Enhancing Concept Localization in CLIP-based Concept Bottleneck Models

Anonymous authors
Paper under double-blind review

Abstract

This paper addresses explainable AI (XAI) through the lens of Concept Bottleneck Models (CBMs) that do not require explicit concept annotations, relying instead on concepts extracted using CLIP in a zero-shot manner. We show that CLIP, which is central in these techniques, is prone to concept hallucination—incorrectly predicting the presence or absence of concepts within an image in scenarios used in numerous CBMs, hence undermining the faithfulness of explanations. To mitigate this issue, we introduce Concept Hallucination Inhibition via Localized Interpretability (CHILI), a technique that disentangles image embeddings and localizes pixels corresponding to target concepts. Furthermore, our approach supports the generation of saliency-based explanations that are more interpretable.

1 Introduction

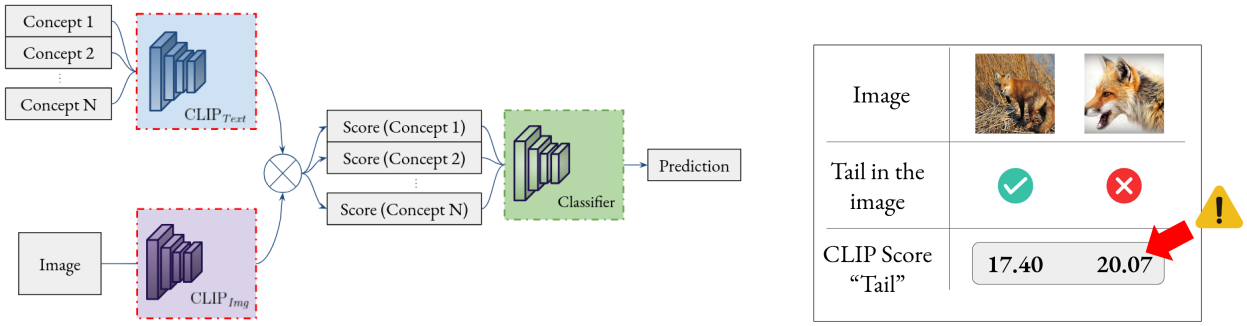
Deep Neural Networks (DNNs) are now used in many areas, including sensitive domains such as medicine and law. In these settings, trust is essential. To build trust, the field of Explainable Artificial Intelligence (XAI) provides tools that help users understand how DNNs make decisions. One important family of methods is *concept-based explanations*. These explanations describe predictions using human-understandable concepts, often expressed as words. For example, a model that classifies an image as a *dog* might rely on concepts such as *fur*, *ears*, *snout*, or *paws*. The ability of a model to represent raw data (e.g., images) as concepts—called *conceptual representation*—is therefore key to creating models that can provide such explanations.

A popular way to use concepts is to embed them directly into the model. This creates an interpretable latent space, where each neuron corresponds to a concept. Models built this way are known as *Concept Bottleneck Models (CBMs)* (Koh et al., 2020; Bennetot et al., 2022). While CBMs improve interpretability by design, they usually require concept annotations during training, which are expensive and difficult to collect.

Recently, contrastive language-image models, such as CLIP (Yan et al., 2023a), have been widely used for tasks like zero-shot classification and open-world recognition. Because CLIP links images and text, researchers have started using it as a free source of concepts for CBMs (Yang et al., 2023; Panousis et al., 2023; Cui et al., 2023). This removes the need for manual annotations, but also introduces a new challenge: the concepts extracted by CLIP may not always reflect what is actually in the image.

A particularly critical challenge for CBMs is the phenomenon of *concept hallucination* (illustrated in Figure 1), wherein concepts are inferred based on contextual cues rather than their actual presence within the image. This issue undermines the foundational hypothesis of CBMs—that the concept bottleneck serves as a faithful conceptual representation of the image content. While prior work has addressed related challenges (Oh & Hwang, 2025; Liu et al., 2024b), our approach distinguishes itself in two key aspects. First, we not only mitigate concept hallucinations but also enhance their localization by explicitly addressing the spatial distribution of activation maps. Second, we extend the applicability of our method to CBMs, thereby offering a tailored solution for improving both the reliability and interpretability of these models.

Our contributions are as follows:



(a) **Principle of a CLIP-based concept bottleneck model.** (b) **CLIP hallucination issue.**

Figure 1: Using the CLIP-score between embedings of input images and predefined concepts, labeling-free concept extraction can be performed, allowing prediction on an interpretable latent space (left). However, CLIP tends to hallucinate the presence of concepts, troubling the localisation of CLIP-based CBMs (right).

- We conduct an extensive statistical analysis to investigate the relationship between the CLIP-score and the localization of concepts. Notably, our findings demonstrate that CLIP-scores fail to accurately represent the actual location of concepts within images.
- Based on this observation, we propose CHILI (Concept Hallucination Inhibition via Localized Interpretability), a method to disentangle the activations of CLIP, and consequently the CLIP-score, distinguishing between object representation, which is related to the physical location of the concept, and contextual representation, which pertains to activations associated with features that do not directly represent the concept but suggest its presence.
- To demonstrate the efficacy of our disentangling method, we employ it as a means to perform image segmentation and binary classification of concepts in spurious situations. We showcase that our method achieves superior results compared to concurrent methods in both tasks.
- We apply CHILI to real-world use cases to construct new, more interpretable CBMs. Our results demonstrate that such an intervention is feasible with only a limited accuracy cost.

2 Related Work

Concept Bottleneck models (CBMs) CBMs constitute a class of models that exploit a conceptual representation of input data, termed the "concept bottleneck," to facilitate inference, thereby enhancing interpretability. While certain studies employ custom datasets featuring concept annotations to construct CBMs (Koh et al., 2020; Díaz-Rodríguez et al., 2022), the emergence of text-image contrastive foundation models has significantly propelled research in this area, enabling the development of CBMs without explicit concept annotations (Yan et al., 2023a; Kazmierczak et al., 2024). Notably, CLIP (Yan et al., 2023a) has become the predominant choice for crafting these CBMs (Kazmierczak et al., 2025).

Neuron interpretation To interpret the behavior of a model post-training, a commonly employed approach involves identifying the role of specific neurons in the process by detecting patterns that induce their activation. Some methods directly display neuron activations in response to designed inputs (Gandelsman et al., 2023). More sophisticated techniques use statistical analysis on a probing dataset to achieve this goal (Shaham et al., 2024; Kalibhat et al., 2023). Alternatively, optimization techniques can be employed to determine the input that maximizes the activation of a given neuron (Olah et al., 2017).

Saliency based explanations To explain image-based decisions, saliency-based explanations—which aim to highlight the most influential regions according to the model—are widely adopted. Among these, model-agnostic approaches such as SHAP Lundberg & Lee (2017), LIME Ribeiro et al. (2016), and RISE Petsiuk

et al. (2018) are particularly popular due to their versatility. These methods analyze model behavior in response to perturbed versions of the input image.

Additionally, the formal structure of deep learning models has enabled alternative approaches for generating saliency maps by directly examining activation patterns Zhou et al. (2016); Gandelsman et al. (2023). Another distinct class of deep neural network (DNN)-based saliency methods leverages gradients to weight activations, as seen in Grad-CAM Selvaraju et al. (2017), FullGrad Srinivas & Fleuret (2019), and HiResCAM Draelos & Carin (2020). However, such gradient-based techniques require differentiable computations, a constraint that does not apply to conceptual representations.

3 Evaluating the concept localization abilities of CLIP

3.1 Preliminaries

First, let us define some notions that we consider essential to describe the experiments we will perform.

Related studies The widespread success of CLIP has spurred significant research effort around its interpretability. Existing studies primarily focus on two issues: bias and spurious feature reliance, and concept hallucination.

The most extensively studied aspect is bias, often examined through image classification tasks. Works such as those by Moayeri et al. (2023b); Zhang et al. (2024) demonstrate accuracy drops on biased datasets, revealing CLIP's reliance on spurious features. Furthermore, Birhane et al. (2021); Hall et al. (2023) show that these biases extend to societal concerns, including gender and racial discrimination. Mitigation approaches include fine-tuning (Alabdulmohsin et al., 2024; Gerych et al., 2024) and activation decomposition (Yeo et al., 2025). Another key challenge is CLIP's tendency to hallucinate text or objects during text-image similarity computations (Oh & Hwang, 2025; Liu et al., 2024b), a phenomenon attributed to the modality gap, where one modality contains more information than the other (Schrodi et al., 2024).

