# Interpretable but Fragile? Robustness of Concept Bottlenecks under Geometric-Semantic Perturbations

**Hanwei Zhang**
Department of Computer Science
Saarland University
Saarbrücken, Germany
zhang@depend.uni-saarland.de

**Tianma Hu**
School of Computer Science and Engineering
Tianjin University of Technology
Tianjin, China
hutianma15299@stud.tjut.edu.cn

**Gaojie Jin**
Department of Artificial Intelligence &
Institute of Artificial Intelligence and Brain Sciences
University of Macau
Macau SAR, China
gaojiejin@um.edu.mo

**Xu Cheng**
School of Computer Science and Engineering
Tianjin University of Technology
Tianjin, China
xu.cheng@ieee.org

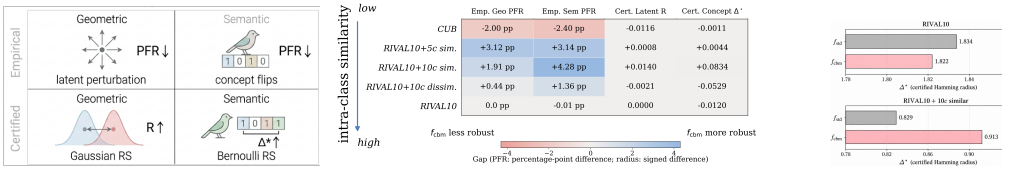
**Ronghui Mu***
Department of Computer Science
University of Exeter
Exeter,UK
r.mu2@exeter.ac.uk

## Abstract

Concept Bottleneck Models (CBMs) are designed to provide interpretable intermediate representations, yet how such bottlenecks affect robustness remains unclear, with existing studies reporting mixed and sometimes contradictory findings. We argue that these discrepancies arise from conflating different robustness notions and perturbation regimes, rather than from fundamental disagreements about CBMs themselves. To disentangle these factors, we introduce a generator-based evaluation framework that enables controlled comparisons between standard classifiers and CBMs under two distinct perturbation types: continuous geometric perturbations in latent space and discrete semantic interventions in concept space. We further evaluate a prototype-based interpretable model, PixPNet, showing that the same perturbation-to-prediction evaluation and certification pipeline extends beyond concept bottlenecks without requiring alignment between heterogeneous internal representations. Within this framework, we evaluate robustness both empirically, via prediction and concept-level sensitivity metrics, and certifiably, using randomized smoothing in latent and concept spaces. Across experiments on CUB and RIVAL10 variants, we reconcile previously conflicting findings by clarifying when, and in what sense, concept bottlenecks do or do not improve robustness. By further analyzing robustness under varying task conditions, including class semantic similarity and concept vocabulary size, we show that interpretability does not inherently confer robustness. Instead, concept bottlenecks shift where and how sensitivity manifests, revealing a nuanced interpretability robustness trade off that depends critically on the perturbation regime and task structure. Together, our results show that interpretability and robustness are distinct objectives: interpretable intermedi-

*corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).




(a) Evaluation framework (b) Robustness gap summary (c) Concrete contrast

Figure 1: **Evaluation framework and representative results.** (a) Evaluation robustness across geometric *vs.* semantic perturbations and empirical *vs.* certified criteria. (b) The heatmap reports representative $f_{cbm}$–$f_{std}$ robustness gaps using $\epsilon = 0.3$, $\tau = 5$, $\sigma = 0.10$, and $\rho = 0.10$ for the four columns, respectively. All gaps are oriented so that positive values indicate greater CBM robustness. Detailed computations in Appendix C.1 (c) Certified concept-radius examples at $\rho = 0.10$ show opposite outcomes across task settings.

ate representations do not uniformly improve robustness, but instead redistribute sensitivity across perturbation spaces and model families.

## 1 Introduction

Concept Bottleneck Models (CBMs) [Koh et al., 2020] are a prominent paradigm of interpretable learning, explicitly structuring predictions around human-interpretable concepts. By exposing intermediate concept representations, CBMs provide semantically grounded explanations, support targeted interventions, and offer a modular design that extends naturally in vision and multimodal settings. These properties make CBMs attractive, particularly in high-stakes and regulated domains such as those governed by the EU AI Act [Hermanns et al., 2024, Parliament and of the EU, 2024], where interpretability is often believed to promote robustness. However, the robustness implications of introducing a concept bottleneck remain underexplored, and existing evidence is mixed.

Prior work reaches conflicting conclusions regarding the robustness of CBMs. Rasheed et al. [2024] report robustness gains from concept bottlenecks, attributing them to the removal of non-essential variation, with robustness evaluated at the level of final class predictions. In contrast, Sinha et al. [2023] argues that concept bottlenecks can reduce robustness: concept representations themselves are often unstable under perturbations, and exposing an explicit concept interface introduces additional attack surfaces. As illustrated in Figure 1(c), we observe this tension directly: under different task settings of the same dataset, evaluating the same standard classifier and its concept bottleneck counterpart can lead to opposite conclusions about whether concept bottlenecks improve robustness.

At first glance, these findings appear irreconcilable. We argue that they instead stem from different notions of robustness and perturbation regimes. In practice, robustness is still most often assessed using small pixel-level perturbations, which lack semantic meaning and fail to reflect changes in underlying concepts. Whether a concept bottleneck improves or degrades robustness depends on *where robustness is measured* (class *vs.* concept level), *how perturbations are applied*, and *the task conditions* under which the model operates. Making these distinctions explicit is key to understanding the role of concept bottlenecks in robust learning. This leads to the central question of this work:

> *How does introducing an interpretable intermediate representation reshape robustness relative to a standard classifier, and how does this depend on the perturbation space and robustness criterion?*

To study this systematically, we adopt a concept bottleneck generator [Kulkarni et al., 2025] that maps latent representations to an explicit concept space, enabling controlled concept-level interventions before image generation. The generator is frozen and shared across classifiers, serving only as an experimental instrument. This design allows us to disentangle two fundamentally different types of perturbations: (i) *geometric perturbations*, which induce continuous variations in the latent embedding space, and (ii) *semantic perturbations*, which directly modify individual concepts. Within this framework, we evaluate robustness *empirically* by measuring changes in class predictions

and intermediate representations, and *certifiably* by deriving guarantees via randomized smoothing. Standard models and their corresponding CBMs share the same backbone, ensuring a fair comparison.

