HARMFUL PARTIAL INTERVENTIONS AND BENIGN LEAKAGE IN CONCEPT-BASED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence systems deployed in high-risk settings increasingly require models that support meaningful human oversight. Concept-based models (CMs) offer a promising framework by enabling users to inspect and intervene on intermediate, human-understandable concepts. However, concept interventions can counterintuitively worsen downstream predictions, a failure mode we formalise as *intervention harm*. We distinguish harm after all concepts have been intervened on from harm under partial interventions, the more likely practical use case. Modern CMs often achieve high accuracy by encoding, or *leaking*, additional task-relevant information in their intermediate concept representations. Recent work argues that such leakage can be benign with respect to intervention harm when concept representations satisfy two conditions: sufficiency and localisation. We prove that these criteria are not sufficient to guarantee that concept interventions improve downstream performance. Across controlled and real-world settings, we find that concept supervision incompleteness associates with increased intervention harm. Importantly, optimising the model when all concepts are intervened on, a previously proposed leakage mitigation objective, reduces endpoint harm but does not reliably prevent harmful partial corrections. Hence, we introduce a complementary monotonic intervention objective that directly penalises increases in task loss along partial intervention trajectories. Combining endpoint and monotonic intervention training substantially reduces the prevalence and magnitude of partial intervention harm while preserving concept and task predictive performance.

1 INTRODUCTION

Artificial intelligence (AI) deployed in high-risk settings is increasingly expected to support interpretability, transparency and meaningful human oversight. Concept-based models (CMs) offer a promising framework by structuring predictions around intermediate, human-interpretable concepts (Koh et al., 2020). Beyond making model reasoning more inspectable, concepts provide an interface through which a human overseer can intervene on the model's reasoning, thereby enabling users to correct erroneous intermediate concept predictions (Shin et al., 2023; Espinosa Zarlenga et al., 2023b). For such models to support meaningful oversight, however, intervening on a model's intermediate concept predictions should have predictable and beneficial effects on downstream task performance.

A fundamental limitation of earlier CMs (e.g., Concept Bottleneck Models (Koh et al., 2020)) is that they require a set of supervised concept labels that completely explain the downstream task. In practice, however, concept sets tend to be *incomplete*, capturing only part of the information needed to predict the task and causing CMs with strict concept bottlenecks to underperform (Mahinpei et al., 2021). Therefore, several works have introduced architectural pathways that explicitly *leak* task-relevant information that is missing from the concept set into the task predictor (Espinosa Zarlenga et al., 2022; Yuksekgonul et al., 2022; Margeloiu et al., 2021; Havasi et al., 2022; Parisini et al., 2025). Although such leakage can improve model accuracy, it *may* undermine the effectiveness of human intervention. When a concept representation contains leaked information, correcting the predicted concept to its ground-truth value may move the model's latent representation in a direction that does not correspond to the human's intended correction (Almudévar et al., 2026). If this intervention worsens the downstream prediction, we refer to the resulting failure mode as *intervention harm*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recently, *benign leakage* was introduced to describe leakage that allows models to retain high concept and task accuracies without displaying intervention harm (Espinosa Zarlenga, 2026). Alongside this notion, *sufficiency* and *localisation* were proposed as criteria for identifying benign leakage. Informally, sufficiency requires the learned concept representations to retain task-relevant information in the input features, while localisation requires assignment of this information to the components of the concept representation that vary under intervention. These conditions motivate approaches for encouraging benign leakage during training: minimising the intervention loss $\mathcal{L}_{\text{int}}$, defined as the expected task loss after all concept representations are intervened on, is, under certain assumptions, equivalent to maximising sufficiency and has been shown to improve task performance under intervention (Espinosa Zarlenga, 2026). However, whether these criteria are sufficient to prevent intervention harm remains to be rigorously established.

Crucially, interventions in real-world scenarios are rarely exhaustive, as CMs may contain hundreds of concepts, and users generally cannot be expected to verify every concept. Hence, interventions are typically *partial*: an overseer may inspect and correct only a small number of concepts considered relevant to the task at hand. However, reliable partial intervention requires a stronger notion of intervenability than that driven by $\mathcal{L}_{\text{int}}$. In practice we want CMs to avoid performance degradation when *any* subset of erroneous concepts is corrected, even when other concepts remain uncorrected. This exposes a distinction between *endpoint intervention harm*, which $\mathcal{L}_{\text{int}}$ minimises, and *partial intervention harm*, which remains understudied, particularly for leaky CMs where benign leakage is encouraged via $\mathcal{L}_{\text{int}}$. In incomplete settings, the former is actually a special case of the latter, since the supervised concept set does not extend to all task-relevant concepts. Therefore, we believe it is critical to better understand the prevalence and consequences of partial intervention harm.

With this in mind, this work develops a theoretical and empirical framework for studying endpoint and, more importantly, partial intervention harm. Specifically, our contributions are: (1) we show that sufficiency and localisation do *not* guarantee that correcting a concept improves task performance; (2) we identify concept-set incompleteness as an important determinant of intervention harm; (3) we distinguish *endpoint* from *partial* intervention harm and show that $\mathcal{L}_{\text{int}}$, while effective at reducing endpoint harm, does not reliably prevent partial harm, especially in incomplete settings; and (4) we introduce $\mathcal{L}_{\text{mono}}$, a complementary objective that penalises increases in task loss along partial intervention paths. Combining $\mathcal{L}_{\text{mono}}$ and $\mathcal{L}_{\text{int}}$ substantially reduces the prevalence and magnitude of partial intervention harm while preserving concept and task predictive performance. Together, these results provide a practical approach to encourage benign leakage under incompleteness while preserving high task and concept accuracy and avoiding intervention harm, even under partial interventions.

2 Related Work

Concept-based Models. Although diverse in architecture and purpose (Knab et al., 2026), most Concept-based Models (CMs) are best framed as Concept Bottleneck Models (CBMs) (Koh et al., 2020). At a high level, a CBM (g, f) is a compositional deep neural network (DNN) formed by a *concept encoder* $g : \mathcal{X} \to \mathcal{C}$ and a *task predictor* $f : \mathcal{C} \to \mathcal{Y}$. Intuitively, the concept encoder $g(\mathbf{x})$ maps input features $\mathbf{x} \in \mathcal{X}$ to a set of *concept representations* $\widehat{\mathbf{c}} \in \mathbb{R}^{k \times d}$ aligned with human-interpretable units of information $\mathbf{c} \in \{0, 1\}^k$ (e.g., colours, textures). The task predictor then uses these representations to predict a downstream task $\widehat{y} = f(\widehat{\mathbf{c}})$. Concept representations can be scalars (i.e., $d = 1$) (Chen et al., 2020; Havasi et al., 2022; Yuksekgonul et al., 2022; Vandenhirtz et al., 2024), or richer high-dimensional vectors (i.e., $d > 1$) (Espinosa Zarlenga et al., 2022; Kim et al., 2023; Xu et al., 2024). For example, *soft CBMs* (Koh et al., 2020) encode each concept c_i using a scalar $\hat{c}_i \in [0, 1]$ representing the probability that c_i is present in $\mathbf{x}$. In practice, all CMs require a concept-annotated training set of triplets $(\mathbf{x}, \mathbf{c}, y)$ to align their intermediate representations with concepts $\mathbf{c}$ useful for a downstream task y. These labels may come from human-annotated datasets (Ramaswamy et al., 2023; Skirzynski et al., 2026), be extracted from a pre-trained model's representations (Alvarez-Melis & Jaakkola, 2018; Yuksekgonul et al., 2022; De Santis et al., 2026), or be discovered using foundation models (Oikarinen et al., 2023; Yang et al., 2023; Rao et al., 2024).

Concept Interventions. A key property of CBMs is that they support *concept interventions*: at test time, an expert can correct mispredicted concept scores, or inject concept-based counterfactuals in a CM's reasoning, by directly modifying a concept representation $\widehat{\mathbf{c}}_i$ so that it is aligned with its corresponding ground-truth concept c_i. Generally, concept interventions are expected to increase

a CM's downstream task accuracy (Koh et al., 2020). However, recent works have shown that not all concept interventions lead to equivalent improvements (Chauhan et al., 2023; Espinosa Zarlenga et al., 2023b), and some may even decrease a model's performance in ill-conditioned CBMs (Koh et al., 2020; Espinosa Zarlenga et al., 2023a). As a result, follow-up research has explored how to improve the impact of interventions by (1) devising mechanisms to propagate interventions (Vandenhirtz et al., 2024; Singhi et al., 2024), (2) learning to reapply previous interventions (Steinmann et al., 2023), or (3) introducing the ability for the CM to explicitly defer concepts to experts (Pugnana et al., 2026).

Information Leakage. Concept representations can encode more information about other concepts or the downstream task than their corresponding ground-truth concepts do, a phenomenon called *information leakage* (Margeloiu et al., 2021; Mahinpei et al., 2021). Intuitively, leakage can be seen as a form of shortcut learning, where the way a CM uses a concept misaligns with how a human expert would use it (Geirhos et al., 2020; Marconato et al., 2023; Enouen & Galhotra, 2026). As such, leakage is more likely to appear when CMs are trained on *incomplete* concept sets that do not fully describe the downstream task. Properly understanding leakage, when it happens, and its consequences has led to a rich literature on the topic. Works in this area vary from proposing ways of measuring leakage (Espinosa Zarlenga et al., 2023a; Makonnen et al., 2025; Parisini et al., 2025), to exploring mechanisms for detecting (Marconato et al., 2022; Espinosa Zarlenga et al., 2025), controlling (Ragkousis & Parbhoo, 2024; Parisini et al., 2026), and mitigating (Havasi et al., 2022; Almudévar et al., 2026) leakage. More relevant to understanding partial intervention harm, Espinosa Zarlenga (2026) recently argued that leakage can be *benign*, in the sense that it does not affect a CM's performance and intervenability, when concept representations are *sufficient* and *localised*. There, the authors argue that these properties can be encouraged by minimising the task loss when all concepts are intervened on, a loss they call $\mathcal{L}_{\text{int}}$.

3 Understanding Intervention Harm

In this paper, we show that, although useful, sufficiency and localisation criteria are *not* sufficient to guarantee no intervention harm when leakage is allowed in a CM's representations. To show this, we first describe our problem setup and then formalise sufficiency, localisation, and intervention harm.

3.1 Problem Setup

Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ denote an n-dimensional feature representation (e.g., image pixels) of a realisation of random variable (r.v.) X. Throughout this paper, we study supervised learning tasks where each sample $\mathbf{x}$ is associated with a task label $y \in \mathcal{Y} \subseteq \mathbb{R}$ (e.g., bird species), which itself is a realisation of a r.v. Y. For each sample $\mathbf{x}$, let $\mathbf{c} \in \mathcal{C} \subseteq \{0,1\}^k$ be a vector of k known binary concept labels relevant to Y (e.g., "has red wing", "has hooked beak"). We assume that each binary concept label $\mathbf{c}_i$ in this vector is a realisation of a r.v. C_i, and use C to represent the r.v. corresponding to the entire known concept set $\mathbf{c}$. Here, we say a concept set $\mathcal{C}$ is *incomplete* when $\mathbf{c}$ cannot fully explain the task y. From an information-theoretic perspective, incompleteness therefore implies $H(Y \mid C) > 0$.

