Are They the Same Picture? Adapting Concept Bottleneck Models for Human-AI Collaboration in Image Retrieval

Vaibhav Balloli[1] , **Sara Beery**[2] , **Elizabeth Bondi-Kelly**[1]
[1]University of Michigan
[2]Massachusetts Institute of Technology
vballoli@umich.edu, beery@mit.edu, ecbk@umich.edu

Abstract

Image retrieval plays a pivotal role in applications from wildlife conservation to healthcare, for finding individual animals or relevant images to aid diagnosis. Although deep learning techniques for image retrieval have advanced significantly, their imperfect real-world performance often necessitates including human expertise. Human-in-the-loop approaches typically rely on humans completing the task independently and then combining their opinions with an AI model in various ways, as these models offer very little interpretability or *correctability*. To allow humans to intervene in the AI model instead, thereby saving human time and effort, we adapt the Concept Bottleneck Model (CBM) and propose `CHAIR`. `CHAIR` (a) enables humans to correct intermediate concepts, which helps *improve* embeddings generated, and (b) allows for flexible levels of intervention that accommodate varying levels of human expertise for better retrieval. To show the efficacy of `CHAIR`, we demonstrate that our method performs better than similar models on image retrieval metrics without any external intervention. Furthermore, we also showcase how human intervention helps further improve retrieval performance, thereby achieving human-AI complementarity.

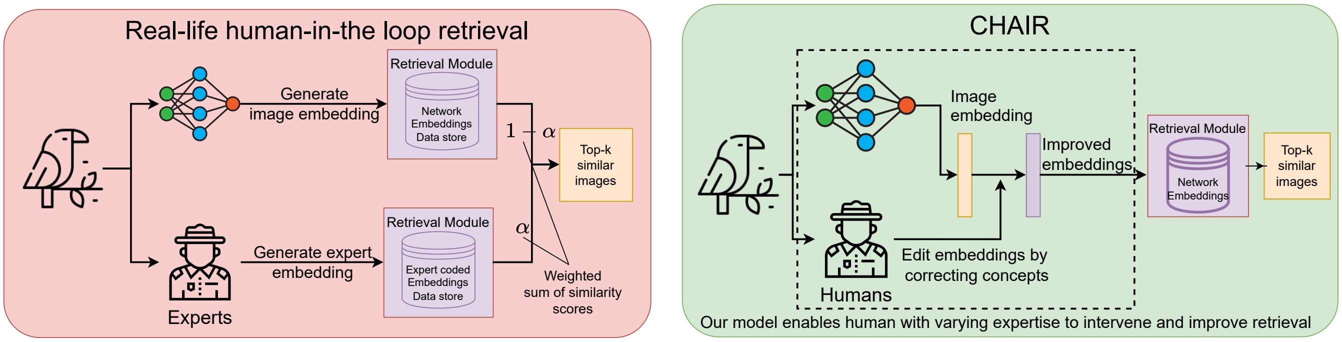
1 Introduction

Recent advances in AI have shown great promise in important, often high-risk domains like healthcare [Peiffer-Smadja *et al.*, 2020; Rajpurkar *et al.*, 2020], wildlife conservation [Kulits *et al.*, 2021; Beery *et al.*, 2019], and reducing misinformation [Mendes *et al.*, 2022]. However, these advances are imperfect, and can lead to harm when deployed. For example, [Beede *et al.*, 2020] reports errors when deploying AI models to detect diabetic retinopathy due to challenging real-world factors like lighting, leading to potential human harms.

Researchers have proposed human-AI collaboration as a promising approach to mitigate the shortcomings of AI models in these domains [De-Arteaga *et al.*, 2020]. For example, prior work in health AI has sought to achieve better accuracy via decision-support tools for clinicians [Peiffer-Smadja *et al.*, 2020] that *assist* humans in decision-making. Human-in-the-loop participatory systems have also been proposed to accurately and robustly categorize wildlife images [Kulits *et al.*, 2021; Miao *et al.*, 2021; Bondi *et al.*, 2022] and to aid fact-checkers [Nguyen *et al.*, 2018; Mendes *et al.*, 2022]. Concept Bottleneck Models (CBMs) are another recent and promising method to facilitate collaboration [Koh *et al.*, 2020]. They allow humans to *interact* with AI models by viewing and manipulating intermediate, high-level concepts (e.g., whether a bird has a blue wing), which are then used for prediction.

While recent works using CBMs have significantly improved performance in classification [Zarlenga *et al.*, 2023], we note that there exist many non-classification application areas that also require human-AI collaboration. For example, ElephantBook [Kulits *et al.*, 2021] is a state-of-the-art elephant re-identification platform – meant to determine which individual elephant is depicted in each image – that adopts a semi-automated human-in-the-loop approach. This approach retrieves and presents to a user the most likely matches to individuals in a known population database for each new elephant sighting. They combine a neural network-based visual similarity ranking with a similarity ranking based on domain expert input via a weighted average (see Figure 1a). However, the weighted average approach requires hyperparameter tuning (α) depending on each user's expertise. Enabling humans to leverage concepts that they already use for retrieval to instead interact with the neural network in the embedding space has the potential to significantly reduce the amount of tuning required for each user while still including human inputs to improve team performance, as illustrated in Figure 1b).