While these studies address general settings, we focus on the specific context of CBMs. This setup presents unique challenges, as concept sets are often highly correlated by design. To our knowledge, the literature lacks a comprehensive evaluation of CLIP's relevance in CBMs, except for the pioneering work by Debole et al. (2025), which assesses the quality of embeddings derived from foundation models. Our study distinguishes itself by addressing concept hallucination in CBMs.

Another underexplored aspect is localization. CBMs implicitly assume that concept representations should not only detect the presence of concepts but also locate them within images. Pre-CLIP CBMs achieved this through backbones trained with localization-aware loss functions (Díaz-Rodríguez et al., 2022; Bennetot et al., 2022). Regarding mitigation methods, Srivastava et al. (2024); Huang & Huang (2024) propose fine-tuning the CLIP backbone to improve localization. The closest work to ours is Yeo et al. (2025), which identifies attention heads responsible for hallucination. However, our method differs in both the identification approach and its application to CBMs.

Class / concept In the context of CBMs, classes refer to the target labels intended for prediction, which are inherently determined by the dataset. Concepts, by contrast, represent a set of interpretable entities—most commonly textual descriptions—that serve as proxies for performing inference. Within CBMs applied to image classification, an image is first represented in terms of these concepts, after which the class prediction is derived from this conceptual representation. In most implementations, concepts correspond to subcomponents of the target label. Two predominant approaches have emerged for defining these concepts. The first one involves prompting large language models: for instance, Yang et al. (2023) extract concepts by querying GPT-3 with prompts such as "describe what the [CLASS NAME] looks like." The second approach leverages dedicated datasets that annotate specific attributes or subparts of the output class present in the image (Díaz-Rodríguez et al., 2022).

3.2 Probing CLIP for Concept Hallucination

To study potential limitations of CLIP-based Concept Bottleneck Models (CBMs), we first need to understand what drives a high CLIP score. In this subsection, we design an experiment to test whether CLIP embeddings reliably reflect the physical presence of concepts in images, or whether they are influenced by contextual or semantic cues.

Datasets We perform experiments on three different datasets: ImageNet (Deng et al., 2009), MonumAI (Lamas et al., 2021), and CUB (Wah et al., 2011).

- **ImageNet:** A large-scale object classification dataset where classes correspond to everyday objects (e.g., *kit fox*), and concepts refer to object parts (e.g., *head*, *tail*, *paw*). Since ImageNet lacks fine-grained part annotations, we extended it with PartImageNet++ (Li et al., 2024).
- **MonumAI:** A dataset focused on monument style classification, where classes are architectural styles and concepts correspond to structural elements such as *arches*, *columns*, or *domes*.
- **CUB:** A fine-grained bird classification dataset, where concepts are visual parts such as *wings*, *beak*, or *tail*.

These datasets cover a wide range of tasks, from generic object recognition to fine-grained classification, making them suitable for evaluating concept detection. A complete list of the concepts used in each dataset is provided in Appendix B.

CLIP Score as a Measure of Concept Detection Given an image I and a text T, CLIP uses an image encoder $M_{\text{img}}(\cdot)$ and a text encoder $M_{\text{text}}(\cdot)$ to project them into a shared embedding space. The similarity between image and text is computed by the cosine similarity:

$$S(I, T) = \langle M_{\text{img}}(I), M_{\text{text}}(T) \rangle \,.$$

This score allows CLIP to match images and text without explicit training on the target dataset, which is why it has become a standard tool for zero-shot classification and concept detection. However, if the score is high even when the concept is absent from the image, it indicates a hallucination problem.

Experimental Setup We construct three subsets of data given two classes c_1 and c_2, and a concept k that is strongly linked to c_1 but absent from c_2:

- Images of class c_1 where concept k is present.
- Images of class c_1 where concept k is absent.
- Images of class c_2, where concept k is always absent by design.

For each triplet (c_1, c_2, k), we sample images randomly and repeat the process 10 times. The full list of triplets is given in the appendix.

The goal of this setup is to test whether CLIP can tell apart the true presence of a concept from its mere semantic association with a class. Concretely, we compute the average CLIP score of each subset with respect to the concept k. We also compute the *failure rate*, defined as the fraction of cases where the subset without the concept receives a higher score than the subset where the concept is actually present.

Results and Analysis Table 1 reports the average CLIP score across all three datasets. The results show that CLIP does not reliably separate the true presence of a concept from its absence. For example, in both MonumAI and CUB, the scores for k-present and k-absent subsets of class c_1 are almost identical, indicating that CLIP relies heavily on class-level associations rather than visual evidence. The high failure rates (40–50%) further confirm that CLIP often assigns higher scores to images without the concept than to those containing it. This demonstrates a significant risk of *concept hallucination*, raising concerns about using CLIP-based embeddings as faithful representations in CBMs.

	c = c_1 *k present*	c = c_1 *k absent*	c = c_2 (*k absent*)	*Fail. Rate*
MonumAI	18.16 ± 2.45	18.09 ± 2.38	16.71 ± 2.75	0.40
CUB	15.58 ± 1.74	15.61 ± 1.65	12.73 ± 2.06	0.50
ImageNet	19.35 ± 1.93	19.18 ± 1.37	14.82 ± 1.80	0.40

Table 1: **Average CLIP score on different setups.** In the first column, k is present in the images. In the second and third ones, k is not present. *Fail. Rate* presents the failure rate, i.e., the fraction of cases where the subset of images that do not possess the desired concept induces a higher score.

4 Disentangling concept representations

4.1 Preliminaries

To address this challenge, we introduce a novel method, CHILI, for disentangling concept localization from concept suggestion within the conceptual representation of images. The primary objective of this approach is to provide users with a conceptual representation that decomposes into distinct components: one related to the object of interest and another one to its surrounding context. By selectively retaining only the object-related component, our method aims to produce a conceptual representation that reduces concept hallucinations.

Notations We now describe in more detail how the image encoder M_{img} of CLIP computes its representations. The encoder is a Vision Transformer (ViT) consisting of L stacked transformer layers, each containing a multi-head self-attention (MSA) block and a multi-layer perceptron (MLP) block. The input image I is first split into a sequence of patches, linearly projected into tokens, and augmented with a special *class token* (denoted by index *cls*). These tokens are processed layer by layer through the transformer. Formally, we denote by $h \in [\![1, H]\!]$: the attention heads, $l \in [\![1, L]\!]$: the transformer layers, $i \in [\![1, N]\!]$: the patch tokens, and Z^l: the residual stream (intermediate representation) at layer l.

The image encoder produces a single vector representation of the image by applying a learned projection P to the final embedding of the class token. In CLIP, this is written as:

$$M_{\text{img}}(I) = P \cdot \left[Z^L\right]_{\text{cls}} ,$$

where $\left[Z^L\right]_{\text{cls}}$ denotes the class token at the final layer.

Unrolling the Transformer. Each layer l of the transformer updates the residual stream by combining the outputs of the MSA and MLP blocks:

$$Z^l = Z^{l-1} + \text{MSA}^l\left(Z^{l-1}\right) + \text{MLP}^l\left(\hat{Z}^l\right) ,$$

where $\hat{Z}^l$ denotes the normalized activations after the attention block.

By expanding this recursion, we can express the final class token representation as a sum of contributions from all layers:

$$M_{\text{img}}(I) = P\left[Z^0\right]_{\text{cls}} + \sum_{l=1}^{L} P\left[\text{MSA}^l\left(Z^{l-1}\right)\right]_{\text{cls}} + \sum_{l=1}^{L} P\left[\text{MLP}^l\left(\hat{Z}^l\right)\right]_{\text{cls}} . \tag{1}$$

Decomposition into Attention Heads. Following the analysis of Elhage et al. (2021); Gandelsman et al. (2023), the image embedding $M_{\text{img}}(I)$ can be written as the sum of contributions from each transformer layer l, each attention head h, and each image token i. Intuitively, let us consider that a transformer layer

has an output linear map (usually denoted W_O^l) that mixes the heads and produces the final MSA vector. Applying W_O^l and then the projection P to the `cls` MSA output gives

$$P\big[\mathrm{MSA}^l(Z^{l-1})\big]_{\mathrm{cls}} = P\Big(W_O^l\big([\mathrm{Head}_{l,1};\ldots;\mathrm{Head}_{l,H}]\big)\Big).$$

For our decomposition it is convenient to view the effect of W_O^l as a linear map applied to each head and then summed. Thus we may write

$$P\big[\mathrm{MSA}^l(Z^{l-1})\big]_{\mathrm{cls}} = \sum_{h=1}^{H}\sum_{i=0}^{N} P\,W_O^{l,h}\big(\alpha_{\mathrm{cls},i}^{l,h}\,v_i^{l,h}\big),$$

where $W_O^{l,h}$ denotes the linear map that extracts the contribution of head h after the usual output projection, $v_i^{l,h} \coloneqq V^{l,h}\big(Z_i^{l-1}\big)$ is the *value* vector produced for token i by head h at layer l, and $\alpha_{\mathrm{cls},i}^{l,h}$ is the attention weight from the class query to token i in head h, layer l.

