Representation Is Not Reliance: Concept-Based Causal Diagnostics of Shortcut Learning in Vision

Anonymous Author(s)
Affiliation
Address
email

Abstract

Vision models systematically exploit predictive but task-irrelevant visual cues as spurious shortcuts. This typically manifests as a severe performance drop when models generalize from in-domain to out-of-domain data. Although such shortcuts range from obvious artifacts to subtle signals, their visual form often remains elusive, even when their presence is statistically evident. To address this blind spot, we propose a diagnostic framework to expose the specific visual patterns that encode such shortcuts, paving the way for targeted removal. Specifically, we apply automated concept discovery to a shortcut-trained model to isolate candidate shortcut concepts and trace how localized visual patterns propagate across layers and become task-predictive. On a recently curated benchmark for investigating spurious correlations, these concepts lose predictive utility under shift despite being predictive under bias. Reusing the same exemplar-defined concepts in a model trained without the shortcut bias shows that these patterns remain separable and spatially coherent, and regain predictive alignment when the shortcut distribution is restored. Across depth, the same localized concepts remain traceable through most layers and weaken only in the final stage, where spatial detail is pooled before classification. Targeted interventions further show that the identified shortcuts are carried by localized visual regions and that weakening them changes predictive reliance, suggesting that shortcut behavior is governed by whether spurious cues are encoded, how they are organized, and when and where they drive prediction.

1 Motivation

Shortcut learning shows that deep models often rely on predictive but causally irrelevant cues [36]. In vision, such cues include background statistics [5], contextual co-occurrences [31], acquisition artefacts [37], and dataset-specific biases [41]. Such shortcuts can perform well when a biased distribution is preserved, yet fail systematically once correlation breaks under shift [18, 14]. Spurious correlations therefore pose a robustness problem not merely of accuracy loss, but of misgeneralization.

Shortcuts emerge when spurious cues are easier to learn than core features, becoming visible in training dynamics [36] and persisting as spatially grounded [8], prediction-relevant [24], yet non-robust features in model representations [22]. Existing diagnostics, however, operate at two incomplete levels. Instance-level analyses expose suspicious prediction-specific evidence [5, 15, 2] but rarely yield reusable shortcut units [12], whereas population-level audits reveal distributional failures [14] but not when recurrent cues drive predictions [42]. Concept-level evidence grounds the missing mechanism-level rationale. Recurrent visual concepts serve as diagnostic units [19, 11], allowing us to test whether encoded shortcut cues [13] causally drive prediction [40].

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

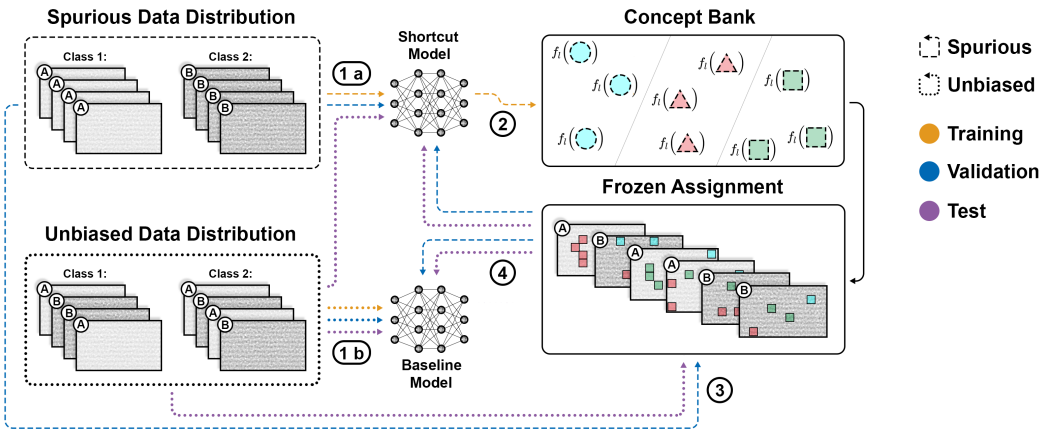


Figure 1: Setup for the diagnostic concept-level shortcut analysis:
(1 a) Shortcut model trained on spurious split and (1 b) baseline model trained on the unbiased split.
(2) Concept bank constructed from the shortcut-model´s activations at a selected intermediate layer.
(3) Fixed concept definitions assign spatial activations from spurious and unbiased diagnostic splits.
(4) Assigned regions define the support for downstream cross-split and cross-model diagnostics.

We address this gap with a causal procedure for isolating and validating shortcut mechanisms in vision models. Figure 1 summarizes the setup. First, we use a shortcut-trained model (1 a) as a deliberate amplifier, creating a regime in which shortcut-driven predictive behavior becomes especially pronounced and its supporting visual cues are more likely to surface as dominant internal regularities. We then derive a concept bank (2) from activations at a selected intermediate layer in the shortcut model's representation space [19], treating these recurrent predictive patterns as contrast agents that renders candidate carriers of the shortcut mechanism explicit, separable, and trackable. Next, the learned concept definitions are kept fixed and reused across diagnostic conditions (3), with spatial activation vectors from spurious and unbiased splits embedded at the same selected layer and assigned whenever they fall within the learned decision region of a concept. This way, concept definitions remain fixed while occurrence is measured beyond the shortcut-training regime. Finally, these fixed assignments provide the basis for downstream diagnostics (4), including cross-model transfer, causal behavior across shifts, architecture-level transport, and targeted region removal. This perspective separates concept structure from predictive importance [40], treats shortcut learning as conditional [3], and enables detection [42], validation [50], and disruption [5] of shortcut mechanisms. Beyond shortcut detection, we ask which patterns emerge under shortcut amplification, transfer beyond the biased regime, become causally effective or collapse, and whether targeted removal weakens them.

2 Background

Concept-level explanations describe model behavior through recurring, human-interpretable visual regularities rather than individual pixels or neurons. In vision, such regularities can be studied by relating intermediate activations to semantic units including object parts, textures, materials, colors, or spatial patterns. Network Dissection [8] measures alignment between unit activation maps and concept segmentation masks. ACE [19] avoids predefined labels by decomposing images into local regions, embedding them at a selected layer, clustering the activations, and summarizing clusters as concept prototypes. Concept Activation Vectors (CAVs) [24] encode concepts as representation-space directions for separability-based validation, while TCAV measures prediction sensitivity along them.

Representation-level analyses reveal whether information is encoded, separable, or geometrically organized inside a model. Linear probes test decodability from intermediate activations [4], while dissection [8] and concept-vector methods [24] expose semantic alignment and concept directions. Causal tracing links this representational evidence to model behavior by testing whether restoring concept-related clean states within an otherwise corrupted forward pass recovers a prediction [34]. In shortcut learning, such tracing reveals learned decision rules that exploit spurious concept–target associations that hold in-distribution yet break under removal, inversion, or decorrelation [18].

3 Related Work

Work on shortcut detection and mitigation has developed complementary routes for identifying or weakening spurious reliance. Robustness benchmarks expose failures under distribution and subpopulation shifts [26], motivating invariance-based methods that seek predictors stable across environments [3]. Training-dynamics analyses detect when spurious features persist or dominate [36], while last-layer retraining reduces spurious dependence by reweighting biased representations on decorrelated data [25, 23]. Difficulties arise when task evidence and shortcut cues are intertwined in representation space, complicating removal and motivating shortcut-carrying pattern isolation [13].

Concept-based methods provide such an interface for interpretation, semantic testing, and intervention. Recent methods extend CAV-style explanations through statistical importance tests [40], automatic concept factorization with local attribution maps [15], and concept bottleneck variants that improve fine-grained visual-concept alignment [51] or impose causal structure on the bottleneck [17]. Others examine drift in separability-optimized concept directions [38], diagnose errors through conceptual counterfactuals [1], or insert concept layers for intervention without explicit concept labels or redesign [9]. Together, these methods establish concepts as reusable units for explanation and intervention.

