Certified Robust Interpretability via Concept-Space Stability under Interventional Proxies

Yanming Cheng **Di Zhang***
Xi'an Jiaotong-Liverpool University
di.zhang@xjtlu.edu.cn

Abstract

A concept-bottleneck classifier can preserve its prediction under small input perturbations while its concept activations change substantially. We construct models with identical logits, and therefore identical representation-blind label certificates, but arbitrarily large concept drift inside the certified ball. CRISP gives a deterministic, closed-form Lipschitz–margin radius $r^\star(x)$ that preserves the label and bounds concept displacement. The radius is tight up to a factor of two among uniform rules using the encoder/head/margin triple. Under ε-concept-faithfulness and a known concept-to-target subgraph, $r^\star$ further certifies the causally meaningful concepts; a matching impossibility result shows that faithfulness cannot be removed. Training relies on *interventional proxies*—Structural Causal Model (SCM) simulators, attribute editors, environment swaps, concept-matched retrieval—which approximate $\mathrm{do}(\cdot)$-interventions on natural images. On synthetic SCMs with observable $c^\star$, we assess sensitivity to noisy intervention signals and counterfactual observations. Radius checks at saved attack points record zero violations, with empirical tightness $r^\star/\hat{r}_{\mathrm{adv}} \approx 0.32$. Concept fidelity (MCC) rises from $[0.37, 0.51]$ for an unconstrained/no-ICA bottleneck to $[0.74, 0.83]$ for CRISP. Four-class SCMs with $12/24$ concepts quantify the trade-off: shallow MLP heads improve MCC while reducing certificate tightness by $46\%/41\%$ relative to linear heads. On frozen ResNet-18 features from Waterbirds, tightness over successful attack searches is 0.104 ± 0.015 across seeds, with zero violations in the radius checks.

1 Introduction

Prediction robustness and concept stability are distinct guarantees. A randomized-smoothing certificate guarantees that no ℓ_2 perturbation of size below $\bar{r}$ flips the *smoothed label*; it says nothing about *why* the label was chosen. A concept-bottleneck model (CBM) routes its prediction through a layer of human-named concepts to expose attribution [Koh et al., 2020], but does not come with a formal stability guarantee on that concept layer. A perturbation can lie inside the smoothed certified ball, preserve the prediction, and still change which concepts support it.

Figure 1 illustrates this with a schematic radiology CBM. An adversary preserves the BENIGN label while shifting its concept attribution from TUMOR-SIZE to CALCIFICATION. An audit based on those concepts would therefore change even though the prediction remains certified.

CRISP makes concept drift part of the certification target. A Lipschitz-bounded encoder controls representation displacement, and the classifier margin determines a ball in which the prediction is preserved. Concept faithfulness and a causal-subgraph restriction connect this representation guarantee to the underlying causal concepts.

*Corresponding author.

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



Figure 1: Failure mode that motivates a concept-space certificate. **(a)** A clean input: the CBM predicts BENIGN driven by the SIZE concept. **(b)** A perturbation x' stays inside an ℓ_2 certified ball (e.g., randomized smoothing [Cohen et al., 2019]) and preserves the label. **(c)** The concept attribution shifts to CALCIFICATION. CRISP additionally bounds concept displacement throughout its certified ball.

Interventional proxies. Training uses an *interventional proxy* oracle: a Structural Causal Model (SCM) simulator, an attribute-conditioned generative editor, a concept-matched retrieval index, or an environment swap. On the synthetic SCM, we evaluate how noise in the intervention signal and in counterfactual observations affects concept fidelity and certified radii.

Contributions.

1. We give a closed-form radius that jointly certifies label preservation and bounded concept drift (Theorem 1). This radius is tight up to a factor of two among uniform rules using the encoder/head/margin triple (Proposition 1).
2. We separate label certification from representation safety by constructing models with identical logits and arbitrarily different concept drift inside the same certified ball (Propositions 2, 3).
3. We connect the radius to causal concepts under concept faithfulness and a known concept-to-target subgraph (Proposition 4). A matching counterexample (Proposition 5) establishes the need for faithfulness.
4. We evaluate radius tightness and concept fidelity on binary and four-class SCMs ($k \in \{3, 6, 12, 24\}$, 5 seeds), quantify the fidelity–tightness trade-off of shallow MLP heads, and evaluate feature-space certification on Waterbirds. Separate task and drift attacks measure objective-dependent empirical flip rates.

2 Related Work

Concept bottlenecks and concept attribution. CBMs route predictions through human-named concepts and support interventions on their values [Koh et al., 2020]. Subsequent work identifies information leakage and imperfect concept alignment [Mahinpei et al., 2021, Margeloiu et al., 2021]; Havasi et al. [2022] address insufficient concept sets and limited concept predictors. Marconato et al. [2023] formalize alignment between human and machine representations; we use "Causal-CBM" as shorthand for a CBM-style instantiation of that framework. CEBaB [Abraham et al., 2022] is a dataset and evaluation framework based on human-edited counterfactual text. CRISP trains with concept-edit pairs and certifies representation stability with a deterministic Lipschitz–margin bound.

Certified adversarial robustness. Randomized smoothing certifies an ℓ_2 ball at a controlled failure probability [Cohen et al., 2019, Salman et al., 2019]; Linear relaxations and interval bounds yield deterministic label certificates [Zhang et al., 2018, Gowal et al., 2018, Wong and Kolter, 2018].

Proposition 3 identifies an information limit of certifiers determined solely by end-to-end label behavior: equivalent label predictors can have different representation drift. CROWN and IBP use layerwise network information; their bounds can also constrain intermediate activations. CRISP uses an encoder bound to control concept drift together with label preservation.

Invariant and distributionally-robust learning. IRM [Arjovsky et al., 2019] and vREx [Krueger et al., 2021] seek stable prediction across environments, while GroupDRO [Sagawa et al., 2020] minimizes worst-group training risk. Rosenfeld et al. [2021] identify failure cases of IRM under distribution shift. L_{ICA} (Section 5.1) uses concept-edit pairs to impose coordinate-level invariance under edits of other concepts.

Causal representation learning. The iVAE family [Khemakhem et al., 2020] and contrastive CRL [von Kügelgen et al., 2021, Brehmer et al., 2022] study identifiability under assumptions on auxiliary variables or paired observations [Schölkopf et al., 2021]. CRISP uses targeted intervention pairs to train concept responses and evaluates coordinate fidelity on synthetic SCMs.

Explanation manipulation and verification. Saliency explanations can be perturbed while keeping model outputs approximately fixed [Dombrowski et al., 2019]. Boopathy et al. [2020] connect prediction-changing attacks to discrepancies in class-specific interpretation maps and develop a defense based on robust interpretation. CRISP certifies drift of a concept layer used by the classifier. Appendix A.4 derives an IBP-based condition for bounding concept drift and preserving the label throughout an ℓ_∞ input ball.

3 Problem Setup

Notation. $X \subset \mathbb{R}^d$ is the input space, $Y \in \{1, \dots, K\}$ the label, $C = (C_1, \dots, C_k)$ human-named concepts. Classifiers factor as $f = \arg\max \circ h \circ g$ with *concept encoder* $g : X \to \mathbb{R}^k$ and a linear *head* $h(z) = Wz + b$ (in the main theorem); we write $\hat{y}(x) = f(x)$ and the decision margin

$$\mathrm{margin}(x) \;=\; h(g(x))_{\hat{y}(x)} \;-\; \max_{y \neq \hat{y}(x)} h(g(x))_y. \tag{1}$$

$\mathcal{B}_r(x)$ denotes the closed ℓ_2 ball of radius r. Admissible perturbed inputs lie in X.

Interventional proxy oracle. An intervention tuple $(x, \tilde{x}, j, \Delta_j)$ contains an original input x, an edited input $\tilde{x} = T_j(x, \Delta_j)$, a target concept j, and a signed edit magnitude Δ_j. The oracle intends to change concept j by Δ_j while holding other causal factors fixed. We allow T_j to be (i) an SCM simulator with $\mathrm{do}(\cdot)$-access, (ii) an attribute-conditioned editor, (iii) concept-matched retrieval, or (iv) environment swap (e.g. Waterbirds). T_j is not assumed to equal a true $\mathrm{do}(\cdot)$-intervention on natural images; sensitivity to noisy proxies is evaluated in Section C.7.

Threat models. The standard ℓ_2 input adversary searches $x' \in X \cap \mathcal{B}_r(x)$ to flip the label. The concept-space adversary

$$\mathcal{A}_{\mathrm{cs}}(x; r) \;=\; \arg\max_{x' \in X \cap \mathcal{B}_r(x)} \|g(x') - g(x)\|_2 \quad \text{s.t. } f(x') \neq \hat{y}(x) \tag{2}$$

maximizes concept drift subject to flipping the prediction; the audit-trail variant $\mathcal{A}_{\mathrm{cs}}^{=}$ replaces the constraint with $f(x') = \hat{y}(x)$. Theorem 1 bounds both because it controls concept drift at every point in $X \cap \mathcal{B}_{r^\star(x)}(x)$.

Assumptions (summary). We index assumptions by the result that needs them; full statements and discussion are deferred to Appendix A.

(A1) Concept-mediated label.
$Y = \psi(C) \oplus \eta$ with $\eta \perp C$. Always in force.

(A-Lip) Encoder Lipschitzness.
$\|g(x) - g(x')\|_2 \leq L_g \|x - x'\|_2$ on X, end to end. With a fixed feature extractor ϕ and $g = g_{\text{head}} \circ \phi$, this composes as $L_g = \mathrm{Lip}(g_{\text{head}}) \cdot \mathrm{Lip}(\phi)$; if only g_{head} is controlled, $r^\star$ becomes a ϕ-feature-space radius (Section 5.2). Needed for Theorem 1.

(A-Lin) Linear head.
$h(z) = Wz + b$. Needed for Theorem 1.

(A-Marg) Positive margin.
$\mathrm{margin}(x) > 0$ at the input of interest. Needed for Theorem 1.

(A-Faith) ε-concept-faithfulness.
$\|g(x) - c^\star(x)\|_\infty \le \varepsilon$, where $c^\star(x)$ is the ground-truth causal-concept vector. Needed for Proposition 4.

(A-Sub) Subgraph restriction.
h depends only on coordinates indexed by a known $S \subseteq [k]$ of causal ancestors of Y. Needed for Proposition 4.

Dependency map. The certificate and causal bridge use the following assumptions:

Result	Assumptions	Guarantee
Certificate (Theorem 1)	A-Lip, A-Lin, A-Marg	Label preservation and bounded learned-representation drift
Conditional bridge (Proposition 4)	Certificate assumptions, A-Faith, A-Sub	Control of the true causal concepts

The certificate applies to smooth or nonsmooth encoders with the stated bounds, independently of training-objective convexity. Exact matrix norms and certified bounds on the candidate ball can tighten the radius (Appendix A.2).

4 Theory

Full proofs are in Appendix A.

4.1 Concept-space certified radius

Theorem 1 (Concept-space certified radius)**.** *Assume (A-Lip), (A-Lin), (A-Marg). Let $W \in \mathbb{R}^{K\times k}$ have rows $w_1, \dots, w_K$. For $x \in X$, suppose $0 < L_g \max_{y\neq\hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_2 < \infty$ and define*

$$r^\star(x) \;:=\; \frac{\mathrm{margin}(x)}{2\,L_g \max\limits_{y\neq\hat{y}(x)} \|w_{\hat{y}} - w_y\|_2}. \tag{3}$$

Then for every $x' \in X \cap \mathcal{B}_{r^\star(x)}(x)$, $f(x') = \hat{y}(x)$.

