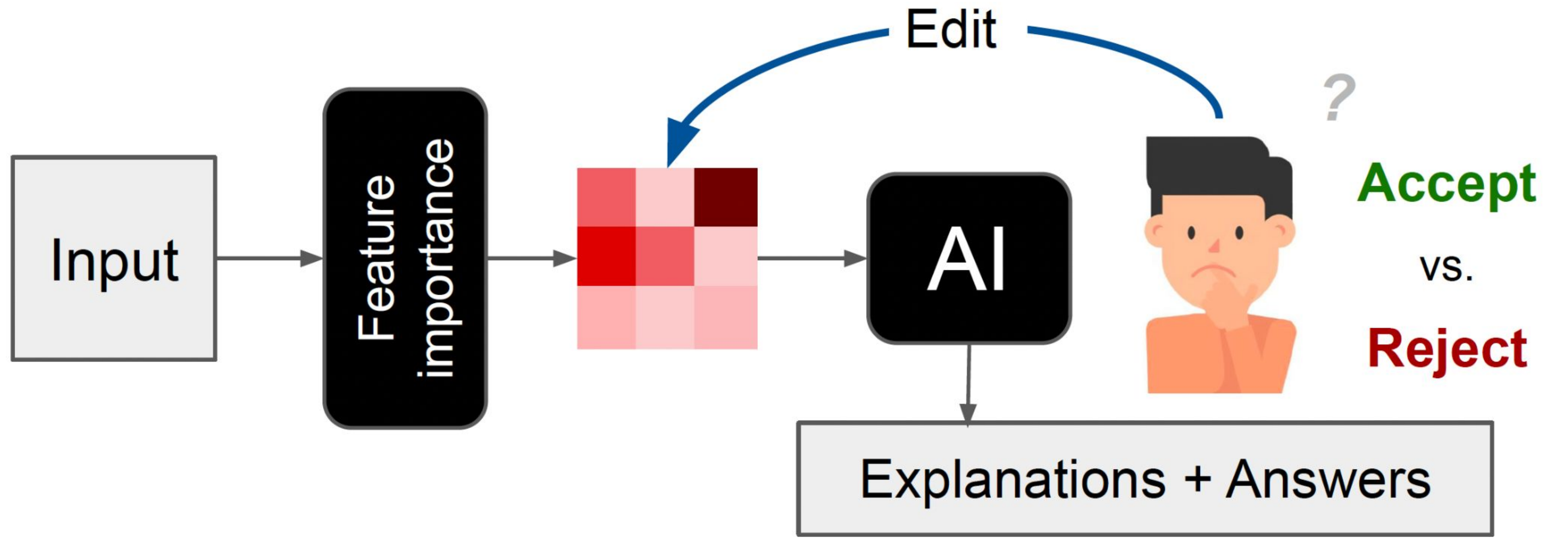
Paper: arxiv.org/pdf/2404.05238
Demo: 137.184.82.109:7080
Code: github.com/anguyen8/chm-corr-interactive

Motivation

- Feature importance and example-based explanations are among the **most popular XAI methods** that offer insights into how a visual classification model makes its predictions.
- However, a **major limitation** of current methods is that they only offer a **static, one-time explanation** of the model prediction.
- There is no way for humans to **provide feedback** to the model (e.g., guide where it should look) and potentially help it perform better, which may also change humans' understanding and decision making with the model.