In this work, we aim to adapt CBMs to facilitate close human-AI collaboration at deployment time for image retrieval tasks, such as ElephantBook. We specifically aim to answer the following research questions: **(RQ1)** How do representations generated by CBMs compare to corresponding traditional models? **(RQ2)** How can we augment CBMs to enable human intervention in image retrieval and classification? **(RQ3)** How can we train these models to incorporate varying levels of expertise? To answer these questions, we analyze how CBMs compare with traditional models for image retrieval, introduce `CHAIR (CBM-Enabled Human-AI Collaboration for Image Retrieval)`, a novel CBM architecture that allows humans to intervene and encode concepts to improve



(a) Typical Human-in-the-loop pipeline for image retrieval.

(b) Our proposed collaborative pipeline for image retrieval.

Figure 1: Our proposed method allows human-AI collaboration in image retrieval by enabling humans to edit embeddings by correcting them through high-level concepts. We also enable flexible intervention to lower the expertise needed to participate.

retrieval, and perform extensive evaluations on how these interventions enable *embedding-level* human-AI collaboration.

2 Background and Motivation

Concept Bottleneck Models: The key promise of CBMs is two-fold: CBMs a) predict *high-level* intermediate concepts, which are then used to predict the final class label, thus improving interpretability, and b) enable humans to intervene and correct these intermediate concepts to improve classification performance, thus providing intervenability. This is achieved in two steps: i) Concept Bottleneck, which predicts the high-level, understandable concepts, and ii) Classifier, which predicts the final class based on the predicted concepts (see Figure 2). Human-AI collaboration is made possible here by allowing humans to *correct* the model by modifying these intermediate concepts. This is shown to increase the overall classification performance on various tasks [Koh *et al.*, 2020].

Image retrieval: Image retrieval is a key component in visual tasks like image re-identification [Wang *et al.*, 2020], wildlife conservation [Kulits *et al.*, 2021], remote sensing [Liu *et al.*, 2020], and visual recommendation systems [Shankar *et al.*, 2017]. Image retrieval requires systems to *fetch* the most relevant images from a database given a query image, where relevance is defined depending on the application. Traditionally, applications using deep learning techniques to perform image retrieval leverage (latent) *embeddings* generated by neural network trained for classification. For a given query image, a *query embedding* is generated using the same neural network to find the top-k nearest embeddings using some distance function [Wan *et al.*, 2014], typically cosine distance. We refer to the images and embeddings that are searched over as *gallery* images and embeddings, respectively. Following previous literature, we utilize the Recall@k metric and RecallAccuracy@k metrics to measure the performance of an image retrieval technique. For a given image and label pair (x_i, y_i), let the top-k retrieved images be $\mathbf{y'}_i \in \mathbf{R}^k$. Then, the Recall@k and RecallAccuracy@k are defined as follows:

$$\texttt{Recall@k} = \frac{\sum_{i=1}^{N} \|y_i == \mathbf{y'}_i\|_\infty}{N} \tag{1}$$

$$\texttt{RecallAccuracy@k} = \frac{\sum_{i=1}^{N} \|y_i == \mathbf{y'}_i\|_0}{N * k} \tag{2}$$

The Recall@k metric evaluates if there exists at least one *accurate* image in the top-k images that were retrieved, whereas the RecallAccuracy@k evaluates the number of accurate images retrieved in the top-K images.

Motivation of our work: Current methods in image retrieval provide little room for human-AI collaboration. Platforms like ElephantBook, a practical, state-of-the-art computer vision system, still require a human-in-the-loop to achieve reliable performance, thus highlighting the importance of human-AI collaboration. Our contributions in this paper leverage CBMs to enable human-AI collaboration in image retrieval tasks. Previous research in CBMs until now focus largely on improving performance by proposing different architectures [Espinosa Zarlenga *et al.*, 2022; Kim *et al.*, 2023; Marconato *et al.*, 2022], modify loss functions [Sheth and Ebrahimi Kahou, 2024], mitigate leakage [Havasi *et al.*, 2022], improve intervenability [Zarlenga *et al.*, 2023; Marcinkevičs *et al.*, 2024] or circumvent requirements of labels [Oikarinen *et al.*, 2023]. In this work, our novelty lies in expanding *intervenability* capabilities of CBMs to tasks like image retrieval, which require capturing *corrected* concepts into the embedding. All the previously mentioned works are orthogonal to the aim of this paper and are complementary in improving the performance of CBMs.

3 Do CBMs Already Work for Image Retrieval?

This section aims to establish the difficulties faced in adopting single-agent (only AI or human) approaches and show how human-AI teams, specifically through CBMs, can potentially resolve some of these issues.

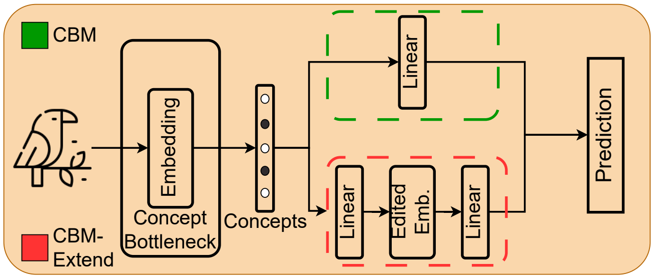


Figure 2: Illustration of the CBM and CBM-Extend, a naive extension of CBM to correct embeddings for retrieval (Edited here refers to capturing human intervention)

Neural networks - irremediable by humans: The key drawback of using neural networks is that when these models make mistakes or inaccurate inferences, their architecture and design allow little room for intermediate human oversight, especially for tasks like image retrieval where the inaccuracies are not obvious until the results are inspected.

