On the Faithfulness of Post-Hoc Concept Bottleneck Models

Laines Schmalwasser[1,2], Jan Blunk[2,3⋆], Niklas Penzel[2], Julia Niebling[1], and Joachim Denzler[2]

[1] Institute of Data Science, German Aerospace Center, Jena, Germany
[2] Computer Vision Group Jena, Friedrich Schiller University Jena, Germany
[3] GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany
✉ Correspondence to laines.schmalwasser@dlr.de

Abstract. Human decision-making interprets the world through high-level concepts, such as recognizing a bird by its belly color. To bridge the gap between opaque deep learning representations and human understanding, Post-Hoc Concept Bottleneck Models (post-hoc CBMs) project latent features onto interpretable concept spaces using auxiliary datasets or vision-language models. However, relying on target task accuracy as the primary measure of post-hoc CBM success obscures whether the learned concepts are semantically meaningful or merely predictive artifacts. For example, random concept projections can achieve competitive accuracy despite being semantically meaningless. In this work, we analyze the learned projections directly and identify two failure cases: First, for concept projections learned from auxiliary data, covariate shifts can lead to unfaithful concept representations for the target task. In particular, we provide an upper bound on the error introduced by this shift. Second, systematic label noise in surrogate concept labels generated by vision-language models leads to unfaithful projections. After formalizing these failure modes, we introduce novel metrics that decouple concept faithfulness from predictive accuracy. Our empirical results across real-world and synthetic benchmarks confirm that these metrics identify unfaithful behaviors that standard accuracy-based evaluation fails to detect[4].

Keywords: Concept Bottleneck Models · Faithfulness · Interpretability

1 Introduction

Human understanding is fundamentally built on concepts [5, 16, 42]. Rather than analyzing raw sensory data, we identify objects through high-level attributes, such as recognizing a specific bird species by its "yellow belly" or "black crown." In contrast, deep neural networks learn opaque, high-dimensional representations that often lack semantic clarity [6,29,33]. To bridge this gap and enforce interpretability, Concept Bottleneck Models (CBMs) were introduced [1, 2, 10, 15, 26, 29, 41, 61, 62].

⋆ Work done while the author was at Computer Vision Group Jena.
[4] Project page: https://posthoc-cbm-faithfulness.github.io/

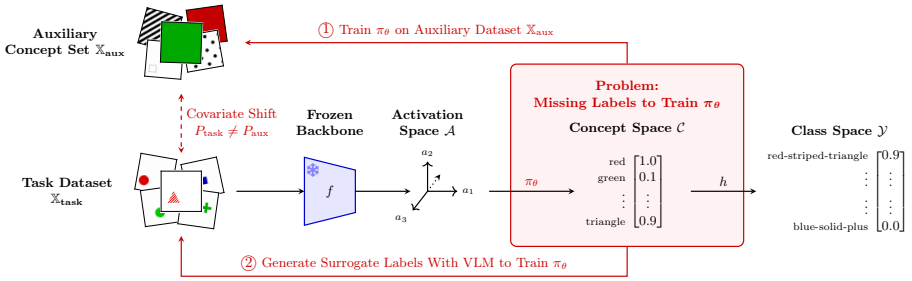


Fig. 1: In the post-hoc CBM setup [44, 57, 64], inputs are mapped to activations using some frozen backbone model f. Then, a learned concept projection π_θ extracts concepts before a final classifier h predicts a class based on these concepts. Unfortunately, concept labels in the downstream task domain are often unavailable. Thus, ① auxiliary concept sets (e.g., Broden [6]) or ② surrogate VLM labels are used [44, 57, 64]. We study these approaches and identify two reasons for potential unfaithfulness of π_θ: ① a covariate shift of the concept set, and ② systematic surrogate label errors.

CBMs constrain the model to first predict a set of human-understandable concepts from the input, and then base the final prediction solely on them, thereby promising transparent decision-making [23, 36, 38].

However, standard CBMs require dense concept annotations for the specific downstream task, which are rarely available in practice [13, 29, 44]. Post-Hoc Concept Bottleneck Models (post-hoc CBMs) circumvent this by freezing a pretrained feature extractor and learning only a lightweight projection from its internal representations to a concept space and final classifier [44, 57, 64] (we use post-hoc CBM to denote this general model class and PCBM for the concrete implementation in [64]). Crucially, post-hoc CBMs tackle the need for in-domain concept labels via two strategies: (1) *Auxiliary Concept Datasets*, where the projection is learned on a separate, richly annotated dataset (e.g., Broden [6]) before application to the target task [64]; or (2) *Surrogate Labels from VLMs*, where a Vision-Language Model (e.g., CLIP [49]) generates surrogate concept labels [44, 45, 51, 57]. Both drastically reduce the annotation burden while maintaining competitive downstream classification accuracy [44, 64].

Unfortunately, evaluating downstream accuracy provides limited insight into the quality of the learned concept projection. Prior work empirically shows that random concepts achieve high performance [40]. In [37], this phenomenon is connected to the Johnson-Lindenstrauss (JL) lemma [24], with the expected error approaching zero as random projections increase [57]. We expand these insights via the smooth manifold JL lemma [4], based on previous findings that neural activations tend to lie on low-dimensional manifolds [3, 48]. We show that under certain assumptions, the original activations can be reconstructed from a modest number of random projections, resulting in high downstream performance.

Consequently, we must investigate the concept projection itself to evaluate whether it faithfully extracts the intended concepts on the target distribution. Here, we analyze the two primary post-hoc CBM training mechanisms (Figure 1).

First, for auxiliary concept datasets, we identify *covariate shift* as the primary source of unfaithfulness. We demonstrate that even if a concept's semantic definition is identical across domains, a geometric shift in the feature space can invalidate the learned projection on the downstream task. We formalize this by providing an upper bound for the error introduced by this shift based on [7] and propose an approximation of it as a practical measure of faithfulness. Second, for methods using VLM-based surrogate labels, the primary problem is *label noise*. However, our analysis shows that unfaithfulness is not solely caused by noise magnitude, but by systematic errors. In other words, random mistakes can cancel each other out, but the surrogate labeling function leads to unfaithfulness when mistakes are consistent, e.g., always predicting "sky" and "clouds" together. To detect this, we measure the surrogate-label error and propose a metric that explicitly quantifies the corresponding systematicity.

We empirically validate our theoretical findings on standard benchmarks (CUB-200 [63] and CIFAR-10/100 [30]) and the synthetic Elements [43] dataset, which provides known ground-truth concepts. Our evaluation includes a range of post-hoc CBMs across standard backbones representing the two main training paradigms: those using auxiliary datasets [64] (e.g., Broden [6]) and those using surrogate labels generated by VLMs (LFCBM [44] and VLG-CBM [57] based on CLIP [49], DINOv3 [55] and Grounding DINO [34]). We first show that random concept projections can achieve competitive accuracy, confirming that downstream accuracy is an insufficient metric for meaningful representations. Then, we directly evaluate the learned concept projections using available ground truth [43], apply our proposed faithfulness metrics, and demonstrate that they successfully identify failures caused by covariate shifts and systematic errors, offering a practical tool for evaluating post-hoc CBMs beyond predictive performance.

