TriHAI: A Combination Enhanced Tri-Mode Deferral for Human-AI Team

Minhui Zhang[1] , **Xuehan Zhao**[1] , **Xin Zhang**[1] , **Jiaqi Liu**[1*] , **Zhiwen Yu**[2,1] and **Bin Guo**[1]
[1]Northwestern Polytechnical University
[2]Harbin Engineering University
{zmh17, xuehan.zhao, zhangxin1}@mail.nwpu.edu.cn, {jqliu, zhiwenyu, guob}@nwpu.edu.cn

Abstract

Learning to Defer operates by deferring AI-uncertain samples to humans, enabling the system to outperform either alone. Some works extend it by introducing human-AI combination to deferral. However, they remain limited to binary choices. Given that AI, humans, and human-AI combination are each indispensable, we extend the binary deferral paradigm to a ternary one. There are two challenges: i) how to effectively select among the three modes, especially distinguishing humans from combination? ii) how to design an enhanced combination without being degraded by unreliable model predictions? To address the challenges, we propose TriHAI, a tri-mode deferral method that routes samples to suitable modes based on normalized confidence scores obtained through a discriminative gating network, and enhances combination mode by fusing model predictions refined by human-guided Conformal Prediction and human predictions via Bayesian theory. Experiments on three datasets show that TriHAI surpasses other human-AI baselines by up to 9.57% in accuracy.

1 Introduction

Over the past decades, AI has made significant strides [Dosovitskiy, 2020]. Although AI-based systems have achieved remarkable results on controlled benchmarks, their reliability remains a concern [Zhang *et al.*, 2020; Koh *et al.*, 2021; Zhou *et al.*, 2022]. For example, in complicated real-world tasks such as medical image diagnosis, even state-of-the-art models can still make hard-to-anticipate errors [Varoquaux and Cheplygina, 2022; Cid *et al.*, 2024]. For these tasks, human-AI collaboration is a promising approach [Wilder *et al.*, 2020; Liu *et al.*, 2024]. Existing research has demonstrated that by leveraging their complementary strengths, human-AI collaboration outperforms either humans or AI alone [Yu *et al.*, 2021; Zhao *et al.*, 2025; Tanno *et al.*, 2025].

As one of the major human-AI collaboration methods, Learning to Defer trains a deferral strategy that decides to predict by AI or defer to humans based on comparing their confidences [Madras *et al.*, 2018; Mozannar and Sontag, 2020; Verma and Nalisnick, 2022]. Further, recent works go beyond simply deferring to humans, pairing deferrals with AI assistance to support humans. Specifically, it is proven that the awareness of AI inability enhances human accuracy [Bondi *et al.*, 2022]. Auxiliary information like AI predictions and explanations augments human decision-making [Tariq *et al.*, 2025; Strong *et al.*, 2024]. In addition, some research [Kerrigan *et al.*, 2021; Singh *et al.*, 2023] maps human predicted labels into soft labels through a confusion matrix and combines them with AI's probabilities via Bayesian theory. This idea has also been incorporated into Learning to Defer [Gupta *et al.*, 2023], yielding a variant of deferral that outputs either the AI's prediction or human-AI combined prediction.

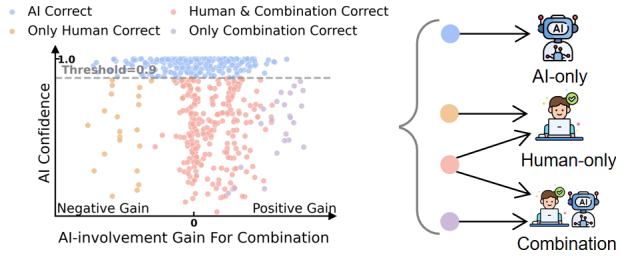


Figure 1: Visualization of prediction correctness across samples on the ImageNet-16H and Chaoyang datasets and the corresponding ideal assigned modes.

Deferring AI-uncertain samples to combination often performs better than humans alone, but this does not hold in some cases. Combination [Kerrigan *et al.*, 2021] gains from the fusion of model and human results with their corresponding confidences. However, when the model provides a low-confidence result while the human provides a high-confidence one, combination performance is degraded by noise from the model. In such cases, the human is correct but combination fails. We validate this statement on real-world datasets, i.e., ImageNet-16H and Chaoyang. As shown in Figure 1, there exist some samples that only the human correctly predicts (marked in yellow). Therefore, an ideal deferral should behave as follows: if the model is confident enough, AI-only is prioritized; otherwise, human involvement is triggered, either via human-only or combination. Some existing works

*Corresponding author

have investigated selecting among these three modes. Zhang et al. [Zhang *et al.*, 2024] directly concatenate human labels with model probabilities, failing to capture fine-grained distributions and leading to suboptimal collaboration. He et al. [He *et al.*, 2025] require expensive concept annotations for their concept-driven selection. We argue that the key issue for mode selection lies in distinguishing human-only mode from combination mode. These two modes incur the same human cost, so the selection should focus on accuracy. Thus, the key issue lies in: when does the model's result degrade the combination, and how to mitigate its effect on combination?

To address the above issue, there are two challenges. First, Combination Degradation: *how to select between human-only and combination modes with higher accuracy?* Existing works typically select between AI-only and human-involved modes based on the model's confidence - an explicit one that can be directly obtained from its probability vector. By contrast, selecting between human-only and combination is more difficult since human confidence is an implicit one that is often unavailable for both. Second, Degradation Mitigation: *how to reduce the model's negative impact on combination and improve the system performance?* The system performance heavily relies on each mode's performance. Since humans and AI are fixed, how to improve the performance of combination is important for the system performance. Unreliable model predictions can introduce noise into combination as illustrated above, and thus the key is to suppress this noise.