Our analysis focuses on generator-based perturbations to enable controlled evaluation of the robustness induced by concept bottlenecks, without claiming universal robustness guarantees. Our results reveal distinct robustness profiles across perturbation regimes: concept bottlenecks can suppress prediction sensitivity in some settings while introducing additional vulnerabilities in others. In particular, we find that the effect of the bottleneck is strongly influenced by the semantic similarity between classes, and the number and relevance of concepts used by the model. These findings provide a unifying explanation for previously conflicting results and clarify the conditions under which concept bottlenecks act as beneficial regularizers versus when they introduce fragility.

Overall, our contributions are threefold:

(i) **A unified robustness perspective for concept bottlenecks.** We reconcile previously conflicting findings on CBM robustness by showing that they often arise from conflating different perturbation spaces and robustness criteria. By distinguishing prediction-level and concept-level sensitivity under geometric and semantic perturbations, as well as empirical, certified, and adversarial robustness, we clarify when concept bottlenecks suppress sensitivity and when they instead introduce additional vulnerabilities. Our results emphasize that interpretability and robustness are distinct objectives, and that concept bottlenecks reshape robustness rather than uniformly improving it.

(ii) **A controlled evaluation and certification framework for geometric and semantic robustness.** We introduce a generator-based framework (Figure 1(a)) that separates continuous latent-space geometric perturbations from discrete concept-level semantic interventions. The framework enables matched comparisons under a common perturbation-to-prediction interface, together with empirical robustness metrics and perturbation-specific randomized-smoothing certificates in latent and concept spaces. Importantly, the evaluation interface does not require different classifier architectures to share or align their internal representations.

(iii) **Empirical and certified analysis of when robustness gains emerge.** Through extensive experiments on CUB-200-2011 and RIVAL-10 variants, we show that concept bottlenecks do not provide an inherent robustness advantage (*e.g.*, Figure 1(b)). Instead, robustness depends strongly on the perturbation regime, robustness criterion, class semantic similarity, and concept vocabulary design. These findings provide practical guidance for designing CBMs that better balance interpretability and robustness. We further validate the broader applicability of the evaluation pipeline on PixPNet, a prototype-based interpretable model, showing that the same generator-defined evaluation and certification procedure can be applied beyond CBMs without requiring alignment between generator concepts and model-specific prototypes.

## 2 Related Work

**Robustness of Concept Bottlenecks.** Early work on interpretable models highlighted connections between interpretability and robustness, with Self-Explaining Neural Networks (SENN) enforcing robustness as the stability of explanations under small perturbations [Sawada and Nakamura, 2022]. More recent studies directly examine robustness in Concept Bottleneck Models (CBMs). Sinha et al. [2023] show that interpretability alone does not guarantee robustness, demonstrating that concept representations can be fragile under adversarial perturbations and that explicit concept interfaces introduce new attack surfaces, revealing a tension between concept-level stability and adversarial robustness. In contrast, Rasheed et al. [2024] find that CBMs, particularly sequentially trained variants, can achieve higher adversarial accuracy than end-to-end models under weak to moderate attacks, suggesting that conceptual bottlenecks may act as an information-filtering regularizer at the prediction level. Related observations beyond CBMs further indicate that highly interpretable representations can remain non-robust [Li et al., 2026], while recent extensions such as language-guided and flexible CBMs improve scalability and adaptability without addressing robustness explicitly [Yu et al., 2025, Du et al., 2026]. Overall, prior work reports seemingly conflicting conclusions, which we argue arise from differences in robustness definitions, perturbation models, and task settings rather than fundamental inconsistencies.

**Randomized Smoothing for Certified Robustness.** Randomized smoothing provides a general framework for certifying robustness by constructing a smoothed classifier through noise injection and

aggregation, yielding probabilistic guarantees against bounded perturbations [Cohen et al., 2019]. While most work applies smoothing in pixel space, recent studies have explored extensions to learned representations or structured perturbations [Hao et al., 2022, Wang et al., 2026, Mu et al., 2023, 2024]. In this work, randomized smoothing is used primarily as a model-agnostic certification and comparison tool. The resulting smoothed predictor can also improve empirical robust accuracy in some adversarial settings, but defense is not the primary objective of our analysis.

**Interpretable intermediate architectures beyond CBMs.** Interpretable intermediate representations are also exposed by post-hoc concept methods [Yuksekgonul et al., 2022] and prototype-based architectures [Zhang et al., 2021, Carmichael et al., 2024, Sicre et al., 2023]. PCBM [Yuksekgonul et al., 2022] constructs concept representations using CAV-style directions, while prototype-based models such as PixPNet [Carmichael et al., 2024] ground predictions in localized parts. Although these models expose different internal representations, our evaluation does not require these representations to be aligned: the generator defines a common external perturbation interface, and robustness is measured at the resulting prediction level.