Our contributions can be summarized as follows:

1. We show that under certain assumptions, random concept projections can achieve competitive performance.
2. We identify and formalize two critical sources of unfaithfulness in the concept projections of post-hoc CBMs: *covariate shifts* when learning with auxiliary concept sets and *systematic label noise* in surrogate supervision.
3. We propose a set of novel faithfulness metrics that evaluate concept alignment independently of the predictive performance, revealing specific failure cases.

2 Related Work

Post-hoc CBM evaluations are complicated by information leakage [18, 36, 38, 53, 59], where models bypass semantic concepts to base predictions on raw backbone activations. Notably, [53] links this phenomenon to the Johnson-Lindenstrauss lemma [24], showing that sufficient random projections preserve pairwise distances for finite sets. We note that the manifold Johnson-Lindenstrauss lemma [4] implies this holds for the complete activation manifold under the manifold hypothesis [3, 48]. Related to this, [57] studies the expected error of random projections for final predictions. Based on these findings, we emphasize the need

to evaluate learned concept projections directly. Crucially, the faithfulness issues we identify persist even in the absence of information leakage.

Other works similarly study CBM and post-hoc CBM limitations [13, 22, 35, 50] and concept faithfulness [31, 32]. While [31] analyzes the faithfulness of unsupervised concept explanations, it explicitly excludes post-hoc CBMs (our focus). For surrogate labels, [32] highlights flaws in GPT-3-derived concept sets [9] for label-free CBMs [44]. We generalize this by providing an alternative mechanism to analyze surrogate-label errors in VLM-based post-hoc CBMs. In concept embedding models [13], the concept bottleneck is constructed from two learned embeddings per concept to address the accuracy-interpretability trade-off. In contrast to our evaluation metrics, [22] studies CBM faithfulness by matching input regions to concepts. Finally, [50] empirically observes issues caused by distribution shifts in concept datasets. We provide an upper bound on the risk of post-hoc CBMs trained on auxiliary data, approximating the resulting generalization error. In contrast, [35] bounds the CBM risk on downstream tasks using the underlying backbone.

3 Method

We consider a standard deep learning setting where a fixed backbone $f : \mathcal{X} \to \mathcal{A} \subseteq \mathbb{R}^d$ maps inputs to a high-dimensional latent activation space. While a classifier $g : \mathcal{A} \to \mathcal{Y}$ trained on top of f may achieve high performance, its decision-making process within the high-dimensional space $\mathcal{A}$ remains opaque. Post-Hoc Concept Bottleneck Models (Post-Hoc CBMs) [44, 57, 64] aim to make this process interpretable by introducing an intermediate concept bottleneck while holding f fixed. Specifically, they learn a concept projection $\pi_\theta : \mathcal{A} \to \mathcal{C}$, parameterized by θ, that maps activations to a K-dimensional space $\mathcal{C} \subseteq \mathbb{R}^K$ of human-understandable concepts (typically $K \ll d$). A subsequent classifier $h : \mathcal{C} \to \mathcal{Y}$ (e.g., a sparse linear layer) predicts target classes based on these concept scores, yielding the final model $h \circ \pi_\theta \circ f$. To construct the bottleneck layer, we initially assume access to a training set $\mathbb{X}_{\text{task}} = \{x_i, c_i\}_{i=1}^N$ sampled i.i.d. from a task distribution $P_{\text{task}}(\mathcal{X}, \mathcal{C})$ over inputs and concept vectors, allowing us to optimize θ to recover these task-relevant concepts.

Definition 1. *We define a concept projection π_θ as* faithful *if it minimizes the expected risk over the true task distribution $P_{\text{task}}(x, c)$:*

$$\arg\min_\theta \mathbb{E}_{(x,c)\sim P_{\text{task}}} \left[\mathcal{L}(\pi_\theta(f(x)), c)\right], \tag{1}$$

for some loss function $\mathcal{L}$ measuring the alignment of the predicted concepts $\pi_\theta(f(x))$ and the ground truth concepts c for an input x.

In this work, we study the post-hoc concept bottleneck framework from this perspective of concept faithfulness, i.e., whether π_θ extracts meaningful concepts after Definition 1. A faithful projection ensures $\pi_\theta(f(x)) = \mathbb{E}_{P_{\text{task}}}[c \mid x]$, recovering the intended concepts in expectation. Note that our formalization of

faithfulness differs from similar notions in information leakage, e.g., [37, 38, 53], and aligns more closely with the informal version in [13].

First, we confirm that evaluating the predictive performance of h is insufficient to assess the faithfulness of the post-hoc CBM in Section 3.1. After establishing that downstream accuracy does not imply a faithful concept projection, we formalize the learning problem of π_θ in Section 3.2 to derive direct faithfulness measures. We identify two primary sources of unfaithfulness: (1) *covariate shifts* from auxiliary training datasets, which we quantify via an upper bound on the generalization error; and (2) *systematic label noise* from VLM-generated surrogate concept labels. For surrogate labels, we show that unfaithfulness requires non-random error structures rather than magnitude alone, and derive a metric to explicitly quantify this systematicity.

3.1 Downstream Performance and Faithfulness

To demonstrate that high post-hoc CBM performance does not imply semantic alignment, we study random projections π_θ without semantic meaning, e.g., independently sampled standard-normal coefficients in a linear layer. Previous work empirically shows that PCBMs [40] using such π_θ can achieve high downstream performance. In [37], the authors note that such projections are geometry-preserving with high probability for finite sets of samples under the Johnson-Lindenstrauss (JL) lemma [24] if the embedding dimension K is large enough. In other words, the complete backbone information is encoded into the concept activation, which is related to the area of information leakage [18, 37, 38, 53].

Here, we first link this to the manifold hypothesis, then derive a concrete mechanism to achieve high downstream performance under additional assumptions. In particular, consider the manifold hypothesis of deep learning [3, 48], i.e., that the activations in neural networks lie on or close to a manifold with an intrinsic dimension of $m \ll d$. If this hypothesis holds, then the manifold version of the JL lemma (Theorem 3.1 in [4]) governs the behavior under random projections. Crucially, it establishes a connection between certain characteristics of this activation manifold and the number of random projections needed to preserve pairwise distances. We make this explicit in Appendix A.1. Nevertheless, if geometry is preserved, the original activations can be recovered [4]. Thus, for a sufficiently complex classifier h (e.g., an MLP [21]) it is possible to learn $g \circ \pi_\theta^{-1}$. In such a case, the meaningless concepts by the random π_θ would be used to recover the original activations, leading to high downstream performance. To make this argument more explicit, we next derive a concrete solution for h under an additional linearity assumption.

Constructive Proof for a Linear Activation Subspace. While the discussion under the manifold hypothesis posits a possible solution for h, we now derive a constructive solution under the assumption that the activation manifold is a linear subspace of $\mathbb{R}^d$. Our concrete assumptions are:

(1) **Linear Subspace:** The activations lie on an m-dimensional subspace $\mathcal{S} \subset \mathbb{R}^d$ with orthonormal basis U, such that $f(x) = a = Uz$ for some latent $z \in \mathbb{R}^m$.