To address these challenges, we propose two solutions: (i) First, we introduce adaptive selection to choose among AI-only, human-only and combination modes. It formulates the mode selection as a ternary decision problem, leveraging a hybrid scoring mechanism that compares intrinsic model confidence to parameterized assessments from a discriminative gating network to identify the suitable decision mode. (ii) Second, we propose human-guided Conformal Prediction to refine model's prediction sets. It calculates the nonconformity scores by incorporating human priors and constructs prediction sets via adaptive group calibration, providing theoretical guarantees on marginal coverage.

In this paper, we propose TriHAI, a novel tri-mode deferral method that adaptively routes samples among AI-only, human-only, and refined human-AI combination modes. Specifically, given a sample, we feed it into a pre-trained classifier to obtain the model prediction and confidence score. Simultaneously, the image is given into a gating network to generate scores for the human and combination modes. Based on the scores, the system routes the sample to one of three modes: outputting the model prediction directly, querying the human, or employing human-guided Conformal Prediction to refine the model's prediction set, and combine it with the human prediction. Our main contributions are as follows:

- **Ternary Deferral**: We propose a tri-mode deferral method integrating AI-only, human-only, and human-AI combination modes to accommodate diverse scenarios.
- **Adaptive Selection**: We design a score-driven selection strategy to choose modes for samples, significantly improving the appropriateness of sample assignment.
- **Human-guided Conformal Prediction**: We propose a refined combination mode that enhances human-AI fusion accuracy by leveraging human-guided Conformal Prediction to suppress model probability noise.
- **Experiment Study**: Experiments on three real-world datasets show that TriHAI outperforms other human-AI baselines by up to 9.57% in overall accuracy.

2 Related Work

2.1 Learning to Defer Frameworks

As a prevalent method of human-AI collaboration, Learning to Defer assigns tasks to either AI or humans. Early works estimate human error or confidence to determine whether to defer [Madras *et al.*, 2018; Raghu *et al.*, 2019]. [Mozannar and Sontag, 2020] then formulates deferral as a differentiable objective and introduces surrogate losses to support end-to-end learning. Subsequent works further refine the surrogate for better joint optimization [Okati *et al.*, 2021] and for calibration to human correctness [Verma and Nalisnick, 2022]. Some works show that surrogate-based approaches may fail in the realizable settings and propose realizable consistent surrogate losses [Mozannar *et al.*, 2023; Mao *et al.*, 2024]. Meanwhile, some works [Mao *et al.*, 2023; Gupta *et al.*, 2023] point out that joint training is expensive since it cannot directly leverage pre-trained classifiers. Consequently, recent studies propose a two-stage approach, which first selects a pre-trained classifier and subsequently learns a separate deferral module [Narasimhan *et al.*, 2022; Gupta *et al.*, 2023; Mohri *et al.*, 2023].

2.2 Enhanced Deferral with Model Output

Some works argue that model output could still be useful to humans, even when the decision is deferred to humans. Research indicates that awareness of model inability enhances human accuracy [Bondi *et al.*, 2022; Jabbour *et al.*, 2025]. Model-predicted probabilities could also support humans [Tariq *et al.*, 2025; Chen *et al.*, 2023], while model-generated explanations offer humans comprehensive guidance [Strong *et al.*, 2024; Strong *et al.*, 2025]. Rather than utilizing model output as support, some research proposes a collaboration method that leverages Bayesian theory to combine the model's probabilities with the human's class-level predictions [Kerrigan *et al.*, 2021; Zou *et al.*, 2024; Gao *et al.*, 2025]. [Gupta *et al.*, 2023] integrates this method into Learning to Defer, which can either make a prediction by model or defer to human-AI combination. Only a few works explore moving beyond the above binary settings, such as [Zhang *et al.*, 2024], which attempts to unify deferral and complementarity. [He *et al.*, 2025] proposes a concept-driven method to choose among three decision modes.

However, most works typically confine the decision space to a rigid binary choice, i.e., AI or humans, and AI or combination, which may exclude the most suitable mode for a given sample. In practice, AI, humans, and human-AI combination are each indispensable. Although a few methods consider a ternary framework, they remain limited by either the direct fusion of human labels with model predictions, which yields suboptimal collaboration, or the reliance on concept annotations that incurs prohibitive costs for ternary selection.