# 3 Methodology

## 3.1 Perturbation Models

Let $g : \mathbb{R}^d \to \mathcal{X}$ denote a pretrained generator that maps a latent noise vector $\mathbf{z} \in \mathbb{R}^d$ to an image $\mathbf{x} = g(\mathbf{z})$. We assume that $g$ admits a decomposition $g = g_2 \circ g_1$, where $g_1$ maps the latent code to an intermediate representation and $g_2$ maps this representation to the image space. Following concept bottleneck generators, we introduce an encoder–decoder pair $(E, D)$ at an intermediate generator layer. The encoder maps the intermediate features $g_1(\mathbf{z})$ to a concept representation $\mathbf{c} = E(g_1(\mathbf{z})) \in \mathbb{R}^{2K}$, containing $K$ concepts, each represented by a pair of state coordinates $(c_i^+, c_i^-)$. The decoder maps this representation back to the generator feature space, yielding $g(\mathbf{z}) = g_2(D(E(g_1(\mathbf{z}))))$. Each concept corresponds to a semantically meaningful attribute, with its binary state indicating the presence or absence of the corresponding attribute. This representation allows us to distinguish two fundamentally different perturbation regimes:

**Geometric perturbations** are applied in latent space by adding noise to the latent code, $\tilde{\mathbf{z}} = \mathbf{z} + \delta, \|\delta\|_2 \leq \epsilon$, which induces a continuous variation along the generator manifold and results in a geometrically perturbed image $\tilde{\mathbf{x}} = g(\tilde{\mathbf{z}})$.

**Semantic perturbations** are applied directly in concept space by modifying the concept vector $\mathbf{c}$. We uniformly sample a subset $S \subseteq \{1, \cdots, K\}$ of size $|S| = \tau$ and, for each $i \in S$, swap the binary representation of corresponding concept from $[c_i^+, c_i^-]$ to $[c_i^-, c_i^+]$, simulating changes in high-level semantic attributes while preserving the underlying geometric structure. To detect concept changes under perturbation, we train binary concept classifiers $C(\mathbf{c}) = (C_1(\mathbf{c}), \ldots, C_K(\mathbf{c})) \in \{0, 1\}^K$ that map the concept representation to binary concept states. We define the concept-flip indicator as

$$\phi_i(\mathbf{c}, \tilde{\mathbf{c}}) := \mathbb{1}\left[C_i(\mathbf{c}) \neq C_i(\tilde{\mathbf{c}})\right]. \tag{1}$$

## 3.2 Statistical Robustness Analysis

Let $f : \mathcal{X} \to \mathcal{Y}$ denote a classifier operating on generated images. We consider two classes of models: *standard classifier* $f_{std}$ directly maps the input image to a label; *concept bottleneck model* is defined as a tuple $f_{cbm} := (r, h)$, where $r : \mathcal{X} \to \mathbb{R}^K$ predicts a concept representation from image and $h : \mathbb{R}^K \to \mathcal{Y}$ maps concepts to the final prediction. This setup enables a controlled comparison of how standard and concept-based classifiers respond to geometric perturbations in latent space and semantic perturbations in concept space. Further methodological details are provided in Appendix.

**Empirical Robustness.** To quantify empirical robustness under geometric and semantic perturbations, we measure changes at the concept and prediction levels using the following metrics:

*1. Concept Flip Rate (CFR)* measures the average number of concepts whose predicted states change under perturbation:

$$\mathrm{CFR} := \mathbb{E}_{(\mathbf{c},\tilde{\mathbf{c}})}\left[\sum_{i=1}^{K} \phi_i(\mathbf{c},\tilde{\mathbf{c}})\right]. \tag{2}$$

*2. Prediction Flip Rate (PFR)* measures the probability that the classifier's predicted label changes under perturbation:

$$\mathrm{PFR}(f) \;:=\; \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}\big[f(\mathbf{x}) \neq f(\tilde{\mathbf{x}})\big]\right];$$

*3. Decision Function Sensitivity (DFS)* quantifies the sensitivity of a concept bottleneck classifier $f_{\mathrm{cbm}}$ by measuring the change in its internal concept representation:

$$\mathrm{DFS} \;:=\; \mathbb{E}_{\mathbf{z}}\left[\big\|r\big(g(\mathbf{z})\big) - r\big(g(\tilde{\mathbf{z}})\big)\big\|_2\right].$$

CFR and DFS characterize concept-level robustness, capturing both true concept flips induced by the generator and the sensitivity of the concept bottleneck classifier, while PFR measures robustness at the prediction level. Together, these metrics allow us to study how concept instability propagates to model predictions.

**Certified Robustness.** We use randomized smoothing to certify robustness in the two perturbation spaces induced by the generator.

**Latent-space smoothing.** For geometric robustness, we apply Gaussian smoothing in latent space:

$$\hat{f}_{\mathrm{G}}(\mathbf{z}) = \arg\max_{y\in\mathcal{Y}} \mathbb{P}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[f(g(\mathbf{z}+\delta)) = y]. \tag{3}$$

**Theorem 3.1** (Latent-space Gaussian certificate [Cohen et al., 2019])**.** *Let $A$ and $B$ be the top and runner-up classes of $\hat{f}_{\mathrm{G}}$ with probabilities $p_A > p_B$. Then $\hat{f}_{\mathrm{G}}$ is constant for all latent perturbations $\|\epsilon\|_2 \le R_{\mathrm{G}}$, where*

$$R_{\mathrm{G}} = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right). \tag{4}$$

**Concept-space smoothing.** For semantic robustness, we sample independent concept flips $\eta_i \sim \mathrm{Bernoulli}(\rho)$ and write $\mathcal{F}(\mathbf{c},\eta)$ for the flipped concept representation. The smoothed classifier is

$$\hat{f}_{\mathrm{C}}(\mathbf{c}) = \arg\max_{y\in\mathcal{Y}} \mathbb{P}_{\eta\sim\mathrm{Bernoulli}(\rho)^K}[f(g_2(D(\mathcal{F}(\mathbf{c},\eta)))) = y]. \tag{5}$$