(2) **Linear Backbone:** The backbone classifier g is linear prior to a softmax, i.e., $g(a) = \text{softmax}(W_g a)$. This reflects the standard post-hoc CBM setup of inserting the concept bottleneck at the penultimate layer, e.g., [29, 50, 64].
(3) **Random Projection:** The projection is defined as $\pi_\theta(a) = \sigma(Pa)$, where $P \in \mathbb{R}^{K \times d}$ is a fixed Gaussian matrix ($K \geq m$) and σ is bijective (e.g., tanh).

Under these assumptions, we can construct a classifier h that exactly recovers the original prediction $g(f(x))$:

$$h(c) := g(a_{\text{rec}}(c)) = \text{softmax}(W_g \cdot [\underbrace{U(PU)^\dagger \sigma^{-1}(c)}_{\text{Reconstructed } a \in \mathcal{A}}]), \tag{2}$$

where $a_{\text{rec}} : \mathcal{C} \to \mathcal{S}$ reconstructs the activations $a = a_{\text{rec}}(\pi_\theta(a))$ and † denotes the Moore-Penrose pseudoinverse [47]. We provide the full derivation in Appendix A.2.

Consequences. Under the discussed assumptions, h does not need to learn a decision based on semantic concepts. Instead, a degenerate solution is to reconstruct the original activations a before applying g. This aligns with [57], who demonstrated that random concept projections achieve zero expected error as the concept dimension K approaches the activation dimension d.

Our analysis adds to this result in two aspects: First, under the manifold hypothesis ($m \ll d$) [3, 48], the manifold JL lemma [4] implies that K need not approach d to enable recovery. Second, beyond bounding the expected error [57], our constructive derivation under additional linearity assumption using [47] demonstrates that exact, sample-wise reconstruction may be possible.

Consequently, if h is optimized rather than fixed, downstream accuracy becomes a poor proxy for concept faithfulness after Definition 1. This necessitates a direct analysis of the learned projection π_θ, which we formalize next.

3.2 Faithfulness of the Learned Concept Projection

Assuming access to a training set $\mathbb{X}_{\text{task}} = \{x_i, c_i\}_{i=1}^N$ for π_θ sampled i.i.d. from the task distribution P_{task}, we minimize the expected risk $J_{\text{task}}(\theta; \mathcal{L})$:

$$J_{\text{task}}(\theta; \mathcal{L}) = \mathbb{E}_{(x,c)\sim P_{\text{task}}} [\mathcal{L}(\pi_\theta(f(x)),\, c)] \tag{3}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\pi_\theta(f(x_i)),\, c_i). \tag{4}$$

To interpret this optimization as a Maximum Likelihood Estimation (MLE), we assume that the projection π_θ defines the parameters of a conditional distribution $p_\theta(\cdot \mid f(x))$ over the concept space. The empirical risk minimization becomes equivalent to MLE if the loss function $\mathcal{L}$ matches the negative log-likelihood of observing the true concept c under this model distribution [8, 17]:

$$\mathcal{L}(\pi_\theta(f(x)),\, c) = -\log p_\theta(c \mid f(x)) + \text{const.} \tag{5}$$

Consequently, the learned projection π_θ represents the optimal estimator for the concept scores distributed according to $P_{\text{task}}(x, c) = P_{\text{task}}(x)P_{\text{task}}(c|x)$. It minimizes Equation (3) and is, therefore, faithful according to Definition 1.

In practice, we identify two key problems that violate this theoretically optimal scenario: First, the dataset used to learn the concept projection may differ from the task we are interested in solving, e.g., [28,52,64]. Intuitively, using an auxiliary concept dataset rather than a task-relevant $\mathbb{X}_{\text{task}}$ leads to covariate shift, i.e., $P_{\text{aux}}(x) \neq P_{\text{task}}(x)$. Second, to address this, prior work aims to assign task-relevant labels to in-domain training data, leading to surrogate labels $\tilde{c}$ with $P_{\text{task}}(\tilde{c}|x) \neq P_{\text{task}}(c|x)$, e.g., [44,57]. Both scenarios are distributional mismatches (infracting Definition 1) while fitting π_θ, which we will detail next.

Covariate Shift of Concept Datasets. Training the projection π_θ requires access to a training set annotated with ground-truth concepts. Because these are rarely available for the downstream task data $\mathbb{X}_{\text{task}}$, we typically train π_θ by minimizing the empirical risk on a smaller auxiliary dataset $\mathbb{X}_{\text{aux}}$ (e.g., Broden [6]) sampled from a distribution P_{aux} [28,64].

However, this introduces the risk of distribution shift ($P_{\text{task}} \neq P_{\text{aux}}$). Intuitively, images of a concept may appear visually different in $\mathbb{X}_{\text{aux}}$ than in $\mathbb{X}_{\text{task}}$, degrading the faithfulness of the learned concept projection after Definition 1. In particular, the optimization objective in Equation (3) changes to minimizing a corresponding J_{aux}. To quantify this effect, we utilize an upper bound for the target generalization error from domain adaptation theory (Theorem 2 in [7]):

$$J_{\text{task}}(\theta; L_1) \leq J_{\text{aux}}(\theta; L_1) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{X}_{\text{task}},\mathbb{X}_{\text{aux}}) + \Omega(\dim_{\text{VC}},N,\delta) + \lambda_{\text{ideal}}. \quad (6)$$

Here, the expected L_1 (absolute) task risk $J_{\text{task}}(\theta; L_1)$ is bounded by the expected auxiliary risk $J_{\text{aux}}(\theta; L_1)$ plus the scaled empirical $\mathcal{H}\Delta\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ [7,27] between the marginal distributions. The term Ω accounts for finite sample estimation and depends on the VC dimension $\dim_{\text{VC}}$, sample size N (assuming $|\mathbb{X}_{\text{aux}}| \approx |\mathbb{X}_{\text{task}}|$), and the desired confidence level δ. λ_{ideal} represents the task's adaptability (the combined error of the optimal hypothesis on both distributions).

In concept probing, we assume this shift is primarily a *Covariate Shift* [54,58], meaning the semantic definition of the concept $P_{\text{aux}}(c|x) = P_{\text{task}}(c|x)$ remains constant while the input marginals $P_{\text{aux}}(x) \neq P_{\text{task}}(x)$ change. Under this assumption, there exists a single optimal decision rule shared across distributions, rendering λ_{ideal} negligible (full formal definitions and derivations in Appendix A.3).

Consequences. When training on in-domain data is infeasible, we can estimate the faithfulness of π_θ after Definition 1 on a given auxiliary dataset via this upper bound (Equation (6)). Assuming a fixed π_θ architecture and equal dataset sizes (rendering Ω constant), and with λ_{ideal} negligible, the empirical divergence $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ becomes the decisive metric for unfaithfulness. Following [7], we approximate $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ by training a domain discriminator to separate $\mathbb{X}_{\text{aux}}$ and $\mathbb{X}_{\text{task}}$. Intuitively, if a simple classifier can distinguish the auxiliary and task data, we expect a larger absolute task error $J_{\text{task}}(\theta; L_1)$, violating Definition 1.