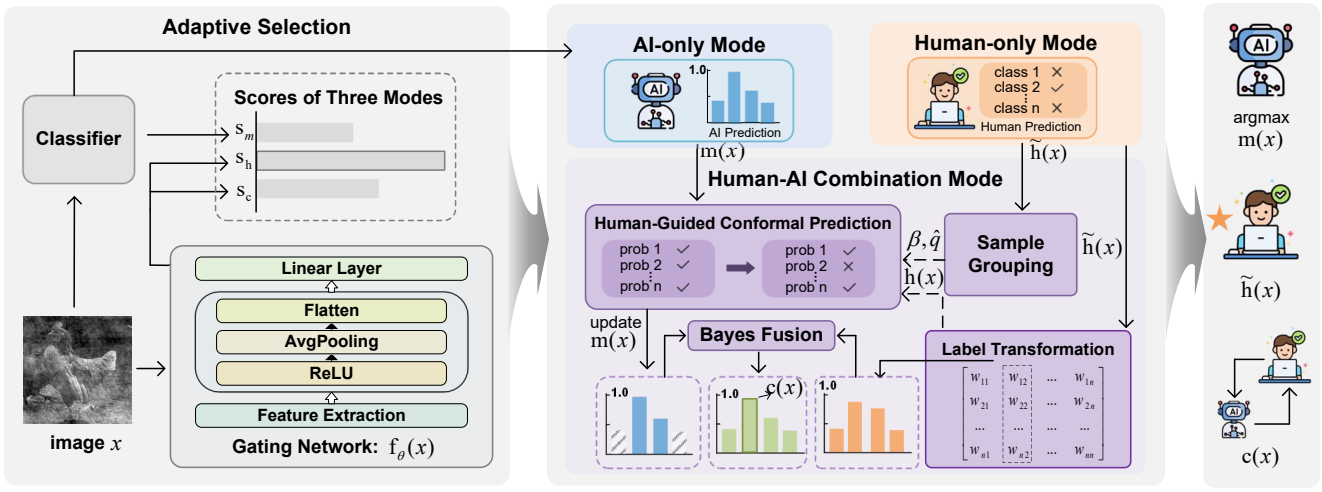


Figure 2: Overview of the proposed TriHAI. Samples are adaptively routed to one of three modes via adaptive selection. Particularly, if the combination mode is selected, we refine the AI prediction set using human-guided Conformal Prediction and integrate it with human predictions via Bayes fusion.

3 Approach

3.1 Problem Formulation

We consider a K-way image classification task defined over an input space $\mathcal{X}$ and a label space $\mathcal{Y} = \{1, \ldots, K\}$. There are two types of decision-makers involved: a model and a human. For an input sample $x \in \mathcal{X}$, a pre-trained model $m : \mathcal{X} \to \mathbb{R}^K$ yields a probability vector $m(x)$; a queryable human, modeled as a decision function $\tilde{h} : \mathcal{X} \to \mathcal{Y}$, provides a hard label $\tilde{h}(x) \in \mathcal{Y}$.

Our goal is to maximize collaboration performance between the human and the AI, thereby improving the overall classification accuracy. Beyond the predictions of the model $m(x)$ and the human $\tilde{h}(x)$, we introduce the third mode: human-AI combination. This mode combines the predictions from both the human and the AI to yield a hybrid result $c(x)$. Therefore, to be more specific, our objective is to design an adaptive decision framework that can flexibly switch among three modes: **AI-only**, **Human-only**, and **Human-AI combination**.

3.2 Implementation of Three Modes

The overall framework is illustrated in Figure 2. In this subsection, we detail the implementation of the three modes.

- **AI-only Mode:** The input x is processed by the pre-trained model m to produce a probability vector $m(x)$. The final prediction is $\arg\max(m(x))$.
- **Human-only Mode:** The input x is deferred to the human, who provides a class label $\tilde{h}(x)$ as the final prediction.
- **Human-AI Combination Mode:** This mode integrates the model output $m(x)$ and the human output $\tilde{h}(x)$ via a specific decision combination strategy to yield the combined prediction result $c(x)$.

Since the AI-only and human-only modes are relatively simple, the following content focuses primarily on the implementation of human-AI combination.

Combination of AI and Human

As confidence is critical for human-AI combination decision, we first map the human's hard label $\tilde{h}(x)$ into soft labels $h(x)$ via confusion matrix ϕ, defined as:

$$\phi_{uv} = p(\tilde{h}(x) = u \mid y = v), \tag{1}$$

where u denotes the human prediction and v denotes the true label. To reduce the variance of maximum likelihood estimation, we follow prior works [Kerrigan *et al.*, 2021] and place a Dirichlet prior on each column of the confusion matrix:

$$\phi_{\cdot v} \sim \text{Dirichlet}(\zeta_v), \tag{2}$$

where $\zeta_v \in \mathbb{R}^K$ is defined by $(\zeta_v)_u = \gamma$ if $u = v$ and $(\zeta_v)_u = \vartheta$ otherwise, assigning higher mass to diagonal entries and lower mass to off-diagonal entries.

Given training data, the posterior confusion matrix is obtained by combining this Dirichlet prior with the empirical counts. The confusion matrix ϕ is then used to map the human's hard label $\tilde{h}(x)$ into soft labels $h(x)$ as follows:

$$h(x) = \phi_{\tilde{h}(x)\cdot}, \tag{3}$$

where $\cdot$ denotes the selection of the row corresponding to the index $\tilde{h}(x)$. Then, we combine the human soft labels $h(x)$ and the model probability vector $m(x)$ based on Bayesian theory, thereby obtaining the final human-AI collaborative decision $c(x)$. Specifically, for the class k, the combined probability $P(y = k|h(x), m(x))$ can be defined as:

$$P(y = k|h(x), m(x)) = \frac{h(x)_k \cdot m(x)_k}{\sum_i h(x)_i \cdot m(x)_i}, \tag{4}$$

where $m(x)_k$ denotes the model's probability of class k, and $h(x)_k$ represents the human's confidence of class k. Finally, the human-AI combination decision $c(x)$ is defined as:

$$c(x) = \arg\max P(y = k|h(x), m(x)). \tag{5}$$

Human-Guided Conformal Prediction

The model output $m(x)$ provides probabilities for all the candidate classes, which may inevitably includes some erroneous classes that exert a negative impact on the final result. We propose human-guided Conformal Prediction (CP) to filter erroneous classes, yielding a refined vector $m^*(x)$ for fusion. Conformal Prediction (CP) [Angelopoulos and Bates, 2021; Sun *et al.*, 2023] is a distribution-free method that outputs prediction sets containing ground-truth labels with a desired marginal coverage guarantee $1-\alpha$. In our tasks, since humans often have a stronger semantic understanding than the model, we incorporate human soft labels $h(x)$ to guide the computation of nonconformity score, which quantifies label disagreement. Considering that $h(x)$ may also be uncertain, we introduce a parameter $1-\beta$ to weight the contribution. It proceeds in three steps, as illustrated below:

Definition of Human-guided Nonconformity Score: For an input sample x with ground-truth label y, we define the human-guided nonconformity score $S(x, y)$ as follows:

$$S(x,y) = -\left(\beta \cdot \log m(x)_y + (1-\beta) \cdot \log h(x)_y\right), \quad (6)$$

where $\beta \in [0,1]$ balances the model probability $m(x)_y$ and human confidence $h(x)_y$ on the true label.