Throughout this paper, we work under the realistic assumption that the set of training concept labels for a CM is incomplete. In this case, we use $\mathbf{u} \in \mathcal{U} \in \{0,1\}^{k'}$ to describe a vector containing all concepts not in $\mathcal{C}$ that are relevant to the task (which we think of as realisations of a r.v. U representing all *unknown* task-relevant concepts). Hence the task outcome is a function of $\mathcal{C} \times \mathcal{U}$. As such, we assume there is an optimal CM $(f^* \circ g^*) : \mathcal{X} \to \mathcal{Y}$ that first predicts the known and unknown concepts $g^*(\mathbf{x}) = (\mathbf{c}, \mathbf{u}) \in \mathcal{C} \times \mathcal{U}$ from the input features $\mathbf{x} \in \mathcal{X}$, and then maps $(\mathbf{c}, \mathbf{u})$ to a downstream task prediction $\widehat{y}$ using a second function $f^*(\mathbf{c}, \mathbf{u}) = \widehat{y}$. We emphasise that, in contrast to this optimal predictor, a real-world CM trained on this task can observe only triples $(\mathbf{x}, \mathbf{c}, y)$ during training, while the corresponding unknown-concept state $\mathbf{u}$ remains unobserved.

Residual and Hybrid CBMs. To overcome the information bottleneck that incompleteness imposes, several works propose introducing residual information pathways that the CM can use to capture information in Y that is not present in the training concepts C (Havasi et al., 2022; Espinosa Zarlenga et al., 2022; Yuksekgonul et al., 2022; Sawada & Nakamura, 2022). A common implementation of this strategy is a Hybrid CBM (Mahinpei et al., 2021), where the output of the concept encoder g is extended so that it produces both a set of aligned representations $\widehat{\mathbf{c}}$ and a set of unaligned, free *residual* representations $\mathbf{z}$. Given the expressivity of this approach, we study leaky CMs in concept-incomplete

settings using Hybrid CBMs as our main object of study. Therefore, from here on, we extend the CBM-based formulation of CMs so that the concept encoder $g(\mathbf{x})$ explicitly outputs both a set of soft concept scalar representations $\widehat{\mathbf{c}} \in [0,1]^k$ and a residual pathway representation $\mathbf{z} \in \mathbb{R}^\mu$. Similarly, we assume the label predictor's input now includes this residual path (i.e., f now maps $\mathcal{C} \times \mathbb{R}^\mu$ to $\mathcal{Y}$). This enables us to work with expressive leaky CMs that can still operate in incomplete settings, while simplifying concept representations to scalars and explicitly encoding the residual path.

3.2 Benign Leakage, Sufficiency, and Localisation

Introducing a residual pathway like that seen in Hybrid CBMs risks losing the model's intervenability (Almudévar et al., 2026), defined as its ability to show monotone increases in task accuracy as one increases the number of concept representations aligned to their ground truth values (Vandenhirtz et al., 2024). Nevertheless, recent work by Espinosa Zarlenga (2026) argues that leakage does not necessarily need to affect a model's intervenability, and can therefore be *benign*, so long as two information-theoretic conditions are satisfied:

Sufficiency. The first claimed condition for benign leakage, termed *sufficiency*, can be described as:

$$\mathrm{I}(Y; Z \mid C) \approx \mathrm{I}(Y; X \mid C), \tag{1}$$

where $I(\cdot)$ is the mutual information, and Z is the r.v. representing the residual representations $\mathbf{z}$. Intuitively, this condition requires that all information about the downstream task Y that is present in the features X but not in the concepts C be encoded in the residual representations $\mathbf{z}$. To capture how closely sufficiency is satisfied, we define the *sufficiency discrepancy* as

$$\Delta_{\mathrm{S}} = \big|\mathrm{I}(Y; X \mid C) - \mathrm{I}(Y; Z \mid C)\big|. \tag{2}$$

Sufficiency holds exactly when $\Delta_{\mathrm{S}} = 0$, and approximately when Δ_{S} is small.

Localisation. The second condition deemed important for leakage to be benign is *localisaton*. For Hybrid CBMs, our main object of study, each ground-truth concept c_i is represented by a single scalar dimension $\widehat{c}_i$, and the model generates a single, explicit residual representation $\mathbf{z}$. Hence, one can think of each concept representation $\widehat{c}_i$ as associated with the same shared block of residual variables $\mathbf{z}$. In this context, localisation can be formally expressed as:

$$\mathrm{I}(Y; C_i \mid C_{-i}) \approx \mathrm{I}\Big(Y; \widehat{C}_i \mid Z, \widehat{C}_{-i}\Big), \tag{3}$$

where $\widehat{C}$ is the r.v. associated with the concept representations $\widehat{\mathbf{c}}$, and the subscript $-i$ denotes all components in a variable except i. Intuitively, this condition requires all information about a known concept c_i to be localised into its corresponding representation $\widehat{c}_i$, rather than spread across the shared residual variables $\mathbf{z}$. Similarly to sufficiency, we define the localisation discrepancy for concept i as

$$\Delta_{\mathrm{L},i} = \big|\mathrm{I}(Y; C_i \mid C_{-i}) - \mathrm{I}(Y; \widehat{C}_i \mid Z, \widehat{C}_{-i})\big|. \tag{4}$$

Localisation holds exactly when $\Delta_{\mathrm{L},i} = 0$, and approximately when $\Delta_{\mathrm{L},i}$ is small.

3.3 Intervention Harm

To better study intervenability under benign leakage, it is helpful to first define intervention harm:

Definition 1 (Intervention harm)**.** Consider a (potentially residual) CM and let $(\widehat{\mathbf{c}}, \mathbf{z})$ denote the model-predicted representation before intervention. For any subset of known concepts $S \subseteq \mathcal{C}$, define the oracle-intervened concept representation $\widehat{\mathbf{c}}^{(S)}$ by the vector of concept representations where

$$\widehat{c}_i^{(S)} := \begin{cases} c_i, & i \in S, \\ \widehat{c}_i, & i \notin S, \end{cases}$$

Here we assume that, if the CM is a residual model, the residual representation $\mathbf{z}$ is held fixed during intervention. Notice that, by construction, $\widehat{\mathbf{c}}^{(\varnothing)} = \widehat{\mathbf{c}}$. For a measurable task loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, define the *intervention task risk* $\mathcal{R}_S$ as

$$\mathcal{R}_S := \mathbb{E}_{(\mathbf{x},\mathbf{c},y)\sim P_{\mathcal{D}}}\Big[L\Big(y, f(\widehat{\mathbf{c}}^{(S)}, \mathbf{z})\Big)\Big], \tag{5}$$

where $P_{\mathcal{D}}$ is the data distribution. Assuming $\mathcal{R}_\varnothing$ is finite, the expected intervention harm associated with correcting the concepts in S is therefore defined as $\eta(S) := \mathcal{R}_S - \mathcal{R}_\varnothing$. We say that intervention on S is *harmful* if $\eta(S) > 0$ and *safe* if $\eta(S) \leq 0$. When $S \subsetneq \mathcal{C}$, $\eta(S)$ describes *partial intervention harm*; when $S = \mathcal{C}$, it describes full, or *endpoint*, harm. By construction, $\eta(\varnothing) = 0$.

3.4 Sufficiency and Localisation Do Not Prevent Intervention Harm

We now consider whether sufficiency and localisation are indeed *sufficient* to prevent intervention harm and thus prescribe benign leakage. We find that for a concept classifier with non-zero predictive error (the setting where concept interventions are warranted), intervention harm can be arbitrarily large even when sufficiency and localisation conditions hold arbitrarily well:

Theorem 1 (Interventions can be arbitrarily harmful under approximate benign leakage)**.** *For every $p \in (0, 1/2)$ and every $n > 0$, there exists a finite data-generating process, concept representations $\widehat{\mathbf{c}} \sim \widehat{C}$, residuals $\mathbf{z} \sim Z$ such that:* ***(i)*** *$\widehat{C}_i$ is misaligned with ground-truth concept C_i with probability p;* ***(ii)*** *localisation holds exactly but sufficiency discrepancy is $\Delta_{\mathrm{S}} = \mathfrak{h}(p)$, where $\mathfrak{h}(p) := -p \log p - (1-p)\log(1-p)$ is the binary entropy;* ***(iii)*** *the prediction of the optimal task classifier f is incorrect with probability p; and,* ***(iv)*** *replacing $\widehat{C}$ by its ground-truth value C, while holding Z fixed, strictly increases expected squared loss:*

$$\mathbb{E}\Big[\big(Y - f(C, Z)\big)^2\Big] - \mathbb{E}\Big[\big(Y - f(\widehat{C}, Z)\big)^2\Big] \geq np > 0. \tag{6}$$

Equivalently, this implies that, for every $\varepsilon > 0$ and every $M > 0$, there exists an example that satisfies

$$\Delta_{\mathrm{S}} < \varepsilon, \qquad \Delta_{\mathrm{L},1} = 0, \qquad \eta(\mathcal{C}) > M. \tag{7}$$

See Appendix A for a proof. Theorem 1 shows that localisation and arbitrarily approximate sufficiency are not sufficient conditions to guarantee benign leakage, and intervention harm can be arbitrarily large in their presence. The key proof construction shows that correcting $\widehat{C}$ while holding Z fixed can change how the task head uses the joint representation, relocating rather than removing prediction error. Crucially, this mechanism extends to partial intervention, because it arises from a concept-correction transition rather than a property of endpoint concept intervention, motivating the study of *partial* intervention trajectories.

4 Minimising Partial Intervention Harm

Recently, Espinosa Zarlenga (2026) showed that, under certain idealised conditions, sufficiency can be maximised by minimising the intervention loss $\mathcal{L}_{\text{int}} = \mathcal{R}_{\mathcal{C}}$, where $\mathcal{R}_{\mathcal{C}}$ is the task risk after ground-truth intervention on all supervised concepts. Minimising $\mathcal{L}_{\text{int}}$ has also been shown to improve model performance following interventions. However, our theoretical results show that sufficiency and localisation are not sufficient to prevent partial intervention harm. Moreover $\mathcal{L}_{\text{int}}$, places no direct constraint on intermediate states in which only a subset of concepts has been corrected. Hence, while it may aid in preventing intervention harm when all concepts are intervened, which we refer to as *endpoint* intervention harm, it cannot be expected to prevent *partial* intervention harm.

As we discuss in greater detail in Section 5, we empirically observe that $\mathcal{L}_{\text{int}}$ alone is not sufficient to prevent partial intervention harm in incomplete settings. Since partial interventions are the expected use case for CMs in practice, it is important to understand what training incentives may eliminate partial intervention harm. Therefore, we argue that a stronger requirement to limit partial intervention harm is *monotonicity*, where correcting an additional concept should not increase task risk. This implies that, for every $S \subseteq \mathcal{C}$ and $i \in \mathcal{C} \setminus S$, we want $\mathcal{R}_{S\cup\{i\}} - \mathcal{R}_S \leq 0$. If this condition holds for every such pair, risk is non-increasing along every sequence of concept corrections; hence, no partial or full interventions can be harmful.