**Theorem 3.2** (Concept-space $\ell_0$ certificate [Lee et al., 2019])**.** *Let $y^\star$ be the top class of $\hat{f}_{\mathrm{C}}$, and let $\underline{p}_{y^\star}$ and $\overline{p}_y$ be conservative confidence bounds on the top and competing class probabilities. For any concept-flip budget $\Delta$, if*

$$\mathcal{L}_\Delta(\underline{p}_{y^\star}) > \mathcal{U}_\Delta(\overline{p}_y), \qquad \forall y \neq y^\star, \tag{6}$$

*then, with probability at least $1-\alpha$ over Monte Carlo estimation, $\hat{f}_{\mathrm{C}}(\tilde{\mathbf{c}}) = y^\star$ for all $\|\mathcal{C}(\tilde{\mathbf{c}}) - \mathcal{C}(\mathbf{c})\|_0 \le \Delta$.*

We report the largest certified semantic radius

$$\Delta^\star(\mathbf{c}) = \max\left\{\Delta : \mathcal{L}_\Delta(\underline{p}_{y^\star}) > \mathcal{U}_\Delta(\overline{p}_y),\ \forall y \neq y^\star\right\}. \tag{7}$$

All confidence bounds use Clopper–Pearson intervals; full transfer-function definitions, proofs, and algorithms are in Appendix.

## 4 Experiments

In this section, we empirically and certifiably evaluate the robustness of $f_{std}$ and $f_{cbm}$ under controlled geometric and semantic perturbations. Due to page limitations, we report a subset of the experimental results in the main paper; the complete results, including additional statistical analyses and backbone comparisons, are provided in Appendix C. Across perturbation regimes, task conditions, and evaluation metrics, our experiments reveal a consistent pattern: *Concept bottlenecks do not induce monotonic robustness improvements; instead, they shift where and how sensitivity manifests.*

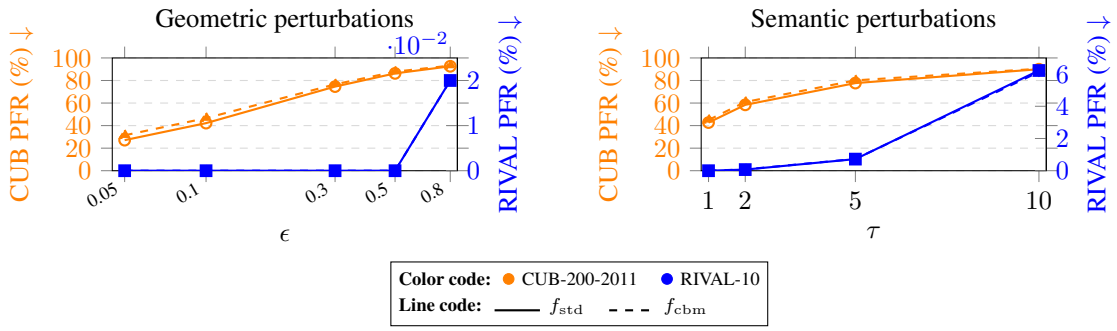



Figure 2: **Empirical robustness under geometric and semantic perturbations.** Orange (left axis): CUB-200-2011; blue (right axis): RIVAL-10. Dataset-specific y-axis scales are used for clarity.

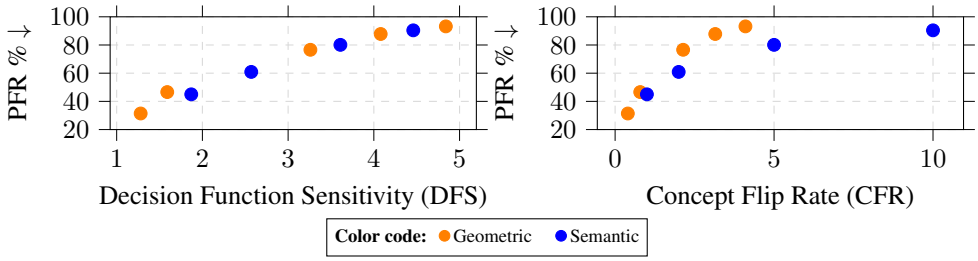



Figure 3: **Prediction sensitivity in concept-level measures on CUB-200-2011.**

### 4.1 Experimental Setup

We conduct experiments on two datasets with rich concept annotations: CUB-200-2011 and RIVAL-10. CUB-200-2011 [Wah et al., 2011] is a fine-grained bird classification benchmark consisting of 11,788 images from 200 species, each annotated with 112 expert-defined visual concepts.[2] RIVAL-10 [Moayeri et al., 2022] aligns CIFAR-10 [Krizhevsky et al., 2009] classes with ImageNet [Krizhevsky et al., 2012] categories and contains approximately 26,000 images annotated with 18 visual attributes and corresponding segmentations.

For CUB-200-2011, we select $K = 15$ concepts by default based on empirical frequency. We adopt a post-hoc concept bottleneck autoencoder (CB-AE) [Kulkarni et al., 2025] built on StyleGAN [Karras et al., 2020] as the generator, and train it on CUB-200-2011 to induce an explicit concept space aligned with these concepts. For RIVAL-10, models are trained on the official training split of 21,098 images and evaluated on the test split of 5,286 images, with all reported results computed on the test set. We adopt BigGAN [Brock et al., 2018] as the base image generator and treat all 18 annotated attributes as concepts by default. For classification, all compared classifiers share the same pretrained backbone to ensure a fair and controlled comparison. For both datasets, we use a ResNet-50 [He et al., 2016] for the standard classifier $f_{std}$ and a ResNet-50 version of the post-hoc concept bottleneck model (PCBM) [Yuksekgonul et al., 2022] for the concept-based classifier $f_{cbm}$.