Calibration: We construct a calibration set $\mathcal{D}_{cal} = \{(x_i, y_i, \tilde{h}(x_i))\}_{i=1}^n$ and compute nonconformity scores $\mathcal{S}_{cal}$ for each sample. To ensure $1-\alpha$ marginal coverage, we set the threshold $\hat{q}$ as the k-th smallest score in $\mathcal{S}_{cal}$, where $k = \lceil (n+1)(1-\alpha) \rceil$.

Prediction: For a new sample x, we refine the prediction set using the calibrated threshold $\hat{q}$:

$$\Gamma(x) = \{k \in \mathcal{Y} \mid S(x,k) \leq \hat{q}\}. \quad (7)$$

We then obtain the refined vector $m^*(x)$ by masking classes outside $\Gamma(x)$ and update the model output:

$$m(x) \leftarrow m^*(x). \quad (8)$$

Sample Grouping

In Equation (6), we employ a parameter $1-\beta$ to weight the human contribution, assigning a lower β (i.e., a larger $1-\beta$) when human judgment is reliable. To quantify this, we compare the statistical accuracy for each class k, derived from the diagonal elements of the human confusion matrix ϕ and the model confusion matrix ϕ^m:

$$Acc_h(k) = \phi_{kk}, \quad Acc_m(k) = \phi^m_{kk}. \quad (9)$$

Then, we partition the label space $\mathcal{Y}$ into two mutually exclusive subsets: the set contains the samples that are easy for the human $\mathcal{Y}_{easy}$ and that are hard for the human $\mathcal{Y}_{hard}$:

$$\mathcal{Y}_{easy} = \{k \in \mathcal{Y} \mid Acc_h(k) > Acc_m(k)\}, \quad (10)$$

$$\mathcal{Y}_{hard} = \mathcal{Y} \setminus \mathcal{Y}_{easy}. \quad (11)$$

We assign each sample x to g_{easy} if the human prediction $\tilde{h}(x)$ belongs to $\mathcal{Y}_{easy}$, and to g_{hard} otherwise. We determine optimal group-specific parameters β_{easy} and β_{hard} via grid search on a validation set $\mathcal{D}_{val}$. We select values within $[0,1]$ that maximize the final fusion accuracy for each group. The complete process for implementing the combination mode is given in Algorithm 1.

Algorithm 1 Implementation of Combination Mode

Require: Calibration set $\mathcal{D}_{cal}$, validation set $\mathcal{D}_{val}$, sample x, miscoverage rate α, step size $\Delta\beta$.
Ensure: Prediction result $c(x)$.

Preparation Compute the confusion matrices for the human and the model. Identify $\mathcal{Y}_{easy}$ and $\mathcal{Y}_{hard}$.
Model Prediction Refinement
Partition $\mathcal{D}_{cal}$ and $\mathcal{D}_{val}$ into two groups.
for g_{easy} and g_{hard} **do**
  for $\beta \in [0, \Delta\beta, ..., 1]$ **do**
    Compute $\mathcal{S}_{cal}$ on $\mathcal{D}_{cal}$ and obtain $\hat{q}$ under α.
    Update $m(x)$ by human-guided CP and evaluate fusion accuracy on $\mathcal{D}_{val}$.
  end for
  Select β with the highest fusion accuracy.
end for
For a new sample x, identify its group g, make human-guided CP and update the model's probability vector $m(x) \leftarrow m^*(x)$.
Human-AI Combination
Compute combined probability $P(y = k \mid h(x), m(x))$ for $k \in \Gamma(x)$ through Equation (4).
Obtain the final prediction $c(x)$ through Equation (5).
return $c(x)$.

3.3 Adaptive Selection Among Each Mode

Next, we address how to adaptively route each sample. We formulate adaptive selection as a ternary decision problem: given an input x, the goal is to select the suitable mode $r \in \{m, h, c\}$ among the AI, the human, and combination.

Specifically, for each instance x, we compute the scores $\mathbf{s}(x) = [s_m(x), s_h(x), s_c(x)]$ that quantify the suitability of each mode. The score for the AI-only mode, i.e., $s_m(x)$, is non-parametric and derived directly from the probability vector $m(x)$, as itself serves as a direct assessment of the model's capability. We formulate it as follows:

$$s_m(x) = \log m_{(1)}(x) - \log m_{(2)}(x), \quad (12)$$

where $m_{(1)}$ and $m_{(2)}$ denote the highest and second-highest probabilities, respectively. In contrast, the scores involving human participation, i.e., $s_h(x)$ and $s_c(x)$, are parameterized and generated by a gating network f_θ:

$$[s_h(x), s_c(x)] = f_\theta(x). \quad (13)$$

During the test phase, the module selects the mode with the highest score as the optimal one $\hat{r}$.

$$\hat{r} = \underset{r \in \{m,h,c\}}{\arg\max} \{s_m(x), s_h(x), s_c(x)\} \quad (14)$$