Building on this interface, concept-level auditing work uses visual concepts, prototypes, or dataset structure to identify bias-inducing evidence. Dataset audits identify classes affected by spurious cues [35], while medical imaging studies show that acquisition signatures [43] and dataset construction artifacts [44] can become predictive shortcuts. Concept-based auditing methods categorize visual concepts into target, context, and bias types [11]. Prototype-based [27] and object-centric methods [6] instead detect shortcut-supporting patch or concept structure for mitigation or balancing without subgroup annotation. Together, these works move concept analysis from generic explanation toward shortcut-specific evidence, showing concepts and prototypes can expose the cues that bias behavior.

Beyond auditing, the closest related work explicitly localizes, corrects, or weakens shortcut mechanisms inside trained models. Concept-aware mitigation identifies unstable concepts and reduces spurious reliance through training [50], while region-aware vision-language methods localize shortcut regions and mitigate them through region-level objectives [42]. Other approaches weaken shortcut behavior through concept-residual modeling [46], subspace separation [21], attribute-level causal regularization [28], generative classification [30], inference-time channel erasure [20], or causal separation of object evidence from misleading context [31]. These methods show that shortcut behavior can be localized, corrected, or weakened once relevant evidence has been identified. We build on this direction by using the shortcut-trained model as a diagnostic amplifier, so exaggerated reliance on the non-causal signal reveals recurrent visual patterns that can be fixed as concept units and validated through transfer, causal reliance, and targeted weakening across model–data settings.

4 Experimental System

4.1 Spurious concept discovery via shortcut amplification

Shortcut-model training for concept amplification. To operationalize spurious behavior at concept level, we train a dedicated *shortcut model* on a distribution in which a known non-causal signal is aligned with the target label. This drives the network to exploit the shortcut, amplifying its footprint in intermediate feature space and making shortcut patterns easier to isolate and cluster. The shortcut-amplification step thus serves as a diagnostic contrast agent. Rather than asking whether a conventionally trained model weakly encodes a shortcut, we construct a setting in which reliance on the non-causal signal becomes maximally visible. Concept discovery is therefore performed on the shortcut model, where shortcut-related structure is most likely to emerge in representation space.

Concept discovery in intermediate feature space. To identify such patterns, we perform ACE-style concept discovery [19] in activation space. We attach a forward hook to a selected intermediate target layer and extract its activation tensor $A \in \mathbb{R}^{B \times C \times H_f \times W_f}$. Each spatial position is treated as a region-level embedding $v_{b,i,j} = A[b, :, i, j] \in \mathbb{R}^C$, yielding a large set of embeddings across locations and images. The target layer is chosen to preserve a spatial activation grid while encoding mid-level visual structure beyond raw pixel statistics. This provides a representation that retains localization and semantics, making it suitable for region-wise concept discovery and localized interventions.

Concept bank construction and cross-model transfer. We convert shortcut-model activation clusters into a fixed concept bank of 100 candidate shortcut prototypes. Each concept is represented by exemplar regions from the shortcut-training split, fixing its visual identity before cross-split and cross-model analysis. For transfer, the same exemplar region IDs are reused when extracting embeddings from the baseline model at the same target layer. Because both models share the same architecture and target-layer grid, these matched regions provide a measurement interface with equivalent receptive fields, allowing concept clouds to be compared across representation spaces without re-clustering. The bank therefore makes shortcut-amplified regularities directly inspectable while preserving shared, reusable concept definitions for assignment, CAVs, TCAVs, and ablations.

We assess cross-model concept preservation with two metrics:

- *CAV separability*: concept recoverability from target-layer activations;
- *centroid drift*: cross-model displacement of the concept centroid.

Together, these metrics separate concept survival from concept preservation. CAV separability tests whether a shortcut-derived concept remains identifiable in target-layer space, while centroid drift tests whether the same exemplar-defined regions occupy a comparable representational location. Concepts with high separability and low drift are retained for subsequent causal concept-level reliance analysis.

4.2 Causal Concept Reliance Analysis

We next quantify decision reliance at target-layer regions, asking which concepts causally and consistently steer the ground-truth margin across data–model combinations via TCAV and ablation.

First, for each concept c, we compute region-wise TCAV using gradients of the ground-truth margin, defined in multiclass form as (i) which in the binary case reduces to (ii):

$$\text{(i)} \quad s(x) = \text{logit}_y(x) - \max_{k \neq y} \text{logit}_k(x) \qquad \text{(ii)} \quad s(x) = \text{logit}_y(x) - \text{logit}_{1-y}(x).$$

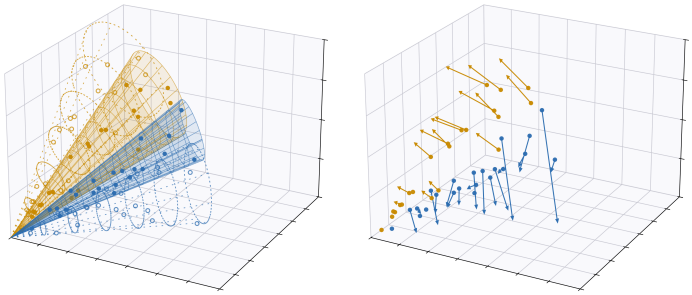
For each image, we compute $\nabla_A s$ and evaluate directional derivatives only at activation locations assigned to concept c. TCAV measures whether infinitesimal increases along the concept direction locally and consistently increase the ground-truth margin on concept-assigned regions. Across models, reusing the same region-anchored concept definition makes TCAV scores comparable. However, TCAV captures directional sensitivity rather than the effect of removing concept-assigned evidence.

The central component is localized activation ablation, implemented as a two-pass replacement. For each image and concept occurrence, we first compute the original margin s_{orig}. In a second forward pass, we replace the activation vectors assigned to concept c with a fixed reference value while leaving other activations unchanged. The causal impact is defined as $\Delta s_{\text{support}} = s_{\text{orig}} - s_{\text{ablated}}$, where $\Delta s_{\text{support}} > 0$ indicates support for the correct-class decision and $\Delta s_{\text{support}} < 0$ indicates suppression.

The regional concept assignment is computed once per model–split at the target layer and then reused for both TCAV and ablation. Both diagnostics therefore operate on the same activation map and spatial coordinates, testing the causal role of concept-assigned regions rather than an image-level or pooled proxy. This four-way design isolates shortcut reliance as a condition-dependent effect, separating concepts that are causally useful only under bias from concepts that remain predictive under unbiased training and evaluation. We interpret concept influence under four conditions:

1. **Spurious split on shortcut model.** Concepts with high positive Δs under the biased distribution are candidates for shortcut-inducing visual cues.
2. **Baseline split on shortcut model.** When the bias is removed, a large influence drop for the same concepts indicates bias-dependent support rather than task-derived evidence.
3. **Baseline split on baseline model.** If these concepts show minimal or near-zero Δs in the unbiased baseline model, they are unlikely to be task-critical.
4. **Spurious split on baseline model.** If the same concepts regain supportive influence once the bias is reintroduced, this indicates inducible shortcut reliance in the same architecture.

Together, these four conditions distinguish shortcut from task-relevant concepts across controlled model–data settings: TCAV captures aggregate effects, while ablation isolates them individually.



Left: Concept Persistence Trajectory (CPT) summarizes concept cohesion and profile stability across layers, shown as directional cones of normalized embeddings around a prototype direction.

Right: Inter-Layer Transport Alignment (ITA) summarizes the consistency of displacement vectors between consecutive layers, indicating whether concept instances share a transformation.

Figure 2: Architecture-level concept transport diagnostics.

4.3 Architecture-Level Concept Transport Statistics

Having quantified causal reliance at the extraction layer, we ask whether those concepts remain stable representational objects across the architecture. Local stability alone does not determine propagation, because concepts may merge with nearby structure or dissolve under later transformations. For persistent concepts, transport has two irreducible levels, the aggregate body and its constituent vectors. At the aggregate level, we ask whether normalized directional mass expands, contracts, or drifts across depth. At the constituent level, we ask whether vectors move coherently or reorganize within that body. Together, they separate aggregate persistence from constituent transport.