We restrict evaluation to generated inputs on which $f_{std}$ and $f_{cbm}$ agree, denoted by $\mathcal{S}_{\text{agree}}$, and use the shared prediction as a reference label.

#### 4.1.1 Empirical robustness results

To quantify variability due to subset selection, we repeat the evaluation on 5 independent random subsets of size 200 drawn uniformly without replacement from the pool (subsets may overlap). For each subset, both $f_{std}$ and $f_{cbm}$ are evaluated on the *same* inputs (paired design), yielding five paired measurements per metric. We then test $H_0 : \Delta = 0$ using a paired two-sided test across the 5 subset-level measurements, where $\Delta$ denotes the per-subset difference between methods. When multiple budgets and metrics are tested, we additionally report Holm–Bonferroni corrected $p$-values in Appendix C.3.

[2] https://www.vision.caltech.edu/datasets/cub_200_2011/

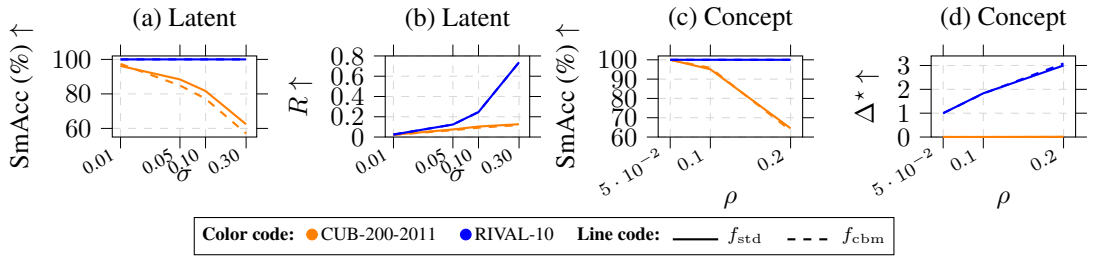



Figure 4: **Certified robustness under latent-space and concept-space randomized smoothing.**

As shown in Figure 2, the prediction flip rate (PFR) increases with perturbation strength. On CUB-200-2011, $f_{\text{cbm}}$ exhibits higher PFR than $f_{\text{std}}$ across all evaluated geometric and semantic perturbation budgets. However, after Holm–Bonferroni correction, only the difference at the semantic budget $\tau = 5$ remains statistically significant. In contrast, on RIVAL-10, both models remain nearly perfectly robust over the same perturbation ranges. As shown in Figure 3, raw concept flip rates (CFR) alone do not fully account for prediction instability. In contrast, decision function sensitivity (DFS) shows a more consistent relationship with PFR, suggesting that instability in the CBM decision representation better explains prediction flips than the number of concept changes. This arises because CBMs primarily rely on task-relevant concepts and are relatively insensitive to perturbations of irrelevant ones, aligning with the intuition that concept bottlenecks can act as a form of regularization under structured semantic noise. Additional analysis of DFS, CFR, and PFR, along with the complete numerical results, is provided in Appendix C.3.

### 4.2 Randomised Smoothing Experiments

Empirical perturbations characterize sensitivity under sampled inputs but do not provide guarantees over all perturbations within a neighborhood. To assess whether these trends persist under worst-case perturbations, we evaluate certified robustness using randomized smoothing in latent ($\ell_2$) and concept ($\ell_0$) spaces. Unless otherwise stated, we use $n = 200$ inputs sampled from $\mathcal{S}_{\text{agree}}$ and estimate smoothed class probabilities with $N = 1000$ Monte Carlo samples. We construct simultaneous one-sided Clopper Pearson confidence bounds with overall confidence level $1 - \alpha = 95\%$, allocating $\alpha/(2C)$ to each bound, where $C$ is the number of classes. For Bernoulli concept-space certification, the bounds are further tightened using the probability simplex constraint. To quantify variability, we repeat the evaluation over five independently sampled subsets of size 200. For latent-space smoothing, we use Gaussian perturbations $\delta \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma \in \{0.01, 0.05, 0.10, 0.30\}$ and report the average certified $\ell_2$ radius $R$ over correctly classified, non abstained samples. For concept space smoothing, each concept is independently flipped with probability $\rho$, and we report the average certified $\ell_0$ concept flip radius $\Delta$ over correctly classified, non-abstained samples. In both settings, smoothed accuracy (SmAcc) counts abstentions as incorrect.

Figure 4 summarizes the results. Under latent-space smoothing, the *$f_{std}$ generally achieves higher SmAcc and larger certified radii than the $f_{cbm}$* on CUB-200-2011, especially as $\sigma$ increases. This indicates that, on generator native inputs, $f_{cbm}$ does not improve certifiable robustness to continuous latent perturbations. On the original RIVAL-10 split, both models are saturated: SmAcc remains at 100% and certified radii are nearly identical across all $\sigma$, reflecting the strong class separability of this benchmark and indicating that samples remain far from decision boundaries in latent space.

Under concept-space smoothing, increasing the flip probability ($\rho$) can reduce SmAcc while increasing the certified ($\ell_0$) radius, highlighting that *smoothed prediction accuracy and certified robustness capture distinct aspects of model stability*. On CUB-200-2011, the $f_{cbm}$ does not consistently improve certified concept robustness: although it slightly improves SmAcc for some noise levels, its certified radius is comparable to or smaller than that of the ResNet baseline. On RIVAL-10, the certificates are again saturated, motivating the harder RIVAL-10 variants studied in section 5.