Faithfulness of Surrogate Label Functions. To avoid potential covariate shifts and the resulting generalization penalty created by training the projection π_θ on a separate auxiliary dataset, it would be ideal to train π_θ directly on the in-domain data $\mathbb{X}_{\text{task}}$ sampled from P_{task} using an oracle concept function $c^* : \mathcal{X} \to \mathcal{C}$ with $\forall (x, c) \in \mathbb{X}_{\text{task}} : c^*(x) = c$. However, because these ground-truth annotations are commonly unavailable in post-hoc settings, we must rely on a surrogate labeling function $\tilde{c} : \mathcal{X} \to \mathcal{C}$ (e.g., derived from VLMs [44, 51, 57]; see Appendix A.4 for examples) to identify the concepts.

Consequently, the optimization objective shifts to minimizing the empirical risk with respect to the surrogate labels:

$$\tilde{J}_{\text{task}}(\theta; \mathcal{L}) = \mathbb{E}_{x \sim P_{\text{task}}} \left[\mathcal{L} \left(\pi_\theta(f(x)),\, \tilde{c}(x) \right) \right]. \tag{7}$$

The bottleneck's faithfulness thus depends on whether the parameters $\tilde{\theta}$ optimized for the surrogate $\tilde{c}$ also minimize the true risk with respect to c^*.

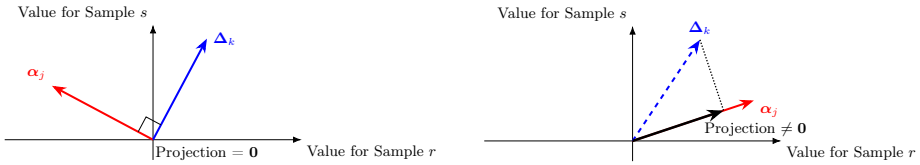
Theoretical Analysis. We analyze this condition by formulating π_θ as a *Multivariate Generalized Linear Model* (GLM) [14, 39]. We assume the conditional distribution of the concepts belongs to the *Multivariate Exponential Dispersion Family (EDF)* [25, 39], which encompasses standard loss settings such as Mean Squared Error (Gaussian) and Cross-Entropy (Bernoulli/Multinomial). In Appendix A.5, we derive the gradient of the *true* objective $J^*_{\text{task}}(\theta; \mathcal{L})$ at the *surrogate* optimum $\tilde{\theta}$. For canonical link functions, this gradient is proportional to the outer product ($\otimes$) of the surrogate error and backbone activations $f(x)$:

$$\nabla J^*_{\text{task}}(\tilde{\theta}; \mathcal{L}) = \mathbb{E}_{P_{\text{task}}}\big[\nabla \mathcal{L}(\pi_{\tilde{\theta}}(f(x)), c^*(x))\big] \propto \mathbb{E}_{P_{\text{task}}}\big[\underbrace{(\tilde{c}(x) - c^*(x))}_{\text{Surrogate Error } \delta(x)} \otimes f(x) \big]. \tag{8}$$

For a faithful post-hoc CBM (i.e., $\tilde{\theta}$ is a global minimum of the objective, see Definition 1), this gradient must be zero. This reveals two distinct mechanisms for faithfulness of π_θ trained to emulate the surrogate label function $\tilde{c}$:

(1) **Surrogate Accuracy:** The surrogate labels have to be accurate, i.e., $\tilde{c}(x) \approx c^*(x)$, causing the surrogate error term $\delta(x) = \tilde{c}(x) - c^*(x)$ to vanish.
(2) **Error Orthogonality:** Further, Equation (8) implies that a model can remain faithful even with noisy surrogate labels if the surrogate error is *orthogonal* to the activation space in expectation. This requires the errors in the surrogate labels to be uncorrelated with the activations $f(x)$.

Geometric Interpretation. Figure 2 visualizes the orthogonality condition. We consider N training samples yielding activation vectors $\boldsymbol{\alpha}_j = [f_j(x_1), \ldots, f_j(x_N)]^\top$ for backbone neuron j, and surrogate discrepancy vectors $\boldsymbol{\Delta}_k = [\delta_k(x_1), \ldots, \delta_k(x_N)]^\top$ for concept k. Equation (8) shows that faithfulness does not strictly require perfect labels ($\boldsymbol{\Delta}_k = \mathbf{0}$). Instead, it is sufficient if the errors $\boldsymbol{\Delta}_k$ are orthogonal to the activations $\boldsymbol{\alpha}_j$ ($\langle \boldsymbol{\Delta}_k, \boldsymbol{\alpha}_j \rangle \approx 0$). In other words, random, unsystematic annotation



(a) Faithful Model (Orthogonal Case) **(b)** Unfaithful Model (Non-Orthogonal Case)

Fig. 2: Visualization of the error orthogonality condition in a 2D sample space. The axes represent two samples r and s from the training set of π_θ. (a) A faithful model is learned if the surrogate label errors of these samples ($\boldsymbol{\Delta}_k$) are orthogonal to their activations ($\boldsymbol{\alpha}_j$), resulting in a zero projection. (b) Non-orthogonal errors have a non-zero projection onto the feature vector, resulting in a non-zero gradient for the true objective.

noise in the surrogate labels can cancel out during training. However, *systematic* errors (such as a VLM consistently predicting the concept "Boat" whenever it detects a "Water" texture, even if no boat is present), create non-orthogonal error components. This results in a non-zero gradient for the true objective, causing the learned $\pi_{\tilde{\theta}}$ to diverge from the ground-truth semantics, violating Definition 1.

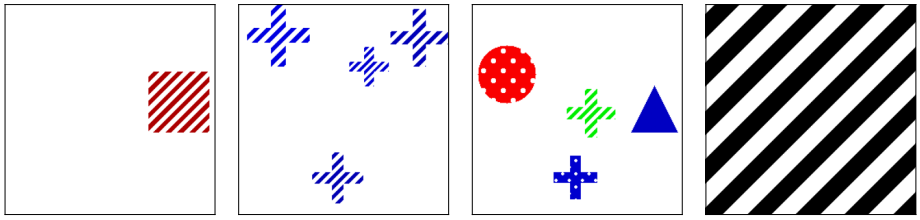
Consequences. We therefore propose two metrics to evaluate the faithfulness of surrogate-based post-hoc CBMs. First, *Surrogate Accuracy* directly measures the raw discrepancy $\delta(x)$. Second, we evaluate the *Alignment of Surrogate Errors and Activations* using the absolute Pearson correlation [46] $\rho_{k,j} = |\text{corr}(\boldsymbol{\Delta}_k, \boldsymbol{\alpha}_j)|$. A correlation of $\rho \approx 0$ indicates orthogonal noise. Conversely, if $|\rho|$ is large and significant (e.g., $p < 0.05$), the surrogate noise is systematically tied to the backbone features. Thus, a post-hoc CBM is unfaithful only if surrogate labels are inaccurate *and* those errors systematically correlate with backbone features.