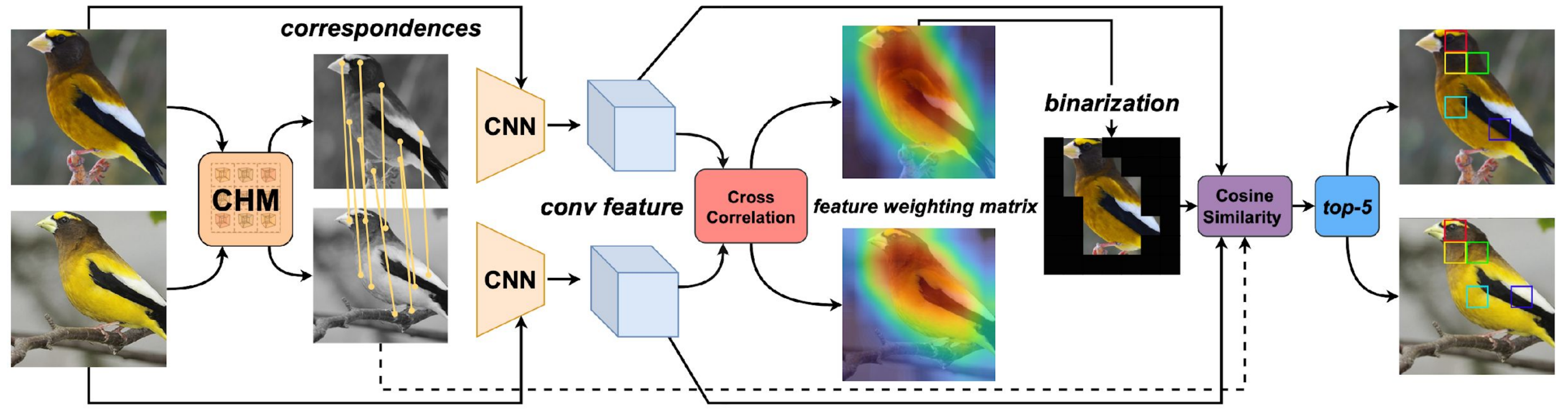
Research Question

- Can **dynamic**, **interactive** explanations improve human-AI team's classification accuracy?



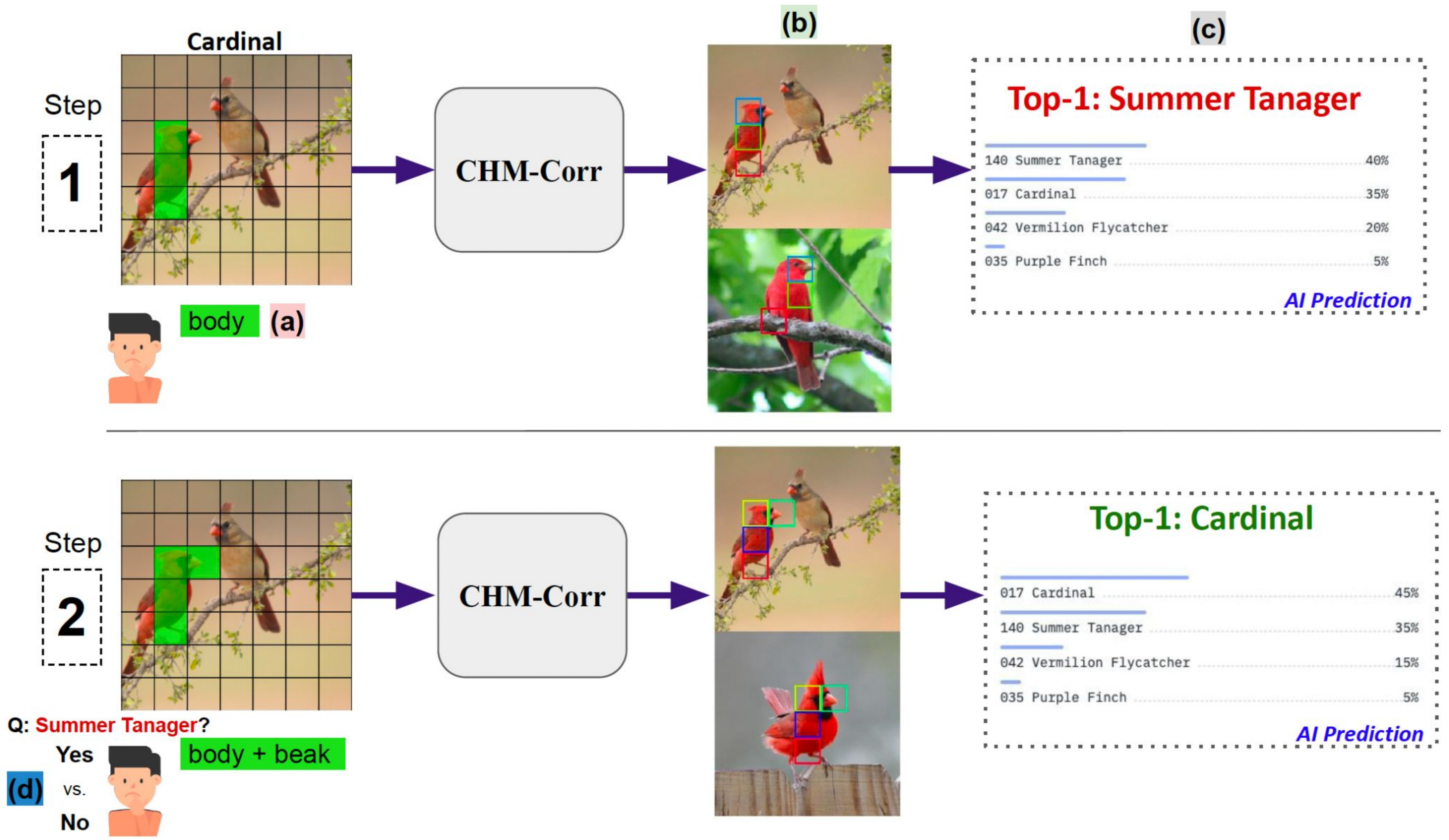
With our proposed method, human users can interactively edit the feature importance explanations and gain insights into *if*, *when*, and *how* the model changes its predictions.