### 4.3 Adversarial Evaluation

We apply the APGD-CE attack [Croce and Hein, 2020] to the entire pipeline, including the generator, for a practical stress test of adversarial robustness rather than a head-to-head robustness comparison.

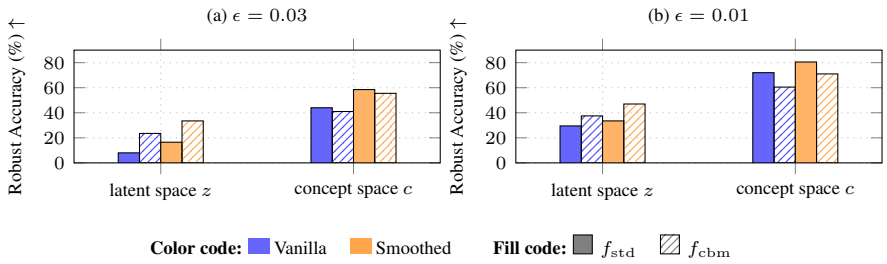



Figure 5: **Adversarial robustness under APGD-CE [Croce and Hein, 2020] attacks on the CUB generated dataset.**

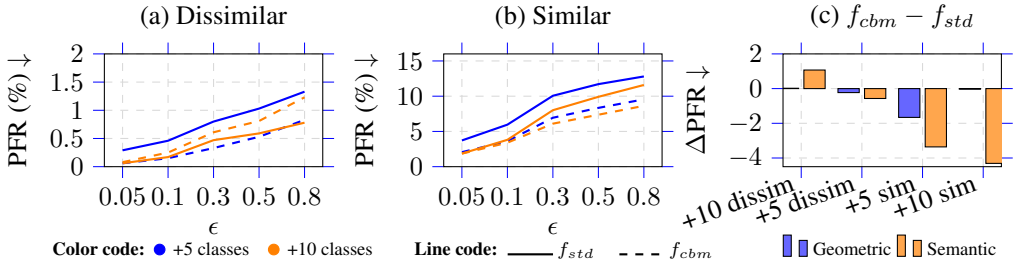



Figure 6: **Effect of class structure and semantic similarity on task-conditional robustness under geometric perturbations.** (a,b) PFR *v.s.* $\epsilon$ for dissimilar and similar class settings. (c) Difference in PFR (denoted as $\Delta$ PFR) between $f_{cbm}$ and $f_{std}$ when $\epsilon = 0.8$. Lower $\Delta$ PFR favors $f_{cbm}$.

As shown in Figure 5, randomized smoothing consistently improves robust accuracy under both latent-space and concept-space attacks, with larger gains at higher attack strengths. While subsection 4.2 indicates that $f_{cbm}$ does not outperform $f_{std}$ under randomized smoothing in either latent or concept space, adversarial evaluations reveal a complementary pattern: $f_{cbm}$ is more robust to latent-space attacks but less robust to concept-space attacks. This reflects the dual role of concept bottlenecks as both a new attack interface and a regularizing mechanism against geometric perturbations, helping resolve prior conflicting conclusions about whether concept bottlenecks introduce robustness.

***Takeaway.*** *(1) Concept bottlenecks do not uniformly improve prediction-level robustness. Although they can exhibit greater stability under certain semantic perturbation settings, this behavior depends on the task configuration and is not consistently observed.(2) Concept bottlenecks introduce a robustness trade-off: they improve robustness against geometric adversarial perturbations while increasing vulnerability to concept-space attacks. Our unified framework explains previously contradictory observations and clarifies how future designs can better exploit the robustness trade-off.*

## 5 Robustness under Task Conditions

The preceding sections show that concept bottleneck robustness varies substantially across datasets, despite using identical model architectures. This suggests that robustness is shaped not only by the presence of a bottleneck, but also by task-level factors. To isolate these effects, we conduct controlled studies on RIVAL-10 by varying (i) class structure and semantic similarity, and (ii) concept vocabulary size, while keeping the backbone and training protocol fixed.

### 5.1 Effect of Class Structure and Semantic Similarity

To study the effect of class similarity on robustness, we augment the original RIVAL-10 classification task with additional classes designed to vary semantic proximity to the original categories. Specifically, we construct two groups of new classes: similar classes, which are semantically close to the original RIVAL-10 categories, and dissimilar classes, which are semantically distant. Using these groups, we define four task configurations by adding 5 (or 10) similar (or dissimilar) classes, to the original

10-class task, yielding classification problems of increasing size and semantic difficulty. Across all configurations, we retain the original 18 concept attributes provided by RIVAL-10, ensuring that only the class structure changes while the concept vocabulary remains fixed. We then evaluate how robustness varies across these controlled settings.

Across all settings, class semantic similarity emerges as the dominant factor shaping robustness differences between $f_{std}$ and $f_{cbm}$. Figure 6 shows that semantic class dissimilarity generally improves robustness, with classifiers exhibiting lower prediction instability when classes are well separated. Across configurations, $f_{cbm}$ tends to match or slightly outperform $f_{std}$ under geometric perturbations and mild semantic changes, but becomes increasingly sensitive under large semantic perturbations, reflecting greater vulnerability at the concept level. The certified robustness results provide a complementary perspective to the empirical robustness evaluations, with differences between the two models depending on the perturbation space and task configuration. In addition, randomized smoothing yields larger robustness gains for $f_{cbm}$ in several settings, suggesting that smoothing can interact favorably with structured concept representations, albeit without providing a uniform robustness advantage.

### 5.2 Effect of Concept Vocabulary Size

We study the effect of concept set size on robustness using the original RIVAL-10 task. Starting from the original set of 18 annotated concepts, we use a large language model to generate an additional 18 concepts that are semantically relevant to these classes. This yields three concept configurations with increasing granularity: 18, 27, and 36 concepts. For each configuration, we evaluate robustness while keeping the classification task fixed, allowing us to isolate the impact of concept vocabulary size on model robustness.