4 Experiments

To validate our theoretical findings, we evaluate the faithfulness of concept projections π_θ trained using various post-hoc CBM methods. First, we demonstrate that downstream performance is insufficient to assess concept bottleneck layers by studying random concept projections. Next, we analyze the learned π_θ directly, leveraging a synthetic dataset with ground-truth concept information. This allows us to measure the concept accuracy and the geometric alignment of the concept bottleneck approaches. Using an optimal ground-truth classifier for the downstream task, we can align these results with downstream performance. Finally, we evaluate our two theoretically identified sources of reduced faithfulness: concept-set covariate shift and systematic surrogate labeling errors.

4.1 Experiment Setup

Datasets and Backbone Models. We validate our theoretical findings on standard real-world benchmarks: CUB-200 [63] using a ResNet18 backbone [19, 60] and



(a) Example of P_{task} **(b)** Example of $P_{\text{near}}^{\forall}$ **(c)** Example of $P_{\text{near}}^{\exists}$ **(d)** Example of P_{OOD}

Fig. 3: Samples of the concept "striped" for each visual set used. In Figure 3a, there is exactly one object with the concept. Figure 3b shows a sample from the backbone domain where all objects share the concept. Figure 3c shows a sample from the training domain and Figure 3d shows a complete out-of-domain example.

CIFAR-10/100 [30] using CLIP ResNet-50 (RN50) [11] as backbone. Additionally, we consider the synthetic *Elements* dataset [43], which includes ground-truth concept annotations.

Elements samples consist of objects defined by three latent attributes: shape, color, and texture (e.g., "red striped square"). To create a realistic shift between backbone pretraining and downstream application, we train the backbone f to identify these objects in a multi-object classification task following [43], whereas the downstream post-hoc CBM task involves classifying images containing only a *single* object. In all experiments, the backbone f remains frozen.

Probing Datasets for Concept Learning. For standard PCBMs [64], we train the concept projection π_θ on different auxiliary distributions, while keeping the semantic concept definitions fixed (visualized in Figure 3):

- **In-Domain:** We train π_θ using samples from the exact downstream distribution P_{task}. For *Elements*, this corresponds to single-object images exhibiting the specific concept (e.g., all images containing "red").
- **Near-Domain:** Concepts are sampled from a distribution that is distinct from, but visually similar to P_{task}. For *Elements*, we consider $P_{\text{near}}^{\forall}$ with multiple objects per image that all share concepts including the target, and $P_{\text{near}}^{\exists}$, where at least one object exhibits the target concept.
- **Out-of-Distribution:** Concepts are learned from a visually distinct distribution P_{OOD} (e.g., abstract patterns) that share the concept label but differ structurally from the downstream task. This represents the domain gap common in post-hoc analysis (e.g., datasets like Broden [6]).

Post-Hoc Concept Bottleneck Models (Post-Hoc CBMs). In our experiments, we train post-hoc CBMs with two strategies: training π_θ on an auxiliary concept dataset and using a VLM to generate surrogate concept labels. Specifically, we utilize a VLM (CLIP [49] and DINOv3 [55], each with a ViT-B/16 [12]) as the backbone f and learn π_θ using the text embedding for each concept [64]. Additionally, we consider Label-Free CBMs (LFCBMs) [44], where surrogate

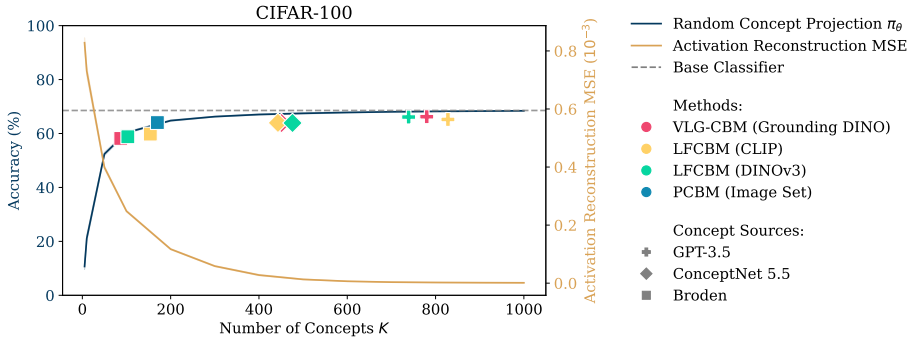


Fig. 4: High downstream accuracy does not imply faithfulness. The dashed line shows the classifier's performance without a bottleneck layer. The downstream classification accuracy of a post-hoc CBM increases with the number of random concept vectors. Additionally, various post-hoc CBMs that use meaningful concept sources only marginally outperform random concepts. Hence, downstream performance does not indicate faithful concept representations, which aligns with our theoretical findings in Section 3.1.

concept labels for P_{task} are generated by computing the similarity between the CLIP-representation [49] of the input image and the text embeddings for the considered concepts. Further, we evaluate Vision-Language-Guided CBMs (VLG-CBMs) [57], which use the open-vocabulary object detector Grounding DINO [34] to generate binary surrogate concept labels for P_{task} for each concept based on the predicted presence of concepts in the image [57]. For all post-hoc CBM variants, the bottleneck size K is determined by the number of concepts derived from the concept source (we consider Broden [6], ConceptNet 5.5 [56], and GPT-3 [9]).

4.2 The Illusion of Classifier Accuracy

Setup. To validate that downstream accuracy is an insufficient metric for concept bottleneck quality (Section 3.1), we follow a similar setup to [40] and build post-hoc CBMs using frozen, standard-normal random projections π_θ of increasing dimension K. For comparison, we train post-hoc CBMs using dominant paradigms from the literature: (1) Auxiliary concept datasets [64], e.g., Broden [6], and (2) VLM surrogate labeling [44, 57] using either the same concepts as in (1) or external sources (ConceptNet 5.5 [56] or GPT 3.5 [9]). In all cases, the downstream classifier h is trained after freezing π_θ (see Appendix B.1 for additional details).

Results. Figure 4 demonstrates that for CIFAR-100 [30] (other datasets in Appendix B.2, increasing the bottleneck dimension K of random projections decreases activation reconstruction error and closes the downstream accuracy gap to a fully supervised, bottleneck-free backbone. Notably, the post-hoc CBMs [44, 57, 64] trained on common concept sources [6, 9, 56] achieve similar performance to random projections of equal size K. This confirms that high downstream

Table 1: Faithfulness vs. Performance. Comparison of post-hoc CBM variants with respect to the ground truth concept labels and direction on the Elements dataset [43]. OOD data degrades faithfulness significantly more than accuracy suggests. Each metric is averaged over the K concepts. To demonstrate faithfulness under error orthogonality, we also train π_θ on ground-truth concept labels while adding 25% random noise.

