SEMANTIC COVERAGE IMBALANCE: A HIDDEN REPRESENTATIONAL BIAS IN VISUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern visual models achieve strong overall accuracy but fail on specific classes. While prior work attributes these failures to class imbalance or dataset bias, recent work has identified a more fundamental cause, **Semantic Coverage Imbalance** (SCI), as uneven coverage of class–concept relations in training data. We ask a complementary question: *how is uneven semantic coverage reflected in the learned representations and behavior of trained visual models?* To address this, we introduce a simple, model-agnostic post-hoc framework that measures semantic representation structure and semantic reliance from concept-based evidence, applicable across datasets, architectures, and concept sources. Across spurious-correlation benchmarks (WaterBirds), attribute-based datasets (CelebA), general object recognition (CIFAR-100), context-shift benchmarks (NICO++), and real-world medical imaging (ISIC-DICM-17K, MILK10k), we find that the model-level effects of uneven semantic coverage vary across architectures and domains, are associated with worst-class performance and generalization gaps, and are not explained by class imbalance or dataset-level bias alone. Under shortcut and context shift, failures depend on *which* semantic concepts a model relies on, not simply on the overall degree of semantic reliance. We further show that post-hoc analysis can identify unstable semantic dependencies that provide useful intervention targets. In medical imaging, however, clinically meaningful concepts remain comparatively stable across domains, suggesting that robust prediction depends on relying on the right semantics rather than uniformly reducing semantic reliance. Overall, our framework offers a principled approach for analyzing and diagnosing SCI as a representation-level bias in visual learning settings.

1 INTRODUCTION

Modern visual models achieve strong average accuracy but fail on specific target classes, rare patterns, and subtle semantic variations (Recht et al., 2019; Taori et al., 2020; Geirhos et al., 2020). These failures persist under class imbalance (Cui et al., 2019), dataset bias (Liu & He, 2025; Zeng et al., 2024), spurious correlations (Choi et al., 2025), and distribution shift (Sagawa et al., 2020; Liu et al., 2021; Koh et al., 2021), where average accuracy remains high but per-class performance degrades sharply. A key but underexplored reason is that models do not rely equally on meaningful semantic evidence across classes: some classes are supported by strong, task-relevant concepts, while others depend on weak, sparse, or inconsistent signals (Kim et al., 2018; Koh et al., 2020; Kulkarni et al., 2025; Bhalla et al., 2024; Oikarinen et al., 2023). Standard evaluation metrics fail to capture this structural imbalance (Sagawa et al., 2020; Hong et al., 2024), obscuring a hidden representational bias that undermines model reliability and robustness.

Prior work has shown that visual models suffer from dataset bias (Zeng et al., 2024; Liu & He, 2025) and exploit spurious concept–class correlations (Choi et al., 2025). These approaches primarily characterize *which* concepts are present or correlated with classes, rather than *how consistently models rely on them across classes*. As a result, they do not explain why models trained on a single dataset still exhibit large class-wise performance disparities.

Recent work formalized *Semantic Coverage Imbalance* (SCI) (Ahammed et al., 2026) as uneven coverage of class–concept relations in training data and proposed a training-time coverage-aware learning approach to mitigate its effects. We ask a complementary question: *how is uneven*

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

semantic coverage reflected in the learned representations and behavior of already-trained visual models? To answer this, we introduce a post-hoc, model-agnostic framework that measures semantic representation structure and semantic reliance from concept-based evidence, without retraining the underlying model.

Because our framework is independent of model architecture and training procedure, it enables systematic comparison across models, datasets, and concept sources. We use it to study whether the model-level effects of uneven semantic coverage are associated with worst-class performance, shortcut learning, and distribution-shift failures, and whether they differ from class imbalance and dataset-level bias. Unlike dataset-bias analyses that study distributional differences (Liu & He, 2025; Zeng et al., 2024) or concept-based methods that identify which concepts are present or correlated with classes (Kim et al., 2018; Choi et al., 2025; Koh et al., 2020), our framework analyzes how trained models represent and use semantic information across classes. Appendix B provides a detailed discussion of related work.

Our proposed framework supports annotated attributes, concept probes, and vision–language models. We evaluate it across spurious-correlation benchmarks (WaterBirds (Sagawa et al., 2020)), attribute-based datasets (CelebA (Liu et al., 2015)), object recognition (CIFAR-100 (Krizhevsky et al., 2009)), context-shift benchmarks (NICO++ (Zhang et al., 2023)), and real-world medical imaging (ISIC-DICM-17K (Ahammed et al., 2025), MILK10k (Philipp et al., 2025)).