We now define the vector contribution coming from token i, head h, layer l after all linear projections:

$$m_{i,l,h} \;\coloneqq\; P\,W_O^{l,h}\big(\alpha_{\mathrm{cls},i}^{l,h}\,v_i^{l,h}\big). \tag{2}$$

Each $m_{i,l,h}$ is a vector in the same embedding space as $M_{\mathrm{img}}(I)$. With this definition we can rewrite the sum of all MSA contributions compactly:

$$\sum_{l=1}^{L} P\big[\mathrm{MSA}^l(Z^{l-1})\big]_{\mathrm{cls}} = \sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i=0}^{N} m_{i,l,h}.$$

Using equation 2 and the expansion above, equation 1 becomes

$$M_{\mathrm{img}}(I) = P[Z^0]_{\mathrm{cls}} + \sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i=0}^{N} m_{i,l,h} + \sum_{l=1}^{L} P\big[\mathrm{MLP}^l(\hat{Z}^l)\big]_{\mathrm{cls}}.$$

The first term $P[Z^0]_{\mathrm{cls}}$ is the projected initial class token; the last sum collects the MLP contributions. Since, by definition, the CLIP score is

$$S(I,T) \;=\; \big\langle M_{\mathrm{img}}(I),\, M_{\mathrm{text}}(T)\big\rangle\,,$$

inserting the expression for $M_{\mathrm{img}}(I)$ gives

$$S(I,T) = \big\langle P[Z^0]_{\mathrm{cls}},\, M_{\mathrm{text}}(T)\big\rangle + \sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i=0}^{N}\big\langle m_{i,l,h},\, M_{\mathrm{text}}(T)\big\rangle + \sum_{l=1}^{L}\big\langle P[\mathrm{MLP}^l(\hat{Z}^l)]_{\mathrm{cls}},\, M_{\mathrm{text}}(T)\big\rangle.$$

Let us define

$$A_{i,l,h}(T) \;\coloneqq\; \big\langle m_{i,l,h},\, M_{\mathrm{text}}(T)\big\rangle,$$

which is the scalar alignment between the head/token contribution and the text embedding. If we collect the small terms (initial class-token projection and the MLP outputs) into a residual ϵ, we obtain the compact decomposition proposed by Elhage et al. (2021); Gandelsman et al. (2023):

$$S(I,T) \;=\; \sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i=0}^{N} A_{i,l,h}(T) \;+\; \epsilon. \tag{3}$$

Here we have $\epsilon \;=\; \big\langle P[Z^0]_{\mathrm{cls}},\, M_{\mathrm{text}}(T)\big\rangle + \sum_{l=1}^{L}\big\langle P[\mathrm{MLP}^l(\hat{Z}^l)]_{\mathrm{cls}},\, M_{\mathrm{text}}(T)\big\rangle$.

Notably, we can represent the contribution of a specific MSA head h and layer l to the score as an attention map by grouping terms $A_{l,i,h}$ by position, which we denote by $A_{l,h} = [A_{l,i,h}]_{i=0}^{N}$. When reshaped, $A_{l,h}$ can represent the heatmap illustrating the patch-wise contributions as a tensor of size N. Finally, we denote by A the summed attention map:

$$A = \sum_{l=1}^{L}\sum_{h=1}^{H} A_{l,h}\,.$$

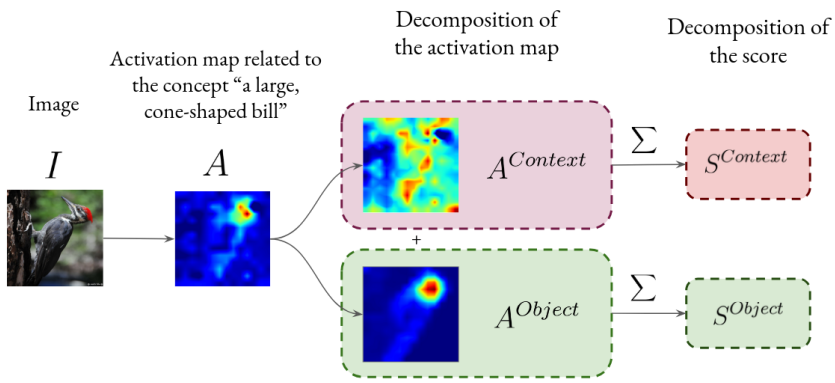


Figure 2: **Decomposition of the activation map**

4.2 Concept Hallucination Inhibition via Localized Interpretability (CHILI) – our method

Our goal is to find a decomposition of $A_{l,h}$ into two terms, respectively representing the activations related to the effective presence of the object in the image, and the activations related to the suggestions of the presence of the concept (see Figure 2):

$$A_{l,h} = A_{l,h}^{\text{Context}} + A_{l,h}^{\text{Object}}. \tag{4}$$

- $A_{l,h}^{\text{Context}}$ is linked with all the content in the image that is not the concept, i.e., locations that do not overlap with the segmentation of the concept.
- $A_{l,h}^{\text{Object}}$ is linked with all the content in the image that is the concept, i.e., locations that do overlap with the segmentation of the concept.

Filtering Pseudo-Register Artifacts The first step of our method is to remove *high-norm tokens*, which are known artifacts of Vision Transformers (ViTs) (Darcet et al., 2024). These tokens act like *pseudo-registers*: they tend to store global information, but their spatial localization on the activation map is not meaningful for interpretability. Following this intuition, we separate this component and denote it as the *pseudo-register part*.

Formally, for each attention map $A_{l,h}$ at layer l and head h, we define the pseudo-register part as the residual after applying a median filter:

$$A_{l,h}^{\text{P. register}} = A_{l,h} - f_m(A_{l,h}),$$

where f_m is a median filter with kernel size 3. This operation removes localized noise while isolating the global, non-informative artifacts.

Remaining activations. After filtering, we assume that the remaining activations in $A_{l,h}$ represent the *spatially-dependent part* of the conceptual representation. In other words:

- some neurons focus on detecting patterns that directly correspond to the object (concept) described by the text T,
- while other neurons detect contextual features that indirectly suggest the presence of T.

Weighting heads and layers. To search for such a decomposition, we assign a weight $w_{l,h}$ to each pair (l, h) according to a score based on the Intersection over Union $\text{IoU}(\cdot,\cdot)$ between a pseudo mask based on activations and ground truth segmentations obtained using a probing dataset:

$$w_{l,h} = \mathop{\mathbb{E}}_{A_{l,h}, G \in \mathcal{D}_{\text{probe}}} \left[1 - e^{-\alpha \ \text{IoU}(h_m(A_{l,h}), G))}\right],$$

with $\mathcal{D}_{\text{probe}}$ the probing dataset containing activations $A_{l,h}$ and ground truth segmentations G that segment the concept, α is a temperature scaling hyperparameter, and

$$h_m(A_{l,h}) = \begin{cases} 1 & \text{if} \quad f_m(A_{l,h}) > \text{mean}(f_m(A_{l,h})), \\ 0 & \text{if} \quad f_m(A_{l,h}) \leq \text{mean}(f_m(A_{l,h})). \end{cases}$$

In our experiments, the probing dataset corresponds to, for each data point of the training set, the selection of a random concept present in the image, and the corresponding mask. Using this weight, we define the following decomposition:

$$\begin{aligned} A_{l,h}^{\text{Object}} &= w_{l,h} f_m(A_{l,h}) \\ A_{l,h}^{\text{Probe}} &= (1 - w_{l,h}) f_m(A_{l,h}) \,. \end{aligned}$$

Decomposition of activations. Once the decomposition is performed for each head and layer, using these weights, we split each activation map into two parts:

$$\begin{aligned} A_{l,h}^{\text{Object}} &= w_{l,h} \cdot f_m(A_{l,h}) \,, \\ A_{l,h}^{\text{Context}} &= (1 - w_{l,h}) \cdot f_m(A_{l,h}) \,. \end{aligned}$$

Here:

- $A_{l,h}^{\text{Object}}$ captures features directly aligned with the concept (object-related),
- $A_{l,h}^{\text{Context}}$ captures features not aligned with the concept, i.e., contextual cues.