CHM-Corr [1]: A Visual Correspondence-based Classifier



CHM-Corr is a **state-of-the-art, ante-hoc explainable classifier** that first predicts patch-wise correspondences between the input and training-set images, and then bases on them to make classification decisions.

CHM-Corr++: An interactive interface that enables machine attention editing



We let users interact with the image classification model (here CHM-Corr [1]) via controlling the attention (selecting patches) the model should focus on (a). Based on the user-guided attention, the model compares the input (GT class: Cardinal) with candidate training examples to simultaneously generate visual-correspondence explanations (b) and predictions (c). The user iteratively observes the dynamic explanations (b) and predictions (c) to understand the model to accept or reject (d) the original top-1 predicted label (here Summer Tanager) – Distinction task [2].

Experimental Results

Explanation type	Static (CHM-Corr)		Dynamic (CHM-Corr++)	
	Overall		Overall	
$\mu \pm \sigma$	72.68 ± 12.36		73.57 ± 10.42	
	AI originally correct	AI originally incorrect	AI originally correct	AI originally incorrect
	85.21 ± 11.82	60.13 ± 18.66	86.79 ± 13.16	59.39 ± 15.51
# of decisions	283	277	443	397
# of submissions	28		42	

Experimental Results

1. **Participants struggled to reject incorrect model predictions**

Evidence: Decision accuracy on correct instances is much higher than that on incorrect ones for both type of explanations: 85.21% vs. 60.13% with static, 86.79% vs. 59.39% with dynamic.

⇨ We need tools that help users detect and reject AI errors better.

2. **The usefulness of interactivity depended on the interaction outcomes**

Evidence 1: **When the model is originally correct 🤖✅: participants' decision accuracy is higher when the model is consistent than not (90.80% vs. 75.21%) – refer to rows (i, ii) below.**

Evidence 2: **When the model is originally incorrect 🤖❌: participants' decision accuracy is lower when the model is consistent than not (52.55% vs. 62.11 → 65.43%) – refer to rows (iii, iv, v) below.**

AI model correctness w.r.t. human interaction	Acc (%)
(i) Originally correct and consistent (always correct)	90.80
(ii) Originally correct and inconsistent (becomes incorrect)	75.21
(iii) Originally incorrect and consistent (always incorrect)	52.55
(iv) Originally incorrect and inconsistent (always incorrect)	62.11
(v) Originally incorrect and inconsistent (becomes correct)	65.43

⇨ Understanding **when users can** and **cannot** help the model be more accurate, and aiding users in the process, would be important directions for future research.

Discussion & Future Works

Why dynamic, interactive explanations may not improve human-AI team's classification accuracy?

Hypothesis #1: AI attention is already sufficient, as the birds 🐧 are well-centered and clearly visible. Changing the task domain, for example, to include complex scenes where AI struggles to focus on the correct pixels, would better show the utility of CHM-Corr++.

Hypothesis 2: CHM-Corr++ is especially helpful when the base CHM-Corr model can classify correctly 🤖✅. Changing the base model to a more accurate AI will likely improve the utility accordingly.

References

[1] Visual correspondence-based explanations improve AI robustness and human-AI team accuracy, **NeurIPS** 2022.

[2] HIVE: Evaluating the Human Interpretability of Visual Explanations, **ECCV** 2022.