Concept Persistence Trajectory (CPT). CPT operationalizes the aggregate-level view of concept transport. Concept instances are defined by the fixed extraction-layer assignments, and spatial coordinates are propagated via center-preserving grid mapping. For each layer ℓ and concept c, we extract feature vectors $\{v_i^{(\ell)}\}_{i=1}^{N_{c,\ell}}$ and unit-normalize them, placing concept instances on the unit hypersphere. CPT thereby isolates norm-invariant persistence of cluster-level directionality.

We then define a concept-specific prototype direction at layer ℓ (i) and measure *cone tightness* through the mean prototype affinity (ii), quantifying angular clustering around it:

$$\text{(i)} \quad \mu_{c,\ell} = \frac{\frac{1}{N_{c,\ell}} \sum_{i=1}^{N_{c,\ell}} v_i^{(\ell)}}{\left\| \frac{1}{N_{c,\ell}} \sum_{i=1}^{N_{c,\ell}} v_i^{(\ell)} \right\|_2} \qquad \text{(ii)} \quad T_c(\ell) = \frac{1}{N_{c,\ell}} \sum_{i=1}^{N_{c,\ell}} \cos\big(v_i^{(\ell)}, \mu_{c,\ell}\big).$$

High values of $T_c(\ell)$ indicate that concept activations form a compact directional cone around a shared prototype, meaning that concept instances remain angularly aligned after normalization.

To measure cross-layer stability of this angular concept structure, we compute for each instance the prototype affinity (iii), summarize the resulting affinity values by a fixed *quantile signature* (iv) rather than direct correlation, and compare adjacent layers using Spearman rank correlation (v):

$$\text{(iii)} \;\; a_i^{(\ell)} = \cos\big(v_i^{(\ell)}, \mu_{c,\ell}\big) \quad \text{(iv)} \;\; q_c(\ell) = Q\Big(\{a_i^{(\ell)}\}_{i=1}^{N_{c,\ell}}\Big) \quad \text{(v)} \;\; P_c(\ell, \ell{+}1) = \rho_s\big(q_c(\ell), q_c(\ell{+}1)\big).$$

This compares the distributional shape of concept membership across layers while reducing artifacts from resolution changes, many-to-one coordinate remapping, and stage-boundary reparameterization.

The final CPT transition score (vi) combines within-layer cone tightness with cross-layer stability:

$$\text{(vi)} \quad \mathrm{CPT}_c(\ell, \ell{+}1) = \frac{T_c(\ell) + T_c(\ell{+}1)}{2} \cdot \max\big(0, P_c(\ell, \ell{+}1)\big).$$

Accordingly, CPT is high when fixed concept instances collectively remain directionally cohesive and their aggregate angular profile stays distributionally stable across the transition. Figure 2 (left) provides the corresponding geometric intuition. Each cone represents normalized embeddings around a concept prototype: cone tightness visualizes directional cohesion, whereas changes in cone orientation indicate prototype drift between layers. CPT combines cone tightness with the stability of the aggregate angular-affinity distribution inside the cone. Thus, even when the overall cone geometry remains similar, reshaping this distribution between layers ℓ and $\ell{+}1$, illustrated by the dashed cone, signals a changed cluster-level membership profile and reduces CPT. In practice, CPT stays stable within stages and often dips at transitions [16], consistent with representational reorganization [39].

Inter-Layer Transport Alignment (ITA). While CPT treats a concept as an aggregate body, ITA captures its constituent-level transport. It asks whether the vectors forming that body undergo a shared transformation or reorganize idiosyncratically between successive layers. For each concept instance and layer pair $(\ell, \ell{+}1)$, we extract the corresponding embeddings and project them into a common low-dimensional space using per-layer IncrementalPCA. Let $z_i^{(\ell)}$ and $z_i^{(\ell+1)}$ denote the resulting vectors. We then define the displacement vector (vii) and its directional coherence (viii) as:

$$\text{(vii)} \quad \Delta z_i^{(\ell)} = z_i^{(\ell+1)} - z_i^{(\ell)} \qquad \text{(viii)} \quad A_c(\ell) = \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{\Delta z_i^{(\ell)}}{\|\Delta z_i^{(\ell)}\|_2} \right\|_2,$$

which is high when displacement vectors point in a common direction and low when they cancel out.

To account for variability in displacement magnitudes, we compute the coefficient of variation (ix) and the resulting magnitude-consistency term (x):

$$\text{(ix)} \quad \mathrm{CV}_c(\ell) = \frac{\sigma\big(\|\Delta z^{(\ell)}\|_2\big)}{\mu\big(\|\Delta z^{(\ell)}\|_2\big) + \varepsilon} \qquad \text{(x)} \quad M_c(\ell) = \frac{1}{1 + \mathrm{CV}_c(\ell)}.$$

The final *Inter-Layer Transport Alignment* score is defined as $\mathrm{ITA}_c(\ell) = A_c(\ell) \cdot M_c(\ell) \in [0, 1]$.

Thus, ITA ≈ 1 indicates shared directional transport with comparable displacement magnitudes, whereas low ITA values indicate heterogeneous or cancelling movements. Figure 2 (right) illustrates this geometry: the orange concept exhibits coherent transport, whereas the blue concept shows fragmented transport through directional cancellation and magnitude dispersion. ITA captures whether instances transform coordinately across layers or shift inconsistently in representation space.

Evaluation scope. The main experiments instantiate the pipeline on SpurBreast with a ResNet-50 backbone and field-strength bias as the controlled shortcut signal. To test whether the diagnostic procedure extends beyond this primary setting, we evaluate three complementary axes: usefulness for debugging via targeted removal of identified shortcut signals in Appendix A, applicability to different domains and shortcut types via industrial visual inspection in Appendix B, and architecture transfer via replication with a lightweight transformer backbone in Appendix C. Together, these extensions assess whether the pipeline is actionable, domain-transferable, and robust across architectures.

5 Results

5.1 Spurious concept discovery under field-strength bias

Backbone and target layer. We instantiate the pipeline with ResNet-50, a standard residual CNN whose blocks form spatial feature hierarchies from low-level primitives to abstract object-level structure [52]. This suits concept-level shortcut analysis because intermediate blocks retain localized activation grids for region-level concept extraction and targeted interventions. We use `resnet.encoder.stages.2.layers.5`, the final residual block of Stage 2, as the target layer because it balances spatial resolution with mid-level abstraction beyond raw pixel statistics [53].

Field-strength bias as a controlled shortcut signal. SpurBreast shows that models trained on biased splits can exploit scanner field strength as a spurious signal [48]. When low- and high-field acquisitions align with opposing class labels, models can infer labels from field strength over tumor morphology, yielding validation accuracy near 100% but collapsing to chance on unbiased test data. Although this establishes field strength as a global shortcut, its *visual patterns* remain unidentified.

MRI physics can produce such acquisition-dependent field-strength signatures, including: (i) differences in signal-to-noise [49] and noise patterns [33], (ii) intensity and contrast shifts from field-dependent relaxation, radio-frequency field inhomogeneity, and bias-field shading [10], and (iii) spatial-frequency differences under high-resolution protocols [33] and subtle reconstruction artefacts such as ringing or interpolation effects in partial-Fourier protocols [29]. These effects are globally distributed, often imperceptible to human observers, yet separable in high-dimensional feature space.

Here, field-strength bias serves as the concrete test case for the general diagnostic procedure introduced above. We therefore train the shortcut model on the spuriously correlated field-strength split [48] and use the amplified shortcut signal to derive the concept bank used throughout the results.