Score decomposition. By combining equation 5, equation 4, and the resummations among tokens:

$$\begin{aligned} S^{\text{Object}} &= \sum_{i=0}^{N} A_i^{\text{Object}} \\ S^{\text{Context}} &= \sum_{i=0}^{N} A_i^{\text{Context}} \,, \end{aligned}$$

we can also disentangle the CLIP score into two interpretable contributions:

$$S(I, T) = S^{\text{Object}} + S^{\text{Context}} + \epsilon \,. \tag{5}$$

The activations $A^{\text{Object}} = \sum_{l=1}^{L} \sum_{h=1}^{H} A_{l,h}^{\text{Object}}$ and $A^{\text{Context}} = \sum_{l=1}^{L} \sum_{h=1}^{H} A_{l,h}^{\text{Context}}$ can thus be interpreted as the token-level contributions to the scores S^{Object} and S^{Context}, respectively, for the given image and text.

5 Experiments

5.1 Concept detection

The most straightforward way to evaluate the efficiency of our method is to evaluate it on a binary object detection task. To do so, we use the same setup as the statistical analysis in Section 3.2. From this setup, we compute the AUROC score in the case where the class present in the image is intended to (column 1 vs column 2 of Table 1) using the different components S (refered as Concept Attention)S^{Object} (refered as CHILI) S^{Object} (refered as CHILI in the table), S^{Context}, $S^{\text{P.register}} = \sum_{i=0}^{N} \sum_{l=1}^{L} \sum_{h=1}^{H} A_{i,l,h}^{\text{P.register}}$ from the decomposition of equation 5, and locate-then-correct (LTC) (Yeo et al., 2025). The results are displayed in Table 2.

First, we observe that the baseline—which corresponds to using the raw CLIP score S—struggles to detect the presence of the concept, thereby reinforcing the findings of Section 3. Regarding the disentangled

	Monumai	*ImageNet*	*CUB*
LTC (Yeo et al., 2025)	0.555	0.566	0.532
Concept Attention (Gandelsman et al., 2023)	0.550	0.495	0.485
Register	0.548	0.495	0.487
CHILI (ours)	**0.587**	**0.596**	**0.533**

Table 2: **Performance comparison across datasets.** Results are shown for different methods (LTC, CHILI, Register, and Concept Attention) on three datasets: Monumai, ImageNet, and CUB. Values represent the mean AUC averaged over the different runs.

components, the S^{Context} component also exhibits poor detection performance. In contrast, the S^{Object} component demonstrates a significantly higher detection capability, as intended by design.

The use of the pseudo register component $S^{\text{P.register}}$ showcases similar performances to using the raw CLIP score, indicating that pseudo registers contain non-localised, hallucination-prone information.

5.2 Object segmentation

To test whether our method can localize concepts in images, we adapt it into a segmentation module. The task is to segment both *classes* and *concepts* from ImageNet across the entire test set. In practice, we evaluate how well the activation maps highlight the relevant pixels using three standard metrics:

- **Pixel accuracy (Acc.)** — percentage of correctly classified pixels,
- **mean Intersection over Union (mIoU)** — overlap between prediction and ground truth masks,
- **mean Average Precision (mAP)** — quality of the predicted mask in terms of precision.

We compare our method with several post-hoc interpretability approaches (i.e., without fine-tuning the model): LRP (Binder et al., 2016), partial-LRP (Voita et al., 2019), rollout (Abnar & Zuidema, 2020), raw attention, Grad-CAM (Selvaraju et al., 2017), Chefer et al. (Chefer et al., 2021), and Concept Attention (Gandelsman et al., 2023). Among them, two are natural baselines for the *concept-level segmentation*:

- **Raw attention:** the penultimate layer of the vision transformer ($M_{\text{img}}(I)$),
- **Concept Attention:** the original, non-disentangled activation map A.

Table 3 reports the results. We highlight the scores of our *Object* component (CHILI) A^{Object}.

Analysis. From the class-level results, we observe that our method outperforms all baselines across all three metrics. In particular:

- Compared to **Concept Attention**, our disentangled *Object* map improves pixel accuracy by +2.0 points (76.78 → 78.79), mIoU by +3.1 points (57.14 → 60.22), and mAP by +2.0 points (82.89 → 84.86).
- The improvement over other classical methods such as Grad-CAM (+8.5 mIoU) or rollout (+19.6 mAP) is even more pronounced.

At the concept level, we also see consistent but smaller gains: about +1 point in accuracy, +1.5 in mIoU, and +0.7 in mAP. These results confirm that isolating the *Object* component leads to cleaner and more accurate localization than using the full, entangled activation map. This demonstrates the interest of CHILI.

Remark. From an XAI perspective, simply providing accurate object segmentations is not enough to build a reliable Concept Bottleneck Model (CBM). In the next section, we show how CHILI can be leveraged to construct a CBM that is not only accurate but also trustworthy.

Method	Pixel Acc. ↑	mIoU ↑	mAP ↑
Class-level segmentation			
LRP	52.81	33.57	54.37
partial-LRP	61.49	40.71	72.29
rollout	60.63	40.64	74.47
raw attention	65.67	43.83	76.05
GradCAM	70.27	44.50	70.30
Chefer et al.	69.21	47.47	78.29
Concept Attention	76.78	57.14	82.89
CHILI (ours)	**78.79**	**60.22**	**84.86**
Concept-level segmentation			
Concept Attention	70.86	46.17	87.65
CHILI (ours)	**71.76**	**47.74**	**88.38**

Table 3: **Segmentation performance on ImageNet.** Results for class-level segmentation (top) and concept-level segmentation (bottom). Higher is better.

6 Applying CHILI to CBMs

6.1 Method

In the previous sections, we showed how CHILI allows us to disentangle object-related and context-related activations, leading to more faithful concept extraction. We now turn to the question of how this disentanglement impacts the performance of Concept Bottleneck Models (CBMs).

Baseline. As a baseline, we consider a standard CLIP-based CBM Yan et al. (2023b), which relies directly on the full CLIP similarity score $S(I, T)$. In contrast, our approach replaces this score with the disentangled *object-only* component S^{Object}, introduced in Section 4.

Since S^{Object} is less affected by contextual bias and concept hallucination, we hypothesize that it can yield a more interpretable CBM, albeit at the possible cost of predictive accuracy.

Evaluation. For each dataset, we compare the classification accuracy of the baseline CBM (using S) with the proposed CBM (using S^{Object}). The results are reported in Table 4.

Method	*Monumai*	*ImageNet*	*CUB*
Baseline CBM (S)	74.67	73,55	65.05
CHILI (ours, S^{Object})	74.34	72,80	64.90

Table 4: **Classification accuracy of CBMs with and without CHILI.** Results are shown for three datasets. Baseline CBM uses the full similarity score S, while our approach uses the disentangled score S^{Object}.

Analysis. Our results show that applying CHILI leads to only a minor decrease in accuracy across datasets, suggesting that it can be a viable strategy for improving CBM interpretability without severely compromising performance. The effect depends on the dataset:

- **CUB:** There is almost no accuracy loss, which suggests that object-related signals dominate the decision process in this dataset.
- **Monumai and ImageNet:** A small drop in accuracy occurs, likely because contextual features (e.g., background or environment) play a role in classification. By removing them, the model becomes less biased but also loses some useful cues.

Figure 3: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Orchard Oriole*

Discussion. Overall, these findings highlight a trade-off:

- On the positive side, CHILI reduces bias and mitigates errors caused by spurious correlations, especially in fine-grained datasets such as CUB.
- On the negative side, filtering out contextual information inevitably discards some predictive features, which can slightly reduce accuracy.