Monotonic Intervention Loss. Inspired by this notion, we introduce the *monotonic intervention loss* $\mathcal{L}_{\text{mono}}$. To understand this loss, consider an *intervention edge* $S \to S \cup \{i\}, S \subsetneq \mathcal{C}, i \in \mathcal{C} \setminus S$, where S is a set of already-intervened concepts and i is an additional concept to be corrected. For each example, define $\ell_S(\widehat{\mathbf{c}}, \mathbf{z}, \mathbf{c}, y) := L(y, f(\widehat{\mathbf{c}}_{\text{sg}}^{(S)}, \mathbf{z}))$, where $\text{sg}(\cdot)$ denotes the stop-gradient operator and $\widehat{\mathbf{c}}_{\text{sg}}^{(S)}$ is constructed from $\text{sg}(\widehat{\mathbf{c}})$. Our loss is then given by:

$$\mathcal{L}_{\text{mono}} = \mathbb{E}_{(S,i)\sim q}\mathbb{E}_{(\mathbf{x},\mathbf{c},y)\sim P_{\mathcal{D}}}\left[\ell_{S\cup\{i\}}\left(g(\mathbf{x}), \mathbf{c}, y\right) - \ell_S\left(g(\mathbf{x}), \mathbf{c}, y\right)\right]_+ \tag{8}$$

where $[u]_+ := \max(u, 0)$, $q(S, i)$ specifies a distribution over intervention edges and we use $\ell_S\left(g(\mathbf{x}), \mathbf{c}, y\right)$ to represent $\ell_S\left(\widehat{\mathbf{c}}, \mathbf{z}, \mathbf{c}, y\right)$ when $g(\mathbf{x}) = (\widehat{\mathbf{c}}, \mathbf{z})$. Thus, $\mathcal{L}_{\text{mono}}$ is zero whenever correcting one additional concept maintains or improves task performance and positive otherwise.

Notice that we stop gradients through $\widehat{\mathbf{c}}$ when optimising $\mathcal{L}_{\text{mono}}$. This prevents the model from reducing the monotonicity penalty by changing the supervised concept predictions themselves, which could trade concept fidelity for intervention compatibility. Gradients are retained through the residual representation $\mathbf{z}$ and task head, allowing the downstream representation and prediction function to adapt so that ground-truth concept corrections are interpreted consistently.

Generating Intervention Trajectories. Computing L_{mono} becomes intractable with a large concept set, as the number of intervention edges increases exponentially with k. Hence, during training we estimate Eq. 8 by Monte Carlo sampling. Specifically, for each minibatch we sample m_{mono} intervention edges independently from $q(S, i)$, where m_{mono} is a training hyperparameter, and average their monotonicity penalties. In our experiments, we sample intervention depth uniformly: first $|S| \sim \text{Uniform}\{0, \ldots, |\mathcal{C}| - 1\}$, then a random concept ordering, taking the first $|S|$ concepts to form S and the next concept as i. Consequently, all intervention depths, from the first concept correction to the final correction, are represented during training.

Integrating $\mathcal{L}_{\text{mono}}$. Our loss can be integrated into the traditional CBM training objective (Koh et al., 2020), extended to include $\mathcal{L}_{\text{int}}$, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{mono}} \mathcal{L}_{\text{mono}}. \tag{9}$$

Here, $\lambda_c, \lambda_{\text{int}}, \lambda_{\text{mono}} \geq 0$ control the contribution of concept prediction, fully intervened task performance, and monotonic partial-intervention performance, respectively. Intuitively, $\mathcal{L}_{\text{int}}$ encourages good performance at the fully corrected endpoint, whereas $\mathcal{L}_{\text{mono}}$ directly penalises harmful transitions between intervention states. As we show next, these two losses complement each other.

5 Experiments

In this section, we empirically show that partial intervention harm can occur even when CMs are trained with $\mathcal{L}_{\text{int}}$. Then, we show that introducing the $\mathcal{L}_{\text{mono}}$ loss defined in Section 4 can significantly mitigate these effects and is a complementary objective mitigating different aspects of harm w.r.t. to $\mathcal{L}_{\text{int}}$. To do this, we explore Hybrid CBMs with varying residual capacity on both synthetic datasets (Section 5.1), where we fully control the data generation process, and real-world datasets (Section 5.2). Across all experiments, we simulate incompleteness by controlling the size of the set of known concepts $\mathcal{C}$ used when training a Hybrid CBM. Appendix B provides details on model configurations, hyperparameters, and datasets.

5.1 Understanding Intervention Harm, Leakage, and Incompleteness

Setup. We first study intervention harm using a controlled linear–Gaussian model adapted from Parisini et al. (2026) (Appendix B.1). In this setup, models are provided with $k = 5$ supervised concepts while the task depends on some additional $k' \in \{2, 5, 10, 15\}$ unobserved concepts. We vary the relative unobserved concept task contribution via a parameter $\omega \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ (higher ω means higher task-relevance of unobserved concepts), and generate datasets using three concept covariance structures (all concepts independent, block independence of observed and unobserved concepts, and random concept covariance). For each configuration, we train Hybrid CBMs with different λ_{int} and λ_{mono}, and with and w/o stop-gradient through $\widehat{\mathbf{c}}$, on five seeds. When studying harm, we exhaustively evaluate all 2^5 intervention states and $5!$ correction orderings.

As unobserved concepts become more important, intervention harm increases. Our results in Fig. 1A, where we show the association of each main model component to observed intervention harm, suggest that greater task dependence on unobserved concepts is associated with increased intervention harm. Specifically, we see that, in adjusted analyses, increasing ω was associated with greater endpoint harm (blue, $\beta = 0.368$, 95% CI 0.304–0.431, $P < 10^{-28}$) and a higher fraction of harmful partial-correction paths (green, $\beta = 0.239$, 95% CI 0.200–0.278, $P < 10^{-32}$), but not with maximum partial harm (orange, $\beta = -0.007$, 95% CI -0.027–0.013, $P = 0.47$). This suggests that incompleteness plays a significant role in the presence of both partial and endpoint intervention harm.

$\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$ are complementary losses. Interestingly, our results in Figs. 1B–D suggest that $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$ control complementary components of harm. Without either objective, both endpoint and partial intervention harm are positive (Fig. 1B–D) and 84% of intervention paths are harmful. Adding $\mathcal{L}_{\text{int}}$ makes endpoint harm negative, but positive partial harm persists. Adding $\mathcal{L}_{\text{mono}}$ to $\mathcal{L}_{\text{int}}$

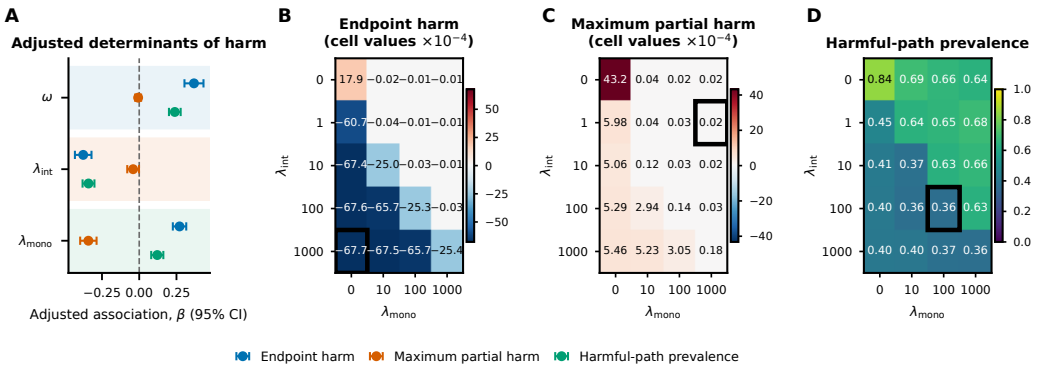

Figure 1: Endpoint ($\mathcal{L}_{\text{int}}$) and monotonic ($\mathcal{L}_{\text{mono}}$) objectives control complementary components of intervention harm. **(A)** Adjusted standardised associations of ω, λ_{int} and λ_{mono} with endpoint harm, maximum partial harm, and harmful-path prevalence. Points show regression coefficients with 95% confidence intervals. **(B)** Mean endpoint harm, **(C)** maximum partial harm, and **(D)** harmful-path prevalence across λ_{int} and λ_{mono}. In these plots, we highlight the cell with the lowest mean.

reduces maximum partial harm dramatically by up to 99.7%, however $\mathcal{L}_{\text{mono}}$ alone cannot reduce endpoint harm as effectively as $\mathcal{L}_{\text{int}}$. Consistent with this complementarity, increasing λ_{int} primarily reduces endpoint harm ($\beta = -0.376$, $P < 10^{-39}$), whereas increasing λ_{mono} primarily reduces maximum partial harm ($\beta = -0.341$, $P < 10^{-33}$). Hence, balancing both objectives provides a strong reduction in general intervention harm.

Intervention-aware losses do not affect predictive performance. Finally, we see that the improvements in intervention harm do not need to come at the cost of either concept accuracy or task accuracy. As discussed in more detail in Appendix B.1.6, with moderate values of λ_{int} and λ_{mono}, the task accuracy changes by at most 0.043%. Moreover, the concept mean squared error (MSE) remains unaffected when $\mathcal{L}_{\text{mono}}$ uses its stop-gradient mechanism (Appendix B.1.6).

Together, these experiments show that concept incompleteness produces distinct endpoint- and partial-intervention failure modes, which can be mitigated by jointly optimising $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$.

5.2 Intervention Harm in Real-World Tasks

Setup. We next test whether our findings above generalise on CUB (Wah et al., 2011), a real-world vision dataset containing 11,788 images from 200 bird species and 312 binary image-level concepts. To study the effect of concept incompleteness in this task, we construct 80 sets containing 50 concepts and quantify the task-relevant information left unexplained by each set using a pre-model incompleteness score, $\Omega_{\mathcal{C}}^{\text{pre}}$ (Appendix B.2). For each set, we train three Hybrid CBMs: one without intervention losses (*baseline*), one trained with $\mathcal{L}_{\text{int}}$ (*endpoint*), and one trained with $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$. Because exhaustively evaluating all 2^{50} intervention states is infeasible, we evaluate all models using the same 512 randomly sampled intervention orderings (see Appendix B.2 for details).

Incompleteness strongly predicts the presence of harm in real-world data. As in our synthetic experiments, our results in Figs. 2A–C show a strong association between dataset incompleteness (captured by $\Omega_{\mathcal{C}}^{\text{pre}}$) and both endpoint and partial-intervention harm. Specifically, we observe that increasing $\Omega_{\mathcal{C}}^{\text{pre}}$ is associated with greater endpoint harm ($\beta = 0.133$, 95% CI 0.084–0.181, $P < 10^{-7}$), maximum partial harm ($\beta = 0.078$, 95% CI 0.036–0.121, $P < 10^{-3}$), and harmful-path prevalence ($\beta = 0.194$, 95% CI 0.134–0.254, $P < 10^{-9}$). More importantly, we see that this harm is strongest when the Hybrid CBM is trained without any intervention-aware loss. As real-world tasks are likely to be incomplete, these results suggest that, without proper conditioning (e.g., $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$ together), CMs are likely to exhibit both partial and endpoint intervention harm in practice.