Expert coding - costly and high barrier of entry: Handcrafted features and coding of retrieval systems require advanced domain knowledge and experience, which raises the barrier of entry for inexperienced users to utilize such systems. Furthermore, these codes often require entire feature sets to achieve perfect retrieval. Any ambiguity in identifying these codes, for example, due to ambiguity in or occlusion of concept-relevant parts of the image, is marked as *Wildcards* in such systems. Since these codes typically lie on a spectrum of easily identifiable to years of expertise needed, we can leverage neural networks to alleviate some of these difficulties while having a variable level of human oversight and help reduce wildcard entries with a collaborative human-AI retrieval system.

Are CBMs the solution? Potentially: Our contributions stem from the observation that high-level concepts are analogous to the codes developed by domain experts to help humans retrieve relevant images. However, adopting these concepts (in the case of CBMs) or expert codes (in the case of humans) to enable collaborative embedding generation for image retrieval is not supported by existing CBM architectures (see Table 1).

Feature/Model	Standard	CBMs	Ideal
Usable for retrieval	✓	✓	✓
Interpretable	✗	✓	✓
Intervenable for classification	✗	✓	✓
Intervenable for retrieval	✗	✗	✓

Table 1: Ideal Model Features

CBMs typically contain a linear layer as a classifier, and the predictions change as a function of predicted concepts and human interventions. Therefore, a straightforward way of achieving intervenability for retrieval would be to extract the latent embeddings for comparison with other images after the concept predictions by adding another linear layer before predicting the final class. This enables capturing human intervention in the embedding space. However, this extension, termed as `CBM-Extend` (illustrated in Figure 2), performs poorly in image retrieval when compared to the latent embeddings used from standard, CBM and HybridCBM [Mahinpei *et al.*, 2021] embeddings with similar base architectures (ResNet-18). (see Figure 3). The significant drop in performance (Recall@k here, as defined in equation 1) is likely due to the lack of generalization with an increasing number of hidden layers (increased capacity), rendering the naive extensions unusable. While section 5 details what `Recall@k` signifies here, it is sufficient to say that there exists a clear gap in abilities and performance, thus giving a clear picture of how representations generated by CBMs compare to corresponding traditional models (**(RQ1)**).

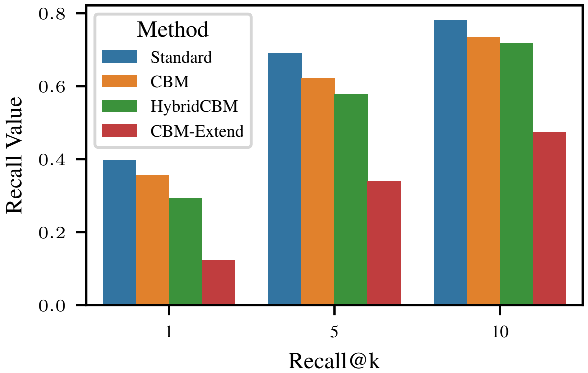


Figure 3: CBM[Koh *et al.*, 2020], HybridCBM[Mahinpei *et al.*, 2021] and Naive CBM extension (Figure 2) have poor retrieval performance when compared to their standard counterpart model

4 Our Proposed Architecture: CHAIR

We have established the importance of enabling human-AI collaboration for tasks beyond classification. Keeping in mind **(RQ2)**, which asks *"How can we augment CBMs to enable human interventions on the representations"* and **(RQ3)** that asks *"How can we enable different levels of expertise for intervention"*, we now present `CHAIR`, a two-stage CBM-like architecture that helps address these questions. This section is organized into two parts: **(a) Architecture** and **(b) Training**.

Architecture: Figure 4 illustrates our proposed `CHAIR` architecture. Let ζ indicate the encoder that generates embeddings $\mathbf{z}$ from the input image $\mathbf{x}$. $\phi(x)$ denotes the concept head that generates the concept vector $\mathbf{c}$ from the input image $\mathbf{x}$ and $\psi(c)$ denotes the classification head that outputs the final class label $\mathbf{y}$. The *Concept Bottleneck* comprises of the encoder and the concept head, that is $\phi(\zeta(\mathbf{x}))$. The final class prediction in a CBM is as follows:

$$\mathbf{y}_{\mathbf{CBM}} = \psi(\phi(\zeta(\mathbf{x}))) \tag{3}$$

We propose a simple two-stage modification to the vanilla CBM architecture that enables *integrating* human interventions to improve retrieval. Specifically, as illustrated in Figure 4, our architecture introduces a **Fusion Head**, whose