Method	Dataset/ Label	Concept Projection π_θ			Downstream Classifier	
		$\text{sim}_\measuredangle$ [↑]	Acc. P_{aux} [↑]	Acc. P_{task} [↑]	Acc. h [↑]	Acc. h^* [↑]
(1) Training on Auxiliary Distributions (Oracle Labels)						
PCBM [64]	P_{task}	0.96±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	$P^{\forall}_{\text{near}}$	0.97±0.01	1.00±0.00	0.70±0.02	1.00±0.00	0.99±0.01
	$P^{\exists}_{\text{near}}$	0.57±0.02	0.82±0.02	0.89±0.01	1.00±0.00	0.85±0.04
	P_{OOD}	0.47±0.01	0.98±0.01	0.77±0.01	1.00±0.00	0.21±0.05
(2) Surrogate Concept Labels (Training Dist.: P_{task})						
LFCBM [44]	25% Label Noise	0.02±0.03	0.75±0.01	0.98±0.02	1.00±0.00	0.98±0.03
	CLIP [49]	0.05±0.02	0.79±0.01	0.56±0.02	0.99±0.00	0.54±0.04
	DINOv3 [55]	0.05±0.01	0.85±0.01	0.58±0.02	1.00±0.00	0.78±0.05
VLG-CBM [57]	G–DINO [34]	0.02±0.02	0.74±0.01	0.85±0.01	1.00±0.01	0.04±0.01

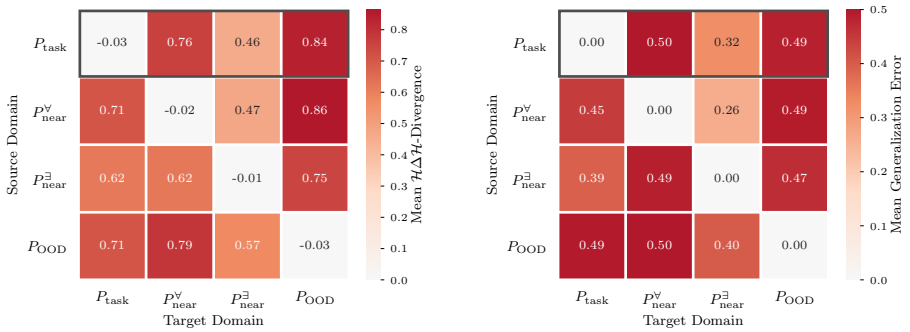
accuracy is insufficient evidence for a meaningful bottleneck, requiring a direct concept projection analysis instead.

4.3 Quantifying Misalignment and Concept Projection Performance

Setup. Having shown that downstream accuracy is deceptive, we now directly measure the faithfulness of π_θ. We evaluate post-hoc CBMs trained on four auxiliary distributions (P_{task}, $P^{\forall}_{\text{near}}$, $P^{\exists}_{\text{near}}$, P_{OOD}; Figure 3), alongside VLM-surrogate models: LFCBM [44] (using CLIP [49] and DINOv3 [55]) and VLG-CBM [57] (using Grounding DINO [34]). We calculate the concept accuracy of π_θ on both its training (P_{aux}) and downstream (P_{task}) distributions, and geometric faithfulness via the cosine similarity between learned concept vectors v_k and target ground-truth directions v_k^*: $\text{sim}_\measuredangle = (v_k \cdot v_k^*)/(\|v_k\|\|v_k^*\|)$ (Appendix B.3). To further determine if h masks identified unfaithfulness, we compare its accuracy against an oracle classifier h^* that uses the optimal rule for mapping ground-truth concepts to classes (cf. independent bottleneck strategy [29]). Though related to information leakage (see Section 2), faithfulness is complementary. Unfaithfulness can occur without leakage and vice versa (analyzed in Appendix B.4).

Results. Table 1 summarizes the results for the Elements dataset [43], where we have access to ground-truth concept labels. We split our analysis into the two primary post-hoc CBM setups: (1) training on auxiliary datasets [64] and (2) using in-domain surrogate concept labels [44, 57].

For (1), all models achieve high accuracy on P_{aux}, but P_{OOD} in particular yields unfaithful target representations, as indicated by the concept accuracy and the geometric alignment $\text{sim}_\measuredangle$. Crucially, the learned h masks these failures with high downstream accuracy. In contrast, the oracle h^* demonstrates the



(a) Average Pairwise $\mathcal{H}\Delta\mathcal{H}$-Divergence Matrix **(b)** Pairwise Generalization Errors

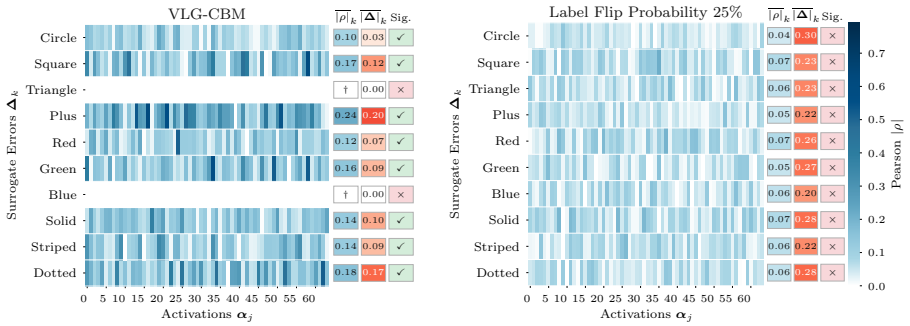
Fig. 5: Visualization of the estimated $\mathcal{H}\Delta\mathcal{H}$-divergence as a proxy for the upper bound of the task domain error. We pairwise compare the four probing datasets for the synthetic Elements downstream task [43], highlighting the actual downstream task in our setup. Note that the colormaps in (a) and (b) use different scales. We find that stronger covariate shifts lead to increased unfaithfulness, whereas lower divergence (i.e., in-domain concept sets) yields improved π_θ.

unfaithfulness. In particular, the model trained on P_{OOD} drops to 21% accuracy, consistent with the expected impact of a severe shift in concept representation.

For (2), LFCBMs [44] with noisy ground-truth labels (25% random noise) predict target concepts accurately. This matches our expectation from Section 3.2 that unsystematic label errors average out during the training of π_θ. In contrast, LFCBMs [44] trained with CLIP [49] and DINOv3 [55] surrogate labels, achieve poor performance in the target distribution, though DINOv3 improves marginally upon CLIP. These observations are confirmed by the learned classifiers h, which show near-perfect accuracy for all settings. However, the oracle h^* can again confirm the underlying unfaithfulness. VLG-CBMs [57] using Grounding DINO [34] exacerbate this: h perfectly solves the task, but h^* beats random guessing ($1/36 \approx 0.028$) only marginally. To explain the representation failures demonstrated in this experiment, we next investigate our theoretically derived sources of unfaithful concept projections π_θ (Section 3.2).

4.4 Measuring the Impact of Covariate Shift

Setup. To evaluate the impact of covariate shifts when training π_θ on an auxiliary dataset, we estimate the $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$-divergence [7, 27] between source and target domains, which serves as an upper bound to the generalization error (Section 3.2). Following [7], we approximate it by training a modified Huber loss discriminator [65] on the backbone activations to separate the distributions, computing $\hat{d}_{\mathcal{H}\Delta\mathcal{H}} \approx 2(1 - 2\epsilon)$ based on the discriminator's classification error ϵ. We then compare these estimates against the actual generalization errors of π_θ.