Sketch. By (A-Lip), $\|g(x') - g(x)\|_2 \le L_g\, r^\star(x)$. By (A-Lin) and Cauchy–Schwarz, $|(w_{\hat{y}} - w_y)^\top (g(x') - g(x))| \le \mathrm{margin}(x)/2$, hence $h(g(x'))_{\hat{y}} - h(g(x'))_y \ge \mathrm{margin}(x)/2 > 0$. Full proof and extensions in Appendices A.1–A.4, including the local encoder bound and self-consistent radius (Appendix A.2). □

Shallow nonlinear heads. For $h(z) = V\sigma(Uz + c) + b$ with coordinatewise 1-Lipschitz σ, Corollary 3 combines operator-product and neuron-wise bounds into a pairwise logit Lipschitz bound $L_h(x)$. For $\mathrm{margin}_h(x) > 0$ and $0 < L_g L_h(x) < \infty$, this gives

$$r^\star_{\mathrm{MLP}}(x) = \frac{\mathrm{margin}_h(x)}{2L_g L_h(x)}, \tag{4}$$

Section 6.2 evaluates this extension.

Proposition 1 (Worst-case bound for uniform radius rules)**.** *For every $L_g, L_W > 0$ and $m > 0$ there exist an L_g-Lipschitz g, a linear head with $\max_{y\neq\hat{y}} \|w_{\hat{y}} - w_y\|_2 = L_W$, and x with $\mathrm{margin}(x) = m$ such that some adversarial $x^\dagger$ at input distance arbitrarily close to $2\,r^\star(x)$ from above flips the label. Hence no uniformly valid Lipschitz–margin radius rule depending only on $(L_g, L_W, \mathrm{margin}(x))$ exceeds $2\,r^\star(x)$; Theorem 1 is tight up to this factor of two.*

Sketch. Take $x_0 = 0$, $g(t) = L_g t$, and $h_{1,2}(z) = \pm(L_W z + m)/2$. The logit gap is $L_g L_W t + m$, so any $t < -m/(L_g L_W)$ flips the label. The boundary lies at distance $m/(L_g L_W) = 2\, r^\star(x_0)$. Full proof in Appendix A.7. □

4.2 Separation of label and representation guarantees

Proposition 2 (Input-space certified $\not\Rightarrow$ concept-space safe)**.** *For every $\bar{r} > 0$ and $M > 0$ there exist a classifier $f = h \circ g$ with $d = 2$, an input x_0, and $\sigma > 0$ such that the Cohen certificate for the Gaussian-smoothed classifier, computed from exact class probabilities, has radius $\bar{r}(x_0) \geq \bar{r}$, yet some $x_0^\dagger$ with $\|x_0^\dagger - x_0\|_2 \leq \bar{r}/M$ satisfies $\|g(x_0^\dagger) - g(x_0)\|_2 \geq 1$.*

Sketch. $g(x) = (x_1, \sin(\omega x_2))$, $f = \operatorname{sign}(x_1)$, $\omega = \pi M/(2\bar{r})$, $x_0 = (\bar{r}, 0)$, $x_0^\dagger = (\bar{r}, \pi/(2\omega))$. Appendix A.9. □

Definition 1 (Representation-blind certificate)**.** Let $\mathcal{F}$ be a class of encoder–head pairs. A label-radius rule $R((g,h),x)$ is *representation-blind on $\mathcal{F}$* if pairs inducing the same end-to-end label distribution at every input receive the same radius. The distributions may be deterministic or induced by a fixed smoothing law; the certified prediction is their most probable class. The Cohen radius computed from exact class probabilities is representation-blind.

Proposition 3 (Generalized separation)**.** *Let $\mathcal{F}$ be a class of encoder–head pairs on $\mathbb{R}^d$, $d \geq 1$, closed under appending an affine encoder coordinate that the head ignores. Let R be a representation-blind label certificate on $\mathcal{F}$. For every $(g,h) \in \mathcal{F}$ with $R((g,h),x_0) \geq \bar{r} > 0$, every $0 < \delta < \bar{r}$, and every $B > 0$, there exist $(\hat{g},\hat{h}) \in \mathcal{F}$ with one additional encoder coordinate and a point $x^\dagger$ such that*

$$\hat{h} \circ \hat{g} = h \circ g, \qquad \|x^\dagger - x_0\|_2 = \delta, \qquad \|\hat{g}(x^\dagger) - \hat{g}(x_0)\|_2 \geq B.$$

The certificate and its prediction throughout the certified ball are unchanged. The construction and a same-dimensional variant are given in Appendix A.10.

Thus a label certificate alone supplies no uniform bound on representation drift over this model class. Theorem 1 adds the encoder bound $\|g(x') - g(x)\|_2 \leq L_g \|x' - x\|_2$ and uses the margin to preserve the label.

4.3 Conditional bridge to causal interpretability

Proposition 4 (Concept-faithful robustness)**.** *Assume the conditions of Theorem 1 together with (A-Faith) and (A-Sub). Then for every $x' \in X \cap \mathcal{B}_{r^\star(x)}(x)$, $f(x')$ is determined by $g(x')_S$, and $\|g(x')_S - c^\star(x')_S\|_\infty \leq \varepsilon$.*

Sketch. (A-Sub) makes $f(x')$ depend only on $g(x')_S$; (A-Faith) bounds the gap to $c_S^\star$; Theorem 1 preserves the prediction. Appendix A.5. □

When (A-Faith) holds at both x and x', the two endpoint errors give the true-concept drift bound

$$\|c^\star(x') - c^\star(x)\|_\infty \leq 2\varepsilon + L_g \|x' - x\|_2. \tag{5}$$

This follows by inserting $g(x')$ and $g(x)$, applying the triangle inequality, and using $\|\cdot\|_\infty \leq \|\cdot\|_2$. If (A-Faith) holds throughout $X \cap \mathcal{B}_{r^\star(x)}(x)$, its true-concept drift is therefore bounded by $2\varepsilon + L_g r^\star(x)$.

Proposition 5 (Faithfulness is necessary)**.** *For every $\eta > 0$ there exist g_1, g_2 satisfying (A-Lip), (A-Lin), (A-Marg), (A-Sub) with $\|g_1 - c^\star\|_\infty \leq \eta$, $\|g_2 - c^\star\|_\infty = \Theta(1)$, yet identical $r^\star$ and identical decisions on $\mathcal{B}_{r^\star}(x)$. No bound on $\|g - c^\star\|$ follows from Theorem 1 alone; a faithfulness-style assumption is unavoidable.*

Sketch. $g_2 := g_1 + c$ for c supported on S, head $(W, b_1 - Wc)$: an isometry that preserves every quantity entering Theorem 1. Appendix A.8. □

Consequences for concept attacks. On $X \cap \mathcal{B}_{r^\star(x)}(x)$, the certificate preserves the label while bounding concept drift. Appendix A.6 collects this audit-trail consequence and the necessary concept displacement for a label flip.

5 Method: CRISP

CRISP combines Interventional Concept Alignment (L_{ICA}) with a Lipschitz-bounded concept encoder. The alignment objective encourages the targeted coordinate to respond to an intervention while keeping other coordinates stable. The encoder bound and classifier margin supply the certificate of Theorem 1.

5.1 Interventional Concept Alignment loss

Given an interventional proxy pair $(x, \tilde{x})$ with target concept index j and signed edit magnitude Δ_j,

$$L_{\mathrm{ICA}}(x, \tilde{x}, j; \Delta_j) \;=\; \underbrace{\big\|g(x)_{-j} - g(\tilde{x})_{-j}\big\|_2^2}_{\text{freeze non-target}} \;+\; \mu \cdot \underbrace{\big[\max\big(0,\; \kappa - \big(g(\tilde{x})_j - g(x)_j\big)\,\mathrm{sign}(\Delta_j)\big)\big]^2}_{\text{hinge on target}}, \tag{6}$$

with $\kappa > 0$ a sign-consistent margin and $\mu > 0$ the hinge weight. The freeze term penalizes changes in non-target coordinates; the hinge term encourages the target coordinate to respond in the given intervention direction with signed displacement at least κ. Appendix B discusses the scope of these pairwise constraints.

Full training objective.

$$L_{\mathrm{CRISP}} \;=\; L_{\mathrm{task}} + \alpha L_{\mathrm{c}} + \beta L_{\mathrm{ICA}} + \gamma L_{\mathrm{Lip}} + \lambda \|W\|_F^2, \tag{7}$$

with L_{task} classification, optional L_{c} per-coordinate concept supervision, L_{Lip} a spectral penalty, and a quadratic penalty on the head weights. The experiments use $L_{\mathrm{task}} + \beta L_{\mathrm{ICA}}$ with architectural spectral normalization and no per-concept supervision: $\beta = 1$ on synthetic SCM and 0.5 on Waterbirds. Optimizer settings appear in Appendix C.

5.2 Lipschitz-bounded concept encoder

The implemented encoder has three affine layers, with GroupSort2 after each hidden layer. GroupSort2 sorts adjacent pairs of activations and is 1-Lipschitz in ℓ_2 [Anil et al., 2019]. Spectral normalization rescales each weight matrix using an estimated leading singular value [Miyato et al., 2018]. For the resulting matrices A_1, A_2, A_3, the product $L_g = \prod_{\ell=1}^{3} \|A_\ell\|_2$ bounds the encoder's Lipschitz constant. The default classifier is affine, $h(z) = Wz + b$; its pairwise row distances determine L_W. Realized encoder bounds are reported in Appendix C.1 and Table 5.

End-to-end Lipschitz scope. On synthetic SCM, g acts on the input vector. On Waterbirds, g acts on frozen ResNet-18 features $\phi(x)$, so $r_\phi^\star$ measures perturbations in that feature space.

6 Experiments

We evaluate binary and four-class SCMs, linear and shallow MLP heads, and frozen-backbone Waterbirds features. All reported binary-SCM drift attacks use random nonzero initialization.

6.1 Binary synthetic SCM

Setup. CRISP (spectrally normalized encoder + GroupSort2 + affine head, 5 seeds) and an unconstrained/no-ICA bottleneck (3 seeds) are trained on k-concept SCMs with $k \in \{3, 6\}$, 5 nuisance background dimensions, and binary label $Y = \mathrm{sign}(w^\top C)$. Evaluation uses 96 fresh samples per seed. These same-configuration reproductions use verified encoder bounds. Table 1 uses distances to successful cross-entropy PGD points from 15-level ℓ_2 sweeps up to 5.0 (30 steps, one zero start). The per-sample analysis in Appendix C.6 combines PGD with APGD-style binary-margin ascent (50 steps, 3 random restarts) [Croce and Hein, 2020].

Drift diagnostic. The implemented attack maximizes $\|g(x+\delta)-g(x)\|_2^2$ over the ℓ_2 ball, dropping the label constraint in both formal threats of Section 3. Binary-SCM runs use 40 PGD steps and one random-direction start at one quarter of the budget. We denote the resulting endpoint label-flip rate by $\mathrm{ASR}_{\mathrm{cs}}$; its denominator is each model's own clean-correct subset. In the two-model budget sweep, the shared budget is mb_s, with b_s the fixed grid-search reference for seed s. Its definition and further settings appear in Appendix C.

Headline. PGD finds flips for 914 of 923 clean-correct CRISP samples. The radius checks in Appendix C.1 record zero violations, with mean tightness 0.31–0.32. Table 1 reports the corresponding MCC gains; the spectral-cap and attack-budget sweeps appear in Appendix C.