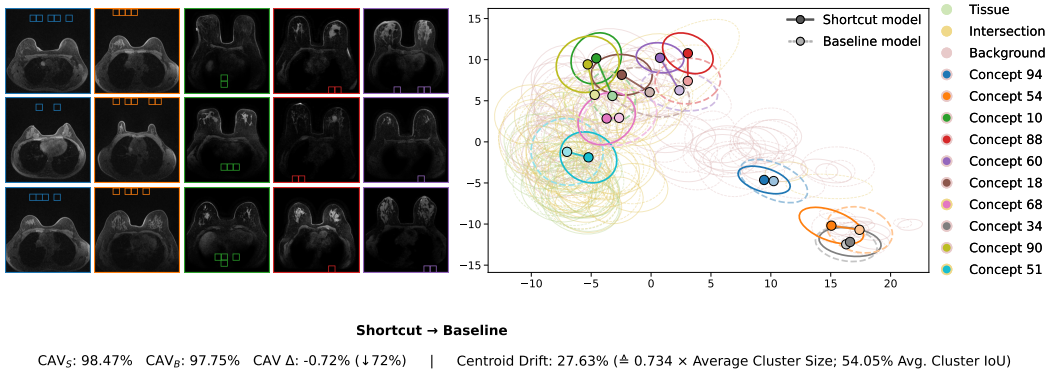


Figure 3: Cross-model concept geometry between shortcut and baseline models. Concepts are defined by shortcut-discovered exemplar regions and re-embedded in the baseline model at the same target layer. The image matrix grounds selected regions. Solid ellipses denote shortcut-model embeddings, dashed ellipses baseline embeddings, and arrows indicate centroid drift. Summary statistics report CAV separability, centroid displacement, cluster sizes and mean cluster IoU.

5.2 Cross-Model Concept Transfer and Representational Stability

To test whether shortcut-derived concepts transfer to the baseline model, we embed the same shortcut-defined regions in both networks at the same target layer. This enables direct activation-space comparison without redefining concept boundaries. Figure 3 combines image-space grounding with representation-space geometry. The left image matrix shows selected concepts as spatially stable anatomical, border, or background regions, while the right panel shows their embedding organization.

Quantitatively, the two metrics separate concept persistence from concept identity. Near-unchanged CAV separability across shortcut and baseline models (98.47% vs. 97.75%, $\Delta_{\text{CAV}} = -0.72\%$) shows that shortcut-defined regions remain recoverable as linearly separable feature-space units in the baseline variant trained on the unbiased split. The concepts therefore remain learned. Centroid drift then asks whether these learned units retain comparable representational locations rather than being redefined. The mean centroid displacement is 27.63%, corresponding to only $0.734\times$ the average cluster size, while shortcut and baseline clouds still retain substantial support overlap (54.05% mean IoU). Hence, concept mass moves less than one within-cluster scale, consistent with bounded displacement rather than relocation. Thus, separability and bounded centroid drift jointly show that the concepts are not merely re-separable, but retain similar representational identity across models. This enables subsequent cross-model causal analysis of shared-concept reliance shifts.

5.3 Causal Shortcut Reactivation Across Data Distributions

The directional sensitivity of the shortcut model on its spurious training distribution ($\text{TCAV}_{\text{SM+SS}} = 52.69\%$) nearly matches that of the baseline model on its unbiased training distribution ($\text{TCAV}_{\text{BM+US}} = 50.72\%$), with $\text{TCAV}_{\Delta} = -1.97\%$. This indicates that aggregate concept-to-logit sensitivity remains highly similar across the full 100-concept bank. However, this full-bank similarity does not identify which particular concepts are spuriously linked. It therefore supports similar overall reliance on the concept set, but not which subset of concepts drives shortcut behavior.

Concept-region ablation resolves this ambiguity at the level of *representation* by extracting the shortcut as a concept subset and tracking its causal behavior across the four evaluation conditions. The left plot in Figure 4 shows that, on the spurious split, the highlighted spurious concepts strongly push the shortcut model toward the true label, often with true label contributions above 0.7–0.8. When the same model is evaluated on the unbiased split, this directional support collapses toward neutrality and in several cases reverses, pushing predictions away from the true class. In the baseline model, the same visually grounded concepts are directionally indifferent on the unbiased split, yet several recover directional support once the spurious split is reintroduced. This pattern indicates latent shortcut readiness rather than simple presence or absence. The shortcut-compatible concept structure remains available, but its directional usefulness changes with the label–environment correlation.

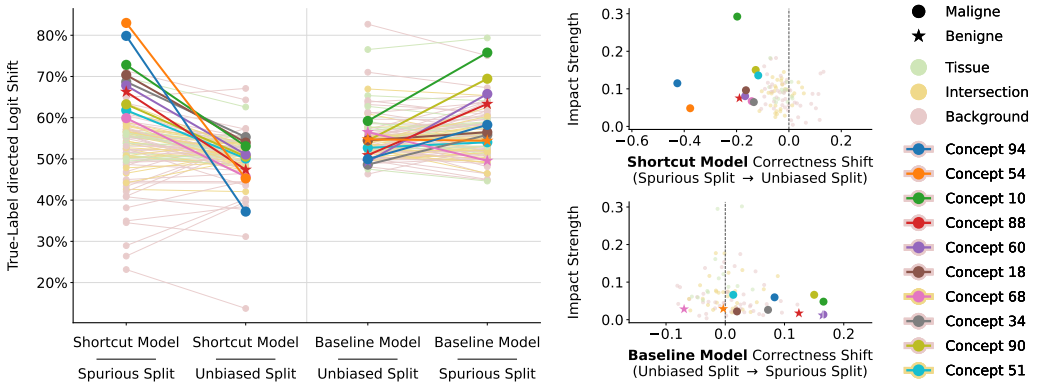


Figure 4: Causal behavior of shortcut concepts under controlled distribution shifts.
Left: Probability that concept interventions shift the ground-truth logit toward the correct class across four model–data evaluation conditions, with bottom summary TCAV scores for the shortcut model on the spurious split, the baseline model on the baseline split, and their mean shift.
Right: Condensed concept-region ablation view for the shortcut model (top) and the baseline model (bottom). The x-axis captures the correctness shift across data splits, while the y-axis measures the strength of the induced logit shift for the highlighted shortcut concepts.

The right plots in Figure 4 address the complementary question of *causal reliance*. They measure impact strength, defined as how strongly the classifier uses the identified spurious concepts to move the logits. Across the highlighted shortcut concepts, impact strength is substantially lower in the baseline model than in the shortcut model. The baseline model has therefore not removed these cues from its internal concept structure, but learned not to route prediction strongly through them. In this sense, the classifier acts as a gate. Under shortcut alignment, the shortcut model relies on the spuriously linked concepts because they function as predictive evidence, whereas the baseline model largely gates them off as non-useful signals. Together, the panels provide causal evidence that shortcut behavior is distribution-conditional rather than tied to a separate shortcut-only representation.

The overall picture is therefore a dissociation between encoded shortcut structure and actual classifier recruitment. TCAV indicates that the concept reservoir remains directionally accessible in both variants, whereas regional ablation shows that reliance on the spurious subset is markedly reduced in the baseline model. Reduced, however, does not mean absent. Even under unbiased training, the baseline model still incorporates the spurious concepts into its logits with non-zero impact strengths of roughly 0.02–0.10. The appendix A shows that even this residual pathway is mechanistically meaningful and can be reduced by targeted interventions on its dominant carriers. Shortcut behavior is therefore best understood not as the existence of a separate shortcut representation, but as the conditional and graded recruitment of still-encoded concept structure. Representation is not reliance.

5.4 Architecture-Level Concept Transport Statistics

Having established concept transferability across models and causal reliance changes across distributions, we next ask how these concepts propagate internally. This analysis tests whether they persist as stable directional structures and whether visually different types follow characteristic transport patterns. We characterize propagation with two statistics: *Concept Persistence Trajectory* (CPT) for directional cluster-level cohesion and *Inter-Layer Transport Alignment* (ITA) for instance consistency.

These statistics first show that the chosen extraction layer captures concept structure traceable through the network. If the discovered concepts were merely local artifacts of that layer, their directional organization would degrade under preceding or subsequent transformations rather than remain coherent. Stable CPT and ITA trajectories therefore support the discovery layer as a meaningful concept snapshot rather than only a fragile layer-specific artifact. Accordingly, CPT remains high through most layers in both models, indicating that normalized directional profiles remain traceable, and declines only in the final ResNet stage, where spatial structure is compressed into global semantic features. ITA remains coherent across layers, indicating structured transport beyond discovery.