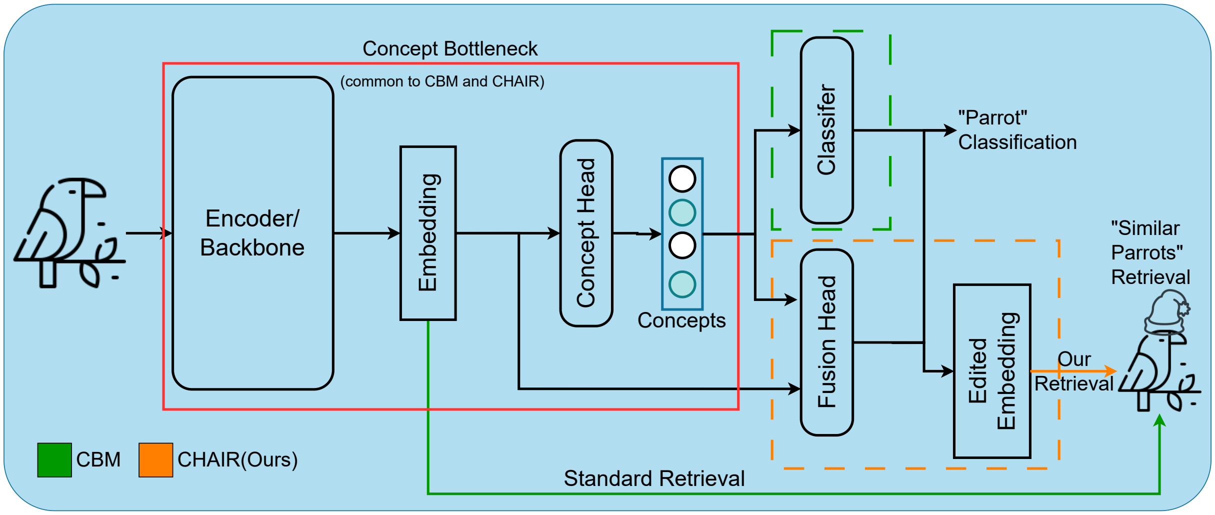


Figure 4: High-level overview of FusionCBM. Our proposed CBM architecture enables *editing* the embeddings using concepts to enable learning better representations and thus improving image retrieval.

output is an additional edited embedding alongside the predicted class. The **Fusion Head** comprises a concept-to-embedding projection layer - a linear layer (ω) that projects the predicted concepts into the same dimensional space as the embedding (z), and a classifier. To incorporate the concepts into the embedding, we add the projection and the previous embedding to give us an *edited* embedding, similar in spirit to a residual connection [He *et al.*, 2016], thus the name Fusion Head. The objective of this *fusion* is to be able to learn a *meaningful* and *better* representation when these concepts are *corrected*, which helps us address **(RQ2)**, that is, enabling concept encoding onto the generated embeddings. Lastly, the classifier (ψ) utilizes this edited embedding for the final classification, thus preserving the same ability as CBMs to incorporate interventions for improving classification.

Training: Given we have an architecture capable of incorporating human intervention in the latent space for image retrieval and classification, we now introduce a two-stage training that learns better embeddings with intervention while maintaining classification accuracy. Note that we adopt the standard classification loss function for image retrieval as used in [Sharif Razavian *et al.*, 2014], which is shown to extend to tasks beyond classification. Furthermore, this helps us understand how our architecture performs against similar models trained on the same loss functions.

Stage 1: From [Sharif Razavian *et al.*, 2014], we know that the embedding z obtained from training on `cross-entropy` like loss functions provides a good baseline embedding to enable image retrieval. Leveraging the fusion head architecture, the learned *edited* embedding encodes information from the already performant embeddings (z) and concepts. Mathematically, the fusion head consists of the $\phi(z)$ to generate concepts $\mathbf{c}$. $\mathbf{c}$ is then projected to the same space as (z) using a simple linear layer $\omega(c)$. Consequently, this projection is added to the original embedding (z) to generate the final edited embedding $\mathbf{z}'$. Our proposed fusion head then utilizes a different classification head $\psi'(z')$ to predict the final class label $\mathbf{y}'$ as follows:

$$\mathbf{y}_{\mathbf{CHAIR}} = \psi'(\zeta(\mathbf{x}) + \omega(\phi(\zeta(\mathbf{x})))) \tag{4}$$

The goal of Stage 1 is to train the concept-to-embedding projection layer ω to enable learning better *edited* embeddings z', which improves classification and image retrieval.

Stage 2: Having now trained the concept-to-embedding projection layer, our model is capable of incorporating corrections in the latent space. Referring back to **(RQ3)**, which aims to include different levels of expertise, we posit that Stage 1 training alone is insufficient to achieve this ability. Hence, we introduce this stage that performs random interventions on a select number of concepts for each mini-batch during training. This random correction of concepts simulates varying levels of expertise while training the *edited* embedding to learn better representations under partial interventions. Note that the concept head here is frozen to preserve the activation values that signify *presence* or *absence* of a concept, similar to [Koh *et al.*, 2020].

Intuition: Revisiting the requirements of a good human-AI collaborative model, the reasoning behind our contribution is three-fold: (a) our architecture helps us retain the same abilities of CBMs to perform classification, (b) our Stage 1 training allows training ω, the concept-to-embedding projection layer, to enable incorporating concepts into the edited embedding (addressing **(RQ2)** and (c) Stage 2 training allows for learning quality embeddings under variable intervention, thus addressing **(RQ3)**.