Table 1: Binary-SCM headline with verified encoder bounds. $\mathrm{ASR}_{\mathrm{cs}}$: drift-endpoint flip rate at $m = 1$. Seed means: CRISP, 5; unconstrained/no-ICA (Unconstr.), 3. Mean $r^\star$ uses all clean-correct samples; attack distance and tightness use found flips. PGD failures: $5/4$ at $k = 3/6$.

k	Method	CA	MCC	$\bar{r^\star}$	$\bar{\hat{r}}_{\mathrm{adv}}$	$\overline{r^\star/\hat{r}_{\mathrm{adv}}}$	Violations	$\mathrm{ASR}_{\mathrm{cs}}$
3	CRISP	0.952	**0.825**	0.310	0.960	0.310	0 / 452	**0.298**
3	Unconstr.	0.972	0.508	–	–	–	–	0.368
6	CRISP	0.971	**0.741**	0.338	1.002	0.325	0 / 462	**0.292**
6	Unconstr.	0.986	0.373	–	–	–	–	0.328

ICA and Lipschitz control. The four-cell ablation in Table 2 separates the roles of the two components. Full, ICA-only, Lip-only, and neither achieve $\mathrm{MCC} = 0.773/0.718/0.399/0.373$, respectively. ICA improves coordinate fidelity in this SCM, while architectural Lipschitz control supplies the encoder bound used for certification.

Table 2: ICA/Lipschitz ablation on synthetic SCM k=6, three-seed means. Drift attacks retain each variant's fixed grid-search budget per seed. "n/a" indicates that this evaluation reports no certified radius for variants without architectural Lipschitz control.

Variant	CA	MCC	$\bar{r^\star}$	$\mathrm{ASR}_{\mathrm{cs}}$
CRISP (ICA + Lip)	0.969	**0.773**	0.349	0.283
ICA-only	0.972	0.718	n/a	**0.267**
Lip-only	0.986	0.399	**0.412**	0.345
Unconstrained/no-ICA bottleneck	0.986	0.373	n/a	0.300

Hinge margin and proxy quality. Table 3 shows the hinge trade-off: raising κ from 0.05 to 0.50 increases MCC from 0.767 to 0.955 with a 1.1-percentage-point accuracy cost. The radius is nearly flat through $\kappa = 0.25$ and rises at 0.50. In the proxy sweeps (Appendix C.7), MCC changes by 0.020 at magnitude-noise level $\sigma_\Delta = 0.50$ and falls from 0.773 to 0.474 at observation-noise level $\sigma_c = 0.50$.

Table 3: Hinge margin sweep on synthetic SCM k=6, three-seed means. The radius checks record zero violations in every cell.

κ	CA	MCC	$\bar{r^\star}$	$\mathrm{ASR}_{\mathrm{cs}}$
0.05	0.969	0.767	0.351	0.286
0.10	0.969	0.773	0.349	0.283
0.25	0.965	0.785	0.355	0.276
0.50	0.958	**0.955**	**0.373**	**0.261**

6.2 Four-class SCM and shallow MLP heads

Two dense orthogonal directions in the concept vector define four quadrant labels. We use $k \in \{12, 24\}$ concepts, eight nuisance variables, and five seeds, comparing affine heads with width-64

one-hidden-layer ReLU heads. Each seed evaluates accuracy and MCC on 512 test points, mean $r^\star$ on all clean-correct points, and attack tightness on the first 128 clean-correct points. Certificates use realized encoder norm products; MLP heads use Equation (4). The 15-level PGD/APGD search uses 50 steps, respectively $1/3$ random starts, and budget 5. All 2,560 CRISP searches find a label flip, with zero violations in the radius checks. Full settings appear in Appendix C.2.

Table 4: Four-class SCM, five-seed mean $\pm$ sample standard deviation. Tightness is the mean per-sample $r^\star/\hat{r}_{\mathrm{adv}}$ on the attacked subset.

k	Head	Accuracy	MCC	Mean $r^\star$	Tightness
12	Linear	0.963 ± 0.007	0.526 ± 0.066	0.140 ± 0.006	0.202 ± 0.009
12	MLP	0.968 ± 0.010	0.588 ± 0.049	0.064 ± 0.001	0.108 ± 0.004
24	Linear	0.957 ± 0.006	0.260 ± 0.009	0.137 ± 0.004	0.203 ± 0.008
24	MLP	0.963 ± 0.008	0.408 ± 0.024	0.071 ± 0.001	0.119 ± 0.004

Linear-head tightness stays near 0.20 as k increases. The MLP improves accuracy by $0.005/0.006$ and MCC by $0.062/0.148$ at $k = 12/24$, while mean radius falls by $55\%/48\%$ and tightness by $46\%/41\%$. The task-only unconstrained/no-ICA ablation has MCC 0.354 ± 0.024 and 0.255 ± 0.006, respectively; at $k = 24$, its MCC and linear CRISP's are nearly equal.

6.3 Waterbirds with a frozen ImageNet backbone

Setup. A frozen ImageNet ResNet-18 [He et al., 2016] maps Waterbirds images [Sagawa et al., 2020] to 512-dimensional features. The encoder g has three spectrally normalized affine layers of width 64, with GroupSort2 between them, followed by an affine classifier. Training uses $L_{\mathrm{task}} + 0.5 L_{\mathrm{ICA}}$, 20 epochs of AdamW, and environment-swap proxy pairs. We evaluate five saved checkpoints using their realized matrix-norm products. Accuracy uses the full 5,794-image test set; attacks use the first 500 test samples per seed, containing 1,980 CRISP clean-correct model–sample pairs in total.

Feature-space attacks. All perturbations act on frozen features in an ℓ_2 ball. Label-radius search combines cross-entropy PGD (20 radii) and APGD-style binary-margin ascent (15 radii), each with 50 steps, 3 random restarts, and maximum budget 5.0. For model comparison, direct-task PGD maximizes cross-entropy; drift PGD maximizes squared representation displacement. Both use 50 steps and 5 random restarts. We record whether any restart's endpoint flips the clean prediction for direct-task PGD, and whether the largest-drift endpoint flips it for drift PGD. Rates use the 1,901 pooled model–sample pairs correctly classified by both models. Each seed's shared budget is m times the mean successful CRISP label-attack radius.

Table 5: Waterbirds feature-space evaluation, mean $\pm$ sample standard deviation over five seeds. Tightness uses the 1,949 clean-correct samples with a found flip; 31 attack failures at budget 5 are reported separately. Comparisons with first-success attack budgets record zero violations.

Quantity	Mean $\pm$ seed SD
Full-test accuracy	0.788 ± 0.020
Full-test worst-group accuracy	0.528 ± 0.024
Realized encoder norm product L_g	1.0007 ± 0.0010
Mean radius $r^\star_\phi$ (clean-correct subset)	0.180 ± 0.029
Tightness $r^\star_\phi/\hat{r}_{\mathrm{adv}}$ (found flips)	0.104 ± 0.015

Results. At $m = 1.5$, direct-task flip rates are 0.736 ± 0.021 for CRISP and 0.781 ± 0.048 for the unconstrained/no-ICA bottleneck. Drift-endpoint flip rates are 0.721 ± 0.028 and 0.471 ± 0.104, respectively. The different orderings reflect the objectives: a point with larger representation displacement need not change the predicted label. Figure 2 reports both budget curves.

7 Discussion

CRISP makes learned-representation stability a certifiable part of the prediction pipeline. The separation results establish why label certification alone leaves concept drift uncontrolled. The encoder bound controls concept displacement, and the margin sets the certified radius. The synthetic ablations show how ICA complements this guarantee by improving coordinate fidelity.

Limitations and next steps. Waterbirds evaluates frozen-backbone features. Pixel-space certification requires end-to-end image bounds; the local construction in Appendix A.2 remains to be implemented. Ground-truth causal concepts are unavailable for measuring A-Faith on these images. For decisions defined by thresholds on individual concepts, the bound in Equation (5) guarantees preservation when smaller than the distance to the nearest threshold. Future image studies will also add compatible external concept-supervised baselines (Appendix C.8).

Acknowledgments and Disclosure of Funding

This work was supported by Xi'an Jiaotong-Liverpool University (XJTLU) Research Development Fund (RDF-24-02-030). The authors declare no competing interests.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning (ICML)*, 2019.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning (ICML)*, 2020.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 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 (ICML)*, 2020.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, 2021.

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

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25(12):1574, 2023. doi: 10.3390/e25121574.

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.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Abbavaram Gowtham Reddy, Benin Godfrey L, and Vineeth N. Balasubramanian. On causally disentangled representations. In *AAAI Conference on Artificial Intelligence*, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

A Proofs for Section 4

A.1 Proof of Theorem 1

Fix $x \in X$ with $\hat{y} = \hat{y}(x)$ and $\mathrm{margin}(x) > 0$ by (A-Marg), with the positive finite denominator assumed in Theorem 1. Let $x' \in X \cap \mathcal{B}_{r^\star(x)}(x)$ be arbitrary. Write $z = g(x)$ and $z' = g(x')$.

Step 1 (encoder contraction). By (A-Lip),

$$\|z' - z\|_2 \;\le\; L_g \, \|x' - x\|_2 \;\le\; L_g \, r^\star(x). \tag{8}$$

Step 2 (margin survival). By (A-Lin), $h(z)_y = w_y^\top z + b_y$. For each $y \neq \hat{y}$, define the pairwise logit difference

$$\Delta_y(u) \;:=\; h(g(u))_{\hat{y}} - h(g(u))_y \;=\; (w_{\hat{y}} - w_y)^\top g(u) + b_{\hat{y}} - b_y.$$

Since $\Delta_y(x) \geq \mathrm{margin}(x)$,

$$\begin{aligned}
\Delta_y(x') \;&=\; \Delta_y(x) \;+\; (w_{\hat{y}} - w_y)^\top (z' - z) \\
&\geq\; \mathrm{margin}(x) \;-\; |(w_{\hat{y}} - w_y)^\top (z' - z)| \\
&\geq\; \mathrm{margin}(x) \;-\; \|w_{\hat{y}} - w_y\|_2 \, \|z' - z\|_2 && \text{(Cauchy–Schwarz)} \\
&\geq\; \mathrm{margin}(x) \;-\; \|w_{\hat{y}} - w_y\|_2 \, L_g \, r^\star(x) && \text{by (8)} \\
&\geq\; \mathrm{margin}(x) \;-\; \frac{\mathrm{margin}(x)}{2} \;=\; \frac{\mathrm{margin}(x)}{2} \;>\; 0,
\end{aligned}$$

where the last line uses the definition of $r^\star(x)$ in (3) with the supremum over y absorbed into $\max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2$.

Step 3 (prediction preservation). $\Delta_y(x') > 0$ for all $y \neq \hat{y}$ means $h(z')_{\hat{y}} > h(z')_y$ for all $y \neq \hat{y}$, hence $\hat{y}(x') = \hat{y}$, i.e. $f(x') = \hat{y} = f(x)$. □

A.2 Certification with a local encoder bound

For $r > 0$ with $\mathcal{B}_r(x) \subseteq X$, define the ball-wise Lipschitz constant

$$L_g(x, r) := \sup_{\substack{u, v \in \mathcal{B}_r(x) \\ u \neq v}} \frac{\|g(u) - g(v)\|_2}{\|u - v\|_2}. \tag{9}$$

Let $\widehat{L}_g(x, r)$ be a finite, nonnegative certified upper bound on $L_g(x, r)$, covering the entire ball.