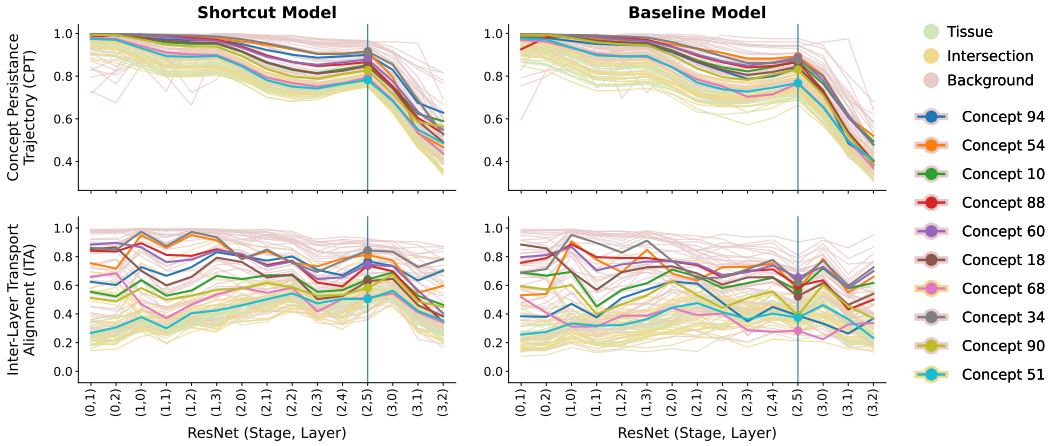


Figure 5: Architecture-level concept transport for shortcut and baseline models. Upper panels show CPT and lower panels ITA. Colored curves denote highlighted shortcut-derived concepts. Remaining concepts are pastel-coded by (non-)anatomical region. The vertical line indicates the hook layer.

Beyond validation, CPT and ITA relate inspected concept types to their transformations in feature space. Figure 5 shows that background concepts follow trajectories distinct from human tissue and tissue–background intersection regions. These trajectories point to differences in semantic content and computational regime. Human-tissue concepts show stronger variation across layers for both statistics, consistent with progressive restructuring rather than simple carry-through, indicating that anatomically grounded evidence is assembled through deeper transformations. In contrast, background concepts remain stable, consistent with earlier access, a cheaper route, and less change.

This interpretation, however, should not collapse into a simple background-versus-tissue dichotomy. The highlighted shortcut concepts are not the most trivially stable background patterns. Rather, they are simpler than tissue-grounded concepts yet sufficiently predictive under the biased split. These results emphasize the logic of shortcut learning: the model need not select the simplest represented pattern, but may route prediction through the simplest sufficiently predictive combination. When a shortcut cue is simpler than the true predictive structure yet label-sufficient, the model may exploit it.

6 Conclusion

This work introduces a *causal diagnostic pipeline* for shortcut learning that goes beyond merely establishing spurious behavior. It decomposes shortcut behavior into the specific visual cues that drive it, separating *representation* as encoded and transferable concept structure from *causal reliance* during prediction. On a field-strength-biased breast-cancer MRI benchmark, shortcut behavior arises from the conditional recruitment of still-encoded concepts rather than from a separate shortcut-only representation. The identified concepts remain transferable across model variants, but their causal role changes with the data distribution and training regime. Additional evidence from targeted shortcut removal, industrial camera-angle transfer, and SegFormer replication supports the pipeline as an actionable, domain-transferable, and architecture-aware concept-level diagnostic. Overall, the findings show that shortcut learning can be governed not only by whether spurious concepts are still encoded, but by when, where, and under which distributions models causally rely on them.

Limitations. Despite the broad empirical scope, limitations remain. First, the evaluation is restricted to two controlled benchmarks and few architectures, so the results should be interpreted as evidence of transfer rather than proof of universal applicability. Second, both shortcut regimes are based on known or strongly suspected spurious factors, enabling clean causal analysis but not exhausting the diversity and entanglement of real-world shortcut behavior. Third, the pipeline depends on concept discovery at a chosen extraction layer, including clustering stability, exemplar assignment, and whether the concept bank captures the dominant shortcut carriers. Finally, targeted debiasing shows that the diagnostic output is actionable, but does not yet constitute a general mitigation framework.

References

[1] A. Abid, M. Yuksekgonul, and J. Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *ICML*, volume 162, 2022.

[2] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju. Openxai: Transparent evaluation of post hoc model explanations. In *NeurIPS*, volume 35, 2022.

[3] K. Ahuja, K. Shanmugam, K. R. Varshney, and A. Dhurandhar. Invariant risk minimization games. In *ICML*, 2020. doi: 10.5555/3524938.3524953.

[4] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL `https://arxiv.org/abs/1610.01644`.

[5] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77, 2022. doi: 10.1016/j.inffus.2021.07.015.

[6] M. R. Arefin, Y. Zhang, A. Baratin, F. Locatello, I. Rish, D. Liu, and K. Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. ICML'24. JMLR.org, 2024. doi: 10.5555/3692070.3692138.

[7] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf`.

[8] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.354.

[9] O. R. Bidusa and S. Markovitch. Concept layers: Enhancing interpretability and intervenability via llm conceptualization, 2025. URL `https://arxiv.org/abs/2502.13632`.

[10] K. Chang, I. Kamel, K. Macura, and D. Bluemke. 3.0-t mr imaging of the abdomen: Comparison with 1.5 t. *Radiographics*, 28, 2008. doi: 10.1148/rg.287075154.

[11] J. Choi, H. Lim, S. Schneider, and J. Choo. Conceptscope: Characterizing dataset bias via disentangled visual concepts. In *NeurIPS*, 2025. URL `https://openreview.net/forum?id=lkmlNHuzY4`.

[12] P. P. S. Dammu and C. Shah. Detecting spurious correlations via robust visual concepts in real and AI-generated image classification. In *XAI in Action: Past, Present, and Future Applications*, 2023. URL `https://openreview.net/forum?id=ewagDhIy8Y`.

[13] R. Darbinyan, H. Harutyunyan, A. Markosyan, and H. Khachatrian. Identifying and disentangling spurious features in pretrained image representations, 06 2023.

[14] A. DeGrave, J. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal, 10 2020.

[15] T. Fel, A. Picard, L. Béthune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. Craft: Concept recursive activation factorization for explainability. In *CVPR*, 2023.

[16] T. Fel, L. Béthune, A. K. Lampinen, T. Serre, and K. Hermann. Understanding visual feature reliance through the lens of complexity. In *NeurIPS*, 2024. URL `https://openreview.net/forum?id=NhqZpst42I`.

[17] G. D. Felice, A. Casanova, F. D. Santis, S. Santini, J. Schneider, P. Barbiero, and A. Termine. Causally reliable concept bottleneck models. In *NeurIPS*, 2025. URL `https://openreview.net/forum?id=UX143QGvb8`.

[18] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), Nov. 2020. doi: 10.1038/s42256-020-00257-z.

[19] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. *Towards automatic concept-based explanations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[20] Q. He, K. Xu, and A. Yao. Eva: Erasing spurious correlations with activations. In *ICLR*, 2025. URL https://openreview.net/forum?id=zKvr0OBouT.

[21] F. Holstege, B. Wouters, N. V. Giersbergen, and C. Diks. Removing spurious concepts from neural network representations via joint subspace estimation. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/holstege24a.html.

[22] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, volume 32, 2019.

[23] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson. On feature learning in the presence of spurious correlations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022. doi: 10.5555/3600270.3603061.

[24] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2017. URL https://api.semanticscholar.org/CorpusID:51737170.

[25] P. Kirichenko, P. Izmailov, and A. G. Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

[26] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

[27] L. Kuhn, S. Sadiya, J. Schlotterer, F. Buettner, C. Seifert, and G. Roig. Efficient unsupervised shortcut learning detection and mitigation in transformers, 2025. URL https://arxiv.org/abs/2501.00942.