Figure 7 shows that changing the concept vocabulary affects the robustness behavior of both $f_{\mathrm{std}}$ and $f_{\mathrm{cbm}}$, but the direction of the effect depends on the perturbation setting. Under geometric perturbations, the original RIVAL-10 setting is nearly saturated. Thus, geometric perturbations provide little evidence that either model is more robust. Semantic concept flips are more diagnostic. With 27 concepts, $f_{cbm}$ is slightly more stable at the largest flip budget ($9.29$ vs. $9.42$ PFR), whereas with 36 concepts the trend reverses ($3.11$ vs. $2.65$ PFR). This suggests that the effect of the concept vocabulary is not monotonic. One possible explanation is that, in the 27 concept setting, some flipped concepts are redundant or weakly used by the $f_{cbm}$ decision head, allowing the model to absorb large semantic flips slightly better. In the 36 concept setting, however, the additional concepts may introduce more correlated or unevenly useful attributes; perturbing them can destabilize the $f_{cbm}$ concept representation without yielding a consistent semantic robustness benefit. The randomized smoothing results further show that certified robustness depends on the perturbation regime and concept configuration, without establishing a consistent advantage for either model. Under latent Gaussian smoothing, both classifiers achieve nearly identical certified radii across the 27- and 36-concept configurations, with their curves largely overlapping. This suggests that, in these highly stable settings, latent-space certification provides limited discrimination between model architectures or concept vocabulary sizes. In contrast, concept-space smoothing reveals more noticeable differences across concept configurations. The 36-concept setting generally achieves larger average certified Hamming radii than the 27-concept setting, while the relative differences between $f_{\mathrm{std}}$ and $f_{\mathrm{cbm}}$ remain small and vary with the smoothing probability. These observations indicate that a larger concept vocabulary can influence certified semantic robustness without consistently benefiting the CBM relative to the standard classifier. Overall, concept vocabulary size changes the measured robustness profile, but does not induce a uniform robustness advantage for $f_{\mathrm{cbm}}$.

***Takeaway.*** *(1) Lower inter-class similarity is generally associated with higher classifier robustness. (2) Under higher class similarity, concept bottlenecks can retain greater prediction stability or smoothed accuracy in certain settings, although this does not consistently translate into larger certified radii. (3) Increasing the number of concepts does not monotonically improve robustness: additional concepts can help when perturbations target redundant attributes, but may also introduce instability when concepts are correlated or weakly grounded.*

**Design Implications for Robust Concept Bottlenecks.** These results indicate that robustness differences between $f_{std}$ and $f_{cbm}$ are primarily governed by task structure. This suggests that concept bottlenecks are particularly beneficial for classification tasks with high inter-class similarity, provided

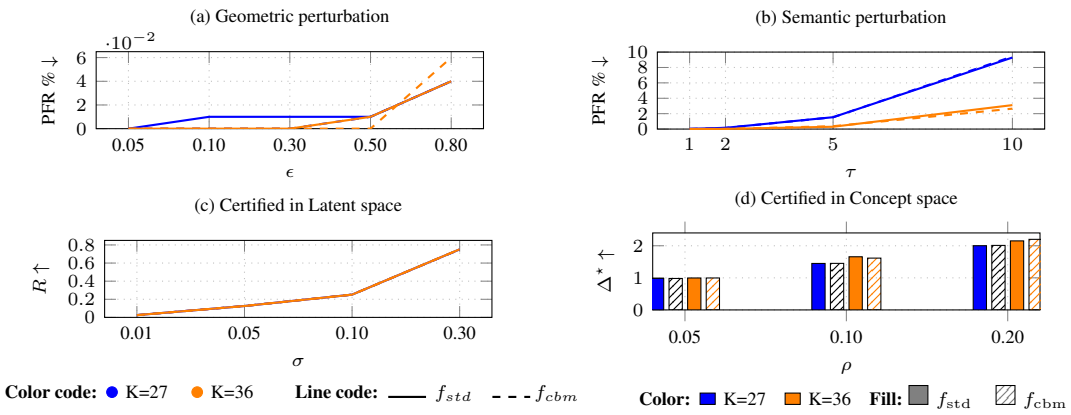



Figure 7: **Impact of concept vocabulary size on robustness.** Concept-vocabulary ablation on RIVAL-10. (a,b) Empirical PFR under geometric latent perturbations and semantic concept flips. (c,d) Certified robustness under Gaussian latent smoothing and Bernoulli concept smoothing.

that the concept vocabulary is chosen carefully. Our findings further highlight that robustness should be treated as a key criterion when selecting and designing concept sets. While concept bottlenecks may reduce prediction sensitivity under certain semantic perturbations, they can also become more vulnerable under larger concept interventions or adversarial semantic attacks. Consequently, improving concept bottleneck robustness requires careful consideration of task structure, concept vocabulary design, and perturbation specific vulnerabilities.

## 6 Limitations

Our analysis relies on generator-native inputs to disentangle geometric and semantic perturbations. While this enables controlled robustness evaluation, results may differ for real-world inputs outside the generator manifold. We focus on post-hoc CBMs with frozen backbones; end-to-end CBMs or alternative concept learning paradigms may exhibit different behaviors. Finally, our conclusions are limited to robustness under structured geometric and semantic perturbations and do not replace standard pixel-space adversarial robustness evaluations.

## 7 Conclusion

We introduced a generator-based framework that disentangles continuous latent/geometric perturbations from discrete concept/semantic interventions and used it to systematically evaluate robustness under empirical, certified, and adversarial criteria. Across CUB-200-2011 and RIVAL-10 variants, we find that concept bottlenecks do not confer uniform robustness improvements. Instead, they reshape where and how sensitivity emerges, with robustness depending strongly on the perturbation space, robustness criterion, task structure, and concept vocabulary design.