Corollary 1 (Local encoder certificate)**.** *Let the head satisfy (A-Lin), let x satisfy (A-Marg), and set $L_W(x) = \max_{y \neq \hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_2$. Every radius $r > 0$ with the certified bound above and*

$$2 L_W(x) \, r \, \widehat{L}_g(x, r) \leq \mathrm{margin}(x) \tag{10}$$

satisfies $f(x') = \hat{y}(x)$ and $\|g(x') - g(x)\|_2 \leq r \widehat{L}_g(x, r)$ for all $x' \in \mathcal{B}_r(x)$.

Proof. The ball-wise bound gives $\|g(x') - g(x)\|_2 \leq \widehat{L}_g(x, r) \|x' - x\|_2 \leq r \widehat{L}_g(x, r)$. For each $y \neq \hat{y}(x)$, the affine logit gap therefore obeys

$$h(g(x'))_{\hat{y}(x)} - h(g(x'))_y \geq \mathrm{margin}(x) - L_W(x) r \widehat{L}_g(x, r) \geq \mathrm{margin}(x)/2 > 0.$$

□

Taking $\widehat{L}_g(x, r) = L_g$ recovers the half-margin condition of Theorem 1. A tighter certified bound on the candidate ball can admit a larger radius through (10).

Constructing the bound. For g continuously differentiable on a neighborhood of the ball, an upper bound on $\sup_{u \in \mathcal{B}_r(x)} \|J_g(u)\|_2$ suffices, by integrating the Jacobian along each segment in the ball. For a continuous piecewise-affine encoder, including ReLU and GroupSort networks, one can instead bound the operator norm of the linear map in every activation region intersecting the ball. Each segment splits into region-wise segments; summing the increment bounds controls the ratio in (9). Certified bounds on possible activation regions or propagation of derivative bounds can supply these upper bounds. A Jacobian at one point or a finite set of sampled slopes does not bound all pairs in the ball.

Self-consistent radius search. Fix a finite search cap $R > 0$ with $\mathcal{B}_R(x) \subseteq X$, and choose certified bounds $\widehat{L}_g(x, r)$ that are nondecreasing in r for $0 < r \leq R$. The exact quantity in (9) is nondecreasing because the balls are nested; the chosen upper-bound function must also have this property for bisection. Then $F(r) = 2 L_W(x) r \widehat{L}_g(x, r)$ is nondecreasing. Set $F(0) = 0$. If $F(R) \leq \mathrm{margin}(x)$, return R. Otherwise, bisect $[0, R]$, retaining the passing lower endpoint and the failing upper endpoint. Return the lower endpoint when the interval width is at most a chosen tolerance $\tau > 0$. This returns a certified radius within τ of the supremum of feasible radii in $[0, R]$.

A.3 Slack-Lipschitz and shallow-MLP-head extensions

Corollary 2 (Slack-Lipschitz)**.** *Suppose g satisfies the weaker bound $\|g(x) - g(x')\|_2 \le (1 + \xi)L_g \|x - x'\|_2$ for some $\xi \ge 0$. Then Theorem 1 holds with $r^\star(x)$ replaced by $r^\star_\xi(x) = r^\star(x)/(1+\xi)$.*

Proof. Replace L_g with $(1 + \xi)L_g$ in Step 2 above. □

Corollary 3 (Shallow-MLP head, combined pairwise bound)**.** *Assume (A-Lip), and let $h(z) = V\sigma(Uz + c) + b$, where $U \in \mathbb{R}^{q\times k}$, $V \in \mathbb{R}^{K\times q}$, and σ is coordinatewise 1-Lipschitz. For the i-th row $u_i^\top$ of U and distinct classes a, d, define*

$$B^{\mathrm{op}}_{ad} = \|V_{a,:} - V_{d,:}\|_2 \|U\|_2, \qquad B^{\mathrm{nw}}_{ad} = \sum_{i=1}^{q} |V_{a,i} - V_{d,i}| \|u_i\|_2,$$
$$B_{ad} = \min\{B^{\mathrm{op}}_{ad}, B^{\mathrm{nw}}_{ad}\}, \qquad L_h(x) = \max_{y\neq\hat{y}(x)} B_{\hat{y}(x),y}. \tag{11}$$

Let $\mathrm{margin}_h(x) = h(g(x))_{\hat{y}(x)} - \max_{y\neq\hat{y}(x)} h(g(x))_y > 0$, and suppose $0 < L_g L_h(x) < \infty$. Then

$$r^\star_{\mathrm{MLP}}(x) = \frac{\mathrm{margin}_h(x)}{2L_g L_h(x)}$$

ensures $f(x') = \hat{y}(x)$ for every $x' \in X \cap \mathcal{B}_{r^\star_{\mathrm{MLP}}(x)}(x)$, with $\|g(x') - g(x)\|_2 \le L_g r^\star_{\mathrm{MLP}}(x)$.

Proof. Write $d_{ad}(z) = h_a(z) - h_d(z)$. Cauchy–Schwarz and the 1-Lipschitz activation give

$$|d_{ad}(z') - d_{ad}(z)| \le \|V_{a,:} - V_{d,:}\|_2 \|U(z' - z)\|_2 \le B^{\mathrm{op}}_{ad} \|z' - z\|_2.$$

Bounding each hidden unit separately also gives

$$|d_{ad}(z') - d_{ad}(z)| \le \sum_i |V_{a,i} - V_{d,i}|\, |u_i^\top (z' - z)| \le B^{\mathrm{nw}}_{ad} \|z' - z\|_2.$$

Both inequalities hold for every pair of representations, so their minimum is a valid bound for that class pair. For any competing class y and x' in the certified ball, (A-Lip) yields

$$|d_{\hat{y}y}(g(x')) - d_{\hat{y}y}(g(x))| \le B_{\hat{y}y} L_g \|x' - x\|_2 \le L_h(x) L_g r^\star_{\mathrm{MLP}}(x) = \mathrm{margin}_h(x)/2.$$

Every competing logit remains below the predicted-class logit by at least $\mathrm{margin}_h(x)/2 > 0$. The representation-drift bound follows from (A-Lip). □

Remark 1 (Interaction with Prop 4). Replacing the linear head with a shallow MLP does not interact with (A-Sub): if U has zero columns outside S, the head depends only on coordinates S, and Proposition 4 carries through with $L_h(x)$ in place of the linear factor.

A.4 Other norms and interval bounds

Let $p, q \in [1, \infty]$, with dual exponent $q^\star$ satisfying $1/q + 1/q^\star = 1$ (including $1^\star = \infty$ and $\infty^\star = 1$). Suppose $\|g(u) - g(v)\|_q \le L_{g,p\to q}\|u - v\|_p$ on X. For an affine head and an input with margin $\mathrm{margin}(x) > 0$, set

$$L_{W,q^\star}(x) = \max_{y\neq\hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_{q^\star}, \qquad r^\star_{p,q}(x) = \frac{\mathrm{margin}(x)}{2L_{g,p\to q} L_{W,q^\star}(x)}.$$

Assume $0 < L_{g,p\to q} L_{W,q^\star}(x) < \infty$ and the closed input ℓ_p ball of this radius lies in X. Every point in that ball preserves the label and has concept drift at most $L_{g,p\to q} r^\star_{p,q}(x)$ in ℓ_q. Indeed, Hölder's inequality bounds every pairwise-logit change by $L_{W,q^\star}(x) L_{g,p\to q}\|x' - x\|_p \le \mathrm{margin}(x)/2$. The attack norm changes this geometry; the ICA objective continues to specify coordinate responses to paired interventions.

Corollary 4 (ℓ_∞ certificate)**.** *Taking $p = q = \infty$ above, write $L^\infty_g = L_{g,\infty\to\infty}$. The certified radius is*

$$r^\star_\infty(x) \;:=\; \frac{\mathrm{margin}(x)}{2L^\infty_g \max_{y\neq\hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_1}.$$

Interval-based certificate. For a candidate radius $r \geq 0$, interval bound propagation (IBP; Gowal et al. 2018) supplies componentwise bounds $\underline{z}(x,r) \leq g(x') \leq \overline{z}(x,r)$ for every $x' \in X$ with $\|x' - x\|_\infty \leq r$. Starting from $[x - r\mathbf{1}, x + r\mathbf{1}]$, an affine layer $v \mapsto Av + c$ propagates an interval $[\underline{v}, \overline{v}]$ as

$$[A^+\underline{v} + A^-\overline{v} + c,\ A^+\overline{v} + A^-\underline{v} + c], \qquad A^+ = \max(A, 0), \quad A^- = \min(A, 0).$$

Elementwise nondecreasing activations act on the two endpoints. For GroupSort, sort the lower and upper vectors within each group; each order statistic is nondecreasing in every input coordinate, so these sorted endpoints enclose the layer output.

At the encoder output, define the center-relative drift bound

$$d_\infty(x,r) := \max\big\{\|\overline{z}(x,r) - g(x)\|_\infty,\ \|g(x) - \underline{z}(x,r)\|_\infty\big\}.$$

Every admissible point in the input ball satisfies $\|g(x') - g(x)\|_\infty \leq d_\infty(x,r)$. Under (A-Lin) and (A-Marg), a radius r satisfying

$$d_\infty(x,r) \max_{y \neq \hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_1 \ \leq\ \frac{\mathrm{margin}(x)}{2}$$

therefore certifies both this representation-drift bound and label preservation on its intersection with X. Indeed, for each $y \neq \hat{y}(x)$,

$$\begin{aligned} h(g(x'))_{\hat{y}} - h(g(x'))_y &= h(g(x))_{\hat{y}} - h(g(x))_y + (w_{\hat{y}} - w_y)^\top (g(x') - g(x)) \\ &\geq \mathrm{margin}(x) - \|w_{\hat{y}} - w_y\|_1 d_\infty(x,r) \geq \mathrm{margin}(x)/2 > 0. \end{aligned}$$

A.5 Proof of Proposition 4

Fix $x' \in X \cap \mathcal{B}_{r^\star(x)}(x)$.

1. By (A-Sub), $w_y^{(j)} = 0$ for $j \notin S$ and every y. Therefore $h(z)_y = (w_{y,S})^\top z_S + b_y$ for every z and y, where z_S denotes the restriction of z to coordinates in S.
2. In particular $f(x') = \arg\max_y \big[(w_{y,S})^\top g(x')_S + b_y\big]$.
3. By (A-Faith), $\|g(x')_S - c^\star(x')_S\|_\infty \leq \varepsilon$.
4. By Theorem 1, $f(x') = \hat{y}(x)$, so the decision is preserved *and* is read off from $g(x')_S$, which is within ε of $c^\star(x')_S$ in ℓ_∞.