[28] A. Kumar, A. Deshpande, and A. Sharma. Causal effect regularization: Automated detection and removal of spurious correlations. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=V5Oh7Aqfft.

[29] H.-H. Lee, D. Novikov, and E. Fieremans. Removal of partial fourier-induced gibbs (rpg) ringing artifacts in mri. *Magnetic Resonance in Medicine*, 86, 07 2021. doi: 10.1002/mrm.28830.

[30] A. C. Li, A. Kumar, and D. Pathak. Generative classifiers avoid shortcut solutions. In *ICLR*, 2025. URL https://openreview.net/forum?id=oCUYc7BzXQ.

[31] R. Liu, H. Liu, G. Li, H. Hou, T. Yu, and T. Yang. Contextual debiasing for visual recognition with causal mechanisms. In *CVPR*, 2022.

[32] W. Lotter. Acquisition parameters influence ai recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-52003-3. URL https://api.semanticscholar.org/CorpusID:272201016.

[33] W. Md. Clinical advantages of 3.0 t mri over 1.5 t. *European journal of radiology*, 65, 02 2008. doi: 10.1016/j.ejrad.2007.11.006.

[34] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. In *NeurIPS*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

[35] M. Moayeri, S. Singla, and S. Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In *NeurIPS*, volume 35, pages 10068–10077, 2022.

[36] N. Murali, A. M. Puli, K. Yu, R. Ranganath, and kayhan Batmanghelich. Beyond distribution shift: Spurious features through the lens of training dynamics. *TMLR*, 2023. URL https://openreview.net/forum?id=Tkvmt9nDmB.

[37] C. Ong Ly, B. Unnikrishnan, T. Tadic, T. Patel, J. Duhamel, S. Kandel, Y. Moayedi, M. Brudno, A. Hope, H. Ross, and C. Mcintosh. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *npj Digital Medicine*, 7, 05 2024. doi: 10.1038/s41746-024-01118-4.

[38] F. Pahde, M. Dreyer, M. Weckbecker, L. Weber, C. J. Anders, T. Wiegand, W. Samek, and S. Lapuschkin. Navigating neural space: Revisiting concept activation vectors to overcome directional divergence. In *ICLR*, 2025. URL `https://openreview.net/forum?id=Q95MaWfF4e`.

[39] R. Rombach, P. Esser, and B. Ommer. Making sense of cnns: Interpreting deep representations their invariances with inns, 2020. URL `https://arxiv.org/abs/2008.01777`.

[40] J. Teneggi and J. Sulam. Testing semantic importance via betting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *NeurIPS*, volume 37, 2024. doi: 10.52202/079017-2435.

[41] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, CVPR '11, USA, 2011. IEEE Computer Society. doi: 10.1109/CVPR.2011.5995347. URL `https://doi.org/10.1109/CVPR.2011.5995347`.

[42] M. Varma, J.-B. Delbrouck, Z. Chen, A. Chaudhari, and C. Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In *NeurIPS*, volume 37, 2024. doi: 10.52202/079017-2614. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/95c6ae3f3393786203a4b6dcb9df1036-Paper-Conference.pdf`.

[43] C. Wachinger, A. Rieckmann, and S. Pölsterl. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67, 2021. doi: https://doi.org/10.1016/j.media.2020.101879. URL `https://www.sciencedirect.com/science/article/pii/S1361841520302437`.

[44] D. Wallis and I. Buvat. Clever hans effect found in a widely used brain tumour mri dataset. *Medical Image Analysis*, 77, 2022. doi: https://doi.org/10.1016/j.media.2022.102368. URL `https://www.sciencedirect.com/science/article/pii/S1361841522000214`.

[45] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. 2024. doi: 10.1109/CVPR52733.2024.02159.

[46] J. Wang, S. Jabbour, M. Makar, M. Sjoding, and J. Wiens. Learning concept credible models for mitigating shortcuts. In *Advances in Neural Information Processing Systems*, volume 35, pages 33343–33356, 2022.

[47] N. Weir, D. Lindenbaum, A. Bastidas, A. Van Etten, S. McPherson, J. Shermeyer, V. Kumar, and H. Tang. Spacenet mvoi: A multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 992–1001, 2019.

[48] J. B. Won, W. De Neve, J. Vankerschaver, and U. Özbulak. Spurbreast: A curated dataset for investigating spurious correlations in real-world breast mri classification. In *MICCAI*, 2026. doi: 10.1007/978-3-032-05325-1_53.

[49] S. Wong, L. Steinbach, J. Zhao, C. Stehling, C. Ma, and T. Link. Comparative study of imaging at 3.0 t versus 1.5 t of the knee. *Skeletal radiology*, 38, 04 2009. doi: 10.1007/s00256-009-0683-0.

[50] S. Wu, M. Yuksekgonul, L. Zhang, and J. Zou. Discover and cure: concept-aware mitigation of spurious correlation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. doi: 10.5555/3618408.3619982.

[51] Y. Xie, Z. Zeng, H. Zhang, Y. Ding, Y. Wang, Z. Wang, B. Chen, and H. Liu. Discovering fine-grained visual-concept relations by disentangled optimal transport concept bottleneck models. 06 2025. doi: 10.1109/CVPR52734.2025.02811.

[52] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization, 2015. URL `https://arxiv.org/abs/1506.06579`.

[53] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-10590-1_53.

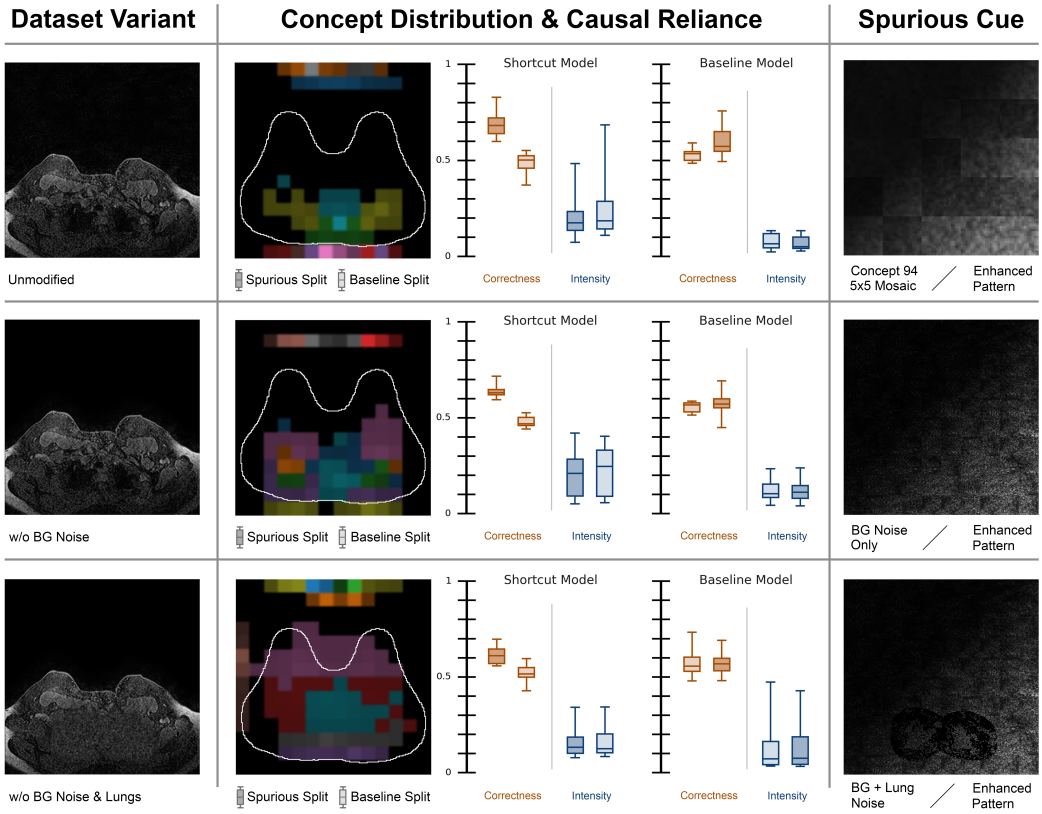


Figure 6: Concept-level analysis under targeted shortcut-removing interventions. Rows correspond to dataset variants. Columns from left to right show: input example, spatial concept attribution, causal reliance statistics across variants and splits, and a visualization of interventions on indicated regions.