The loss function of the gating network is given below:

$$\mathcal{L} = \mathcal{L}_{\text{route}} + \lambda \mathcal{L}_{\text{rank}}, \quad (15)$$

where λ is a hyperparameter used to balance the regularization strength. The first term aims to maximize the probability that the optimal mode is selected. It is defined as follows:

$$\mathcal{L}_{\text{route}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{r \in \{m,h,c\}} w_r(x_i) \log P(r|x_i), \quad (16)$$

where N is the total number of train samples, and $P(r|x)$ is the probability that r mode is selected. It can be obtained through the following formula:

$$P(r|x) = \frac{\exp(s_r(x))}{\sum_{k\in\{m,h,c\}} \exp(s_k(x))}. \tag{17}$$

The equation normalizes the scores across the three modes. Then, the target weights $w_r(x)$ are assigned dynamically based on mode characteristics. For the AI-only mode ($r = m$), the weight serves as a binary indicator activated strictly upon correct prediction:

$$w_m = \mathbb{I}(\hat{y}_m = y). \tag{18}$$

Conversely, for the other two modes involving human participation, i.e., $r \in \{h, c\}$, the weights are calculated as follows:

$$w_r = \Lambda_r(\delta) \cdot \bar{C}_r, \tag{19}$$

where

$$\begin{aligned} \Lambda_r(\delta) &= \sigma(\kappa_r(\delta - \eta_r)), \\ \bar{C}_r &= 1 - (c_p \mathbb{I}(\hat{y}_r \neq y) + c_b). \end{aligned} \tag{20}$$

The weight comprises two terms. $\Lambda_r(\delta)$ utilizes the sigmoid function σ to act as a differentiable switch governed by uncertainty $\delta = 1 - m_{(1)}$, where η_r sets the activation threshold and κ_r controls sharpness. Simultaneously, $\bar{C}_r$ models human's utility, applying a base cost c_b and a heavy penalty c_p for errors via an indicator function.

The second term of total loss enforces the correct relative ordering between the scores of the human (s_h) and that of combination (s_c):

$$\mathcal{L}_{\text{rank}} = \mathbb{E}\left[\mathbb{I}_{\{c\succ h\}}\psi(s_h - s_c) + \mathbb{I}_{\{h\succ c\}}\psi(s_c - s_h)\right], \tag{21}$$

where $\mathbb{I}_{\{a\succ b\}}$ indicates samples where mode a is correct while mode b is incorrect, and $\psi(z) = \log(1 + e^z)$ is the softplus function used to smoothly penalize score inversions.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the proposed method on three datasets.

ImageNet-16H [Steyvers *et al.*, 2022]: This dataset contains 1,200 ImageNet test images from 16 classes, each corrupted with phase noise at $\omega \in \{80, 95, 110, 125\}$ (4,800 noisy images). Model predictions come from diverse pretrained networks, and human predictions are obtained by randomly sampling one from six annotations per image. We utilize the $\omega = 125$ subset, as other levels lack the 10% accuracy gap necessary to establish scenarios between humans and AI.

Chaoyang [Zhang *et al.*, 2024]: This dataset contains colon slides collected from Chaoyang Hospital. It has 6,160 images across four classes: normal, serrated lesion, adenocarcinoma, and adenoma. Model predictions are generated by three distinct trained models. Human labels are derived from ground truth annotations provided by professional doctor.

CIFAR-100 [Hemmer *et al.*, 2022]: This dataset contains 60,000 images categorized into 100 subclasses and 20 superclasses. Model predictions are obtained from two separately trained models, while human labels are artificially generated. Following prior works, we create synthetic humans proficient in 2 to 18 superclasses. A human predicts perfectly within its proficient superclasses; otherwise it randomly outputs an incorrect superclass label. This yields nine humans, from which we select two with accuracies of about 70% and 80%.

Baselines. We compare our proposed TriHAI with the following baselines belonging to three categories.

- **Learning to defer.** OVA-Surrogate [Verma and Nalisnick, 2022] and Realizable-Surrogate [Mozannar *et al.*, 2023], which jointly learn classifier and deferral through two consistent surrogate losses, respectively; Two-Stage Deferral [Mao *et al.*, 2023], which trains the classifier and deferral module sequentially.
- **Human-AI combination.** CM [Kerrigan *et al.*, 2021], which fuses human and AI predictions via Bayesian combination; CP-CM, which further refines AI predictions via human-guided CP.
- **Deferral-Combination hybrid methods.** C2D [Gupta *et al.*, 2023], which combines Learning to Defer with Bayesian human-AI combination; LECODU [Zhang *et al.*, 2024], which allows choosing among model prediction, human prediction, or human-AI collaboration.

Implementation Details. We establish scenarios across three datasets covering AI-superior, comparable, and human-superior cases, defining superiority as an accuracy gap exceeding 10%. For ImageNet-16H, we directly use three pretrained models: VGG-19 [Simonyan and Zisserman, 2014], GoogLeNet [Szegedy *et al.*, 2015], and AlexNet [Krizhevsky *et al.*, 2012]. For the other datasets, we split the data to train backbone models. For Chaoyang, 4,724 images are used to train NSHE [Zhu *et al.*, 2021], NF-Net [Cao *et al.*, 2020], and Co-teaching [Han *et al.*, 2018]. By pairing each of these models with the human, we establish three distinct scenarios for each of above two datasets. For CIFAR-100, 50,000 images are used to fine-tune AlexNet and ResNet-18 [He *et al.*, 2016], which are then paired with two simulated humans to form four scenarios, including one AI-superior, one human-superior, and two comparable scenarios. The remaining data from Chaoyang, CIFAR-100, and ImageNet-16H are reserved for training and validating our method.