□

A.6 Concept displacement and the audit-trail consequence

Proposition 6 (Sufficient hardness against $\mathcal{A}_{\mathrm{cs}}$)**.** *Under the conditions of Theorem 1, any $x' \in X$ with $f(x') \neq f(x)$ satisfies*

$$\|g(x') - g(x)\|_2 \ \geq\ \frac{\mathrm{margin}(x)}{2\ \max_{y \neq \hat{y}(x)} \|w_{\hat{y}} - w_y\|_2}.$$

Proof. If $f(x') \neq f(x)$ then the Step-2 chain in Theorem 1 must fail for some $y \neq \hat{y}$, i.e. there exists y with

$$\|w_{\hat{y}} - w_y\|_2 \cdot \|g(x') - g(x)\|_2 \ \geq\ \Delta_y(x) \ \geq\ \mathrm{margin}(x).$$

Using the supremum-over-y definition in (3) gives

$$\|g(x') - g(x)\|_2 \ \geq\ \frac{\mathrm{margin}(x)}{\max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2} \ \geq\ \frac{\mathrm{margin}(x)}{2\ \max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2},$$

which is the claim of Proposition 6. □

Corollary 5 (Audit-trail threat $\mathcal{A}_{\mathrm{cs}}^{=}$)**.** *Under the conditions of Theorem 1, every $x' \in X \cap \mathcal{B}_{r^\star(x)}(x)$ satisfies $f(x') = f(x)$ and $\|g(x') - g(x)\|_2 \leq L_g\, r^\star(x)$: the label-preserving audit adversary $\mathcal{A}_{\mathrm{cs}}^{=}$ cannot silently shift attribution by more than $L_g\, r^\star$ without leaving the certified ball.*

The corollary follows directly from the encoder bound and Theorem 1.

A.7 Proof of Proposition 1 (uniform radius rules)

Formalization. Call a function $R : (0,\infty)^3 \to [0,\infty)$ a *uniform Lipschitz–margin radius rule* if, for every input dimension $d \geq 1$, every concept dimension $k \geq 1$, every number of classes $K \geq 2$, every L_g-Lipschitz $g : \mathbb{R}^d \to \mathbb{R}^k$, every affine head $h(z) = Wz + b$, and every input x with $L_W = \max_{y \neq \hat{y}(x)} \|w_{\hat{y}(x)} - w_y\|_2 > 0$ and $m = \mathrm{margin}(x) > 0$, it guarantees $f(x') = \hat{y}(x)$ for all $x' \in \mathcal{B}_{R(L_g, L_W, m)}(x)$. Theorem 1 supplies the rule $R^{\mathrm{CRISP}}(L_g, L_W, m) := m/(2L_gL_W)$.

Witness classifier. Fix any $L_g, L_W, m > 0$. Set $d = k = 1$, $K = 2$, $x_0 = 0$, and

$$g(t) = L_g t, \qquad h_1(z) = \frac{L_W}{2} z + \frac{m}{2}, \qquad h_2(z) = -\frac{L_W}{2} z - \frac{m}{2}.$$

The logit gap is $h_1(g(t)) - h_2(g(t)) = L_gL_Wt + m$. Thus $\hat{y}(x_0) = 1$, $\mathrm{margin}(x_0) = m$, and the input-space boundary is $t^* = -m/(L_gL_W)$. For every $\epsilon > 0$, the point $x^\dagger = t^* - \epsilon$ has logit gap $-L_gL_W\epsilon < 0$, so it flips the label at distance

$$|x^\dagger - x_0| = \frac{m}{L_gL_W} + \epsilon = 2R^{\mathrm{CRISP}}(L_g, L_W, m) + \epsilon.$$

Upper bound on uniform rules. If $R(L_g, L_W, m) > m/(L_gL_W)$, choose $0 < \epsilon < R(L_g, L_W, m) - m/(L_gL_W)$. The witness then flips inside the ball certified by R, a contradiction. Hence

$$R(L_g, L_W, m) \;\leq\; \frac{m}{L_g\, L_W} \;=\; 2\, R^{\mathrm{CRISP}}(L_g, L_W, m) \qquad \forall\, (L_g, L_W, m) \in (0,\infty)^3. \tag{12}$$

□

Remark 2 (Half-margin choice). The proof of Theorem 1 gives the residual logit gap $h(g(x'))_{\hat{y}} - h(g(x'))_y \geq m - L_gL_W\|x' - x\|_2$. Consequently, each rule $R_q = q\, m/(L_gL_W)$ with $0 < q < 1$ preserves a gap of at least $(1-q)m > 0$ on its closed ball. CRISP chooses $q = 1/2$, retaining half the original margin. At $q = 1$, the witness reaches a tie at the boundary.

A.8 Proof of Proposition 5 (faithfulness necessity)

A translation in concept space, compensated by a shift in the head bias, preserves (A-Lin) and every quantity in Theorem 1: L_g, $\max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2$, and $\mathrm{margin}(x)$.

Setup. Fix $\eta > 0$ and an encoder $g_1 : \mathbb{R}^d \to \mathbb{R}^k$ together with a linear head (W, b_1) satisfying (A-Lip), (A-Lin), (A-Marg), (A-Sub), and (A-Faith) with parameter η; i.e. $\|g_1(x) - c^\star(x)\|_\infty \leq \eta$ for every x in the input support. Let $\hat{y}_1(x) := \arg\max_y (W g_1(x) + b_1)_y$ and let $r_1^\star(x)$ denote the radius of Theorem 1 at input x.

Construction. Pick any constant vector $c \in \mathbb{R}^k$ with $\|c\|_\infty = C$ for a free parameter $C > 0$, supported on the coordinate set S of (A-Sub) (so $c_j = 0$ for $j \notin S$). Define

$$g_2(x) \;:=\; g_1(x) + c, \qquad b_2 \;:=\; b_1 - Wc, \tag{13}$$

with the same weight matrix W.

Step 1: end-to-end decisions and $r^\star$ agree exactly. For every $x \in \mathbb{R}^d$,

$$W\, g_2(x) + b_2 \;=\; W(g_1(x) + c) + b_1 - Wc \;=\; W\, g_1(x) + b_1,$$

so $\arg\max$-based predictions agree pointwise: $\hat{y}_2(x) = \hat{y}_1(x)$. In particular $\mathrm{margin}_2(x) = \mathrm{margin}_1(x)$. The Lipschitz constant is also preserved, since translation by c is an isometry: for any x, x',

$$\|g_2(x) - g_2(x')\|_2 \;=\; \|g_1(x) - g_1(x')\|_2 \;\leq\; L_g\, \|x - x'\|_2.$$

The weight gap $\max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2$ is identical because W is unchanged. Plugging these equalities into (3): $r_2^\star(x) = r_1^\star(x)$ *exactly*, for every x.

Step 2: (A-Sub) is preserved. The head matrix W is untouched and the shift c is supported on S, so the end-to-end dependence $y \mapsto W g_2(x)$ is still supported only on coordinates in S. (A-Sub) holds for the pair (g_2, W) identically.

Step 3: (A-Faith) for g_2 is violated by $\Theta(C)$. For any x in the support,

$$\begin{aligned} \|g_2(x) - c^\star(x)\|_\infty &= \|g_1(x) + c - c^\star(x)\|_\infty \\ &\geq \|c\|_\infty - \|g_1(x) - c^\star(x)\|_\infty \\ &\geq C - \eta. \end{aligned}$$

Hence choosing $C := \eta + \Theta(1)$ gives $\|g_2 - c^\star\|_\infty \geq \Theta(1)$ uniformly, independent of η.

Step 4: conclusion. The pair (g_1, g_2) shares every quantity Theorem 1 depends on—same Lipschitz constant, same weight-gap, same margin, same end-to-end decisions on every input, hence the same certified radius $r^\star$ and the same certified ball $\mathcal{B}_{r^\star}(x)$. Yet g_1 is η-faithful while g_2 is $\Theta(1)$-unfaithful. No proposition whose hypotheses are only (A-Lip, A-Lin, A-Marg, A-Sub)—which is to say, no proposition provable from Theorem 1's assumptions alone—can distinguish the two. Therefore (A-Faith) or an equivalent faithfulness assumption is necessary for any bridge result to causal attribution. □

A.9 Proof of Proposition 2 (formal separation)

The construction is two-coordinate: the classifier label depends on a smooth coordinate while the concept encoder embeds a high-frequency coordinate that the head ignores. This decouples the smoothed label certificate from concept-space drift in closed form.

Proof. Fix $\bar{r} > 0$ and $M > 0$. Let the input dimension be $d = 2$ and the concept dimension be $k = 2$. Define

$$g(x_1, x_2) := \big(x_1,\, \sin(\omega x_2)\big), \qquad \omega := \frac{\pi M}{2\,\bar{r}}. \tag{14}$$

Use the binary linear head $W \in \mathbb{R}^{2\times 2}$ with $W_{1,:} = (1, 0)$, $W_{2,:} = (-1, 0)$, $b = 0$, so that $f(x) = \arg\max_y (Wg(x))_y = \operatorname{sign}(x_1)$. (The concept-2 coordinate is encoded by g but *not used* by the classifier; this asymmetry is the audit-trail-corruption scenario of the main-text radiology example, in which a concept that an auditor would inspect is not load-bearing for the decision.) Let $x_0 := (\bar{r}, 0)$.

Step 1 (Cohen certificate at x_0 is $\geq \bar{r}$). Since f depends only on x_1, the smoothed classifier $\bar{f}_\sigma(x) := \arg\max_y \mathbb{P}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}[f(x+\xi) = y]$ obeys

$$p_\sigma(x) := \mathbb{P}_\xi[f(x+\xi) = +1] = \mathbb{P}[\xi_1 > -x_1] = \Phi\big(\tfrac{x_1}{\sigma}\big),$$

and Cohen's certificate at x_0 from this exact probability [Cohen et al., 2019, Thm. 1] is

$$r_{\text{Cohen}}(x_0) = \sigma\,\Phi^{-1}\big(p_\sigma(x_0)\big) = \sigma\,\Phi^{-1}\big(\Phi(\bar{r}/\sigma)\big) = \bar{r}$$

for every $\sigma > 0$. Thus $\bar{f}_\sigma(x_0 + \delta) = +1$ whenever $\|\delta\|_2 < \bar{r}$.

Step 2 (concept-space drift at distance $\bar{r}/M$). Set $x_0^\dagger := (\bar{r},\, \pi/(2\omega))$. By direct computation using $\omega = \pi M/(2\bar{r})$:

$$\|x_0^\dagger - x_0\|_2 = \frac{\pi}{2\omega} = \frac{\bar{r}}{M},$$

so $x_0^\dagger$ lies inside $\mathcal{B}_{\bar{r}/M}(x_0)$. The concept drift is

$$\|g(x_0^\dagger) - g(x_0)\|_2 = \big\|\big(0, \sin(\pi/2) - \sin(0)\big)\big\|_2 = |1 - 0| = 1.$$

The base and smoothed labels at $x_0^\dagger$ are both $+1$, as at x_0, because its first coordinate is $\bar{r}$. Thus this perturbation realizes $\mathcal{A}_{\text{cs}}^{=}$; for $M > 1$ it lies strictly inside the Cohen-certified region.

Step 3 (combining). Choose $\bar{r}$ and M from the statement, set $\omega = \pi M/(2\bar{r})$, choose any $\sigma > 0$, and pick $x_0 = (\bar{r}, 0)$, $x_0^\dagger = (\bar{r}, \pi/(2\omega))$. Steps 1 and 2 verify both parts of the claim with quantitative constants in closed form, for arbitrary $\bar{r} > 0$ and $M > 0$. □ □

Remark 3 (Relation to the main-text statement). Proposition 2 as stated in Section 4.2 allows the concept drift bound to be any fixed constant; we stated 1 for concreteness. Scaling g's second coordinate by any $\alpha > 0$ rescales both drift and Lipschitz constant proportionally; the separation ratio M is invariant.