A Proof of Concept: Targeted Removal of Spurious Field Strength Signals

Figure 6 traces the shortcut mechanism across progressively stronger interventions, from its initial localization in background-dominated concepts to its attenuation after background removal and its near-elimination after replacing the lung region with synthetic tissue-like texture patterns.

Initial Observation (Unmodified Data). In the unmodified setting, the shortcut model shows tight coupling between predictive performance and concept reliance under distribution shift. This pattern is supported by the concept attribution map, where $7/10$ dominant concepts lie outside the average tissue mask and cluster in peripheral dark regions. The extracted concept visualizations show that these concepts correspond primarily to low-frequency structure in the black background (e.g., Concept 94), rather than anatomically meaningful tissue. The causal reliance statistics show the same mechanism from a complementary angle: the shortcut model shows a pronounced drop in correctness and concept intensity on the baseline distribution, while the baseline model recovers predictive utility on the shortcut-corrupted split. Together, these observations identify a transferable, dataset-specific cue represented by a small set of dominant background concepts causally exploited as a shortcut.

Intervention I: Background Noise Removal. Removing background noise weakens the shortcut mechanism but does not eliminate it. Spatially, the dominant concepts shift away from the image periphery and toward the interior, indicating reduced dependence on background-localized structure. However, the activations remain strongly concentrated around the lung region, suggesting residual reliance on structured noise and geometric cues such as lung contours. The causal plots show corresponding attenuation of the U-shaped shortcut signature, with reduced concept intensity shifts and weaker performance transfer across distributions. The intervention reduces the original artifact but mostly redistributes reliance toward remaining non-anatomical carriers rather than eliminating it.

Table 1: Variant-wise performance under shortcut-preserving & shortcut-removing perturbations. Values are reported as mean ± standard deviation over 10 training runs across both models.

Variant	Validation		Test	
	Acc	F1	Acc	F1
Field Strength				
Unmodified	96.70 ± 0.36	96.65 ± 0.38	55.08 ± 0.45	50.08 ± 0.93
w/o BG Noise	95.14 ± 0.67	95.10 ± 0.68	56.88 ± 0.52	53.59 ± 1.08
w/o BG & Lungs	92.75 ± 0.68	92.65 ± 0.70	57.83 ± 0.87	55.68 ± 1.71
BG Noise Only	82.03 ± 0.75	81.21 ± 0.94	57.27 ± 0.77	55.30 ± 0.84
BG & Lung Noise	91.60 ± 0.52	91.52 ± 0.51	55.00 ± 0.66	51.19 ± 1.27
Baseline Large				
Unmodified	75.95 ± 0.51	77.14 ± 1.07	78.52 ± 0.53	80.10 ± 0.77
w/o BG Noise	75.69 ± 0.47	76.92 ± 0.75	78.38 ± 0.34	79.88 ± 0.57
w/o BG & Lungs	75.55 ± 0.39	76.95 ± 0.76	78.57 ± 0.26	80.37 ± 0.58
BG Noise Only	68.33 ± 0.33	70.59 ± 0.81	67.94 ± 0.44	69.90 ± 0.95
BG & Lung Noise	70.07 ± 0.35	72.44 ± 0.77	70.57 ± 0.41	72.43 ± 0.60

Intervention II: Lung Region Removal. Adding lung removal disrupts this residual mechanism substantially more. The dominant concepts redistribute toward the remaining tissue and visibly extend into the breast region itself, indicating that the shortcut model is forced away from dataset-specific artifacts and toward anatomically meaningful structure. This qualitative shift is matched by a near-collapse of the U-shaped causal pattern. The shortcut model no longer shows the same distribution-sensitive concept dependence, and the baseline model no longer benefits on the shortcut-corrupted split. Together, these results indicate that the shortcut was not merely background-localized, but jointly supported by background and lung-associated structure. Removing both largely breaks it.

Quantitative Validation. Table 1 quantitatively confirms this progression across all variants. For the shortcut model, stepwise removal of shortcut-bearing regions reduces validation performance, consistent with diminished access to the spurious field-strength signal, while test performance improves from 50.08 to 55.68 F1, reflecting greater robustness under unbiased evaluation. The noise-only controls clarify the role of the removed regions more directly: background noise alone partly restores shortcut behavior, whereas background plus lung noise restores it more strongly, approaching the unmodified validation baseline under the spurious distribution. By contrast, the baseline model remains nearly unchanged under the shortcut-removing variants, but degrades substantially on the noise-only controls, indicating that the targeted removals primarily suppress shortcut exploitation rather than useful tissue-based signal. The table therefore mirrors the figure-level mechanism: removing spurious carriers trades in-distribution shortcut gains for out-of-distribution robustness.

Summary. Taken together, the figure and table support a causal progression: shortcut reliance is first concentrated in background-dominated concepts, then redistributed toward lung-associated structure after background suppression, and finally displaced toward anatomically meaningful tissue after both background and lung cues are removed. This targeted removal of spurious carriers weakens shortcut-dependent validation gains and improves robustness to field-strength acquisition artifacts.

B Domain Transfer to Industrial Visual Inspection

B.1 Spurious concept discovery under camera-angle bias

Camera-angle bias as a controlled shortcut. To test transfer beyond medical imaging, we construct **SpurReal-IAD**, a SpurBreast-style benchmark on Real-IAD, a large-scale industrial anomaly detection dataset [45]. Using its multi-view setup, we inject a controlled shortcut by mapping a subset of camera angles predominantly to damaged and the remaining views to non-damaged samples. Models trained on this split can therefore exploit viewpoint-related cues over actual defect structure.

Image-recording-related acquisition artefacts. Camera angle can produce acquisition-dependent signatures unrelated to defect semantics. Multi-view imagery studies show that viewing-angle changes induce perspective distortion, alter effective resolution and lighting, and affect object visibility through occlusion, shadows, and reflections [47]. Bias-controlled recognition benchmarks therefore randomize background, rotation, and viewpoint to suppress trivial predictive correlations [7], while medical-imaging evidence shows that acquisition parameters, including view position, can influence AI predictions [32]. In industrial inspection, such view-dependent image-formation effects can therefore become globally predictive when spuriously aligned with labels.

Extraction-layer concept geometry. Representations of concept regions derived from camera-angle-biased training remain highly separable after transfer to the unbiased model, but their geometry changes substantially at the extraction layer. Small changes in CAV separability indicate that the same concept regions remain accessible, whereas reduced concept-cloud overlap and larger centroid drift indicate a changed activation-space organization. Thus, camera-angle information is not removed from the representation, but encoded differently. This goes beyond simple concept persistence, since the concept bank still transfers while the transferred concepts occupy a restructured feature space.

Representation–reliance analysis. Causal ablation identifies what this restructuring means for prediction. The concepts with the strongest correctness drop are region-labeled as object concepts, indicating that the camera-angle shortcut is carried by view-dependent object appearance rather than by a detached acquisition cue. Across the distribution shift, these regions move from strong label-aligned evidence toward weakened shortcut usefulness. The same regions can still affect the logits in the unbiased model, but their impact is less tied to camera-angle alignment. The important distinction is therefore between using object regions and using object regions as shortcut evidence.

Architecture-level trajectories. The transport statistics support this interpretation across depth. Object and intersection concepts attain lower CPT and ITA values than simpler background-related concepts, indicating stronger transformation across the network. This pattern is consistent with the visual nature of the shortcut, because camera angle changes the inspected object itself and affects shape, boundary, visibility, and surface information. The post-target-layer decline is consistent with the compression pattern already observed in SpurBreast, while the lower object-region trajectories suggest that the industrial shortcut involves more strongly transformed surface evidence.