Contributions. We introduce a simple, post-hoc, model-agnostic framework to diagnose how uneven semantic coverage is reflected in the representations and behavior of trained visual models (§2). First, we show that these model-level effects vary across models and domains and are associated with worst-class performance and generalization gaps beyond class imbalance and dataset bias (§3.1-§3.3). Second, we show that failures under distribution shift depend on *which* semantic concepts a model relies on, rather than on semantic representation structure alone (§3.4). Finally, we show that post-hoc analysis can identify unstable semantic dependencies as candidate intervention targets, while effective correction still benefits from classifier-aware adaptation (§3.5). In medical imaging, comparatively stable clinical concepts further show that semantic reliance must be interpreted in relation to the concepts being used.

2 Characterizing the Model-Level Effects of Semantic Coverage Imbalance

In this section, we develop a post-hoc framework for characterizing how uneven semantic coverage is reflected in trained visual models. We first introduce the problem setup (§2.1) and define class-level semantic profiles (§2.2). We then formulate model-level diagnostics of semantic representation structure and semantic reliance (**??**), introduce quantitative measures to analyze them (§2.3), and conclude with a theoretical analysis of the framework (§2.4).

2.1 Problem Formulation

Setup. We consider a multi-class classification task with input space $\mathcal{X}$ and labels $y \in \{1, \dots, C\}$. Given an input $x \in \mathcal{X}$, a trained model learns a representation $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and predicts $\hat{y} = h_\theta(f_\theta(x))$. To analyze the semantic content of the learned representation, we consider a set of K predefined concepts, such as visual attributes, textual descriptors, or domain-specific features. A concept function $g : \mathbb{R}^d \to [0,1]^K$ maps the representation to a semantic evidence vector:

$$\mathbf{z}(x) = g(f_\theta(x)) \in [0,1]^K. \tag{1}$$

We treat SCI as a property of uneven class–concept coverage in the training data and use $\mathbf{z}(x)$ to study how this uneven semantic coverage is reflected in an already-trained model.

Concept Evidence. For each concept $k \in \{1, \dots, K\}$, $z_k(x) \in [0,1]$ denotes the semantic evidence associated with concept k for input x. The concept function g can be instantiated using (i) *linear concept probes* on frozen representations, (ii) *vision–language* similarity scores, such as CLIP (Radford et al., 2021), with cosine similarity rescaled to $[0,1]$, or (iii) *domain-specific descriptor models*, such as MONET (Kim et al., 2024) for medical imaging. Thus, $\mathbf{z}(x)$ provides a common semantic interface for analyzing trained models across different architectures, datasets, and concept sources.

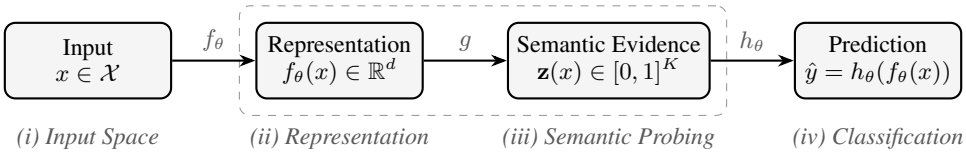

Figure 1: **Framework Overview.** Given input x, a representation $f_\theta(x)$ is extracted and mapped to a semantic evidence vector $\mathbf{z}(x) = g(f_\theta(x)) \in [0,1]^K$ via the concept function $g(\cdot)$. Classification proceeds via $h_\theta(\cdot)$. SCI is measured between the representation and semantic probing stages (dashed region), without modifying the underlying model.

Framework Overview. Figure 1 summarizes the post-hoc pipeline. A fixed trained model produces the representation $f_\theta(x)$ and prediction $h_\theta(f_\theta(x))$, while the concept function g maps the representation to semantic evidence $\mathbf{z}(x)$. This separation between representation, semantic probing, and prediction allows us to analyze how semantic information is represented and used across classes without modifying or retraining the underlying model. The resulting class-level semantic evidence forms the basis of our model-level characterization of the effects of uneven semantic coverage.

2.2 Class-Level Semantic Profiles

Given the semantic evidence vector $\mathbf{z}(x) \in [0,1]^K$ (Eq. 1), we define the *class semantic profile* $\boldsymbol{\mu}_c$ for class c as

$$\boldsymbol{\mu}_c := \mathbb{E}_{x \sim \mathcal{D}_c}[\,\mathbf{z}(x)\,] \in \mathbb{R}^K, \qquad \bar{\boldsymbol{\mu}} := \frac{1}{C}\sum_{c=1}^{C} \boldsymbol{\mu}_c, \tag{2}$$

where $\mathcal{D}_c$ denotes the data distribution conditioned on $y = c$. Each entry

$$\mu_{c,k} := \mathbb{E}[\,z_k(x) \mid y = c\,] \tag{3}$$

denotes the mean semantic evidence for concept k within class c.