A.10 Proof of Proposition 3 (representation-blind impossibility)

Proof of Proposition 3. Fix $(g, h) \in \mathcal{F}$, x_0, $0 < \delta < \bar{r}$, and $B > 0$ as in the statement. Choose a unit vector $u \in \mathbb{R}^d$ and define

$$\psi(x) = \frac{B}{\delta} u^\top (x - x_0), \qquad \hat{g}(x) = (g(x), \psi(x)), \qquad \hat{h}(z, s) = h(z).$$

The closure assumption places $(\hat{g}, \hat{h})$ in $\mathcal{F}$. For an affine head $h(z) = Wz + b$, this extends W by a zero column and leaves b unchanged. For every $x \in \mathbb{R}^d$, $\hat{h}(\hat{g}(x)) = h(g(x))$, so the complete logits and base predictions agree. They also induce identical label distributions under the fixed smoothing law at every input. Hence

$$R((\hat{g}, \hat{h}), x_0) = R((g, h), x_0) \geq \bar{r}.$$

Set $x^\dagger = x_0 + \delta u$. Then $\|x^\dagger - x_0\|_2 = \delta < \bar{r}$, while $\psi(x_0) = 0$ and $\psi(x^\dagger) = B$. Orthogonality of the added coordinate gives

$$\|\hat{g}(x^\dagger) - \hat{g}(x_0)\|_2^2 = \|g(x^\dagger) - g(x_0)\|_2^2 + B^2 \geq B^2.$$

The prediction certified by R is preserved at $x^\dagger$ by the certificate's label-preservation guarantee. □

Fixed concept dimension. For an affine head, the dimension can be preserved when there is a unit vector $v \in \mathbb{R}^k$ with $(w_y - w_1)^\top v = 0$ for all y, and the model class is closed under the encoder replacement below with the head unchanged. This direction exists whenever $k > \operatorname{rank} D$, where D has rows $w_y - w_1$; in particular, $k \geq 2$ suffices for binary classification. With $P = I - vv^\top$ and the same ψ, set

$$\hat{g}(x) = Pg(x) + \psi(x)v, \qquad \hat{h} = h.$$

Since $(w_y - w_1)^\top P = (w_y - w_1)^\top$, all pairwise logit differences, base predictions, and noise-marginalized label distributions are unchanged. The two components of the encoder increment are orthogonal, giving

$$\|\hat{g}(x^\dagger) - \hat{g}(x_0)\|_2^2 = \|P(g(x^\dagger) - g(x_0))\|_2^2 + B^2 \geq B^2.$$

When $Wv = 0$, the complete logits are preserved as well.

B Scope of interventional alignment

The two terms of L_{ICA} act on the supplied intervention pairs. The freeze term penalizes changes outside the target coordinate, and the hinge term penalizes signed target responses below κ. Their effect on coordinate fidelity is evaluated in the ICA ablation (Table 2), hinge sweep (Table 3), and proxy-noise sweeps (Section C.7).

For both synthetic SCMs and natural images, the causal bridge in Proposition 4 assumes (A-Faith) on the admissible certified region. The true-concept drift bound in Equation (5) uses this condition at both endpoints. MCC and calibrated residuals describe empirical alignment on sampled data; the residual calculation is specified in Section C.7.

C Experimental details

C.1 Binary-SCM pipeline

SCM definition. `code/toy_scm/scm_3node.py` exposes `SCMConfig` with `n_concepts` $\in \{3, 6\}$, `n_bg`$= 5$, `input_dim` $= 16$. Concepts and background coordinates are exogenous i.i.d. $\mathcal{N}(0, 1)$. Observations are $X = AC + BX_{\mathrm{bg}} + \epsilon$, where $[A\ B]$ has orthonormal columns and

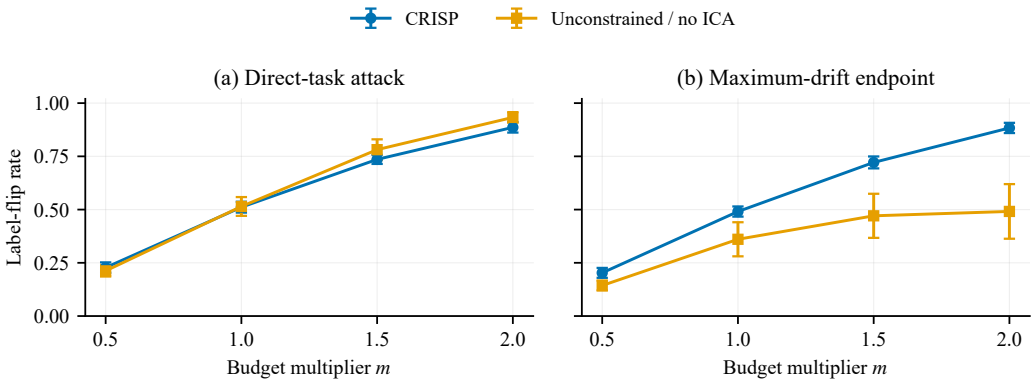

Figure 2: Waterbirds attacks on frozen ResNet-18 features. Left: direct-task PGD flip rate, taking any flipped restart endpoint. Right: flip rate of the endpoint with largest representation drift across restarts. Both use 50 steps, 5 restarts, and the common-clean-correct subset (1,901 pooled model–sample pairs). The shared budget is m times each seed's mean successful CRISP label-attack radius. Error bars show sample standard deviations across five seeds.

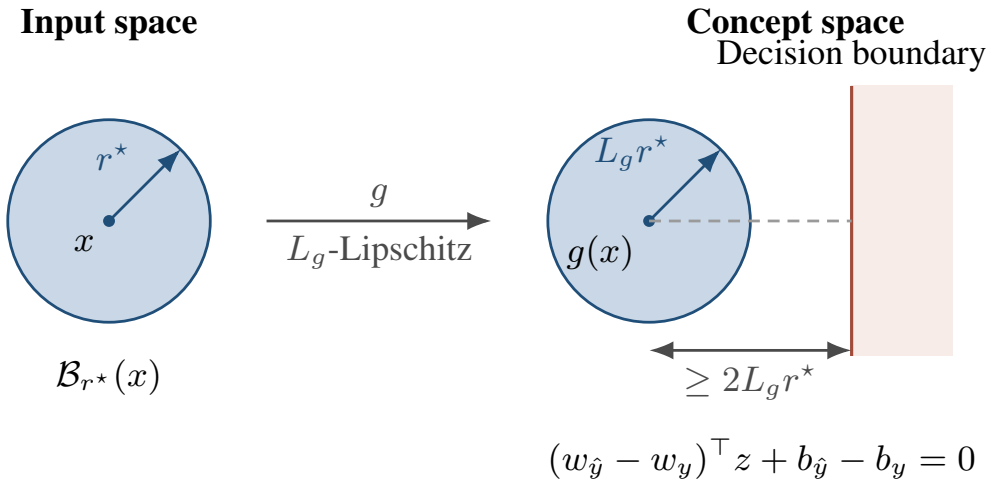

Figure 3: Geometry of Theorem 1. The image of $X \cap \mathcal{B}_{r^\star(x)}(x)$ under g is contained in $\mathcal{B}_{L_g\, r^\star(x)}(g(x))$. Each $\hat{y}$-versus-y decision hyperplane is at least $\mathrm{margin}(x)/L_W = 2L_g\, r^\star(x)$ from $g(x)$, where $L_W = \max_{y \neq \hat{y}} \|w_{\hat{y}} - w_y\|_2$; equality is drawn here. Thus the enclosing concept ball retains a logit gap of at least $\mathrm{margin}(x)/2$.

independent observation noise is $\epsilon \sim \mathcal{N}(0, 0.05^2 I_{16})$. The rendering is injective in C for fixed background and noise realization. Labels follow $Y = \mathrm{sign}(w^\top C)$, with $w = (1, -1, 0.5)$ for $k = 3$; for $k = 6$, three standard-normal weights are appended, drawn once per seed. Ground-truth interventions $\mathrm{do}(C_j := C_j + \Delta_j)$ share the original background and noise realization, so the synthetic proxy oracle T_j is exact.

Architectures. CRISP *encoder.* Spectrally-normalized linear layers with GroupSort2 activations, `hidden_dim` $= \max(32, 8k)$, 3 affine layers and 2 GroupSort2 activations. Certification uses products of verified matrix spectral-norm upper bounds. Across the five headline seeds, these products range over $[1.00004, 1.01119]$ for $k = 3$ and $[1.00017, 1.00994]$ for $k = 6$. *Head.* Affine, 2 classes, without normalization. *Unconstrained/no-ICA ablation.* Same encoder width and depth but with `ModelConfig.constrained=False` (no spectral normalization, ReLU in place of GroupSort; `ica_weight=0`).

Training. Adam, $\mathrm{lr} = 10^{-3}$, batch 128, 500 optimizer steps per seed, no learning-rate schedule or weight decay. The loss is $L_{\mathrm{task}} + L_{\mathrm{ICA}}$ (freeze 1, hinge 1, $\kappa = 0.1$); concept labels are used for evaluation, not direct supervision. The five-seed headline uses seeds 0–4; the budget sweep and ablations use seeds 0–2.

Evaluation and drift initialization. 96 held-out samples per seed. The headline uses cross-entropy PGD over 15 budget levels up to 5.0 (30 steps, step size 0.1, one zero start). The per-sample evaluation adds APGD-style ascent on the binary logit margin (50 steps, 3 random restarts) on the same budget grid. The latter uses momentum and adaptive step sizes, following the APGD recipe [Croce and Hein, 2020]; the code names it APGD-DLR. Each attack retains its first successful endpoint; the combined analysis takes the smaller successful endpoint distance. The drift diagnostic uses 40 projected normalized-gradient steps of size 0.05 on $\|g(x + \delta) - g(x)\|_2^2$, with $\|\delta\|_2 \le m b_s$ for $m \in \{0.5, 1, 1.5, 2, 3\}$. Here b_s is the original per-seed mean first-success PGD grid budget, retaining 5 for failed searches. These fixed references are stored separately from the successful endpoint distances. It starts in a random direction at radius $0.25 m b_s$, since the drift gradient vanishes at $\delta = 0$. All reported binary-SCM drift attacks use this nonzero initialization. In the two-model budget sweep, both models use CRISP's b_s. The four-cell ablation and parameter sweeps use each variant's own reference at $m = 1$. Each rate uses that model's own clean-correct subset and records endpoint flips relative to the clean prediction. MCC via Hungarian matching on the $|\mathrm{corr}|$ matrix between learned coordinates and ground-truth concepts.

Radius checks. Binary-SCM radii use saved models from same-configuration reproductions. Matrix-norm upper bounds are checked by exact positive-definiteness tests, and radii are rounded down using exact margins of the saved real-valued networks. The reported $\hat{r}_{\mathrm{adv}}$ is an upper bound on the distance to a saved prediction-flipping endpoint. Radius means use all clean-correct samples; distance and tightness means use successful searches, first within each seed and then across seeds. Failures are counted separately using saved success masks. Violations count $r^\star > \hat{r}_{\mathrm{adv}}$ on successful searches. Four-class and Waterbirds checks retain their first-success grid budgets and tolerance 10^{-6}.