Importantly, there is no guarantee that reducing hallucinations will improve accuracy. However, from an XAI perspective, prioritizing faithfulness and interpretability is crucial, and our results suggest that CHILI can achieve this while keeping performance reasonably close to the baseline.

6.2 Explanations

Once the model is trained, we can leverage the activations produced to build visual explanations that gather the name of the most important concept and their location on the image. We explain the process below.

We extract the concept representation using CHILI. We then apply DeepSHAP (Lundberg & Lee, 2017) to the model, with a key distinction: rather than computing SHAP values at the image level, as is conventional for image classification tasks, we perform the computation at the concept level. This approach allows us to quantify the importance of each concept in the inference of the target label. Finally, for the top five concepts identified by DeepSHAP, we visualize their contributions as heatmaps derived from their corresponding activations, A^{Object}. Examples of these explanations are presented in Figure 3 and Appendix C.

7 Limitations and discussion

In this work, we studied the ability of CLIP to focus on patterns located in the object designated by the textual encoding to produce an inference and proposed a way to disentangle the activations of the model without fine-tuning. We want to discuss here the limitations of such a procedure.

Layer/head decomposition We base our approach on a layer/head decomposition to achieve the disentangling. Such an assumption, as noted by Gandelsman et al. (2023), neglects indirect effects, i.e., potential interactions from previous layers on deeper ones. Additionally, we assume that each position in the layer/head pairs can be attributed to pseudo-register, context, or object (or at least can be more easily separated by doing so).

Other factors of a high CLIP score We voluntarily focus on the impact of the object's presence on the increase or decrease of the CLIP score. However, being a complex model, the factors that influence high CLIP scores are multifactorial. For example, the proportion of salient features (Darcet et al., 2024) or the text prompt (Zhou et al., 2022) also influences its output. Such factors are a notable reason why our disentangling does not completely eliminate the failures in the experiments of Section 3.2. For example, the images from the case $c = c_1; k$ *absent* have by construction many more close-ups than the case $c = c_1; k$ *present*, inducing perturbation towards higher scores. It is also evident that CLIP suffers from numerous biases (Moayeri et al., 2023b) that can influence the score in either a decreasing or increasing manner.

Role of pseudo registers One aspect of our disentangling is the presence of high-norm artifacts (Darcet et al., 2024), which we refer to as pseudo-registers. The reason we dedicated a special part to them in our decomposition is that their role is ambiguous: since they do not seem to exhibit spatial coherence with the image, it is difficult to determine whether they store information about the object or the concept.

8 Conclusion

In this paper, we examined the limitations of using CLIP as a concept extractor. Through statistical analysis, we identified challenges associated with correlating high scores with the localization of concepts in images, particularly in cases where the presence of a concept is merely suggested. To address this issue, we propose a method that factorizes the embedding space into components related to the object, the context, and pseudo-registers. Empirical results demonstrate that our disentangling approach can partially eliminate the contextual aspects of conceptual representation thereby advancing towards more localization-focused CLIP-based concept bottleneck models. In addition, while a probing dataset is required to compute the calibration performed in our method, our approach does not require any additional training of the model.

However, several limitations are present in our study. Primarily, we were unable to achieve complete disentanglement of the activations. This shortcoming can be attributed to multiple hypotheses we adopted in our experimental setup. Notably, we neglected second-order effects and assumed that attention heads are not polysemantic, an assumption that is somewhat reductive.

In this paper, we limit ourselves to CLIP based on ViT backbones. This restriction is motivated by the fact that this paradigm has become the gold standard for most zero-shot CBMs released in recent years (Yang et al., 2023; Yan et al., 2023a; Kazmierczak et al., 2024). However, we plan to extend these analyses to other overlooked backbones, such as those based on ResNet, and models, such as LLaVa. The goal is twofold: first, this opens the way to post-hoc disentangling on other CBMs; secondly, it allows us to study the similarities and differences across various embedding networks, potentially leading to more interpretable zero-shot CBMs.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*, 2024.

Adrien Bennetot, Gianni Franchi, Javier Del Ser, Raja Chatila, and Natalia Diaz-Rodriguez. Greybox xai: A neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowledge-Based Systems*, 258:109947, 2022.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37: 84298–84328, 2024.

Yequan Bie, Luyang Luo, Zhixuan Chen, and Hao Chen. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 773–783. Springer, 2024.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pp. 63–71, 2016.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander Toshev, Jonathon Shlens, Ruoming Pang, and Yinfei Yang. Stair: Learning sparse text and image representation in grounded tokens. *arXiv preprint arXiv:2301.13081*, 2023.

Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *European Conference on Computer Vision*, pp. 183–201. Springer, 2024.

Jihye Choi, Jayaram Raghuram, Yixuan Li, Suman Banerjee, and Somesh Jha. Adaptive concept bottleneck for foundation models. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Yan Cui, Shuhong Liu, Liuzhuozheng Li, and Zhiyuan Yuan. Ceir: Concept-based explainable image representation learning. *arXiv preprint arXiv:2312.10747*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

Nicola Debole, Pietro Barbiero, Francesco Giannini, Andrea Passerini, Stefano Teso, and Emanuele Marconato. If concept bottlenecks are the question, are foundation models the answer? *arXiv preprint arXiv:2504.19774*, 2025.

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*, pp. 248–255, 2009.

Natalia Díaz-Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco Herrera. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, 79:58–83, 2022.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Jessica Echterhoff, An Yan, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. Driving through the concept gridlock: Unraveling explainability bottlenecks in automated driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7346–7355, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Javier Fumanal-Idocin, Javier Andreu-Perez, Oscar Cordón, Hani Hagras, and Humberto Bustince. Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *IEEE Transactions on Fuzzy Systems*, 32(4):1915–1926, 2023.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502, 2024.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023.

Lijie Hu, Songning Lai, Wenshuo Chen, Hongru Xiao, Hongbin Lin, Lu Yu, Jingfeng Zhang, and Di Wang. Towards multi-dimensional explanation alignment for medical classification. *Advances in Neural Information Processing Systems*, 37:129640–129671, 2024.

Xinyi Huang and Long-Kai Huang. Model editing for clip with unknown spurious correlations in visual encoder. *openreview*, 2024.

Neha Kalibhat, Shweta Bhardwaj, C. Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *International Conference on Machine Learning*, pp. 15623–15638, 2023.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. CLIP-QDA: An explainable concept bottleneck model. *Transactions on Machine Learning Research Journal*, 2024.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. Explainability and vision foundation models: A survey. *Information Fusion*, 122:103184, 2025.

Injae Kim, Jongha Kim, Joonmyung Choi, and Hyunwoo J Kim. Concept bottleneck with visual concept filtering for explainable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 225–233. Springer, 2023.

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*, pp. 5338–5348, 2020.

Alberto Lamas, Siham Tabik, Policarpo Cruz, Rosana Montes, Álvaro Martinez-Sevilla, Teresa Cruz, and Francisco Herrera. Monumai: Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification. *Neurocomputing*, 420:266–280, 2021.

Xiao Li, Yining Liu, Na Dong, Sitian Qin, and Xiaolin Hu. Partimagenet++ dataset: Scaling up part-based models for robust recognition. In *European Conference on Computer Vision*, pp. 396–414, 2024.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, Yang Feng, and Zuozhu Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis. *arXiv preprint arXiv:2307.01981*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. *arXiv preprint arXiv:2410.03176*, 2024b.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, pp. 25037–25060. PMLR, 2023a.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriosity rankings: sorting data to measure and mitigate biases. *Advances in Neural Information Processing Systems*, 36:41572–41600, 2023b.

Thomas Norrenbrock, Marco Rudolph, and Bodo Rosenhahn. Q-senn: Quantized self-explaining neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21482–21491, 2024.

Hongseok Oh and Wonseok Hwang. Vision-encoders (already) know what they see: Mitigating object hallucination via simple fine-grained clipscore. *arXiv preprint arXiv:2502.20034*, 2025.

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

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Konstantinos P Panousis, Dino Ienco, and Diego Marcos. Coarse-to-fine concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:105171–105199, 2024.

Konstantinos Panagiotis Panousis, Dino Ienco, and Diego Marcos. Sparse linear concept discovery models. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 2767–2771, 2023.