$\mathcal{L}_{\text{int}}$ is not sufficient to avoid partial intervention harm. Our results in Fig. 2A show that training with $\mathcal{L}_{\text{int}}$ alone reduced mean endpoint harm from 0.0198 to −0.0019. However, mean maximum partial harm remained 0.0054, and, crucially, 91.2% of sampled correction paths contained a harmful intermediate state (Fig. 2B–C). When looking at representative intervention trajectories

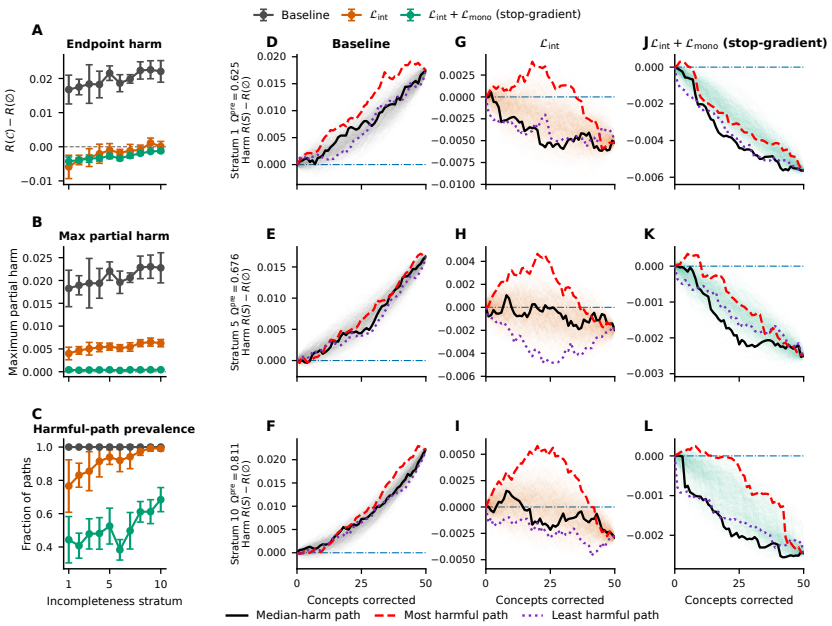

Figure 2: $\mathcal{L}_{\text{mono}}$ suppresses harmful partial intervention trajectories on CUB. **(A)** Mean endpoint harm, **(B)** maximum partial harm, and **(C)** harmful-path prevalence across ten concept incompleteness strata (higher means higher incompleteness Ω_C^{pre}). We show Hybrid CBMs trained (1) w/o intervention-aware losses (*baseline*) , (2) with $\mathcal{L}_{\text{int}}$, and (3) with $\mathcal{L}_{\text{int}} + \mathcal{L}_{\text{mono}}$. Points show means, and bars show 95% confidence intervals. **(D–L)** We show partial-intervention trajectories from **(D, G, J)** low-, **(E, H, K)** intermediate-, and **(F, I, L)** high-incompleteness strata. Columns show intervention paths for different modalities: **(D–F)** baseline, **(G–I)** $\mathcal{L}_{\text{int}}$ only, and **(J–L)** $\mathcal{L}_{\text{int}} + \mathcal{L}_{\text{mono}}$. Faint lines show all 512 sampled intervention paths, with the median, most harmful, and least harmful paths highlighted. Harm is measured relative to the observational state, $R(S) - R(\varnothing)$ (i.e., higher is worse).

across different incompleteness levels, shown in Figs. 2D–I, we see that $\mathcal{L}_{\text{int}}$ yields a harmless fully corrected state while still exhibiting partial intervention harm across several intermediate paths.

In contrast, Figs. 2B–C show that adding $\mathcal{L}_{\text{mono}}$ (1) reduces mean maximum partial harm to 0.0004, a 92.6% reduction relative to training with $\mathcal{L}_{\text{int}}$ alone, and (2) reduces harmful-path prevalence from 91.2% to 51.4%. In matched comparisons, the reductions are 0.0050 in maximum partial harm (95% CI 0.0047–0.0053, $P < 10^{-49}$) and 39.9 percentage points in harmful-path prevalence (95% CI 37.5–42.2, $P < 10^{-47}$). Corresponding trajectories show that $\mathcal{L}_{\text{mono}}$ substantially suppresses harmful intermediate excursions (Figs. 2G–L). All of these results suggest that, as in our synthetic task, $\mathcal{L}_{\text{int}}$ cannot guarantee no partial intervention harm (or benign leakage). Therefore, optimising across different intervention trajectories, as $\mathcal{L}_{\text{mono}}$ does, brings benefits beyond those brought by $\mathcal{L}_{\text{int}}$.

Model performance is preserved across training regimes. We further see that introducing intervention-aware losses has little effect on task accuracy: mean top-1 accuracy is 77.94%, 78.21%, and 78.11% for baseline, $\mathcal{L}_{\text{int}}$-only, and $\mathcal{L}_{\text{int}} + \mathcal{L}_{\text{mono}}$ models, respectively (see Appendix Fig. 7A). Although accuracy is slightly lower with combined training than with $\mathcal{L}_{\text{int}}$ alone (paired difference -0.10 percentage points, 95% CI -0.16 to -0.04, $P = 0.0024$), it remained above baseline. Greater incompleteness is also associated with modestly reduced task accuracy ($\beta = -0.186$, 95% CI -0.347–-0.025, $P = 0.024$; Appendix Fig. 7A). In terms of concept predictive performance, we see that adding $\mathcal{L}_{\text{mono}}$ reduces the concept MSE of endpoint training from 0.1558 to 0.1461 (paired reduction 0.0097, 95% CI 0.0087–0.0107, $P < 10^{-30}$). Nevertheless, this concept MSE remains above the baseline value of 0.1410 (Appendix Fig. 7B).

Taken together, these results confirm, on real-world data, that concept incompleteness is associated with endpoint and partial intervention harm, and that these are distinct failure modes. More importantly, however, our results strongly suggest that $\mathcal{L}_{\text{int}}$ reduces endpoint harm, whereas $\mathcal{L}_{\text{mono}}$ directly suppresses harmful partial interventions, suggesting both are required to encourage benign leakage.

6 Discussion and Conclusion

Connecting Theoretical and Empirical Results. Our results establish two broader findings. First, we prove that sufficiency and localisation can hold even when interventions produce arbitrarily large intervention harm. Second, we empirically complement these results by showing, in both synthetic and real-world tasks, that preventing endpoint harm via $\mathcal{L}_{\text{int}}$ is insufficient to ensure safe partial interventions. To address this, we introduce $\mathcal{L}_{\text{mono}}$, a loss that mitigates partial intervention harm by minimising the CM's loss across distinct intervention trajectories. Contrary to previous intervention-aware losses that simulate training interventions (Espinosa Zarlenga et al., 2023b), our loss avoids updating the concept representations (thereby focusing on adapting the label predictor to reduce partial harm), and does not require training an expensive sampling model. Importantly, our results show that $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{mono}}$ capture complementary aspects of harm: while $\mathcal{L}_{\text{int}}$ reduces harm from total concept corrections, $\mathcal{L}_{\text{mono}}$ reduces the prevalence and magnitude of harmful *intervention paths*.

Partial Interventions and Out-of-Distribution Concept States. Sufficiency and localisation constrain how task-relevant information is represented, but not how the CM's task predictor responds to modifications of its input representations. This distinction is key in concept incompleteness, where the CM must rely on information represented outside the supervised concept set to maintain task performance. There, a CM may learn dependencies between the concept and residual representations that are disrupted when intervention changes $\widehat{\mathbf{c}}$ without updating $\mathbf{z}$, driving the joint representations $(\widehat{\mathbf{c}}, \mathbf{z})$ out-of-distribution (OOD). Crucially, such OOD intermediate states can occur even in completeness: intervening on a single concept without intervening on correlated or dependent concepts could generate an OOD concept vector $\widehat{\mathbf{c}}$ and hamper the predictor's performance. This distributional shift is key to understanding why intervention harm is more likely as incompleteness increases, and why $\mathcal{L}_{\text{int}}$ struggles to eradicate partial intervention harm. As such, $\mathcal{L}_{\text{mono}}$ provides a mechanism for a CM to learn to better deal with potential OOD intermediate latent spaces, thereby significantly reducing partial intervention harm.

Lastly, our results suggest a connection between intervention harm and shortcut learning (Geirhos et al., 2020; Marconato et al., 2023). If interventions increase task loss, this suggests the predictor has learned dependencies on the *particular* joint distribution of predicted concepts and residual information, which may include errors in $\widehat{\mathbf{c}}$. This behaviour is consistent with shortcut learning where the predictor may exploit potentially spurious correlations between $\widehat{\mathbf{c}}$ and $\mathbf{z}$ (or within $\widehat{\mathbf{c}}$) that are predictive on the training distribution, but brittle under intervention. We hypothesise that $\mathcal{L}_{\text{mono}}$ mitigates this behaviour by exposing the predictor during training to diverse combinations of correct and erroneous concept predictions, thereby reducing its reliance on brittle dependencies and encouraging predictions that remain robust across interventions.

Limitations and Future Work. We bring forth three limitations of our work. First, intervention trajectories in $\mathcal{L}_{\text{mono}}$ are randomly sampled, generating a disconnect between how we train them and how they are used afterwards (where they may be intervened according to some policy). This motivates future work in exploring harm and monotonic training under varying policies. Second, some partial intervention trajectories may result in implausible (i.e., OOD) intermediate states that could cause partial harm, even if full interventions would not be harmful. Hence, future work could bias sampling in $\mathcal{L}_{\text{mono}}$ toward plausible states. Finally, sampling paths with equal probability may fail to cover distinct but meaningful regions of the intervention lattice. Therefore, future research can explore how to more efficiently sample paths that cover most important regions of the intervention lattice.

Conclusion. In this paper, we argued that intervention harm, and in particular partial intervention harm, is an understudied failure mode of CMs, directly undermining a key use case of these models. We proved that *sufficiency* and *localisation*, criteria previously suggested to encourage so-called *benign leakage*, cannot guarantee that concept interventions improve downstream performance, and showed, across synthetic and real-world tasks, that concept-set incompleteness is associated with both endpoint and partial intervention harm. Moreover, we found that $\mathcal{L}_{\text{int}}$, a previous objective to encourage benign leakage, does not reliably prevent harmful partial corrections. To address this, we introduced $\mathcal{L}_{\text{mono}}$, a complementary objective that substantially reduces the prevalence and magnitude of partial intervention harm while largely preserving concept and task performance. Overall, our results suggest that whether leakage is benign should be judged not only by the fully intervened endpoint, but also by considering all partial intervention paths an overseer may take.

AI USE STATEMENT

In this work, we used generative AI tools to: help develop theoretical models and conceptual frameworks, refine hypotheses, design and provide feedback on research methodology and experiments, support interpretation of results, and to assist in the formulation and writing of mathematical proofs. We did not use generative AI tools to generate synthetic data sets, implement methods, provide critical ingredients for proving mathematical claims, clean or reformat datasets, perform qualitative or thematic data analysis or assist with translation. Additionally, we used generative AI tools to create and edit software code and draft parts of the manuscript. We have reviewed all AI-assisted work substantially: we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by multiple authors, all mathematical proofs and claims were reviewed and corrected by multiple authors and all LLM-generated text was reviewed and edited by multiple authors. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates the safety and reliability of human intervention in concept-based models. The experiments use synthetic data and the established `CUB` benchmark and do not involve human participants, personal data, or sensitive attributes. No new datasets are collected or released. We have considered the potential for both beneficial and unintended uses of the proposed methodology and believe that explicitly characterising intervention harm contributes to the responsible development and evaluation of human-interpretable AI systems.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of our theoretical results in the Appendix. Detailed experimental methods are provided for both the synthetic and `CUB` experiments, including data generation and preprocessing, dataset partitions, model architectures, training objectives, optimisation procedures, hyperparameters, intervention procedures, experimental designs, evaluation metrics, and statistical analyses. Random seeds and the procedures used to construct and match experimental conditions are specified where applicable. We additionally provide anonymised source code sufficient to reproduce the experiments, analyses, and figures reported in this work as supplementary material.