Ablation-cell standard deviations. For Table 2 (three-seed means; source `p2_ica_lip_20260424_172622`), sample standard deviations across seeds (divisor 2) are: CRISP σ_{CA}=0.010, σ_{MCC}=0.099, $\sigma_{\mathrm{ASR_{cs}}}$=0.038; ICA-only σ_{CA}=0.030, σ_{MCC}=0.162, $\sigma_{\mathrm{ASR_{cs}}}$=0.025; Lip-only σ_{CA}=0.006, σ_{MCC}=0.021, $\sigma_{\mathrm{ASR_{cs}}}$=0.067; unconstrained/no-ICA σ_{CA}=0.006, σ_{MCC}=0.027, $\sigma_{\mathrm{ASR_{cs}}}$=0.072. The radius check finds zero violations on 562 successful searches: 278 for CRISP and 284 for Lip-only; one CRISP search fails. For the hinge, magnitude-noise, and observation-noise sweeps, failure counts in table order are $(1, 1, 1, 1)$, $(1, 1, 1, 1)$, and $(1, 1, 1, 1, 12)$, respectively.

C.2 Four-class SCM and head comparison

Data and training. For $k \in \{12, 24\}$, independent standard-Gaussian concepts and eight nuisance variables are mixed by an orthogonal matrix into $k + 8$ observed dimensions, with additive Gaussian noise of standard deviation 0.05. Two dense orthonormal vectors $u, v \in \mathbb{R}^k$ define class weights $(u + v, u - v, -u + v, -u - v)/\sqrt{2}$; the largest concept logit determines the four-class label.

Counterfactual pairs share nuisance and noise while changing one concept coordinate. All four classes occur in each evaluation set.

The encoder has three affine layers, hidden width $\max(64, 4k)$, and a k-dimensional output. CRISP uses spectral normalization and two GroupSort2 activations. Heads are affine or a width-64 one-hidden-layer ReLU MLP, both with biases. Adam uses learning rate 10^{-3}, batch size 256, and 1,500 steps. The loss is $L_{\text{task}} + L_{\text{ICA}}$ with $\mu = 1$, hinge margin 0.1, and the coordinate-summed freeze term of Equation (6). Seeds 0–4 share task-data and evaluation streams across variants. The unconstrained/no-ICA ablation uses ReLU, unnormalized encoder layers, and task supervision. Additional ICA-only and Lip-only variants are evaluated at $k = 12$: together with the three variants at both widths, the saved experiment has 40 runs.

Certificate and attack statistics. Accuracy and MCC use all 512 test points per run. Mean $r^\star$ uses all correctly classified points; attack metrics use the first 128 correctly classified points for each model. Encoder bounds are products of realized matrix spectral norms. Their five-seed mean $\pm$ sample SD is 1.014 ± 0.007 and 1.011 ± 0.010 for the linear/MLP CRISP models at $k = 12$, and 1.015 ± 0.008 and 1.013 ± 0.004 at $k = 24$. The MLP head uses Equation (11), taking the maximum only over competitors of the current prediction.

Label-radius search tests 15 budgets $5/15, 2(5/15), \ldots, 5$. At each budget, cross-entropy PGD uses 50 steps and one random start, and APGD-DLR uses 50 steps and 3 random restarts. The estimate is the smaller first-success budget from the two attacks. All 5,120 attacked model–sample pairs across the 40 runs find flips, with zero failures and zero violations in the radius checks; the linear/MLP CRISP comparison comprises 2,560 of these pairs. Tightness averages per-sample ratios within each run and then reports the mean and sample SD across five seeds (divisor 4).

Direct-task and representation-drift attacks use 50 steps, 5 random-ball starts, and a shared budget $\min\{5, 1.5\bar{r}_{\text{adv}}^{\text{CRISP-linear}}\}$ for each (k, seed). Direct-task PGD maximizes cross-entropy and records a flip if any restart endpoint changes the clean prediction. Drift PGD maximizes squared representation displacement and selects the endpoint with largest drift. Table 6 reports the direct-task comparison on each model's own clean-correct subset.

Table 6: Four-class SCM: concept fidelity and direct-task flip rates at the shared budget. Values are five-seed mean $\pm$ sample standard deviation.

k	Variant	MCC	Direct-task flip rate
12	CRISP, linear	0.526 ± 0.066	0.894 ± 0.023
12	CRISP, MLP	0.588 ± 0.049	0.894 ± 0.027
12	Unconstrained/no-ICA	0.354 ± 0.024	0.894 ± 0.017
24	CRISP, linear	0.260 ± 0.009	0.891 ± 0.010
24	CRISP, MLP	0.408 ± 0.024	0.891 ± 0.010
24	Unconstrained/no-ICA	0.255 ± 0.006	0.892 ± 0.010

C.3 Compute

The local workstation was equipped with an NVIDIA RTX 4060. The binary reproductions used CPU PyTorch 2.11.0, 16 threads; training the 93 model instances took about 504 seconds. The corrected five-checkpoint Waterbirds re-evaluation took approximately 339 seconds on its recorded CUDA device, using cached features. This timing excludes the original CRISP training and feature extraction.

The 40 saved four-class runs record a total of approximately 17.5 minutes of wall-clock time using CUDA, including training and attack evaluation.

C.4 Reproduction

From the code-package root, regenerate the binary statistics and figures with

```
python binary_verified/paper_results.py
python code/scripts/plot_saved_synthetic.py
```

The `binary_verified/` directory contains 93 saved model instances, per-sample results, norm checks, and training/attack scripts. Its README maps each table and figure to its source; historical outputs are preserved separately. Waterbirds training is implemented in `crisp-exp/scripts/train_waterbirds.py`; the corrected re-evaluation is in `rebuttal/experiments/run_waterbirds_corrected.py`. The saved five-seed outputs used here are in `rebuttal/experiments/results_waterbirds_corrected/`. The four-class experiment uses `rebuttal/experiments/run_multiclass_rebuttal.py`; the results reported in Tables 4 and 6 are saved in `rebuttal/experiments/results_submitted_objective/`.

C.5 Waterbirds training and evaluation

We use Waterbirds v1.0 [Sagawa et al., 2020] through WILDS, with its official training, validation, and test splits. RGB images are resized directly to 224×224, scaled to $[0, 1]$, and normalized with channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. A frozen Torchvision ResNet-18 with `IMAGENET1K_V1` weights and its classifier removed produces the 512-dimensional features.

The encoder dimensions are $512 \to 64 \to 64 \to 64$, with GroupSort2 between affine layers, followed by a $64 \to 2$ affine classifier. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and 20 epochs; checkpoints are selected by validation worst-group accuracy. Environment-swap pairs target coordinate 0 by changing background while matching bird type, or coordinate 1 by changing bird type while matching background. The remaining coordinates receive the freeze term. The task-only unconstrained/no-ICA ablation uses unnormalized affine layers with ReLU at the same width and depth, trained for 20 epochs without per-concept supervision.

For certification, the encoder bound is the product of the three realized weight-matrix spectral norms. The full-test accuracy and worst-group accuracy appear in Table 5. On the fixed attack subset of the first 500 test points, the corresponding values are 0.792 ± 0.010 and 0.555 ± 0.030.

Label-radius search uses the PGD/APGD settings in Section 6.3. Failed searches are stored separately and excluded from tightness means. The comparison attacks use uniform-in-volume ℓ_2-ball starts, projected normalized steps of size $2\varepsilon/50$, and 5 restarts. Direct-task PGD maximizes cross-entropy with the clean label; drift PGD maximizes squared encoder displacement without a label constraint. Table 7 gives the denominators. For each seed, tightness is the mean of per-sample $r^\star_\phi / \hat{r}_{\mathrm{adv}}$ over found flips; Table 5 then averages these five seed means and reports their sample standard deviation (divisor 4).

Table 7: Waterbirds denominators. Each seed uses the first 500 test points for attacks. "Common correct" is the subset correctly classified by both CRISP and the unconstrained/no-ICA model.

Seed	CRISP correct	Found flips	Failures at 5	Common correct
0	399	399	0	373
1	402	389	13	387
2	397	386	11	385
3	391	386	5	375
4	391	389	2	381
Total	1,980	1,949	31	1,901

C.6 Synthetic-SCM extras

Label-radius refinement. The appendix combines PGD and APGD on the same models and samples as Table 1. Their union finds flips for all $457/466$ clean-correct samples at $k = 3/6$. Figure 5 pools the per-sample ratios using the smaller successful endpoint distance. Arrays are stored in `binary_verified/extended/appendix/`. For $k = 6$, sample-bootstrap 95% BCa intervals (10,000 resamples) are $[0.316, 0.333]$ for PGD ($n = 462$) and $[0.316, 0.332]$ for PGD/APGD ($n = 466$).

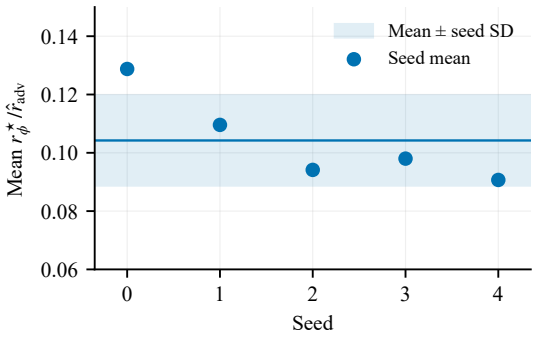

Figure 4: Waterbirds tightness, computed on each seed's found flips. Dots show seed means; the line and shaded band show the mean and sample standard deviation across seeds, 0.104 ± 0.015. Table 7 lists successes and failures.

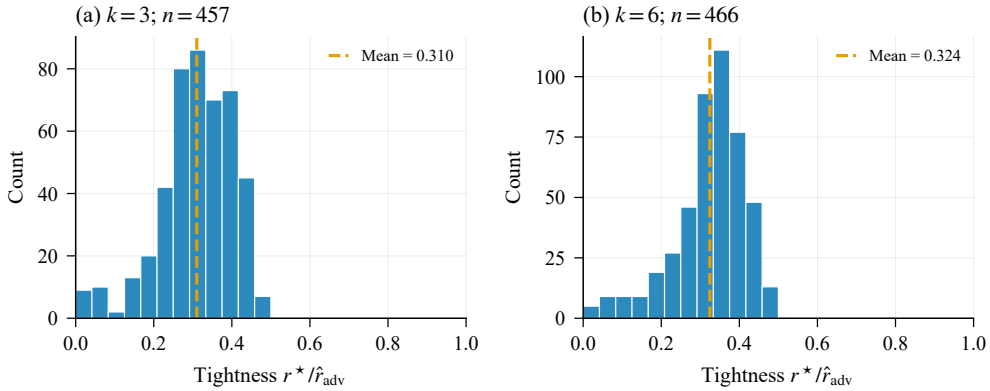

Figure 5: Per-sample tightness $r^\star(x)/\hat{r}_{\text{adv}}(x)$ on the synthetic SCM, five seeds. Left: k=3 (n=457, mean 0.310). Right: k=6 (n=466, mean 0.324). Distances use the smaller successful PGD/APGD endpoint distance.

Drift-endpoint flip-rate sweep. Table 8 reports ASR_{cs} versus attack budget mb_s, $m \in \{0.5, 1.0, 1.5, 2.0, 3.0\}$. The gap between the unconstrained/no-ICA ablation and CRISP peaks at m=1.5.