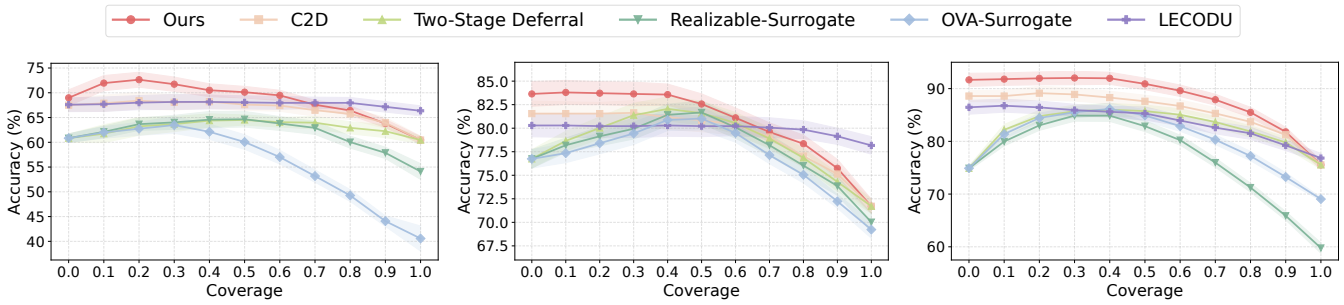
4.2 Overall Performance

We evaluate the accuracy of our method and the baselines on ImageNet-16H, Chaoyang, and CIFAR-100 datasets. The results in Table 1 show that our method outperforms all the baselines. Specifically, compared to humans or AI alone, our method achieves at least 7.17% improvement in General setting, which underscores the necessity of human-AI collaboration. Additionally, our method outperforms CP-CM. As AI-only, human-only, and CP-CM constitute the three foundational decision modes within our approach, it demonstrates that adopting any single mode is not optimal and therefore we should make adaptive selection among the three modes.

Next, we conduct ablation studies to evaluate human-guided CP and adaptive selection. First, we remove human-guided CP, and instead directly combine human and AI predictions via Bayesian rule. Experimental results indicate that

Method	ImageNet-16H ($\omega = 125$)			Chaoyang			CIFAR-100		
	General	AI-Superior	Human-Superior	General	AI-Superior	Human-Superior	General	AI-Superior	Human-Superior
Human	60.83±0.86	60.83±0.86	60.83±0.86	76.74±0.98	76.74±0.98	76.74±0.98	74.90±0.40	70.73±0.25	79.97±0.33
AI	60.46±0.75	70.98±0.55	49.72±0.62	71.71±0.75	86.98±0.86	58.60±0.59	75.55±0.81	81.10±0.49	69.57±0.66
CM	67.50±1.22	72.22±1.20	64.44±1.24	81.55±1.32	89.53±1.12	75.35±1.40	88.59±1.03	90.03±0.76	88.37±0.81
CP-CM	68.43±1.35	74.44±1.22	65.00±1.43	83.41±1.42	90.23±1.28	76.84±1.56	91.15±1.30	91.20±0.90	89.61±0.99
OVA-Surrogate	61.22±1.37	-	-	80.42±1.13	-	-	85.78±0.99	-	-
Realizable-Surrogate	61.96±1.02	-	-	81.49±1.07	-	-	85.03±1.10	-	-
Two-Stage Deferral	62.68±1.44	71.11±1.28	62.77±1.31	81.72±1.28	88.63±1.33	75.70±1.45	85.06±1.05	87.20±0.99	88.03±1.01
C2D	67.86±1.51	73.88±1.49	62.22±1.45	81.50±1.15	89.30±1.19	72.40±1.28	88.69±1.11	89.50±1.33	87.56±1.23
LECODU	67.96±1.34	75.46±1.35	62.50±1.33	80.37±1.25	86.51±1.43	74.65±1.31	86.46±1.09	81.50±1.22	72.83±1.15
Ours	**70.79**±1.21	**75.95**±1.30	**65.27**±1.27	**83.91**±1.25	**91.86**±1.25	**77.42**±1.21	**91.93**±1.04	**92.01**±1.11	**91.42**±1.15
w/o CP	68.60±1.06	74.45±1.37	65.00±1.56	82.79±1.15	90.45±1.23	76.17±1.39	89.60±1.14	91.20±1.05	90.57±1.17
w/o Adaptive Selection	67.70±0.65	72.69±1.23	63.61±1.13	81.02±0.61	88.00±1.15	75.58±1.40	89.62±0.18	89.58±0.38	90.20±0.37

Table 1: Accuracy comparison across the three datasets. "General" denotes the comprehensive dataset, encompassing scenarios where the AI is superior, inferior, or comparable to the human. "AI-superior" and "human-superior" refer to specific sub-datasets where the accuracy gap between the AI and the human exceeds 10% in favor of the respective agent.



(a) ImageNet-16H (ω=125) (b) Chaoyang (c) CIFAR100

Figure 3: Accuracy vs coverage (proportion of samples not deferred) plots across three datasets showcasing the behavior of our method and baselines. To achieve different levels of coverage, we sort the deferral scores for each method on the test set and vary the threshold. In our method, we define the deferral score as the maximum of human-only and combination scores, deferring samples to the corresponding mode.

this ablation leads to a significant drop in accuracy, particularly on CIFAR-100 with a decrease of 2.33%. Second, for adaptive selection, we employ a threshold-based strategy: samples are routed to the AI-only mode if model confidence exceeds τ_{AI}, to the human-only mode if below τ_h, and to combination otherwise. This ablation leads to a clear accuracy drop, highlighting the importance of adaptive selection.