Cristiano Patrício, Luis F Teixeira, and João C Neves. Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In *2024 IEEE international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. Chatgpt-powered hierarchical comparisons for image classification. *Advances in neural information processing systems*, 36:69706–69718, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. *arXiv preprint arXiv:2404.07983*, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

Divyansh Srivastava, Ge Yan, and Lily Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094, 2024.

Abhinav Kumar Thakur, Filip Ilievski, Hông-Ân Sandlin, Zhivar Sourati, Luca Luceri, Riccardo Tommasini, and Alain Mermoud. Multimodal and explainable internet meme classification. *arXiv preprint arXiv:2212.05612*, 2022.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia Vogt. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:51787–51810, 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pp. 142–160. Springer, 2024.

Yifan Wu, Yang Liu, Yue Yang, Michael S Yao, Wenli Yang, Xuehui Shi, Lihong Yang, Dongjun Li, Yueming Liu, Shiyi Yin, et al. A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. *Nature communications*, 16(1):3504, 2025.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*, 2023.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3090–3100, 2023a.

An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023b.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19187–19197, 2023.

Wei Jie Yeo, Rui Mao, Moloud Abdar, Erik Cambria, and Ranjan Satapathy. Debiasing clip: Interpreting and correcting bias in attention heads. *arXiv preprint arXiv:2505.17425*, 2025.

J Yow, Neha Priyadarshini Garg, Manoj Ramanathan, Wei Tech Ang, et al. Extract–explainable trajectory corrections from language inputs using textual description of features. *arXiv preprint arXiv:2401.03701*, 2024.

Miao Zhang, Ben Colman, Ali Shahriyari, Gaurav Bharaj, et al. Common-sense bias discovery and mitigation for classification tasks. *arXiv preprint arXiv:2401.13213*, 2024.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022.

A Use of CLIP in CBMs

Table 5 presents the foundation models used in the CBMs highlighted in the study of Kazmierczak et al. (2025). Note that the list is not exhaustive.

Title	PFM used
STAIR (Chen et al., 2023)	CLIP
Chat GPT XAI (Liu et al., 2023)	CLIP
ARTxAI (Fumanal-Idocin et al., 2023)	CLIP
Explanable meme classification (Thakur et al., 2022)	CLIP
Label free CBM (Oikarinen et al., 2023)	CLIP
LaBo (Yang et al., 2023)	CLIP
Learning Concise (Yan et al., 2023a)	CLIP
Sparse CBM (Panousis et al., 2023)	CLIP
CBM with filtering (Kim et al., 2023)	CLIP
Robust CBM (Yan et al., 2023b)	CLIP
Hierarchichal CBM (Panousis et al., 2024)	CLIP
ChatGPT CBM (Ren et al., 2023)	CLIP
Skin lesion CBM (Patrício et al., 2024)	CLIP
R-VLM (Xu et al., 2023)	CLIP
CEIR (Cui et al., 2023)	CLIP
Concept Gridlock (Echterhoff et al., 2024)	CLIP
SpLiCE (Bhalla et al., 2024)	CLIP
MMCBM (Wu et al., 2025)	CLIP
XCoOp (Bie et al., 2024)	CLIP
CLIP-QDA (Kazmierczak et al., 2024)	CLIP
Text-To-Concept (Moayeri et al., 2023a)	CLIP
LLM-Mutate (Chiquier et al., 2024)	Llama2+CLIP
VAMOS (Wang et al., 2024)	BLIP-2
Q-SENN (Norrenbrock et al., 2024)	CLIP
ExTraCT (Yow et al., 2024)	CLIP+BERT
Adaptative CBM (Choi et al., 2024)	CLIP
Stochastic CBM (Vandenhirtz et al., 2024)	CLIP
Med-MICN (Hu et al., 2024)	CLIP
VLG-CBM (Srivastava et al., 2024)	CLIP

Table 5: **PFM usage in CBMs.**

B Datasets

ImageNet The first dataset we use is ImageNet (Deng et al., 2009), that provides annotation of images into 1000 classes. The dataset having not concept-level annotations natively, we used the PartImageNet++ dataset (Li et al., 2024), which provides semantic segmentation annotations for different images of ImageNet. The scenarios used in our study are detailed in Table 6.

Scenario	c_1	c_2	k
1	Tiger_cat	Gondola	tail
2	Bolete	Stole	lamellae
3	LesserPanda	BlackSwan	paw
4	ModelT	Turnstile	wheel
5	Plunger	CommonIguana	handle
6	AnalogClock	Goldfish	dial
7	Fly	Strawberry	wing
8	Barracouta	Barbell	fin
9	ComputerKeyboard	Convertible	key
10	FountainPen	HowlerMonkey	ink_cartridge

Table 6: **List of ImageNet runs with respective triplets classes c_1, c_2, and concepts k.**

Monumai Monumai (Lamas et al., 2021) is a specialized dataset containing images of monuments. It is composed of 908 images. Each image is annotated accordingly to the overall structure that corresponds to the class, and the architectural features that corresponds to the concepts. There are 15 concepts and 4 classes available. The scenarios used in our study are detailed in Table 7.

Scenario	c_1	c_2	k
1	hispanic-muslim	baroque	lobed_arch
2	baroque	renaissance	porthole
3	baroque	gothic	broken_pediment
4	baroque	renaissance	solomonic_column
5	gothic	hispanic-muslim	pointed_arch
6	renaissance	baroque	serliana
7	gothic	baroque	trefoil_arch
8	baroque	renaissance	rounded_arch
9	gothic	renaissance	ogee_arch

Table 7: **List of Monumai runs with respective triplets classes c_1, c_2, and concepts k.**

CUB CUB (Wah et al., 2011), is a dataset dedicated to the classification of birds, with 200 classes corresponding to species. Concept level, localised annotations are also not available natively. To obtained such annotation, we used the procedure of VLG-CBM (Srivastava et al., 2024) that uses GroundingDino (Liu et al., 2024a) to localize concepts as bounding boxes. The scenarios used in our study are detailed in Table 8.

Scenario	c_1	c_2	k
1	Orchard Oriole	Least Auklet	long tail
2	Red headed Woodpecker	Bay breasted Warbler	long pointed beak
3	Worm eating Warbler	Chuck will Widow	yellowish belly
4	Whip poor Will	Rock Wren	brown or grayish body
5	House Sparrow	Belted Kingfisher	brown streaks on the chest
6	Herring Gull	Worm eating Warbler	black wingtips
7	Ring billed Gull	Red breasted Merganser	white body with gray wings
8	Red bellied Woodpecker	Red breasted Merganser	white front
9	Golden winged Warbler	Geococcyx	white belly
10	Pied Kingfisher	Vermilion Flycatcher	black back

Table 8: **List of CUB runs with respective triplets classes c_1, c_2, and concepts k.**