REFERENCES

Antonio Almudévar, José Miguel Hernández Lobato, and Alfonso Ortega. There was never a bottleneck in concept bottleneck models. In *International Conference on Learning Representations*, volume 2026, pp. 114146–114165, 2026.

David Alvarez-Melis and Tommi S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks, December 2018. URL `http://arxiv.org/abs/1806.07538`. arXiv:1806.07538 [cs.LG].

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive Concept Bottleneck Models, April 2023. URL `http://arxiv.org/abs/2212.07430`. arXiv:2212.07430 [cs.LG].

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence*, 2(12):772–782, December 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00265-z. URL `http://arxiv.org/abs/2002.01650`. arXiv:2002.01650 [cs.LG].

Antonio De Santis, Schrasing Tong, Marco Brambilla, and Lalana Kagal. Learning concept bottleneck models from mechanistic explanations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Eric Enouen and Sainyam Galhotra. Debugging Concept Bottleneck Models through Removal and Retraining, March 2026. URL `http://arxiv.org/abs/2509.21385`. arXiv:2509.21385 [cs.CV].

Mateo Espinosa Zarlenga. In Defense of Information Leakage in Concept-based Models, June 2026. URL http://arxiv.org/abs/2606.10669. arXiv:2606.10669 [cs.LG].

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Zohreh Shams, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, and Mateja Jamnik. Towards robust metrics for concept representation evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11791–11799, 2023a.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023b.

Mateo Espinosa Zarlenga, Gabriele Dominici, Pietro Barbiero, Zohreh Shams, and Mateja Jamnik. Avoiding leakage poisoning: Concept interventions under distribution shifts. *arXiv preprint arXiv:2504.17921*, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic Concept Bottleneck Models, June 2023. URL http://arxiv.org/abs/2306.01574. arXiv:2306.01574 [cs.LG].

Patrick Knab, David Steinmann, Christian Bartelt, Kristian Kersting, Bernt Schiele, Thomas Seidl, Udo Schlegel, and Wolfgang Stammer. What's in the bottle? a survey and roadmap of concept bottleneck models. *Transactions on Machine Learning Research*, 2026.

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

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

Mikael Makonnen, Moritz Vandenhirtz, Sonia Laguna, and Julia E. Vogt. Measuring Leakage in Concept-Based Methods: An Information Theoretic Approach, July 2025. URL http://arxiv.org/abs/2504.09459. arXiv:2504.09459 [cs.LG].

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretable, leak-proof concept-based models. *Advances in neural information processing systems*, 35:21212–21227, 2022.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. *Advances in Neural Information Processing Systems*, 36:72507–72539, 2023.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-Free Concept Bottleneck Models, June 2023. URL http://arxiv.org/abs/2304.06129. arXiv:2304.06129 [cs.LG].

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D MacArthur, and Christopher RS Banerji. Leakage and interpretability in concept-based models. *arXiv preprint arXiv:2504.14094*, 2025.

Enrico Parisini, Christopher J Soelistyo, Ahab Isaac, Alessandro Barp, and Christopher RS Banerji. Concept frustration: Aligning human concepts and machine representations. *arXiv preprint arXiv:2603.29654*, 2026.

Andrea Pugnana, Riccardo Massidda, Francesco Giannini, Pietro Barbiero, Mateo Espinosa Zarlenga, Roberto Pellungrini, Gabriele Dominici, Fosca Giannotti, and Davide Bacciu. Deferring concept bottleneck models: Learning to defer interventions to inaccurate experts. *Advances in Neural Information Processing Systems*, 38:67989–68044, 2026.

Angelos Ragkousis and Sonali Parbhoo. Tree-based leakage inspection and control in concept bottleneck models. *arXiv preprint arXiv:2410.06352*, 2024.

Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10932–10941. IEEE, 2023.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

Yoshihide Sawada and Keigo Nakamura. Concept Bottleneck Model with Additional Unsupervised Concepts, February 2022. URL `http://arxiv.org/abs/2202.01459`. arXiv:2202.01459 [cs.CV].

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International conference on machine learning*, pp. 31504–31520. PMLR, 2023.

Nishad Singhi, Jae Myung Kim, Karsten Roth, and Zeynep Akata. Improving intervention efficacy via concept realignment in concept bottleneck models. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

Julian Skirzynski, Harry Cheon, Shreyas Kadekodi, Meredith Stewart, and Berk Ustun. Measuring what matters: Synthetic benchmarks for concept bottleneck models. *arXiv preprint arXiv:2606.04326*, 2026.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. *arXiv preprint arXiv:2308.13453*, 2023.

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

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *International Conference on Learning Representations*, volume 2024, pp. 6944–6966, 2024.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197. IEEE, 2023.

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

A Proof of Theorem 1

We first restate the theorem more expansively:

Theorem 1 (Interventions can be arbitrarily harmful under approximate benign leakage)**.** *For every $p \in (0, 1/2)$ and every $n > 0$, there exists a finite data-generating process, concept representations $\widehat{\mathbf{c}} \sim \widehat{C}$, residuals $\mathbf{z} \sim Z$ such that:*

1. *$\widehat{C}_i$ is misaligned with ground-truth concept C_i with probability p.*
2. *Localisation holds exactly but sufficiency discrepancy is $\Delta_{\mathrm{S}} = \mathfrak{h}(p)$, where $\mathfrak{h}(p) := -p \log p - (1-p)\log(1-p)$ is the binary entropy.*
3. *The task prediction of the optimal task classifier f is incorrect with probability p.*
4. *Intervening by replacing $\widehat{C}$ by its ground-truth value C, while holding Z fixed, strictly increases expected squared loss:*

$$\mathbb{E}\Big[\big(Y - f(C, Z)\big)^2\Big] - \mathbb{E}\Big[\big(Y - f(\widehat{C}, Z)\big)^2\Big] \geq np > 0. \tag{10}$$

Equivalently, this implies that, for every $\varepsilon > 0$ and every $M > 0$, there exists an example that satisfies

$$\Delta_{\mathrm{S}} < \varepsilon, \qquad \Delta_{\mathrm{L},1} = 0, \qquad \eta(\mathcal{C}) > M. \tag{11}$$

Proof. To prove the theorem, it suffices to construct a finite generating process, representation and task head such that the stated conditions hold.

A random variable B is called a *fair coin* if

$$B \sim \mathrm{Bernoulli}(1/2).$$

Let C and U be independent fair coins, and let

$$E \sim \mathrm{Bernoulli}(p), \qquad p \in (0, 1/2),$$

be independent of (C, U). Here, C is the known concept, U is an omitted task-relevant factor, and $E = 1[C \neq \hat{C}]$ indicates whether the learned concept predictor makes an error. Fix any $n > 0$ and set

$$b := \sqrt{n+1} > 1.$$

Define

$$X = (C, U, E), \qquad Y = C + bU. \tag{12}$$

The four possible pairs (C, U) give the four distinct values

$$(0,0) \mapsto 0, \qquad (1,0) \mapsto 1, \qquad (0,1) \mapsto b, \qquad (1,1) \mapsto b+1.$$

Because $b > 1$, these values are distinct. Hence the optimal predictor $f^* : \mathcal{C} \times \mathcal{U} \to \mathcal{Y}$ is a bijection from $\{0,1\}^2$ onto $\{0, 1, b, b+1\}$, so both C and U are deterministic functions of Y.

We assume the residual satisfies $Z = \widehat{C} \oplus U$ where $\oplus$ denotes XOR (addition mod 2) so that

$$\widehat{C} = C \oplus E, \qquad Z = \widehat{C} \oplus U. \tag{13}$$

Notice $\widehat{C}$ is a fair coin. Indeed, for every $a \in \{0, 1\}$,

$$\Pr(\widehat{C} = a) = \sum_{e \in \{0,1\}} \Pr(E = e) \Pr(C = a \oplus e) = \frac{1}{2} \sum_{e \in \{0,1\}} \Pr(E = e) = \frac{1}{2}.$$

In addition, $\widehat{C}$ and Z are independent. Indeed, because U is a fair coin independent of $\widehat{C}$, since the latter is a deterministic function of random variables independent of U, for any $a, z \in \{0, 1\}$,

$$\begin{aligned} \Pr(\widehat{C} = a, Z = z) &= \Pr\big(\widehat{C} = a,\, \widehat{C} \oplus U = z\big) \\ &= \Pr\big(\widehat{C} = a,\, U = a \oplus z\big) \\ &= \Pr(\widehat{C} = a) \Pr(U = a \oplus z) \\ &= \Pr(\widehat{C} = a) \frac{1}{2}. \end{aligned} \tag{14}$$

Moreover,

$$\Pr(Z = z) = \sum_{a \in \{0,1\}} \Pr(\widehat{C} = a, Z = z) = \frac{1}{2}.$$

Hence

$$\Pr(\widehat{C} = a, Z = z) = \Pr(\widehat{C} = a) \Pr(Z = z),$$

which proves

$$\widehat{C} \perp Z. \tag{15}$$

In particular,

$$H(\widehat{C} \mid Z) = H(\widehat{C}) = \log 2. \tag{16}$$

For localisation, recall that U is a deterministic function of Y. Since

$$\widehat{C} = Z \oplus U,$$

$\widehat{C}$ is therefore a deterministic function of (Y, Z). Thus

$$H(\widehat{C} \mid Y, Z) = 0.$$

Combining this with the independence established above gives

$$\begin{aligned} I(Y; \widehat{C} \mid Z) &= H(\widehat{C} \mid Z) - H(\widehat{C} \mid Y, Z) \\ &= \log 2. \end{aligned} \tag{17}$$

Similarly, because C is a deterministic function of Y,

$$H(C \mid Y) = 0,$$

and hence

$$I(Y; C) = H(C) - H(C \mid Y) = H(C) = \log 2. \tag{18}$$

Therefore

$$\Delta_{\mathrm{L},1} = \Big| I(Y; C) - I(Y; \widehat{C} \mid Z) \Big| = 0. \tag{19}$$

Thus localisation holds exactly.