Mechanistic synthesis. Together, these results indicate that the diagnostic pipeline does more than detect a shortcut. It characterizes how the shortcut is represented, where its causal carriers lie, and whether their use remains distribution-fragile. In SpurBreast, the shortcut is mainly expressed through simpler acquisition-related cues whose representation remains comparatively stable while reliance changes. In SpurReal-IAD, the shortcut is entangled with object appearance, since camera angle changes the visual evidence on the inspected surface and biased training turns these changes into label evidence. Under unbiased training, these regions remain available and causal, but their geometry and shortcut alignment change. The appendix therefore extends the main claim: shortcut correction can involve reduced reliance on preserved cues or reorganization of meaningful visual evidence.

C Architecture Transfer to SegFormer

To test whether the proposed diagnostic route transfers beyond convolutional backbones, we repeat the analysis with SegFormer as a lightweight transformer-based architecture. SegFormer is well suited because its hierarchical encoder preserves spatial organization while replacing ResNet's residual convolutional processing with transformer-style token mixing. Since both SpurBreast and SpurReal-IAD require image-level binary decisions rather than dense outputs, we fine-tune SegFormer for image classification under the same label conventions, splits, and evaluation protocol as ResNet.

For SegFormer, we use `segformer.encoder.block.2.1`, the final block of the third encoder stage, as the target layer for concept-bank extraction. This layer provides a close counterpart to the ResNet extraction layer `resnet.encoder.stages.2.layers.5` in architectural depth and spatial granularity. For 224×224 inputs, both layers operate at output stride 16 and yield a 14×14 activation or token grid, balancing semantic abstraction with sufficient spatial resolution for region-level concept assignment. Thus, the SegFormer target layer preserves the central design criterion of the ResNet

study. Concepts are extracted after substantial feature abstraction, while the representation remains localized enough for concept assignment, transport analysis, and targeted causal intervention.

Cross-architecture mechanistic comparison. The SegFormer replication shows that architecture transfer does not imply an identical shortcut mechanism. Across both SpurBreast and SpurReal-IAD, shortcut-derived concepts remain linearly accessible in SegFormer, but the geometry of this access differs sharply from ResNet. Whereas ResNet preserves a meaningful portion of local activation-space support and exposes clearer concept-level causal carriers, SegFormer shows a strong dissociation between direction and support. Concept directions remain almost perfectly aligned across shortcut and baseline variants, while concept-cloud overlap nearly vanishes and centroid drift becomes substantially larger. Thus, transferred concept identity in SegFormer survives primarily as a stable token-space axis rather than as a spatially stable concept neighborhood. This also changes the interpretation of causal ablation. The weaker and more compressed effects at `segformer.encoder.block.2.1` should not be read as shortcut removal or higher robustness. They indicate that shortcut evidence is less concentrated in the highlighted local cells and may be distributed, reparameterized, or aggregated later. The transport statistics support this reading: SegFormer forms coherent concept axes around the selected block, but later encoder processing can reconfigure the supporting token geometry, so within-model traceability remains compatible with cross-model support relocation. The architecture-transfer study therefore strengthens the model-agnosticity claim precisely because the diagnostic pipeline transfers without imposing the ResNet mechanism. In a convolutional hierarchy, shortcuts appear as more localized causal pathways, whereas in SegFormer the same concept families remain accessible but become directionally stable, geometrically relocated, and less locally decisive.

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 introduction state the paper's contribution as a methodological diagnostic framework for shortcut learning that separates encoded representation from causal reliance. The claims are supported by the reported experiments on SpurBreast, targeted proof-of-concept interventions, domain transfer to SpurReal-IAD, and transformer-based replication, while the limitations section constrains the generality of these findings.

   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: The paper discusses limitations including the restricted number of datasets and architectures, the use of controlled shortcut regimes, dependence on concept discovery at selected hook layers, and sensitivity to clustering, exemplar assignment, and diagnostic metric choices. It also clarifies that the proof-of-concept intervention study illustrates possible actionability but does not establish a general mitigation framework.

   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: NA

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The mathematical content defines diagnostic quantities such as CAV separability, TCAV-style directional influence, CPT, ITA, centroid drift, and ablation effects, which are used as empirical measurement tools.

   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: The paper describes the methodological framework, including shortcut-trained diagnostic models, ACE-style concept discovery, TCAV-based reliance estimation, cross-model concept transfer, CPT and ITA definitions, and activation ablation/patching protocols. Although code is not released, the paper provides the dataset construction logic, mathematical definitions, target-layer choices, and evaluation procedures needed to reproduce the central experimental claims.

   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 paper does not release code. The main contribution is a methodological diagnostic framework built around the use of shortcut-trained models to reveal underlying spurious concepts, rather than a new software package. The core components build on established methods such as ACE and TCAV, mathematically specified CPT and ITA scores, and broadly applicable activation ablation/patching procedures.

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: The experimental sections describe the datasets, shortcut regimes, train/validation/test logic, model backbones, hook layers, concept-bank construction, diagnostic metrics, and evaluation protocols. These details are sufficient to understand how the reported results were obtained and how the methodological framework is instantiated.

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: No

   Justification: The paper reports diagnostic values and matched comparisons across shortcut-trained and baseline models, but it does not provide full multi-seed error bars, confidence intervals, or statistical significance tests for all experiments. This is acknowledged because repeating the full pipeline, including model training, concept discovery, CAV estimation, TCAV scoring, CPT/ITA computation, and ablation across seeds, is computationally expensive.

   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: No

   Justification: The paper does not provide a complete compute-resource accounting for each experiment, including runtime, memory usage, and total compute. Most experiments were conducted on an RTX 3080, while SpurReal-IAD training was conducted on an RTX A5000-class GPU, but these details are not systematically reported in the paper.

   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 research uses existing datasets and model architectures under their respective terms and does not involve human subjects, crowdsourcing, private user data, or deployment of high-risk systems. The paper focuses on diagnostic methods for identifying shortcut reliance and discusses responsible interpretation of the resulting explanations.

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 paper discusses positive impacts such as improved transparency, debugging, and targeted analysis of shortcut reliance in vision systems, including safety-relevant medical and industrial contexts. It also discusses potential risks, including overconfidence in diagnostic outputs, incorrect interpretation of concept-level explanations, and the possibility that detected shortcuts are mistaken for proof of model safety without further domain validation.

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: NA

Justification: The paper does not release high-risk pretrained models, generative models, scraped datasets, or dual-use assets that require special release safeguards. The work is focused on post-hoc diagnostic analysis of shortcut reliance in controlled experimental settings.

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: The paper cites the original sources for the datasets, model architectures, pretrained weights, and methodological foundations used in the experiments, including ACE, TCAV, SpurBreast, Real-IAD, ResNet-50, and SegFormer. The paper uses these assets for research purposes and follows the corresponding dataset and model-use terms where applicable.

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: NA

Justification: The paper does not release new standalone assets. SpurReal-IAD is used as an experimental derived benchmark constructed from Real-IAD through a described camera-angle shortcut protocol, and the proof-of-concept dataset modifications are exploratory interventions used to illustrate possible mitigation behavior rather than released dataset assets.

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: NA

Justification: The paper does not involve crowdsourcing, user studies, annotation studies, participant recruitment, or research with human subjects. All experiments are conducted on existing image datasets and model outputs.

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: NA

Justification: The paper does not involve human subjects research, participant interaction, crowdsourcing, surveys, interviews, or behavioral data collection. Therefore, IRB or equivalent approval is not applicable.

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: Yes

Justification: LLM assistance was used for writing, editing, LaTeX refinement, and limited implementation support, including brainstorming and iteratively refining diagnostic metric formulations. The conceptual framing, methodological choices, experimental design, code validation, empirical analysis, result interpretation, and all final scientific claims were made and verified by the authors. LLMs were not used as an experimental component, data source, evaluator, or autonomous research agent.

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.