C Additional samples

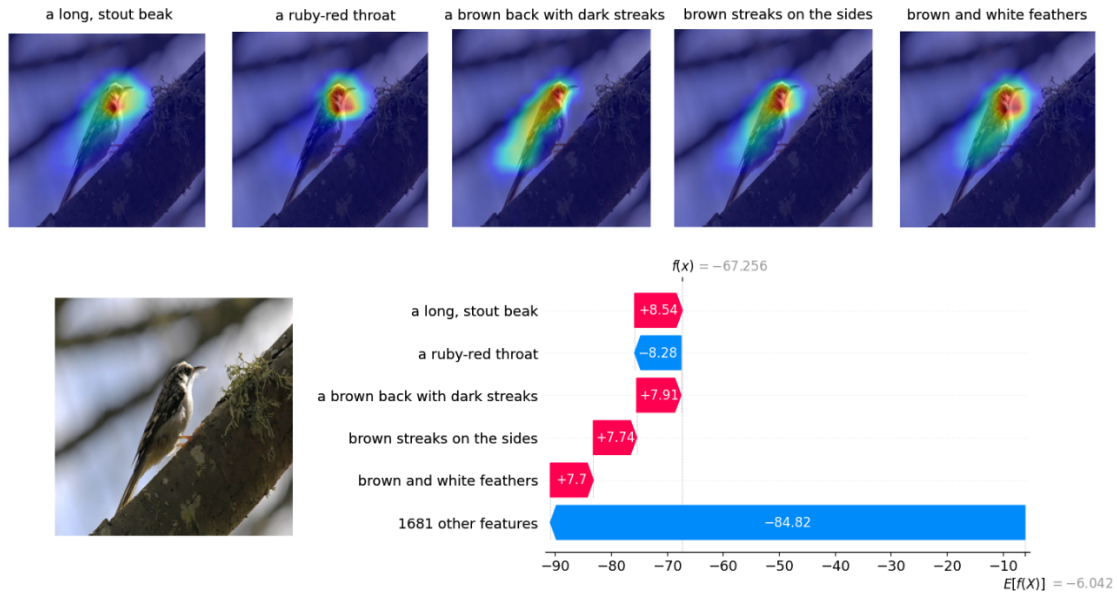


Figure 4: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Brown Creeper*

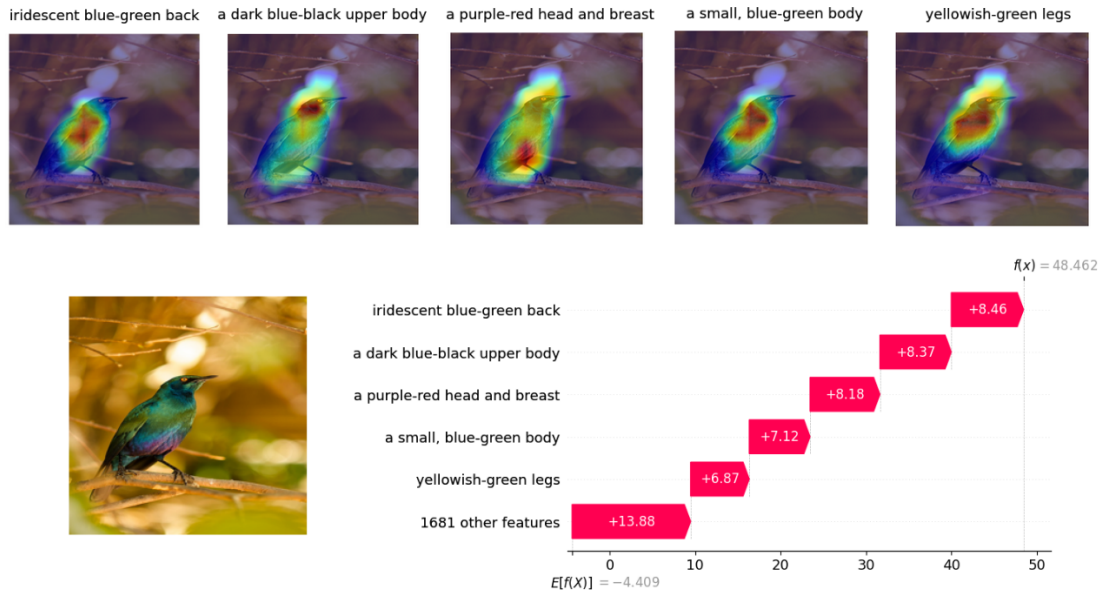


Figure 5: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Cape Glossy Starling*

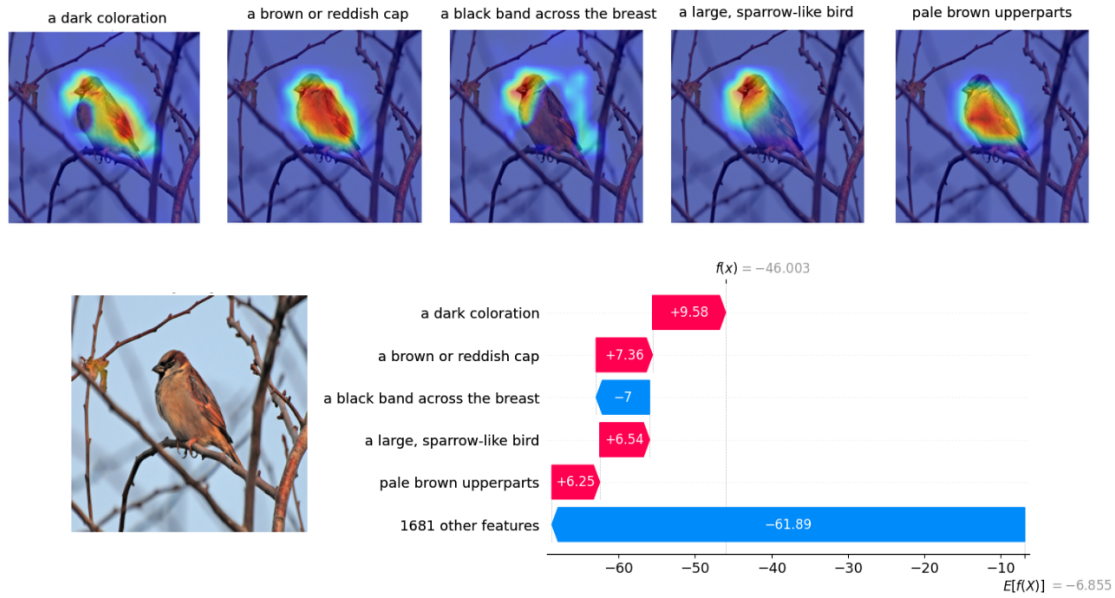


Figure 6: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *House Sparrow*

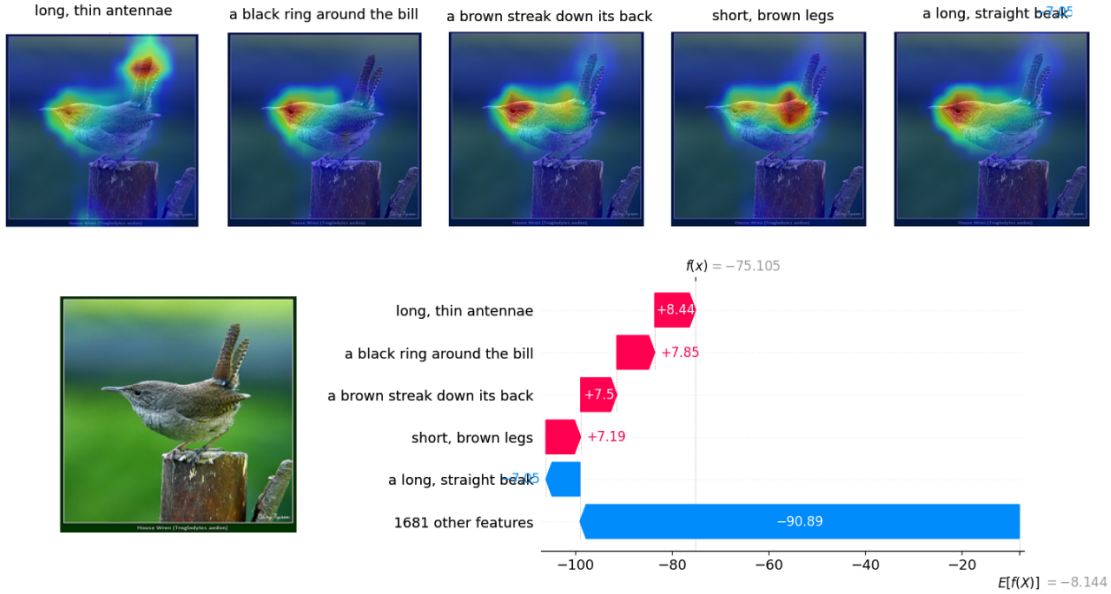


Figure 7: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *House Wren*

Figure 8: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Red Bellied Woodpecker*