We adopt the two training modes outlined in [Koh *et al.*, 2020]: `Sequential` (shortened as `Seq` hereafter), where the concept and classification heads are trained separately, and `Joint` where the heads are trained all at once. Note that these training modes mainly differ in `Stage 1` training in the loss function. Algorithms 1 and 2 outline training our proposed `CHAIR` model for both modes. More specifically, *class_loss* and *concept_loss* correspond to the

Algorithm 1 Training `CHAIR` model

```
Input: Training data: $\mathbf{D} = [(x, c, y)]$, mode
Initialize: $\phi, \omega, \psi, \zeta$ =pre-trained,$p_{int}$
Stage-1
    for $(x_i, c_i, y_i)$ in D do
        $z_i \leftarrow \zeta(x_i)$                          ▷ Get Embedding
        $c'_i \leftarrow \phi(z_i)$                          ▷ Get Concepts
        $loss = concept\_loss(c_i, c'_i)$
        if $mode == "sequential"$ then
            $z'_i = z_i + \omega(c'_i)$                      ▷ Get edited embedding
            $y'_i = \psi(z'_i)$
            $loss \mathrel{+}= class\_loss(y_i, y'_i)$
        loss.backward()
Stage-2
    Freeze: $\zeta, \phi$
    if $mode == "sequential"$ then
        reset_weights($\omega$)
    $c_{int} \leftarrow$ intervention_values($\zeta, \phi, D$)
    for $(x_i, c_i, y_i)$ in D do
        $p_i \leftarrow$ torch.rand(1)                       ▷ Partial intervention
        $\hat{c}_i \leftarrow$ concept_intervention($z_i$,$\phi$,$c_i$,$p_i$,$c_{int}$)
        $z"_i \leftarrow \omega(\hat{c}_i) + z_i$
        $y"_i \leftarrow \psi(z"_i)$
        $loss = class\_loss(y_i, y"_i)$
        loss.backward()
return $\zeta, \phi, \omega, \psi$
```

same loss functions utilized in the original CBM training. While we observe the same performance pattern for the `sigmoid` activation function as noted by [Koh *et al.*, 2020], all of the results we report here are from using the `ReLU` activation and `cross-entropy` loss for *class_loss* and individually for each *concept_loss*. Furthermore, the `intervention_values` function enables calculating the activation values for concepts that indicate their *presence* or *absence*. We utilize the training data to calculate these values, similar to [Koh *et al.*, 2020].

Intervention: Intervention in Stage 2 is performed by sampling from a uniform distribution, since we do not assume any level or category of expertise from the humans in the human-AI team. Therefore, all evaluations with interventions also assume *expertise* on a percentage of the available concepts sampled uniformly.

5 Results

5.1 Datasets and Evaluation

We conduct experiments on two real-world datasets (similar to [Espinosa Zarlenga *et al.*, 2022]), the Caltech-UCSD-Birds-200-2011 (CUB) dataset [Wah *et al.*, 2011] and Large-scale CelebFaces Attributes dataset (CelebA) [Liu *et al.*, 2015] to demonstrate the effectiveness of our proposed architecture. Specifically, we utilize the CUB dataset to measure performance in both classification and image retrieval tasks, while the CelebA dataset is used for classification only. The CUB dataset comprises $\mathbf{n} = 11,788$ bird images belonging to a total of 200 possible species. There are 112 binary concepts

Algorithm 2 Stage 2 Intervention functions

```
Function intervention_values($\zeta, \phi, D$):
    Initialize: $c_{int}^{max} = [], c_{int}^{min} = []$
    $x, c, y = D$                                            ▷ Training data
    $z \leftarrow \zeta(x)$
    $c' \leftarrow \phi(z_i)$
    for $c'_i$ in $c$ do
                ▷ activations when concept is present and absent
        $c_{int_i}^{max} \leftarrow$ top-95-percentile(c'i)
        $c_{int_i}^{min} \leftarrow$ bottom-5-percentile(c'i)
    $c_{int} = [c_{int}^{max}, c_{int}^{min}]$
    return $c_{int}$
Function concept_intervention($z$, $\phi$, $c$, $p$, $c_{int}$):
    $c' \leftarrow \phi(z)$
    $idx =$ torch.randperm(c)$[: p * len(c)]$
    $c_{max} = c\_int[0][idx]$                ▷ Simulates partial correction
    $c_{min} = c\_int[1][idx]$
    $c'[idx] =$ torch.where($c[idx] == 1, c_{max}, c'$)
    $c'[idx] =$ torch.where($c[idx] == 0, c_{min}, c'$)
    $\hat{c} \leftarrow c'$
    return $\hat{c}$
```

associated with each image. We follow the same experimental setup for the classification task as detailed in [Koh *et al.*, 2020], where the dataset is divided into training, validation, and testing. Furthermore, we follow the data split established in the literature for image retrieval. The training data here consists of bird images from the first 100 classes. The images from the next 100 unseen classes are then used to create the *gallery*, which is used to retrieve images and measure performance. In contrast, each image in the CelebA dataset consists of 40 concept labels, and we utilize 1000 classes for this classification task. We note that the CelebA dataset was not used in evaluating retrieval, as previous works measure retrieval based on binary concepts, which deviates from the main target application of our work. We use ResNet-18 [He *et al.*, 2016] pre-trained on ImageNet [Deng *et al.*, 2009] for all experiments and train on a single `NVIDIA Tesla V100 16GB` GPU. [1]