Within our controlled evaluation setting, these results clarify when concept bottlenecks suppress irrelevant variation and when they introduce additional sensitivity through the semantic interface. In particular, class semantic similarity and concept granularity emerge as important design factors governing this trade-off. Our findings reinforce that interpretability and robustness are distinct objectives: concept bottlenecks should not be treated as an implicit robustness mechanism, but as a modeling choice whose robustness consequences depend on how concepts are defined and used.

Finally, although CBMs provide our primary analytical setting, the additional PixPNet experiments show that the same generator-defined perturbation-to-prediction evaluation and randomized-smoothing certification pipeline can be applied to a prototype-based interpretable classifier without requiring alignment between heterogeneous internal representations. This broader applicability suggests a path toward evaluating robustness across a wider family of interpretable architectures while preserving the controlled perturbation framework developed here.

**Acknowledgments** This work was supported by The Royal Society Grant (Ensuring Trustworthy AI: Robustness Certification for Large Language Models)[Reference RGS\R2\252444]. This work also received support from DFG under grant No. 389792660 (TRR 248), and No. 547583482. GJ is supported by the University of Macau under grants MYRG-SRG2026-00033-FIC. This work was supported by the National Natural Science Foundation of China (NSF) (T2422015, 62306212) and the Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project (25JJJJC0009).

## References

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Zachariah Carmichael, Suhas Lohit, Anoop Cherian, Michael J Jones, and Walter J Scheirer. Pixel-grounded prototypical part networks. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4756–4767. IEEE, 2024.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pages 5948–5955, 2023.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7920–7929, 2020.

Xingbo Du, Qiantong Dou, Lei Fan, and Rui Zhang. Flexible concept bottleneck model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20932–20940, 2026.

Zhongkai Hao, Chengyang Ying, Yinpeng Dong, Hang Su, Jian Song, and Jun Zhu. Gsmooth: Certified robustness against semantic transformations via generalized randomized smoothing. In *International Conference on Machine Learning*, pages 8465–8483. PMLR, 2022.

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

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

Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Sarah Sterz, and Hanwei Zhang. Ai act for the working programmer. In *International Conference on Bridging the Gap between AI and Reality*, pages 74–98. Springer, 2024.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Hui Xiong, Jingfeng Zhang, and Di Wang. Editable concept bottleneck models. *arXiv preprint arXiv:2405.15476*, 2024.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

Aya Abdelsalam Ismail, Julius Adebayo, Hector Corrada Bravo, Stephen Ra, and Kyunghyun Cho. Concept bottleneck generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15273–15283, 2022.

Gaojie Jin, Sihao Wu, Jiaxu Liu, Tianjin Huang, and Ronghui Mu. Enhancing robust fairness via confusional spectral regularization. In *International Conference on Learning Representations*, volume 2025, pages 52945–52966, 2025.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 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*, pages 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable generative models through post-hoc concept bottlenecks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8162–8171, 2025.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Aaron J Li, Suraj Srinivas, Usha Bhalla, and Himabindu Lakkaraju. Evaluating adversarial robustness of concept representations in sparse autoencoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5940–5957, 2026.

Mark Huasong Meng, Guangdong Bai, Sin Gee Teo, Zhe Hou, Yan Xiao, Yun Lin, and Jin Song Dong. Adversarial robustness of deep neural networks: A survey from a formal verification perspective. *IEEE Transactions on Dependable and Secure Computing*, 2022.

Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19087–19097, 2022.

Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15046–15054, 2023.

Ronghui Mu, Leandro Soriano Marcolino, Yanghao Zhang, Tianle Zhang, Xiaowei Huang, and Wenjie Ruan. Reward certification for policy smoothed reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21429–21437, 2024.

European Parliament and Council of the EU. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), 2024. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689`.

Bader Rasheed, Mohamed Abdelhamid, Adil Khan, Igor Menezes, and Asad Masood Khatak. Exploring the impact of conceptual bottlenecks on adversarial robustness of deep neural networks. *IEEE Access*, 12:131323–131335, 2024.

Yoshihide Sawada and Keigo Nakamura. C-senn: Contrastive self-explaining neural network. *arXiv preprint arXiv:2206.09575*, 2022.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11040, 2024.

Ronan Sicre, Hanwei Zhang, Julien Dejasmin, Chiheb Daaloul, Stéphane Ayache, and Thierry Artières. Dp-net: learning discriminative parts for image recognition. In *2023 IEEE international conference on image processing (ICIP)*, pages 1230–1234. IEEE, 2023.

Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15127–15135, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent M Berry, and Yogatheesan Varatharajah. Evaluating latent space robustness and uncertainty of eeg-ml models under realistic distribution shifts. *Advances in Neural Information Processing Systems*, 35:21142–21156, 2022.

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

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1645–1653, 2021.

Zixia Wang, Gaojie Jin, Jia Hu, and Ronghui Mu. Clucert: Certifying llm robustness via clustering-guided denoising smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37998–38006, 2026.

Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, and Changsheng Xu. Language guided concept bottleneck models for interpretable continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14976–14986, 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Hanwei Zhang, Luo Cheng, Rui Wen, Yang Zhang, Lijun Zhang, and Holger Hermanns. Sl-cbm: Enhancing concept bottleneck models with semantic locality for better interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38093–38101, 2026.

Xing-Xing Zhang, Zhen-Feng Zhu, Ya-Wei Zhao, and Yao Zhao. Prototype learning in machine learning: A literature review. *Journal of Software*, 33(10):3732–3753, 2021.