Table 8: ASR_{cs} at matched budget, mean over 3 seeds. Gap is unconstrained/no-ICA − CRISP.

k	m	CRISP	Unconstr.	Gap	Notes
3	0.5	0.191	0.197	+0.006	
3	1.0	0.318	0.368	+0.050	
3	1.5	**0.325**	**0.457**	**+0.132**	gap peaks
3	2.0	0.451	0.518	+0.066	
3	3.0	0.418	0.525	+0.107	
6	0.5	0.151	0.190	+0.039	
6	1.0	0.287	0.328	+0.041	
6	1.5	**0.308**	**0.444**	**+0.135**	gap peaks
6	2.0	0.427	0.458	+0.031	
6	3.0	0.420	0.447	+0.027	

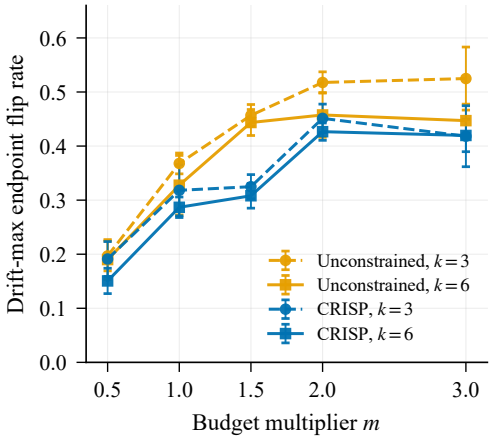

Figure 6: Synthetic-SCM label-flip rate at the endpoint of drift maximization vs. matched fixed-reference budget mb_s, for $k \in \{3, 6\}$. The unconstrained ablation removes both the spectral cap and L_{ICA}. Markers show the five evaluated budgets; error bars show standard deviations over three seeds (divisor 3). Mean values are listed in Table 8.

Table 9: Per-factor decomposition of $r^\star = \mathrm{margin}/(2L_g L_W)$ on a three-factor SCM with six learned coordinates (hidden width 32, 1,500 training steps, 5 seeds, 128 test samples/seed). $L_g^{\mathrm{nom}} = c^3$ is the nominal cap product; $r^\star$ uses each model's verified upper bound. L_g^{local} is a random-spherical-probe slope estimate; L_W is the maximum pairwise head-row norm. PGD uses 20 budgets up to 5, 40 steps of size 0.05, and a zero start. Distance and tightness use found flips.

c	$\bar{\mathrm{margin}}$	L_g^{nom}	L_g^{local}	$\bar{L}_W$	$\bar{r^\star}$	$\bar{\hat{r}}_{\mathrm{adv}}$	$\overline{r^\star/\hat{r}_{\mathrm{adv}}}$
0.25	0.084	0.016	0.0095	6.56	0.412	0.882	**0.393**
0.50	0.728	0.125	0.073	7.37	0.394	0.865	**0.384**
0.75	1.846	0.422	0.221	6.33	0.345	0.836	**0.367**
1.00	2.635	1.000	0.481	5.63	0.235	0.828	**0.278**

$L_g^{\mathrm{nom}}/\overline{L_g^{\mathrm{local}}} \in [1.65, 2.08]$; zero violations on found flips.
Found/failed in cap order: 570/28, 597/22, 587/34, 573/49.

C.7 Ablation breakdowns

Proxy-magnitude noise. We inject zero-mean Gaussian noise on the intervention magnitude Δ_j that is fed to the ICA hinge (the counterfactual $\tilde{x}$ still uses the true Δ_j), sweeping $\sigma_\Delta \in \{0, 0.1, 0.25, 0.5\}$ on synthetic k=6 (three seeds, source `p2_proxy_20260425_195753`). The freeze term in L_{ICA} is independent of Δ. MCC stays within 0.002 of its baseline for $\sigma_\Delta \le 0.25$; at $\sigma_\Delta = 0.5$, clean accuracy is 0.979 and MCC is 0.753. Table 10 also reports intervention-direction consistency (CF): the fraction of 256 held-out counterfactual pairs per seed for which $\mathrm{sign}(g_j(\tilde{x}) - g_j(x)) = \mathrm{sign}(\Delta_j)$ at the targeted coordinate j, averaged over three seeds.

Table 10: Proxy-magnitude noise sensitivity on synthetic SCM k=6, three-seed means. The radius checks record zero violations in every cell.

σ_Δ	MCC	CF	$\bar{r^\star}$	$\mathrm{ASR}_{\mathrm{cs}}$
0.00 (ground truth)	**0.773**	**0.815**	0.349	**0.283**
0.10	0.773	0.818	0.351	0.287
0.25	0.771	0.811	0.350	0.290
0.50	0.753	0.798	0.341	0.351

Concept alignment under observation noise. We Hungarian-match learned to ground-truth coordinates, fit a per-dimension affine rescaling, and report $\varepsilon_{\mathrm{meas}}(x) := \|z_{\mathrm{aligned}}(x) - c^\star_{\mathrm{matched}}(x)\|_\infty$.

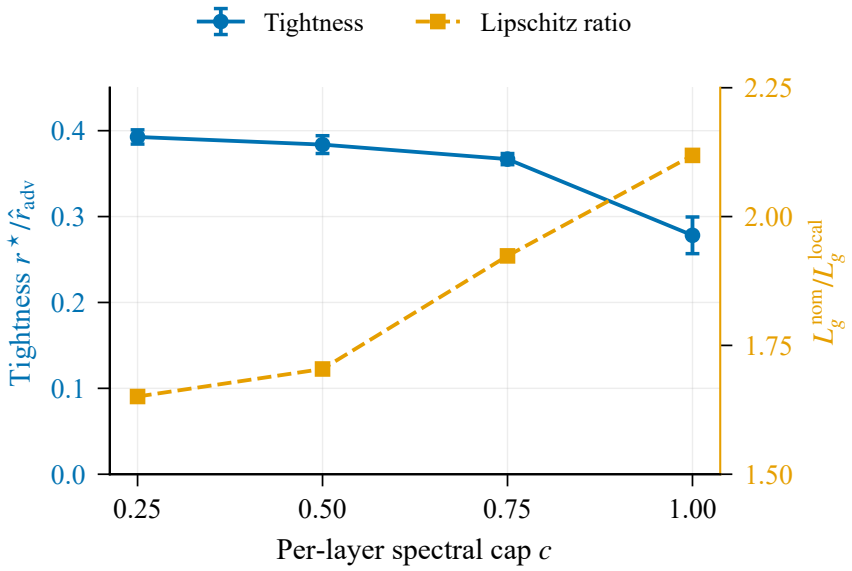

Figure 7: Tightness $r^\star/\hat{r}_{\text{adv}}$ vs. per-layer spectral cap c with three SCM factors and six learned coordinates (5 seeds, error bars are standard deviations across seeds, divisor 5). Tightness rises from 0.278 to 0.393 as c decreases from 1.00 to 0.25. The right axis shows the mean per-seed ratio $L_g^{\text{nom}}/L_g^{\text{local}}$.

Matching, affine fitting, and residual evaluation use the same held-out sample batch. The mean residual summarizes alignment after calibration; (A-Faith) concerns the uncalibrated encoder g throughout the region where the causal guarantee is applied. We inject Gaussian noise $\mathcal{N}(0, \sigma_c^2 I)$ on the counterfactual observation $\tilde{x}$ before the encoder. Sweeping $\sigma_c \in \{0, 0.05, 0.10, 0.25, 0.50\}$ (three seeds, source `p2_eps_20260426_004604`) gives the sensitivity of coordinate fidelity, radius, and drift-endpoint flips to proxy observation noise in Table 11.

Table 11: Concept alignment under counterfactual observation noise on synthetic SCM k=6, three-seed means. $\bar{\varepsilon}_{\text{meas}}$ is the sample mean of the L_∞ residual after Hungarian matching and per-dimension affine rescaling fitted on the same evaluation batch.

σ_c	$\bar{\varepsilon}_{\text{meas}}$	MCC	$\bar{r^\star}$	ASR_{cs}
0.00	**1.04**	**0.773**	**0.349**	**0.283**
0.05	1.05	0.767	0.347	0.294
0.10	1.09	0.744	0.331	0.339
0.25	1.28	0.618	0.290	0.356
0.50	1.43	0.474	0.241	0.393

Hinge margin κ. Table 3 reports the sweep over $\kappa \in \{0.05, 0.10, 0.25, 0.50\}$ on synthetic SCM $k = 6$ (three-seed source `p2_kappa_20260424_172332`). The headline runs of Table 1 use $\kappa = 0.10$.

C.8 Future natural-image evaluation

The planned follow-up extends the frozen-feature Waterbirds study to end-to-end image encoders and pixel-space certification. It requires a certified bound covering the backbone and concept head, and intervention pairs from attribute editing, matched retrieval, or controlled rendering.

Candidate datasets are CUB-200-2011 [Wah et al., 2011], with the 112 binary attributes and class-level denoising of Koh et al. [2020]; Waterbirds [Sagawa et al., 2020], with bird-type and background

interventions; and CANDLE [Reddy et al., 2022], whose six factors are object shape, colour, size, rotation angle, light position, and scene.

The comparison would cover ERM, IRM, vREx, GroupDRO, CBM, causal bottlenecks, randomized smoothing, and CROWN-IBP [Zhang et al., 2020]. Evaluation would measure accuracy, concept fidelity and accuracy after concept interventions [Havasi et al., 2022], certified radius, and direct-task attack success at matched budgets. Label-preserving concept drift would be evaluated with the clean-label constraint enforced during optimization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the concept-space certificate, separation from label-only guarantees, conditional causal bridge, and empirical scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 describes the current evaluation scope and the next steps for pixel-space certification and external concept-fidelity comparisons.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions (A-Lip), (A-Lin), (A-Marg), (A-Faith), (A-Sub), (A1) are declared in Section 3. Theorem 1 and the propositions on tightness, separation, and the causal bridge have proof sketches in Section 4 and full proofs in Appendix A. The concept-displacement and audit-trail consequences are collected in Appendix A.6.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix C specifies the binary and four-class SCM generators, architectures, training, and evaluation. Appendix C.5 gives the executed Waterbirds setup, including data splits, image preprocessing, backbone weights, and attack denominators. The end-to-end image study in Appendix C.8 is a future protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The prepared code archive contains the binary-SCM, four-class, and Waterbirds scripts, saved results, and setup instructions. Public access to this updated archive is pending.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 reports the binary, four-class, and Waterbirds configurations; Appendix C provides full hyperparameters, optimizer, batch size, seeds, and evaluation budget.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: CRISP binary-SCM headline results use five seeds; binary baselines and ablations use three. The four-class and Waterbirds experiments use five seeds. Appendix C reports standard deviations and the specified bootstrap intervals. Seeds govern training and synthetic-data generation; Waterbirds uses fixed data splits.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.3 reports the workstation hardware, the binary-SCM environment and training time, and CUDA timings for the corrected Waterbirds evaluation and 40 four-class runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The empirical studies use procedurally generated SCMs and Waterbirds images. They involve no human-subject experiments or privacy-sensitive data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The certificate supports auditing concept-based predictions by bounding representation drift. Confusing stability of learned coordinates with causal validity could create misplaced trust; Section 3 specifies the faithfulness and subgraph conditions for the causal interpretation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work evaluates classifiers using generated SCM samples and the established Waterbirds benchmark. It introduces no dataset or generative model requiring release safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Waterbirds v1.0 [Sagawa et al., 2020], built from CUB-200-2011 [Wah et al., 2011] and Places images, is used for non-commercial research under its dataset terms. The WILDS loader uses the MIT license; Torchvision code uses BSD-3-Clause. Appendix C.5 identifies the pretrained backbone weights. Torchvision documents their separate model-use terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code package documents the binary-SCM, four-class, and Waterbirds experiments through a README, dependency instructions, experiment scripts, and saved results. The original binary-SCM code retains its MIT license. Appendix C.4 lists execution commands and entry points.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects; IRB approval is not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the core methodology. Any use for text editing or formatting does not impact the scientific contribution and per NeurIPS policy does not require declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.