5.2 Retrieval Performance

Figure 5 shows how `CHAIR` performs against a standard ResNet-18 model and CBM. A typical CBM with the same underlying ResNet-18 architecture performs slightly worse than the standard ResNet model, which aligns with the behavior noted by [Koh *et al.*, 2020] in classification. Our model provides 15-20 % $Recall@k$ improvement for $k = [1, 5, 10]$, indicating that our model has learned better-unedited embeddings as opposed to the embeddings from the standard and CBM models. Furthermore, we randomly select $p\%$ of concepts for every image in the *gallery* and each image during *query* to investigate how intervention improves retrieval and evaluate the $Recall@k$ performance, as shown in Figure

[1] Code (written in Python using PyTorch [Paszke *et al.*, 2019]) and instructions to reproduce these results can be found here: https://github.com/realize-lab/CHAIR

6. We observe that as the number of interventions (corrections) increases, the retrieval performance improves, eventually reaching near-perfect $Recall@k$. A key observation that emerges from this plot is that when all the predicted concepts are set to their correct values, the method achieves nearly perfect recall when $k = 10$.

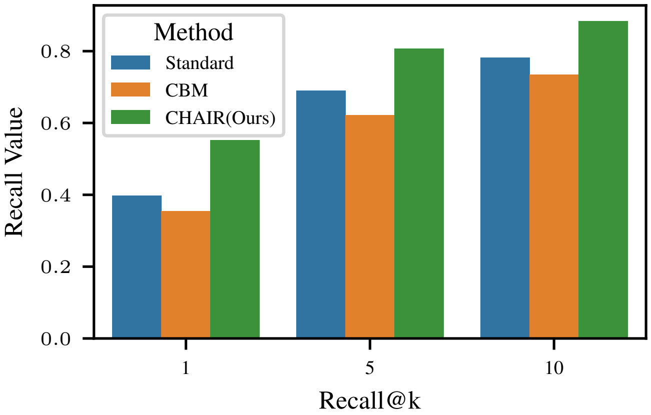


Figure 5: Comparison of the baseline retrieval performance (without any intervention, if possible) of the standard ResNet model, vanilla CBM, and the proposed `CHAIR` model.

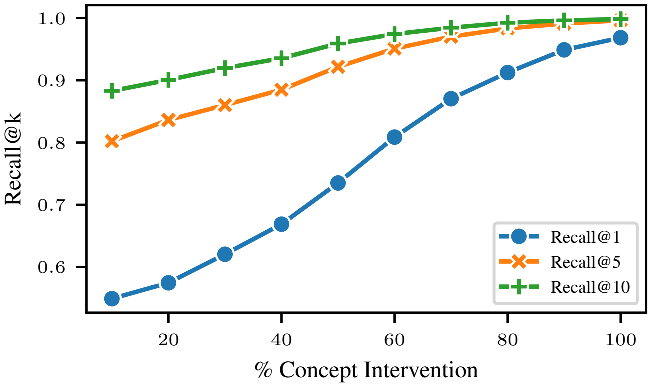
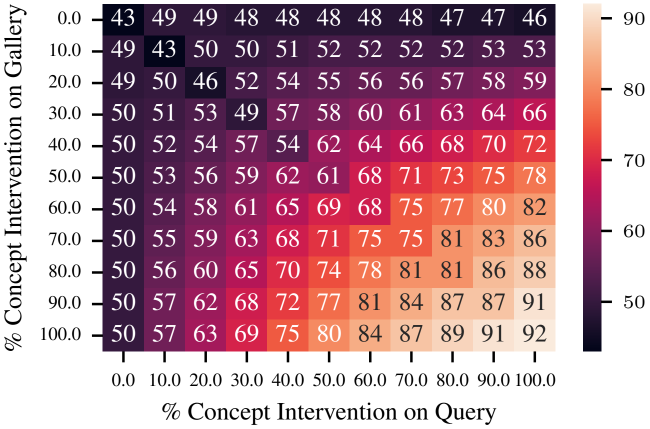


Figure 6: **Intervention vs `Recall@k` for `CHAIR`** *Accurate* interventions on an increasing subset of concepts helps extract better quality representations, thus improving retrieval performance.

In the real world, it is possible that the labelers constructing the gallery embeddings could be less experienced than the users during query time (especially when the gallery is crowdsourced). Hence, we demonstrate how different levels of intervention in the query and gallery images impact the accuracy of our proposed retrieval architecture. Figure 8 shows a heatmap of % interventions on query images and gallery images on the x and y axes, respectively, and the `RecallAccuracy@10` value for each intervention pair. We observe that even when the gallery is constructed with no intervention, any amount of human intervention during query time helps improve retrieval performance. Notably, this trend emerges across all interventions during query time when the interventions of the gallery embeddings are kept fixed. Note that we do not plot a similar heatmap for `Recall@k`, since our method already performs well without any intervention.