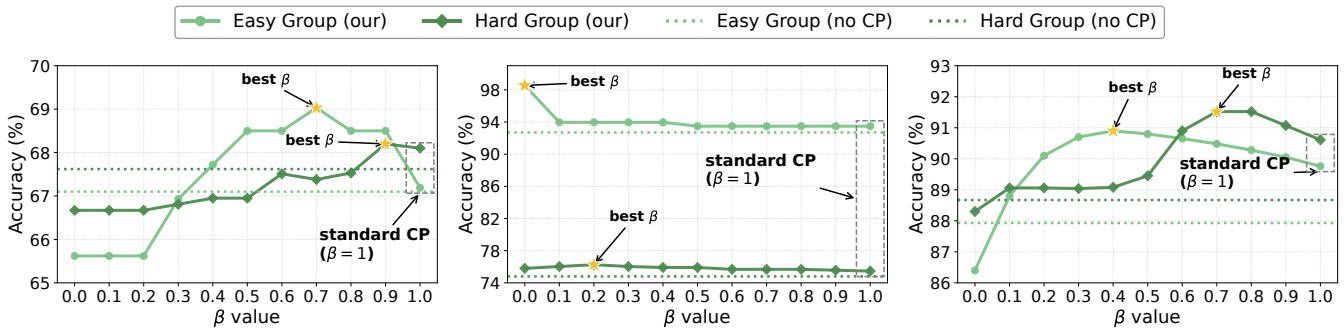
Intuitively, deferring to combination outperforms deferring to humans alone, as it integrates human knowledge and model predictions. However, the conclusion doesn't always hold. Specifically, in human-superior settings, the results reveal that C2D (select AI or combination) frequently underperforms the Learning to Defer baselines, such as Two-Stage Deferral (select AI or humans). This counterintuitive result suggests that, in some scenarios, relying on humans alone can be better. Therefore, it's important to include all three modes.

4.3 Accuracy and Coverage

We analyze the relationship between accuracy and coverage, defined as the proportion of samples predicted by AI, to evaluate system performance under manpower constraints. Notably, as coverage increases, the demand for human manpower decreases accordingly. We generate accuracy-coverage curves by sorting test samples using deferral scores and sweeping a threshold to vary coverage from 0 to 1. In our method, we define the deferral score as the maximum of the human-only and combination scores, and defer each sample to the mode that attains this maximum. As shown in Figure 3, our method achieves superior accuracy compared to the baselines across most coverages. Additionally, when coverage is 0, meaning the model is completely uninvolved, our approach (humans or combination) outperforms the C2D (combination). This demonstrates that assigning all samples to combination mode is not always the optimal strategy.

4.4 Evaluation on Human-AI Combination Mode

We also investigate the impact of the parameter β on the accuracy of combination mode across sample groups of varying difficulty (i.e., easy vs. hard). It is worth noting that when $\beta = 1$, our method reduces to standard CP. As illustrated in Figure 4, our proposed human-guided CP, starred at the optimal β values for peak combination accuracy, consistently outperforms standard CP ($\beta = 1$) across all experimental settings, demonstrating the effectiveness of integrating hu-



(a) ImageNet-16H (ω=125) (b) Chaoyang (c) CIFAR100

Figure 4: Human-AI combination mode accuracy across varying β values for easy and hard sample groups. Peak accuracies (marked by stars) are consistently achieved when $\beta < 1$, outperforming the standard CP baseline ($\beta = 1$).

Dataset	Mode	Precision	Recall	F1 score
ImageNet-16H (ω=125)	AI-only	0.902	0.885	0.893
	Human-only	0.880	0.820	0.848
	Combination	0.845	0.905	0.874
Chaoyang	AI-only	0.801	0.760	0.780
	Human-only	0.902	0.835	0.867
	Combination	0.860	0.920	0.889
CIFAR-100	AI-only	0.972	0.883	0.925
	Human-only	0.850	0.781	0.814
	Combination	0.842	0.890	0.865

Table 2: Precision, recall, and F1 score of the selection on three modes across three datasets. It is evaluated against the ground truth that prioritizes the AI-only mode when it is correct or when all modes fail; otherwise, the ground truth is assigned to any successful human-involved mode.

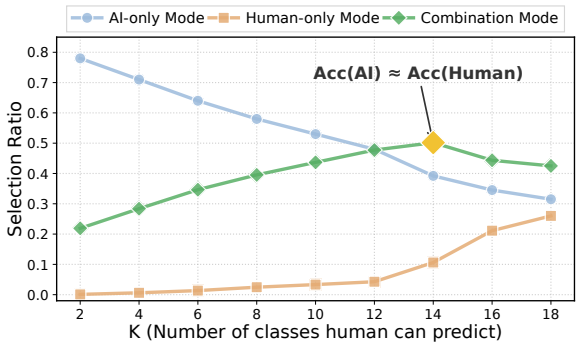


Figure 5: Selection ratios of the AI-only, human-only, and combination modes across varying levels of human expertise on CIFAR-100.

man guidance. We observe a lower optimal β for the easy group compared to the hard group. As $1 - \beta$ controls the human's weight (Equation (6)), this supports our assumption that easy samples require higher reliance on humans. Notably, on Chaoyang dataset, we observe a significant accuracy gap between easy and hard groups. It stems from near-perfect human performance (over 98%) on two categories, which sharply elevates the easy group's accuracy.

4.5 Evaluation on Adaptive Selection

To evaluate the suitability of our mode selection, we first define the ground truth by prioritizing the AI-only mode (if correct or if all modes fail) over successful human or combination modes. Subsequently, we evaluate the performance using precision, recall, and F1-score. As presented in Table 2, our method achieves high F1 scores across all three datasets. The result demonstrates that our method accurately assigns samples to their appropriate modes. AI-only and human-only modes exhibit high precision but low recall, unlike combination. This indicates a cautious strategy that defaults ambiguous samples to combination. Future work will focus on the switching boundaries among these modes.

We further examine how the selection ratios of the three modes change under different numbers of classes humans can predict K. We use CIFAR-100 with a fixed ResNet-18 model and nine synthetic humans of varying expertise. As shown in Figure 5, as the human's capability gradually increases, the selection ratio of the human-only mode increases, while the selection ratio of the AI-only mode decreases. The pre-trained model ResNet-18 achieves an accuracy of about 72%. When K reaches 14, corresponding to an accuracy of about 70%, human and model capabilities become comparable, and combination mode attains its highest selection ratio.