For sufficiency, note

$$I(Y; Z \mid C) = H(Y \mid C) - H(Y \mid Z, C) = \log 2 - \mathfrak{h}(p),$$

where

$$\mathfrak{h}(p) := -p \log p - (1 - p) \log(1 - p)$$

is the binary entropy, since

$$H(Y \mid C) = H(U) = \log 2$$

and

$$H(Y \mid Z, C) = H(U \mid Z) = H(U \mid U \oplus E) = \mathfrak{h}(p).$$

On the other hand,

$$I(Y; X \mid C) = H(U \mid C) - H(U \mid X, C) = \log 2 - 0.$$

Thus the sufficiency condition holds:

$$\begin{aligned} \Delta_{\mathrm{S}} &:= I(Y; X \mid C) - I(Y; Z \mid C) \\ &= \mathfrak{h}(p). \end{aligned} \tag{20}$$

We now analyse the risk and intervention harm of the optimal predictor. Under square loss, the optimal predictor $\mathcal{C} \times \mathcal{Z} \to \mathcal{Y}$ is the conditional expectation

$$f^*(c, z) = \mathbb{E}[Y \mid \hat{C} = c, Z = z] = c(1 - 2p) + p + b(z \oplus c).$$

Indeed, $Y = C + bU$, and

$$\mathbb{E}[U \mid \hat{C} = c, Z = z] = \mathbb{E}[Z \oplus \hat{C} \mid \hat{C} = c, Z = z] = z \oplus c,$$

while using the fact that $C \perp Z$ (which follows from essentially the same proof as that of $\hat{C} \perp Z$), we find

$$\mathbb{E}[C \mid \hat{C} = c, Z = z] = \mathbb{E}[c \oplus E \mid \hat{C} = c] = p + c(1 - 2p).$$

Notice the optimal predictor f^* outputs values in $\mathcal{Y}^* = \{p, 1-p, p+b, b+1-p\}$ which do not belong to $\mathcal{Y} = \{0, 1, b, b+1\}$ as $p \in (0, 1/2)$. By matching the output of f^* to the closest element in $\mathcal{Y}$, we obtain the task label predictor

$$f(c, z) := c + b(c \oplus z). \tag{21}$$

In other words, f is the projection of the optimal predictor f^* onto $\mathcal{Y}$, obtained by identifying $\mathcal{Y}^* \sim \mathcal{Y}$, which is used to predict labels.

Now, since $\widehat{C} \oplus Z = U$,

$$f(\widehat{C}, Z) = Y \quad \Longleftrightarrow \quad \widehat{C} = C \quad \Longleftrightarrow \quad E = 0,$$

and both the concept prediction and the task prediction are incorrect with probability p.

In addition,

$$f(\widehat{C}, Z) - Y = \widehat{C} - C. \tag{22}$$

Because $C, \widehat{C} \in \{0, 1\}$, this difference has squared value 1 exactly when $\widehat{C} \neq C$, or equivalently when $E = 1$. Therefore

$$\big(f(\widehat{C}, Z) - Y\big)^2 = \mathbb{1}_{\{E=1\}} = E. \tag{23}$$

It follows that the squared-error risk is

$$R_{\text{obs}} := \mathbb{E}\Big[\big(Y - f(\widehat{C}, Z)\big)^2\Big] = \mathbb{E}[E] = p. \tag{24}$$

We next verify the benign-leakage criteria.

We now perform the usual known-concept-only concept correction

$$(\widehat{C}, Z) \longmapsto (C, Z). \tag{25}$$

This intervention replaces the predicted concept by its ground-truth value but leaves the realised residual Z exactly unchanged. In particular,

$$Z_{\text{after}} = Z_{\text{before}} = \widehat{C} \oplus U = C \oplus E \oplus U.$$

The input supplied to the task head therefore changes as follows:

$$\underbrace{f(\widehat{C}, Z)}_{\widehat{C} + b(\widehat{C} \oplus Z)} \quad \longmapsto \quad \underbrace{f(C, Z)}_{C + b(C \oplus Z)}. \tag{26}$$

Although Z itself is unchanged, its decoding changes because the task head now XORs Z with C rather than with $\widehat{C}$. Specifically,

$$\begin{aligned} C \oplus Z &= C \oplus (C \oplus E \oplus U) \\ &= E \oplus U. \end{aligned} \tag{27}$$

Thus the corrected prediction is

$$f(C, Z) = C + b(E \oplus U). \tag{28}$$

If $E = 0$, then $\widehat{C} = C$, so the intervention changes neither the concept coordinate nor the prediction. In this case,

$$E \oplus U = U, \qquad f(C, Z) = C + bU = Y.$$

If $E = 1$, then the intervention correctly changes $\widehat{C} = 1 - C$ to C, but the fixed residual remains

$$Z = \widehat{C} \oplus U.$$

Consequently,

$$E \oplus U = 1 \oplus U = 1 - U,$$

so the omitted factor decoded by the task head changes from U to $1 - U$. Therefore

$$f(C, Z) - Y = \begin{cases} 0, & E = 0, \\ b(1 - U) - bU = b(1 - 2U), & E = 1. \end{cases} \tag{29}$$

Since $U \in \{0, 1\}$,

$$\big(1 - 2U\big)^2 = 1.$$

It follows pointwise that

$$\big(f(C, Z) - Y\big)^2 = \begin{cases} 0, & E = 0, \\ b^2, & E = 1, \end{cases} = b^2 E. \tag{30}$$

Taking expectations now gives, explicitly,

$$\begin{aligned} R_{\text{corr}} &:= \mathbb{E}\Big[\big(Y - f(C, Z)\big)^2\Big] \\ &= \mathbb{E}[b^2 E] = b^2 \Pr(E = 1) = b^2 p. \end{aligned} \tag{31}$$

Here, it is worth highlighting how correcting a concept differs from improving the task head: before corrections, we had $f(\hat{C}, Z) - Y = \hat{C} - C$, so that better concept predictions result in better predictions overall, while after intervention we see from above that now $f(C, Z) - Y$ only depends on U, and produces greater risk. Thus, intervention did not remove an error, but rather worsened the risk by relocating it.

Combining this with $R_{\text{obs}} = p$, we obtain

$$\eta(\mathcal{C}) := R_{\text{corr}} - R_{\text{obs}} = (b^2 - 1)p = np > 0. \tag{32}$$

The intervention harm is also positive if we use the optimal predictor f^* directly (rather than its projection f), in which case

$$\eta(\mathcal{C}) = R_{\text{corr}} - R_{\text{obs}} = \big(p^2 + b^2 p\big) - \big(p - p^2\big) = 2p^2 + np > np > 0.$$

Finally, let $\varepsilon > 0$ and $M > 0$. Since $\mathfrak{h} \to 0$ as $p \downarrow 0$, we may first choose $p \in (0, 1/2)$ sufficiently small that

$$\Delta_{\text{S}} = \mathfrak{h} < \varepsilon.$$

We may then choose $n > M/p$ and set $b = \sqrt{n + 1}$. For these choices,

$$\Delta_{\text{L},1} = 0, \qquad \eta(\mathcal{C}) \geq np > M.$$

Thus arbitrarily accurate satisfaction of the sufficiency criterion, together with exact localisation, places no finite upper bound on the harm caused by ground-truth concept correction. □

B Experiment Details

B.1 Synthetic Experiments

B.1.1 Data-Generating Process

We adapted the known–unknown concept construction introduced by Parisini et al. (2026). Each synthetic observation contained $k_K = 5$ supervised concepts and k' additional unobserved concepts:

$$\boldsymbol{\chi} = (\mathbf{c}, \mathbf{u}) \in \mathbb{R}^{5+k'}.$$

The complete concept vector was sampled as

$$\boldsymbol{\chi} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}).$$

We evaluated three covariance conditions. In the *diagonal* condition, $\mathbf{B} = \mathbf{I}$. To construct the other two conditions, we generated a random symmetric positive-definite correlation matrix. The *correlated* condition used this complete matrix. The *block-independent* condition retained its within-known

and within-unknown covariance blocks but set the cross-covariance blocks to zero. The covariance construction was matched across these conditions for each seed and value of k'.

Inputs were generated through a noisy linear map:

$$\mathbf{x} = \mathbf{A}\boldsymbol{\chi} + \boldsymbol{\epsilon}_x, \qquad \boldsymbol{\epsilon}_x \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{I}),$$

where $\mathbf{A} \in \mathbb{R}^{64\times(5+k')}$ had independently sampled Gaussian entries and we set $\sigma_x = 4$.

A Gaussian task-weight vector $\boldsymbol{\psi} = (\boldsymbol{\psi}_K, \boldsymbol{\psi}_U)$ was sampled once for each synthetic world. For a specified incompleteness parameter ω, the known and unknown components were scaled as

$$\widetilde{\boldsymbol{\psi}}_K = (1-\omega)\boldsymbol{\psi}_K, \qquad \widetilde{\boldsymbol{\psi}}_U = \omega\boldsymbol{\psi}_U.$$

Binary labels were then generated by

$$y = \mathbb{1}\left\{\widetilde{\boldsymbol{\psi}}_K^\top \mathbf{c} + \widetilde{\boldsymbol{\psi}}_U^\top \mathbf{u} + \epsilon_y > 0\right\}, \qquad \epsilon_y \sim \mathcal{N}(0, \sigma_y^2).$$

We set $\sigma_y = 4$. Thus, ω controls the relative weighting of known and unknown task directions.

Each synthetic world contained $N = 8{,}000$ observations. A seeded random split assigned 75% to development and 25% to held-out testing. The development observations were further divided into 80% training and 20% validation subsets, giving $4{,}800/1{,}200/2{,}000$ observations.

B.1.2 Hybrid CBM and Interventions

The hybrid CBM contained a linear concept encoder

$$\widehat{\mathbf{c}} = g_\phi(\mathbf{x}) \in \mathbb{R}^5,$$

a five-dimensional linear residual encoder

$$\mathbf{z} = h_\psi(\mathbf{x}) \in \mathbb{R}^5,$$

and a linear binary task head

$$y = f_\theta([\widehat{\mathbf{c}}, \mathbf{z}]).$$

Continuous synthetic concepts were supervised using mean-squared error, and binary task prediction used binary cross-entropy with logits.

For intervention set $S \subseteq \{1, \ldots, 5\}$, we constructed

$$\widehat{c}_j^S = \begin{cases} c_j, & j \in S, \\ \widehat{c}_j, & j \notin S. \end{cases}$$

The residual representation was held fixed. After training, each intervention state S was evaluated on the same held-out test set. We defined its empirical test risk as the mean task loss

$$\mathcal{R}_S = \frac{1}{n_{\text{test}}} \sum_i \ell\big(f_\theta([\widehat{\mathbf{c}}_i^S, \mathbf{z}_i]), y_i\big),$$

and intervention harm was

$$\eta(S) = \mathcal{R}_S - \mathcal{R}_\varnothing.$$

B.1.3 Training Objective

Models were trained using

$$\mathcal{L} = \frac{\mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{mono}} \mathcal{L}_{\text{mono}}}{1 + \lambda_c + \lambda_{\text{int}} + \lambda_{\text{mono}}}, \qquad \lambda_c = 1.$$

The endpoint loss $\mathcal{L}_{\text{int}}$ was the task loss after all five concept predictions had been replaced by ground truth.

For an intervention edge $S \to S \cup \{j\}$, the monotonic penalty was

$$\big[\ell_{S\cup\{j\}} - \ell_S\big]_+ .$$

m_{mono} edges were sampled per minibatch. For each edge, intervention depth was sampled uniformly from $\{0, \ldots, 4\}$, followed by a random permutation of the five concepts. The first concepts in this permutation defined S, and the next concept defined j.

When stop-gradient was enabled, the predicted concept representation was detached only within $\mathcal{L}_{\text{mono}}$. The monotonic loss could therefore update the residual encoder and task head, but could not directly update the concept encoder. The ordinary concept and task losses continued to train the concept encoder.

Models were optimised using Adam with learning rate 10^{-3}, batch size 600, and at most 300 epochs. Training continued for at least 50 epochs and stopped after 30 validation epochs without an improvement of 10^{-5}. The checkpoint with the lowest complete validation objective was retained.