(a) Systematic Surrogate Error Introduced by Grounding DINO

(b) Unsystematic (Random) Surrogate Error

Fig. 6: Systematic vs. Unsystematic Surrogate Errors. Unfaithful projections occur when surrogate errors are both large ($\overline{|\boldsymbol{\Delta}|}_k$) and highly correlated ($\overline{|\rho|}_k$) with activations $\boldsymbol{\alpha}_j$. (a) Surrogates from Grounding DINO [34] exhibit systematic, correlated errors, causing unfaithful projections. (b) Random unsystematic label noise produces no correlation. If the VLM perfectly predicts a concept ($\overline{|\boldsymbol{\Delta}|}_k = 0$), correlation is undefined (†).

Results. As shown in Figure 5 (with additional details and results in Appendix B.5), empirical divergences align closely with actual π_θ generalization errors. The out-of-distribution domain P_{OOD} consistently yields the highest divergences and generalization errors. Conversely, matched source and target domains (on the diagonal) yield near-zero empirical divergences, confirming the high PCBM [64] performance observed in Table 1 (slightly negative values occur if the discriminator's accuracy falls marginally below 0.5). Among mismatched domains, $P^{\exists}_{\text{near}}$ and $P^{\forall}_{\text{near}}$ show the lowest divergence, reflecting the latter being a subdistribution of the former. Overall, we find a strong correlation (> 0.9) between the estimated $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ and actual generalization errors, demonstrating that $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ is an effective proxy for the unfaithfulness introduced by a covariate shift of the auxiliary concept set. This is highly practical because it can be estimated entirely without ground-truth concept labels for the downstream task.

4.5 Measuring Systematic Surrogate Label Errors

Setup. Following Section 3.2, we analyze the faithfulness of post-hoc CBMs trained with surrogate labels based on the surrogate error $\boldsymbol{\Delta}_k$ for concept k using two metrics: (1) error magnitude (mean absolute error), and (2) error orthogonality (absolute Pearson correlation [46] $|\rho_{k,j}|$ between surrogate label errors and backbone activations $\boldsymbol{\alpha}_j$). Because calculating $\boldsymbol{\Delta}_k$ requires ground truth concepts, we evaluate on the Elements dataset [43]. We compare surrogate labels generated with Grounding DINO [34, 57] against a baseline with unsystematic label noise obtained by randomly flipping a portion of the ground-truth concept labels (see Appendix B.6 for CLIP [49], DINOv3 [55], and CUB-200 [63]).

Results. Figure 6a shows these correlations alongside per-concept summary statistics $\overline{|\rho|}_k$ and $\overline{|\boldsymbol{\Delta}|}_k$ (calculated as the absolute average over per-activation scores, using a Holm-Bonferroni correction [20] at $p = 0.05$ to determine significance). As established in Equation (8), optimizing π_θ yields unfaithful representations when both values are large. Unlike the random-noise baseline, where the correlations effectively vanish, VLM surrogate errors are clearly systematic and correlated with specific activations. Qualitatively, we find that Grounding DINO [34] struggles to positively detect these concepts, which we attribute to the domain shift between its natural-image pretraining and our synthetic setup (geometric shapes on white backgrounds). Because the VLM label noise is non-random, training on it can cause π_θ to learn unfaithful projections (see (2) in Table 1).

While faithfulness is preserved for concepts that are labeled perfectly by the VLM (e.g., "Triangle" and "Blue" for Grounding DINO [34]), achieving purely random, uncorrelated surrogate errors may be difficult in many real-world scenarios. Since we identify systematic surrogate label errors as a source of unfaithfulness, correlation can be used to rank imperfect surrogate labeling functions by the severity of this failure mode. For P_{task}, the average ρ across concepts is lowest for Grounding DINO [34] (0.142), followed by CLIP [49] (0.152) and DINOv3 [55] (0.155), making Grounding DINO the preferred surrogate labeling function under this criterion.

5 Conclusions

While post-hoc CBMs promise transparent decision-making, current evaluations often conflate predictive performance with conceptual alignment. As our theoretical analysis based on [4] shows, even non-semantic random projections achieve competitive accuracy. Consequently, downstream task performance is an uninformative metric for concept bottleneck quality, requiring the direct inspection of the concept projection π_θ. Doing so reveals two primary sources of unfaithfulness in standard post-hoc CBM training. First, relying on auxiliary datasets may introduce a covariate shift that can invalidate learned concepts in the target domain. Based on [7], we provide an upper bound on the generalization error and a practical metric to measure it. Second, we demonstrate that systematic label errors introduced by VLM surrogate supervision lead to unfaithfulness. By formalizing these issues and validating them across real-world and synthetic benchmarks, we establish an evaluation framework for post-hoc CBM unfaithfulness.

These insights open several avenues for future research. Our formalization of distribution shifts suggests the need for explicit domain adaptation techniques within the post-hoc CBM framework to align feature spaces, rather than just labels. Furthermore, the presence of systematic VLM label noise requires the development of de-biasing mechanisms that disentangle correlated errors during surrogate supervision. Ultimately, we provide the theoretical and practical tools necessary to investigate learned concept projections.

References

1. Almudévar, A., Hernández-Lobato, J.M., Ortega, A.: There Was Never a Bottleneck in Concept Bottleneck Models. In: The Fourteenth International Conference on Learning Representations (2026) 1
2. Alukaev, D., Kiselev, S., Pershin, I., Ibragimov, B., Ivanov, V., Kornaev, A., Titov, I.: Cross-Modal Conceptualization in Bottleneck Models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5241–5253 (2023) 1
3. Ansuini, A., Laio, A., Macke, J.H., Zoccolan, D.: Intrinsic dimension of data representations in deep neural networks. Advances in Neural Information Processing Systems **32** (2019) 2, 3, 5, 6
4. Baraniuk, R.G., Wakin, M.B.: Random Projections of Smooth Manifolds. Found. Comput. Math. **9**(1), 51–77 (2009) 2, 3, 5, 6, 15
5. Barsalou, L.W.: Perceptual symbol systems. Behav. Brain. Sci. **22**(4), 577–660 (1999) 1
6. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network Dissection: Quantifying Interpretability of Deep Visual Representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6541–6549 (2017) 1, 2, 3, 7, 10, 11
7. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**(1-2), 151–175 (2010) 3, 7, 13, 15
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006) 6
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020) 4, 11
10. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'23/IAAI'23/EAAI'23, AAAI Press (2023) 1
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009) 10
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021) 10
13. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. Advances in Neural Information Processing Systems **35**, 21400–21413 (2022) 2, 4, 5
14. Fahrmeir, L., Lang, S.: Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. J. R. Stat. Soc. Ser. C **50**(2), 201–220 (2001) 8