% Concept Intervention on Gallery	0.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0
0.0	43	49	49	48	48	48	48	48	47	47	46
10.0	49	43	50	50	51	52	52	52	52	53	53
20.0	49	50	46	52	54	55	56	56	57	58	59
30.0	50	51	53	49	57	58	60	61	63	64	66
40.0	50	52	54	57	54	62	64	66	68	70	72
50.0	50	53	56	59	62	61	68	71	73	75	78
60.0	50	54	58	61	65	69	68	75	77	80	82
70.0	50	55	59	63	68	71	75	75	81	83	86
80.0	50	56	60	65	70	74	78	81	81	86	88
90.0	50	57	62	68	72	77	81	84	87	87	91
100.0	50	57	63	69	75	80	84	87	89	91	92

Figure 8: Heatmap of $RecallAccuracy@10$ (%) for variable levels of intervention in the gallery and query time

5.3 Importance of Stage 2

Recall that Stage 2 was introduced to help improve performance under partial intervention. To measure the impact of Stage 2, we calculate the recall values of both the training modes (`Seq` and `Joint`, as defined in Section 4) perform with and without Stage-2 training for each subset of intervention in Figure 9. This figure clearly shows the benefit acquired when Stage 2 training is employed, especially in the 0-80% range, where we obtain 5%-25% improvement in recall depending on the intervention. Note that the `Seq.` performance here without Stage 2 implies the classification head is trained on predicted concept activation values without any intervention.

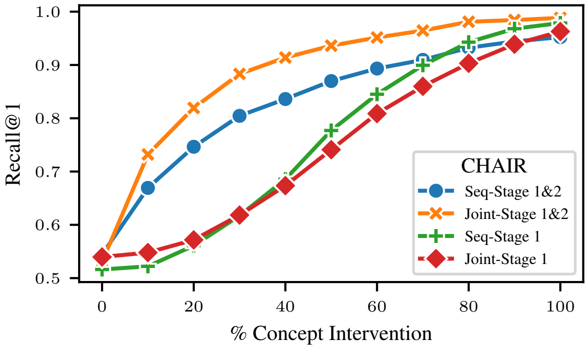


Figure 9: Stage 2 training allows better performance when the intervention is incomplete.

5.4 Quality of Edited Representations

In order to visualize the representations at each intervention level, we use t-Distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique that helps visualize high-dimensional data, which in our case are the embeddings that are used for retrieval. We extract embeddings for each image present in our test data (all the classes in the

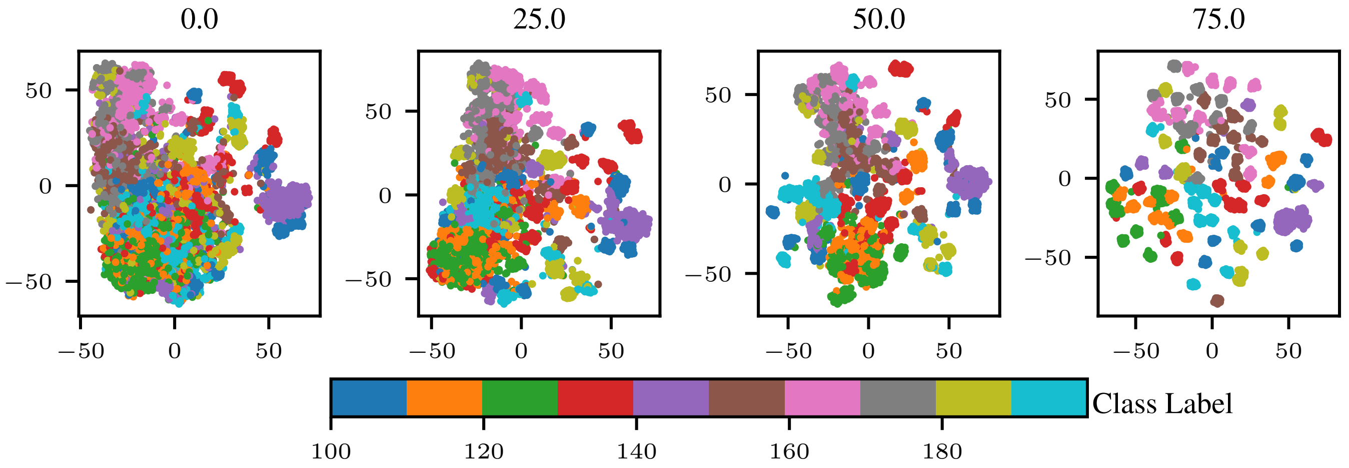


Figure 7: t-SNE visualization of the embeddings generated at increasing levels of intervention (%level indicated at the top of each sub-plot)

test data are unseen during training) and apply t-SNE to reduce them to 2 dimensions. We then plot these reduced embeddings, as shown in Figure 7 for different levels of intervention. With accurate intervention on increasing random subsets of the concepts, the clusters for each class become more distinct. These visual representations also help provide evidence to support how the recall performance under no intervention is much lower when compared to performance under intervention.

5.5 Classification Performance

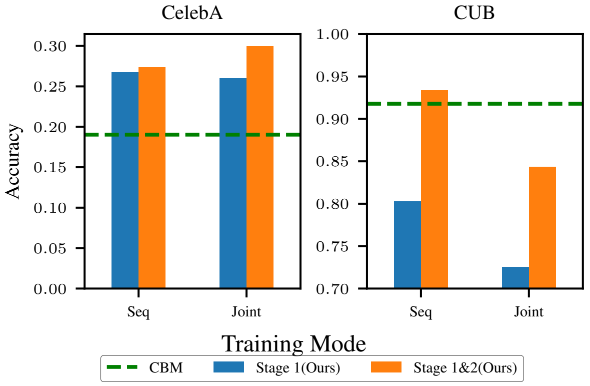


Figure 10: Comparing our proposed model with different stages of training against CBM in classification tasks with maximum intervention. Our model consistently outperforms CBM in classification.