B.1.4 Experimental Design

We crossed

$$\begin{aligned} k' &\in \{2, 5, 10, 15\}, \\ \omega &\in \{0.10, 0.25, 0.50, 0.75, 0.90\}, \\ \lambda_{\text{int}} &\in \{0, 1, 10, 100, 1000\}, \\ \lambda_{\text{mono}} &\in \{0, 10, 100, 1000\}, \end{aligned}$$

with the three covariance conditions and seeds $\{0, 1, 2, 3, 4\}$ and set $m_{\text{mono}} = 5$. Both stop-gradient settings were evaluated whenever $\lambda_{\text{mono}} > 0$; one canonical setting was retained when $\lambda_{\text{mono}} = 0$. The resulting factorial contained $10{,}500$ models.

To explore the impact of varying m_{mono} we fixed $\lambda_{\text{int}} = 10$, $\lambda_{\text{mono}} = 100$, the block-independent covariance condition, and stop-gradient enabled. We varied

$$m_{\text{mono}} \in \{1, 2, 5, 10, 20, 40\}$$

for $k' \in \{5, 10\}$, all five values of ω, and five seeds, yielding 300 additional models.

B.1.5 Evaluation and Statistical Analysis

All $2^5 = 32$ intervention states and all $5! = 120$ correction orderings were evaluated. Endpoint harm was

$$\eta(\mathcal{C}) = \eta(\{1, \ldots, 5\}),$$

and maximum partial harm was

$$H^{\max}_{\text{partial}} = \max_{\varnothing \subset S \subset \{1,\ldots,5\}} \eta(S).$$

For adjusted analyses, outcomes and continuous predictors were standardised. Regularisation weights were transformed as $\log_{10}(1 + \lambda)$. Additive ordinary least-squares models included k', ω, $\log_{10}(1 + \lambda_{\text{int}})$, $\log_{10}(1 + \lambda_{\text{mono}})$, and covariance condition. Standard errors were clustered by synthetic world, defined by seed, k', and covariance condition. The stop-gradient contrast was estimated only among models with $\lambda_{\text{mono}} > 0$.

B.1.6 Robustness and ablation analyses for synthetic experiments

We performed additional analyses to assess the robustness of the synthetic results and the effects of key modelling choices. Multivariable analyses confirmed the complementary effects of the two intervention objectives: increasing λ_{int} was primarily associated with reduced endpoint harm, whereas increasing λ_{mono} was primarily associated with reduced maximum partial harm (Fig. 3). Importantly, these reductions in harm were achieved with little change in predictive performance: across moderate regularisation weights ($\lambda_{\text{int}} \in \{1, 10, 100\}$ and $\lambda_{\text{mono}} \in \{10, 100\}$), task accuracy differed by at most 0.043 percentage points from the unregularised model. The effects on intervention harm were also robust across concept covariance structures (Fig. 4B–D). Stopping gradients from $\mathcal{L}_{\text{mono}}$ through the concept encoder mitigated a significant increase in concept prediction error (Fig. 4A), while the benefit of sampling additional monotonic edges largely saturated at $m_{\text{mono}} = 5$ (Fig. 5).

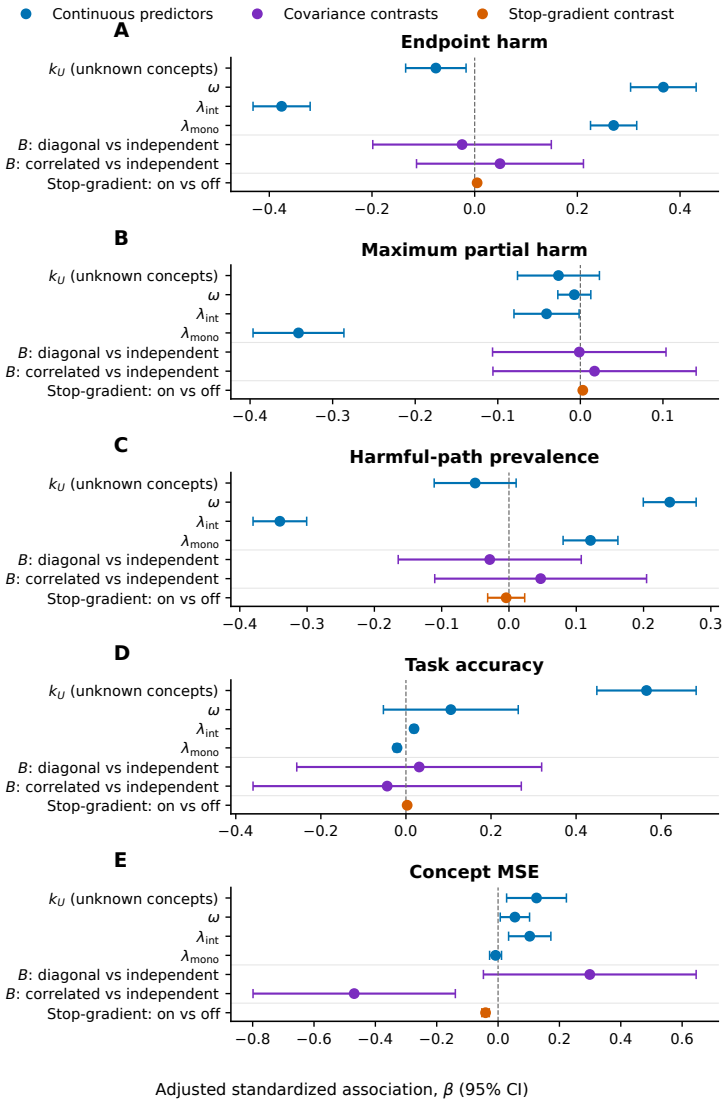

Figure 3: Adjusted determinants of intervention harm and model performance. Adjusted standardised associations with **(A)** endpoint harm; **(B)** maximum partial harm; **(C)** harmful-path prevalence; **(D)** task accuracy; and **(E)** concept mean-squared error. Points show regression coefficients and bars show 95% cluster-robust confidence intervals across 60 matched synthetic worlds. Outcomes and continuous predictors were standardised. Covariance contrasts use the block-independent construction as the reference; the stop-gradient contrast was estimated among models with $\lambda_{\text{mono}} > 0$.

B.2 Real-world Data Experiments

B.2.1 Representations, Concepts, and Data Partition

CUB contains 11,788 images from 200 bird species and 312 binary image-level attributes. We encoded each image using the pretrained CLIP ViT-B/16 image encoder and L2-normalised the resulting 512-dimensional embedding. CLIP parameters were fixed throughout.

We used one class-stratified split for the complete experiment: 7,488 images for training, 1,926 for validation, and 2,374 for held-out testing. Embeddings were standardised using feature-wise training-set means and standard deviations, with the same transformation subsequently applied to

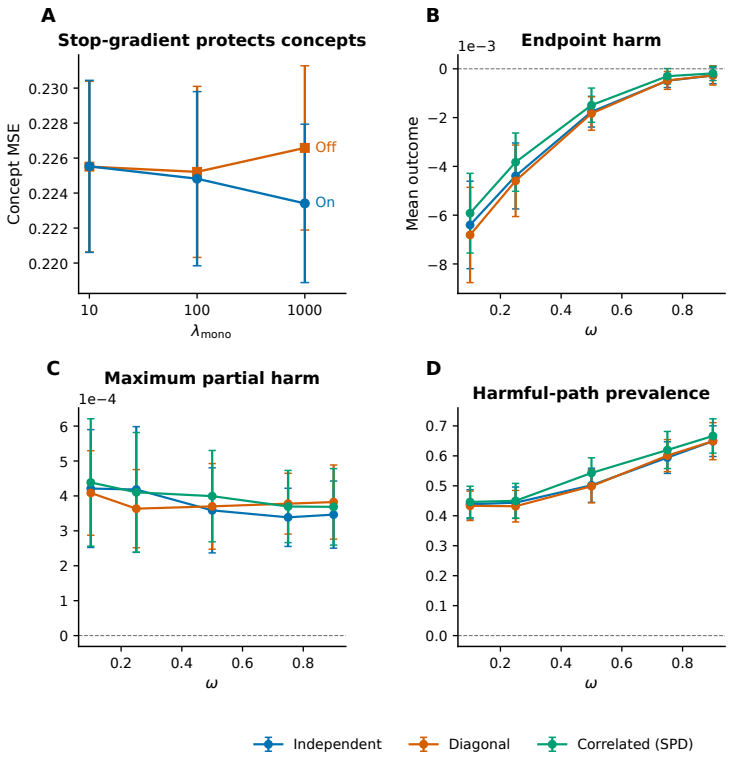

Figure 4: Robustness to stop-gradient and concept covariance structure. **(A)** Concept mean-squared error with and without stopping gradients from $\mathcal{L}_{\text{mono}}$ through the concept encoder. Values were first averaged within each of 60 matched synthetic worlds over λ_{int} and ω; points and bars show means and 95% confidence intervals across worlds. **(B–D)** Endpoint harm, maximum partial harm, and harmful-path prevalence as functions of ω under block-independent, diagonal, and correlated positive-definite covariance structures. Stop-gradient-enabled models were averaged over the regularisation grid within each matched seed–k' unit. Points and bars show means and 95% confidence intervals across 20 units per covariance condition and value of ω.

the validation and test data. Test images were not used for concept-set selection, training, or early stopping.

B.2.2 Pre-model Concept Incompleteness

To vary concept incompleteness while holding concept-set size fixed, we generated 210 candidate sets containing 50 attributes each. We first fitted a linear 200-class probe using all 312 attributes in the training partition and ranked attributes by the L2 norm of their class coefficients. This ranking was used only to diversify candidate generation.

We sampled 30 unique concept sets at 7 importance biases

$$b \in \{-6, -3, -1, 0, 1, 3, 6\}.$$

Let $r_j \in [0, 1]$ denote the normalised importance rank of attribute j, with zero denoting the most highly ranked attribute, we defined the attribute sampling weight as

$$p_j \propto \exp(-br_j).$$

Positive biases therefore favoured more task-important attributes, negative biases favoured less important attributes, and $b = 0$ produced uniform sampling.