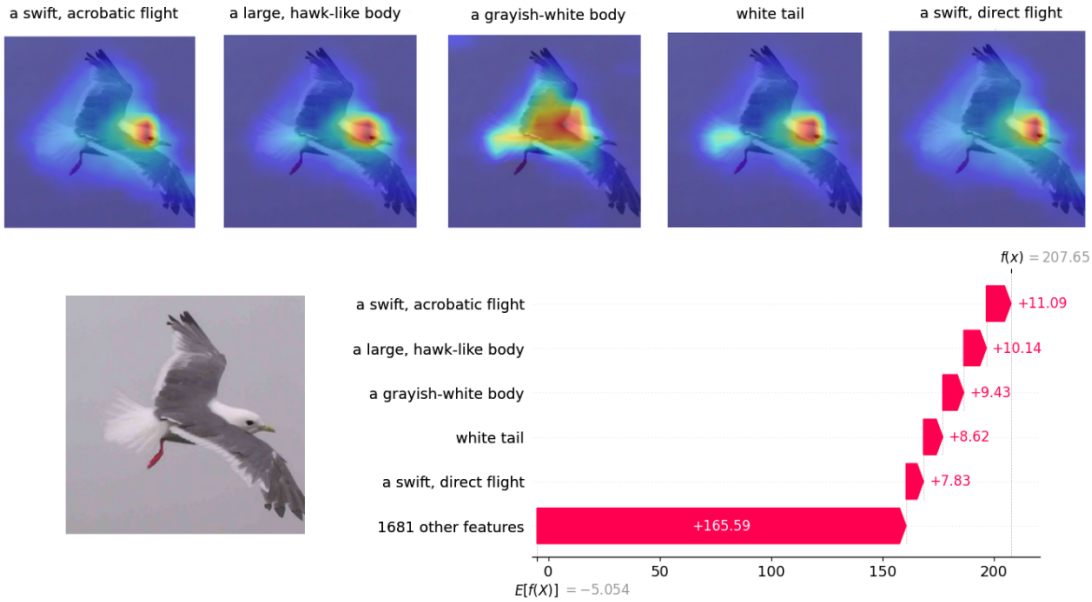


Figure 9: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Red Legged Kittiwake*

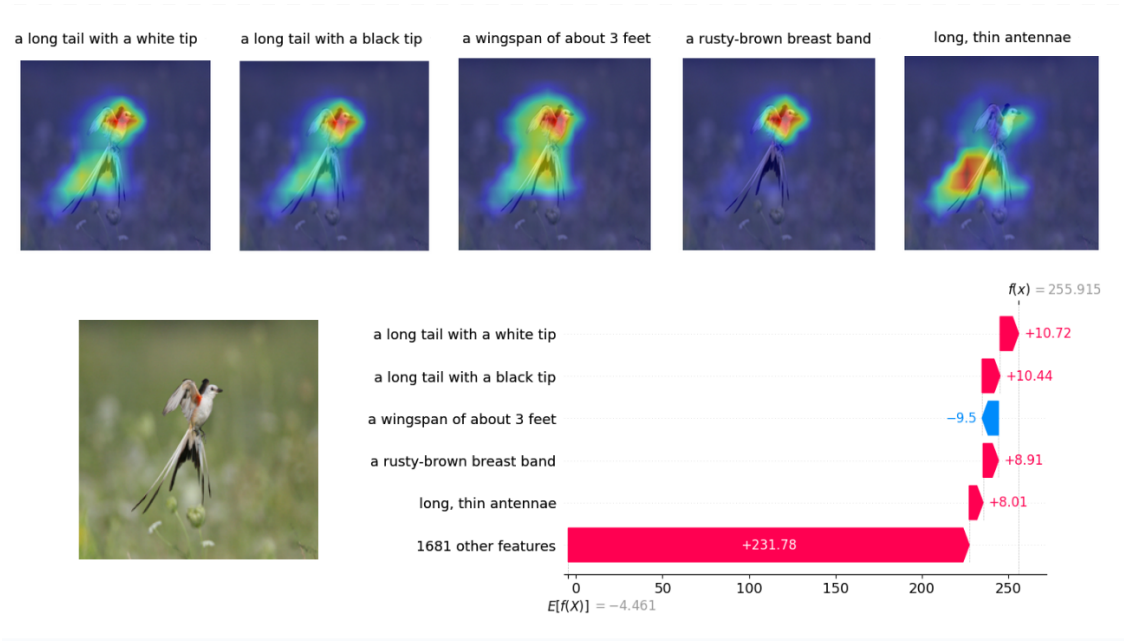


Figure 10: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Scissor Tailed Flycatcher*

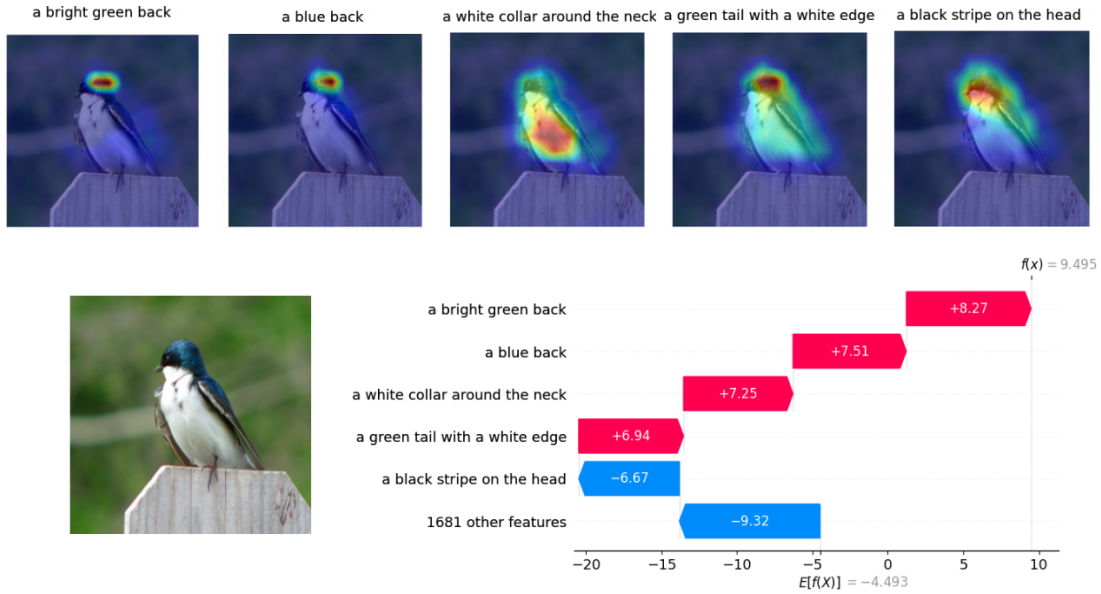


Figure 11: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Tree Swallow*

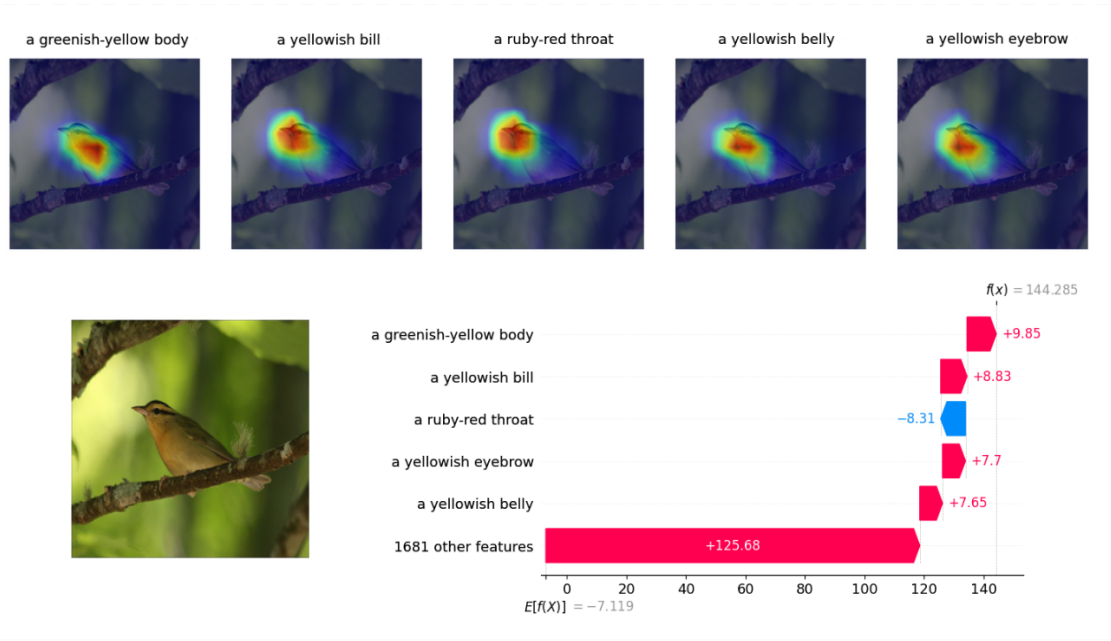


Figure 12: **Example of explanation produced by the intervention of CHILI in a CBM.** On the bottom left, the input image. On the bottom right, the SHAP values. Target label: *Worm Eating Warbler*