As referred to in **(RQ2)**, our model must be able to match the performance of CBMs in classification and allow intervenable retrieval so that practitioners do not have a tradeoff when selecting the architectures. Figure 10 compares the classification performance of both sequential (*Seq*) and *joint* training modes under perfect concept intervention against the best-performing classifying CBM on both the CUB and the CelebA datasets. It is evident that our proposed CHAIR model outperforms the standard best-performing CBM in these real-world datasets. Furthermore, this plot shows how Stage 2 helps improves classification performance, thus showing its importance beyond improving retrieval under partial intervention.

6 Conclusion and Future Work

In this work, we first establish that CBMs, a model that allows for human-AI collaboration, underperform when compared to standard neural networks on retrieval tasks. We propose CHAIR, a modification of both CBM architecture and training strategy, which can a) incorporate human input in the form of concept correction for image retrieval, b) allow varying levels of human input and expertise, and c) significantly improve retrieval performance while maintaining similar classification performance to vanilla CBMs. Furthermore, we show that the quality of the embeddings generated with some corrections performs better than the alternatives through t-SNE plots and improved retrieval performance. Finally, it is evident that while both the Seq and Joint training modes perform equally well with a high level of intervention, we recommend choosing the joint training mode for image retrieval-related tasks. Our work enables the expansion of CBMs to domains that move beyond classification, further improving human-AI collaboration in impactful applications such as wildlife population monitoring.

We envision future work in this area along multiple frontiers: a) incorporating probabilistic concepts, label prediction, and embeddings to capture the uncertainty of the prediction process better [Kim *et al.*, 2023; Li *et al.*, 2021], b) learning when and how to defer to humans that are interfacing with CBMs to achieve the best possible complementary performance [Bondi *et al.*, 2022; Mozannar and Sontag, 2020], and c) conducting human studies to understand effective strategies of presenting CBM information to achieve complementarity.

Ethical Statement

Our contributions and future work involve including and complementing human efforts. Hence, we strongly advocate rigorously testing with all stakeholders before deploying CHAIR-like models.

Acknowledgements

We want to thank the reviewers for their insightful comments and suggestions. We also thank the e-Health and Artificial Intelligence (e-HAIL) initiative and the Advanced Research Computing (ARC) at the University of Michigan, Ann Arbor, for their support and for providing computing resources and services, respectively.

References

[Beede *et al.*, 2020] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[Beery *et al.*, 2019] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

[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.

[De-Arteaga *et al.*, 2020] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Espinosa Zarlenga *et al.*, 2022] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability tradeoff. *Advances in Neural Information Processing Systems*, 35, 2022.

[Havasi *et al.*, 2022] Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

[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.

[Kim *et al.*, 2023] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[Kulits *et al.*, 2021] Peter Kulits, Jake Wall, Anka Bedetti, Michelle Henley, and Sara Beery. Elephantbook: A semi-automated human-in-the-loop system for elephant re-identification. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 88–98, 2021.

[Li *et al.*, 2021] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[Liu *et al.*, 2020] Yishu Liu, Liwang Ding, Conghui Chen, and Yingbin Liu. Similarity-based unsupervised deep transfer learning for remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7872–7889, 2020.

[Mahinpei *et al.*, 2021] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[Marcinkevičs *et al.*, 2024] Ričards Marcinkevičs, Sonia Laguna, Moritz Vandenhirtz, and Julia E Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv preprint arXiv:2401.13544*, 2024.

[Marconato *et al.*, 2022] Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretable, leak-proof concept-based models. *Advances in Neural Information Processing Systems*, 35:21212–21227, 2022.

[Mendes *et al.*, 2022] Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. Human-in-the-loop evaluation for early misinformation detection: A case study of covid-19 treatments. *arXiv preprint arXiv:2212.09683*, 2022.

[Miao *et al.*, 2021] Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, 2021.

[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.

[Nguyen *et al.*, 2018] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 189–199, 2018.

[Oikarinen *et al.*, 2023] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Peiffer-Smadja *et al.*, 2020] Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595, 2020.

[Rajpurkar *et al.*, 2020] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ digital medicine*, 3(1):115, 2020.

[Shankar *et al.*, 2017] Devashish Shankar, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344*, 2017.

[Sharif Razavian *et al.*, 2014] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[Sheth and Ebrahimi Kahou, 2024] Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Jul 2011.

[Wan *et al.*, 2014] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166, 2014.

[Wang *et al.*, 2020] Zheng Wang, Xin Yuan, Toshihiko Yamasaki, Yutian Lin, Xin Xu, and Wenjun Zeng. Re-identification= retrieval+ verification: Back to essence and forward with a new metric. *arXiv preprint arXiv:2011.11506*, 2020.

[Zarlenga *et al.*, 2023] Mateo Espinosa Zarlenga, Katherine M Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *arXiv preprint arXiv:2309.16928*, 2023.