5 Conclusion

In this paper, we propose TriHAI, a method designed to optimize decision-making by adaptively routing samples among AI-only, human-only, and human-AI combination modes. Through an adaptive selection approach, we dynamically identify the suitable decision mode for each sample to maximize system performance. Furthermore, we employ human-guided Conformal Prediction to suppress model probability noise, thereby enhancing the accuracy of human-AI combination mode. Experimental results on three datasets demonstrate that our proposed method consistently outperforms existing human-AI collaboration methods. This approach effectively addresses the limitations of binary deferral strategies in complex scenarios, offering a robust solution for advanced human-AI teaming.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62372381).

References

[Angelopoulos and Bates, 2021] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[Bondi *et al.*, 2022] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5286–5294, 2022.

[Cao *et al.*, 2020] Zhantao Cao, Guowu Yang, Qin Chen, Xiaolong Chen, and Fengmao Lv. Breast tumor classification through learning from noisy labeled ultrasound images. *Medical Physics*, 47(3):1048–1057, 2020.

[Chen *et al.*, 2023] Guanting Chen, Xiaocheng Li, Chunlin Sun, and Hanzhao Wang. Learning to make adherence-aware advice. *arXiv preprint arXiv:2310.00817*, 2023.

[Cid *et al.*, 2024] Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57, 2024.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gao *et al.*, 2025] Wei Gao, Jintian Feng, Mengqi Wei, Rui Zou, and Jianwen Sun. Towards a multi-granulated statistical framework for human–machine collaboration in image classification. *IEEE Transactions on Multimedia*, 27:1625–1636, 2025.

[Gupta *et al.*, 2023] Sumeet Gupta, Shweta Jain, Shashi Shekhar Jha, Pao-Ann Hsiung, and Ming-Hung Wang. Take expert advice judiciously: Combining groupwise calibrated model probabilities with expert predictions. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pages 956–963. IOS Press, 2023.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2025] Chengbo He, Bochao Zou, Junliang Xing, Jiansheng Chen, Yuanchun Shi, and Huimin Ma. Decode: Defer-and-complement decision-making via decoupled concept bottleneck models. *arXiv preprint arXiv:2505.19220*, 2025.

[Hemmer *et al.*, 2022] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-ai teams: building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948*, 2022.

[Jabbour *et al.*, 2025] Sarah Jabbour, David Fouhey, Nikola Banovic, Stephanie D Shepard, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. On the limits of selective ai prediction: a case study in clinical decision making. *arXiv preprint arXiv:2508.07617*, 2025.

[Kerrigan *et al.*, 2021] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[Liu *et al.*, 2024] Jiaqi Liu, Fengming Zhang, Xin Zhang, Zhiwen Yu, Liang Wang, Yao Zhang, and Bin Guo. hmcodetrans: Human–machine interactive code translation. *IEEE Transactions on Software Engineering*, 50(5):1163–1181, 2024.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

[Mao *et al.*, 2023] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in Neural Information Processing Systems*, 36:3578–3606, 2023.

[Mao *et al.*, 2024] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable h-consistent and bayes-consistent loss functions for learning to defer. *Advances in Neural Information Processing Systems*, 37:73638–73671, 2024.

[Mohri *et al.*, 2023] Christopher Mohri, Daniel Andor, Eunsol Choi, and Michael Collins. Learning to reject with a fixed predictor: Application to decontextualization. *arXiv preprint arXiv:2301.09044*, 2023.

[Mozannar and Sontag, 2020] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

[Mozannar *et al.*, 2023] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

[Narasimhan *et al.*, 2022] Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35:29292–29304, 2022.

[Okati *et al.*, 2021] Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

[Raghu *et al.*, 2019] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Singh *et al.*, 2023] Sagalpreet Singh, Shweta Jain, and Shashi Shekhar Jha. On subset selection of multiple humans to improve human-ai team accuracy. In *Proceedings of the 2023 international conference on autonomous agents and multiagent systems*, pages 317–325, 2023.

[Steyvers *et al.*, 2022] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

[Strong *et al.*, 2024] Joshua Strong, Qianhui Men, and Alison Noble. Towards human-ai collaboration in healthcare: Guided deferral systems with large language models. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

[Strong *et al.*, 2025] Joshua Strong, Qianhui Men, and J Alison Noble. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28413–28421, 2025.

[Sun *et al.*, 2023] Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36:80324–80337, 2023.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Tanno *et al.*, 2025] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.

[Tariq *et al.*, 2025] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. A2c: A modular multi-stage collaborative decision framework for human–ai teams. *Expert Systems with Applications*, 282:127318, 2025.

[Varoquaux and Cheplygina, 2022] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.

[Verma and Nalisnick, 2022] Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022.

[Wilder *et al.*, 2020] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

[Yu *et al.*, 2021] Zhiwen Yu, Qingyang Li, Fan Yang, and Bin Guo. Human–machine computing. *CCF Transactions on Pervasive Computing and Interaction*, 3(1):1–12, 2021.

[Zhang *et al.*, 2020] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–41, 2020.

[Zhang *et al.*, 2024] Zheng Zhang, Wenjie Ai, Kevin Wells, David Rosewarne, Thanh-Toan Do, and Gustavo Carneiro. Learning to complement and to defer to multiple users. In *European Conference on Computer Vision*, pages 144–162. Springer, 2024.

[Zhao *et al.*, 2025] Xuehan Zhao, Jiaqi Liu, Xin Zhang, Zhiwen Yu, and Bin Guo. Activehai: Active collection based human-ai diagnosis with limited expert predictions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 4282–4290, 2025.

[Zhou *et al.*, 2022] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

[Zhu *et al.*, 2021] Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE transactions on medical imaging*, 41(4):881–894, 2021.

[Zou *et al.*, 2024] Rui Zou, Sannyuya Liu, Yawei Luo, Yaqi Liu, Jintian Feng, Mengqi Wei, and Jianwen Sun. Balancing humans and machines: A study on integration scale and its impact on collaborative performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17628–17636, 2024.