15. Galliamov, K., Kazmi, S.M.A., Khan, A., Rivera, A.R.: Concepts' Information Bottleneck Models. In: The Fourteenth International Conference on Learning Representations (2026) 1
16. Gardenfors, P.: Conceptual spaces: The geometry of thought. MIT press (2004) 1
17. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016) 6
18. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing Leakage in Concept Bottleneck Models. Advances in Neural Information Processing Systems **35**, 23386–23397 (2022) 3, 5
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 9
20. Holm, S.: A Simple Sequentially Rejective Multiple Test Procedure. Scand. J. Stat. **6**(2), 65–70 (1979) 15
21. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989) 5
22. Huang, Q., Song, J., Hu, J., Zhang, H., Wang, Y., Song, M.: On the concept trustworthiness in concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21161–21168 (2024) 4
23. Jacovi, A., Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205 (2020) 2
24. Johnson, W.B., Lindenstrauss, J., et al.: Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**(189-206), 1 (1984) 2, 3, 5
25. Jørgensen, B.: Exponential Dispersion Models. J. R. Stat. Soc. Ser. B **49**(2), 127–162 (1987) 8
26. Kazmierczak, R., Berthier, E., Frehse, G., Franchi, G.: CLIP-QDA: An Explainable Concept Bottleneck Model. Transact. Mach. Learn. Res. (2024) 1
27. Kifer, D., Ben-David, S., Gehrke, J.: Detecting Change in Data Streams. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. p. 180–191. VLDB '04, VLDB Endowment (2004) 7, 13
28. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (2018) 7
29. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept Bottleneck Models. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5338–5348. PMLR (2020) 1, 2, 6, 12
30. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Tech. rep. (2009) 3, 10, 11
31. Kumar, S., Ahuja, N.: Measuring the (Un) Faithfulness of Concept-Based Explanations. arXiv preprint arXiv:2504.10833 (2025) 4
32. Lai, S., Hu, L., Wang, J., Berti-Equille, L., Wang, D.: Faithful Vision-Language Interpretation via Concept Bottleneck Models. In: The Twelfth International Conference on Learning Representations (2024) 4
33. Lipton, Z.C.: The Mythos of Model Interpretability. Commun. ACM **61**(10), 36–43 (2018) 1
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with Grounded

Pre-training for Open-Set Object Detection. In: Computer Vision – ECCV 2024. p. 38–55. Lecture Notes in Computer Science (2024) 3, 11, 12, 13, 14, 15
35. Luyten, M.R., van der Schaar, M.: A theoretical design of concept sets: improving the predictability of concept bottleneck models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024) 4
36. Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., Pan, W.: Promises and Pitfalls of Black-Box Concept Learning Models. arXiv preprint arXiv:2106.13314 (2021) 2, 3
37. Makonnen, M., Vandenhirtz, M., Laguna, S., Vogt, J.E.: Measuring Leakage in Concept-Based Methods: An Information Theoretic Approach. In: ICLR 2025 Workshop: XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge (2025) 2, 5
38. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do Concept Bottleneck Models Learn as Intended? ICLR 2021 Workshop on Responsible AI (2021) 2, 3, 5
39. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Monographs on statistics and applied probability, 2 edn. (1989) 8
40. Midavaine, N., Go, G.H.T., Canez, D., Simion, I., Chatterji, S.: [Re] On the Reproducibility of Post-Hoc Concept Bottleneck Models. Trans. Mach. Learn. Res. (2024) 2, 5, 11
41. Moayeri, M., Rezaei, K., Sanjabi, M., Feizi, S.: Text-To-Concept (and Back) via Cross-Model Alignment. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 25037–25060. PMLR (2023) 1
42. Murphy, G.: The Big Book of Concepts. MIT Press (2002) 1
43. Nicolson, A., Schut, L., Noble, A., Gal, Y.: Explaining Explainability: Recommendations for Effective Use of Concept Activation Vectors. Trans. Mach. Learn. Res. (2025) 3, 10, 12, 13, 14
44. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free Concept Bottleneck Models (2023) 2, 3, 4, 7, 8, 10, 11, 12, 13
45. Oikarinen, T., Weng, T.W.: Clip-dissect: Automatic description of neuron representations in deep vision networks. In: The Eleventh International Conference on Learning Representations (2023) 2
46. Pearson, K.: VII. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philos. Trans. R. Soc. A. (187), 253–318 (12 1896) 9, 14
47. Penrose, R.: On best approximate solutions of linear matrix equations. In: Math. Proc. Camb. Philos. Soc. vol. 52, pp. 17–19. Cambridge University Press (1956) 6
48. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The Intrinsic Dimension of Images and Its Impact on Learning. In: International Conference on Learning Representations (2021) 2, 3, 5, 6
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021) 2, 3, 10, 11, 12, 13, 14, 15
50. Ramaswamy, V.V., Kim, S.S., Fong, R., Russakovsky, O.: Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10932–10941 (2023) 4, 6

51. Sampat, S.K., Patel, M., Yang, Y., Baral, C.: Help Me Identify: Is an LLM+ VQA System All We Need to Identify Visual Concepts? arXiv preprint arXiv:2410.13651 (2024) 2, 8
52. Schmalwasser, L., Penzel, N., Denzler, J., Niebling, J.: Fastcav: Efficient computation of concept activation vectors for explaining deep neural networks. In: International Conference on Machine Learning (ICML) (2025), `https://fastcav.github.io/` 7
53. Schoen, R., Abeloos, B., Herbin, S.: Measuring and Addressing Information Leakage in Concept Bottleneck Models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 624–632 (2025) 3, 5
54. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference **90**(2), 227–244 (2000) 7
55. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025) 3, 10, 12, 13, 14, 15
56. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. p. 4444–4451 (2017) 11
57. Srivastava, D., Yan, G., Weng, T.W.: VLG-CBM: Training Concept Bottleneck Models with Vision-Language Guidance (2024) 2, 3, 4, 6, 7, 8, 11, 12, 13, 14
58. Sugiyama, M., Kawanabe, M.: Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. The MIT Press (2012) 7
59. Sun, A., Yuan, Y., Ma, P., Wang, S.: Eliminating information leakage in hard concept bottleneck models with supervised, hierarchical concept learning. arXiv preprint arXiv:2402.05945 (2024) 3
60. Sémery, O.: imgclsmob: Deep learning networks. `https://github.com/osmr/imgclsmob` (2024), GitHub repository, accessed February 2026 9
61. Tan, A., Zhou, F., Chen, H.: Explain via Any Concept: Concept Bottleneck Model with Open Vocabulary Concepts. In: Computer Vision – ECCV 2024. p. 123–138. Lecture Notes in Computer Science, Springer-Verlag (2024) 1
62. Vandenhirtz, M., Laguna, S., Marcinkevičs, R., Vogt, J.E.: Stochastic Concept Bottleneck Models (2024) 1
63. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200 (2010) 3, 9, 14
64. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023) 2, 3, 4, 6, 7, 10, 11, 12, 14
65. Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the Twenty-First International Conference on Machine Learning. Association for Computing Machinery (2004) 13