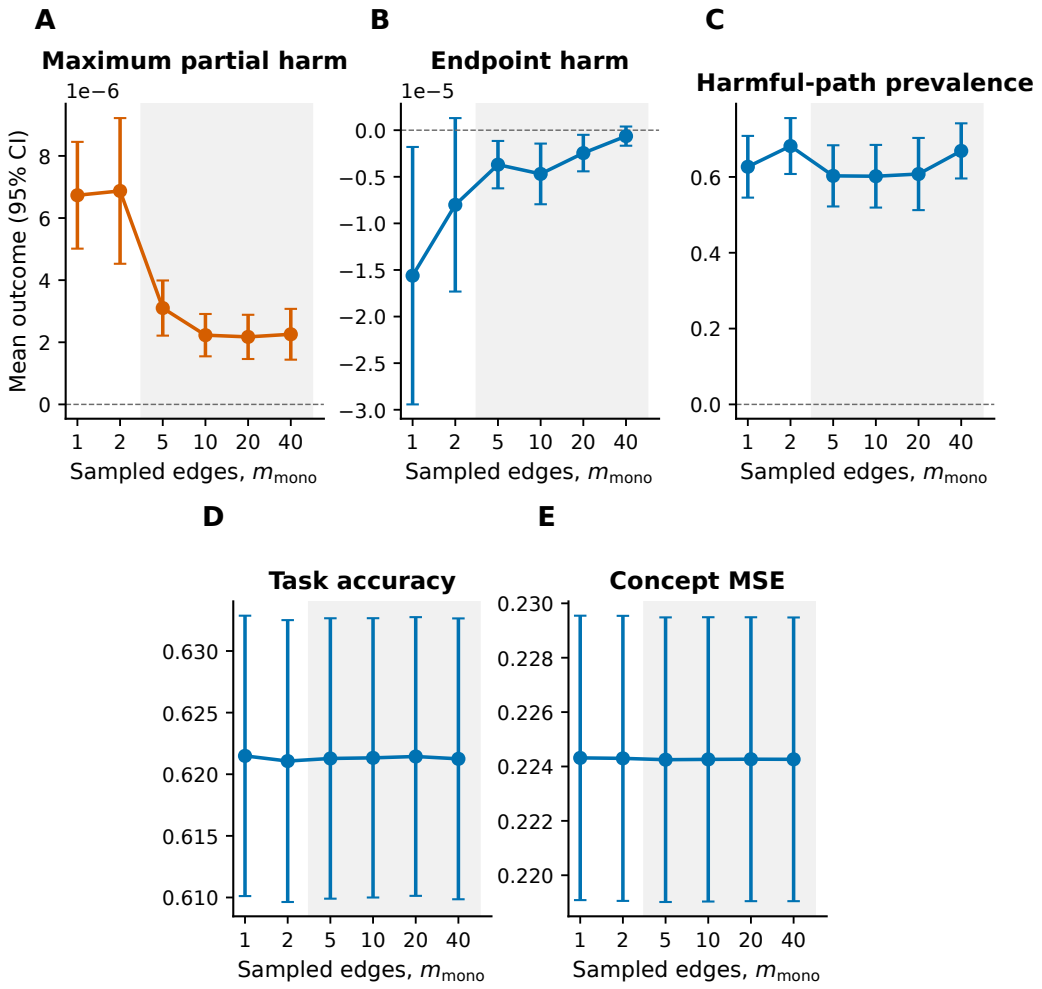

Figure 5: Effect of the number of monotonic edges sampled per minibatch. Mean **(A)** maximum partial harm; **(B)** endpoint harm; **(C)** harmful-path prevalence; **(D)** task accuracy; and **(E)** concept mean-squared error as the number of sampled monotonic edges m_{mono} was varied. The ablation fixed $\lambda_{\text{int}} = 10$, $\lambda_{\text{mono}} = 100$, block-independent concept covariance, and stop-gradient. Points and bars show means and 95% confidence intervals across 50 models spanning two values of k', five values of ω, and five seeds. Shading denotes $m_{\text{mono}} \geq 5$, beyond which the reduction in maximum partial harm largely plateaued.

For each candidate concept set $\mathcal{C}$, we fitted a separate linear 200-class classifier using its 50 ground-truth concepts and defined pre-model incompleteness as

$$\Omega_{\mathcal{C}}^{\text{pre}} = \frac{\widehat{H}(Y \mid C)}{\widehat{H}(Y)}.$$

The numerator was the classifier's best validation cross-entropy and the denominator was the validation cross-entropy of a null classifier using fitting-set class frequencies. This is a normalised probe-based measure rather than a literal information-theoretic fraction; higher values indicate that the selected concepts leave more task-relevant information unexplained.

Candidates were divided into ten quantile strata according to $\Omega_{\mathcal{C}}^{\text{pre}}$, and eight sets were sampled without replacement from each stratum. The resulting 80 sets spanned $\Omega_{\mathcal{C}}^{\text{pre}} = 0.597$–$0.863$, were fixed before hybrid-model training, and were shared across all three training regimes (Fig. 6).

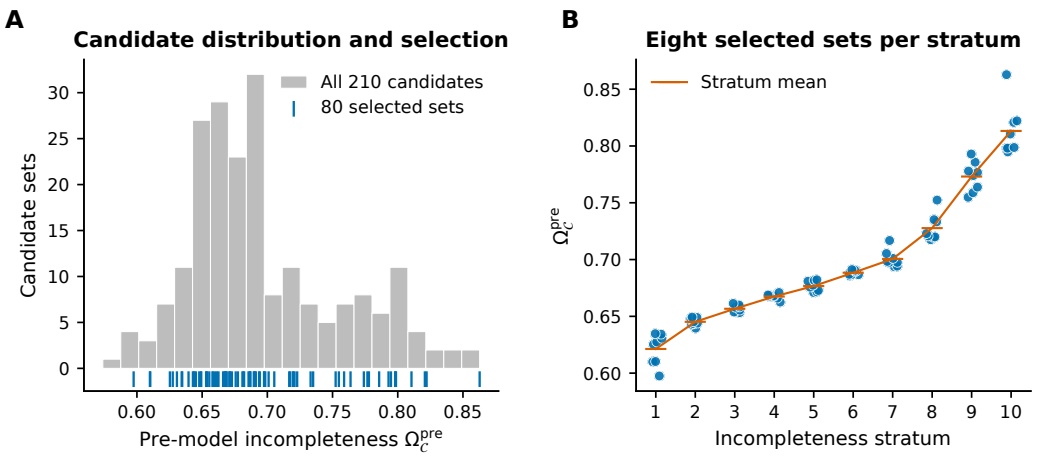

Figure 6: Construction of fixed-size concept sets spanning pre-model incompleteness. **(A)** Distribution of Ω_C^{pre} across 210 candidate sets containing 50 concepts each. Blue marks identify the 80 sets selected for model training. **(B)** Pre-model incompleteness of the eight selected sets within each of ten quantile strata; orange lines show stratum means. Higher Ω_C^{pre} indicates that more task-relevant information remained unexplained by the selected concepts.

B.2.3 Hybrid CBM and Matched Training

For each concept set, the hybrid CBM contained a linear concept encoder, a 32-dimensional linear residual encoder, and a linear 200-class task head:

$$\widehat{\mathbf{c}} = \sigma(g_\phi(\mathbf{x})) \in [0,1]^{50}, \qquad \mathbf{z} = h_\psi(\mathbf{x}) \in \mathbb{R}^{32},$$
$$\mathbf{y} = f_\theta([\widehat{\mathbf{c}}, \mathbf{z}]).$$

Concept supervision used binary cross-entropy and observational task training used multiclass cross-entropy, with $\lambda_c = 1$.

For every concept set, we trained:

Training regime	λ_{int}	λ_{mono}
baseline	0	0
endpoint	1	0
endpoint + monotonic, stop-gradient	1	100

giving $80 \times 3 = 240$ models. The three models for a concept set shared their initial parameters.

The normalised training objective was Models were trained using

$$\mathcal{L} = \frac{\mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{mono}} \mathcal{L}_{\text{mono}}}{1 + \lambda_c + \lambda_{\text{int}} + \lambda_{\text{mono}}}, \qquad \lambda_c = 1.$$

The endpoint loss was calculated after replacing all predicted concepts with their ground-truth values.

For monotonic training, we set $m_{\text{mono}} = 5$. Intervention depth was sampled uniformly from $\{0, \ldots, 49\}$, and a random concept permutation defined the current set S and next corrected concept i. The monotonic penalty was

$$\left[\ell_{S \cup \{j\}} - \ell_S\right]_+ .$$

Gradients from this term were stopped through predicted concepts but retained through the residual encoder and task head.

Models were optimised using AdamW with learning rate 10^{-3}, weight decay 10^{-5}, and batch size 512. Training continued for at most 300 epochs, with a minimum of 50 epochs and patience of 30 validation epochs at an improvement threshold of 10^{-5}. The parameters minimising the full validation objective were retained.

B.2.4 Intervention Evaluation

For an intervention set S, selected predicted concepts were replaced by their image-level ground-truth values while the residual representation was held fixed. All three matched models for a concept set were evaluated using the same 512 random concept-correction orderings.

At state S, test risk was

$$\mathcal{R}_S = -\frac{1}{n_{\text{test}}} \sum_i \log p_\theta(y_i \mid \hat{\mathbf{c}}_i^S, \mathbf{z}_i),$$

and harm was $\eta(S) = \mathcal{R}_S - \mathcal{R}_\varnothing$. We measured endpoint harm after all 50 concepts had been corrected, the maximum harm observed among the intermediate states of the 512 sampled paths, and the fraction of sampled paths containing at least one intermediate state with harm above 10^{-8}. The reported CUB maximum is therefore a maximum over sampled paths rather than all 2^{50} possible intervention states.

We additionally recorded observational top-1 accuracy, fully intervened accuracy, and concept MSE.

B.2.5 Statistical Analysis

For each outcome, the outcome and Ω_C^{pre} were standardised across the complete 240-model design. We fitted

$$Y_z = \beta_0 + \beta_\Omega(\Omega_C^{\text{pre}}) + \beta_1 \mathbb{1}(\text{endpoint}) + \beta_2 \mathbb{1}(\text{endpoint} + \text{monotonic}) + \epsilon,$$

with the baseline regime as reference. Standard errors were clustered by concept-set identifier, preserving the matched three-model design. This additive model estimates the average association with incompleteness and each intervention-training regime's adjusted difference from baseline. The results are shown in Fig. 7.

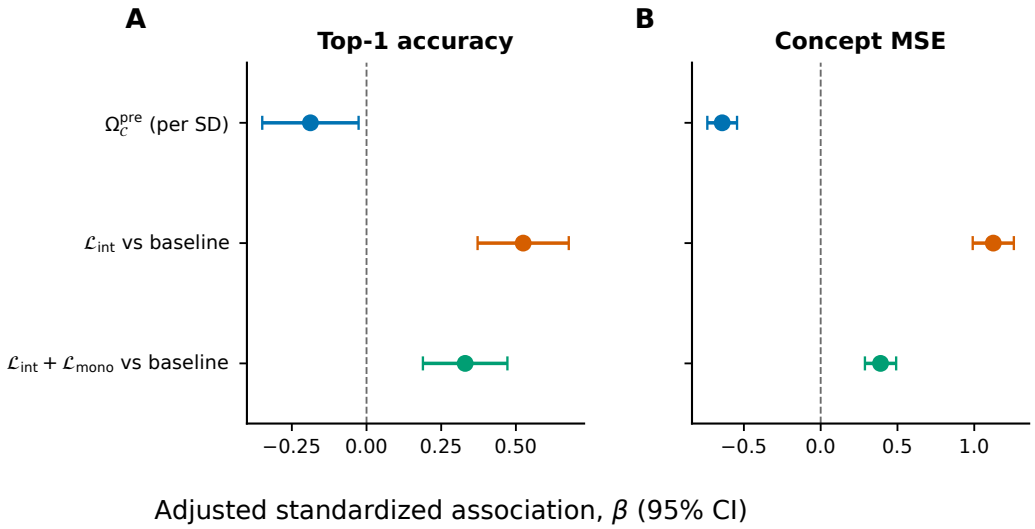

Figure 7: Adjusted associations with model performance and concept quality on CUB. Adjusted standardised associations with **(A)** top-1 accuracy and **(B)** concept mean-squared error. Each panel shows the association with a one-standard-deviation increase in pre-model incompleteness, Ω_C^{pre}, and the adjusted differences between endpoint training or endpoint-plus-monotonic training and the baseline regime. Points show regression coefficients and bars show 95% cluster-robust confidence intervals; dashed vertical lines denote no association. Outcomes and Ω_C^{pre} were standardised. Models included all 240 runs across 80 matched concept sets, with baseline as the regime reference and standard errors clustered by concept-set identifier.