The profile $\boldsymbol{\mu}_c$ summarizes the semantic evidence associated with class c in the learned representation. Comparing these profiles across classes allows us to examine whether semantic information is represented similarly or unevenly by the trained model. Because each profile is computed within a class and $\bar{\boldsymbol{\mu}}$ weights classes equally, these comparisons are not directly determined by class frequency.

These class-level profiles provide the basis for our post-hoc characterization of how uneven semantic coverage is reflected in trained models. We next distinguish two complementary model-level properties: *Semantic Imbalance* (SI), which characterizes variation in semantic representation structure across classes, and *Semantic Reliance* (SR), which characterizes how strongly model predictions depend on that semantic evidence.

2.3 Quantifying Semantic Imbalance and Semantic Reliance

To characterize how uneven semantic coverage is reflected in a trained model, we consider two complementary model-level properties: *Semantic Imbalance* (SI), which captures variation in semantic representation structure across classes, and *Semantic Reliance* (SR), which captures how strongly prediction outcomes are associated with semantic evidence.

Semantic Imbalance (SI). Using the class semantic profiles defined in Eq. 2, we quantify SI as

$$\mathcal{I}_{\mathrm{SI}} := \frac{1}{C}\sum_{c=1}^{C} \|\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}\|_2^2 . \tag{4}$$

$\mathcal{I}_{\mathrm{SI}}$ measures the variation of class semantic profiles around their global mean. A larger value indicates greater variation in semantic representation structure across classes, while $\mathcal{I}_{\mathrm{SI}} = 0$ indicates identical class semantic profiles.

Semantic Reliance (SR). SI describes semantic representation structure, but does not indicate how strongly semantic evidence is associated with prediction outcomes. For each class–concept pair (c, k),

we define

$$\rho_{c,k} := \left|\mu^{+}_{c,k} - \mu^{-}_{c,k}\right|, \qquad \mathcal{I}_{\mathrm{SR}} := \frac{1}{CK}\sum_{c=1}^{C}\sum_{k=1}^{K}\rho_{c,k}, \tag{5}$$

where

$$\mu^{+}_{c,k} := \mathbb{E}[\, z_k(x) \mid y = c, \hat{y} = y\,], \qquad \mu^{-}_{c,k} := \mathbb{E}[\, z_k(x) \mid y = c, \hat{y} \neq y\,]. \tag{6}$$

Thus, $\rho_{c,k}$ measures how much the semantic evidence for concept k differs between correct and incorrect predictions within class c. A larger value indicates a stronger association between that concept and prediction success or failure. $\mathcal{I}_{\mathrm{SR}}$ summarizes this association across all class–concept pairs.

SR measures the degree of semantic reliance observed post-hoc, but does not determine whether the relied-upon concepts are stable, task-relevant, or spurious under distribution shift. We examine this distinction later using concept instability and the Unified Reliance Score (URS).

Joint SI–SR Characterization. SI and SR capture complementary model-level properties. SI measures variation in semantic representation structure across classes, whereas SR measures how strongly semantic evidence is associated with prediction outcomes. Two models can therefore have similar SI but substantially different SR. We analyze models jointly in the SI–SR space to distinguish differences in semantic structure from differences in semantic reliance. We validate this characterization empirically in Section 3. Appendix D provides alternative norm choices and empirical estimation details.

2.4 Theoretical Analysis

We characterize the two-dimensional structure of SCI, SI and SR, and derive their implications for class-level risk and robustness under distribution shift. Our analysis is post-hoc: concept scores are extracted from learned representations and are used to diagnose model behavior. Proofs are provided in Appendix C.

Let $\rho_{c,k}$ denote the class-specific reliance on concept k for class c (Eq. 5). We define the average semantic reliance of concept k as $\mathrm{SR}_k := \frac{1}{C}\sum_{c=1}^{C}\rho_{c,k}$. Intuitively, SR_k measures how strongly prediction outcomes depend on concept k across classes.

Assumption 2.1 (Semantic concentration)**.** For each class c, the post-hoc concept vector $\mathbf{z}(x) = g(f_\theta(x))$ satisfies $\mathbf{z}(x) \mid y{=}c$ being sub-Gaussian with mean $\boldsymbol{\mu}_c$ and shared scale σ^2.

Assumption 2.2 (Semantic relevance)**.** The dominant directions of $\boldsymbol{\mu}_c$ correspond to task-relevant factors.

Assumption 2.3 (Post-hoc semantic faithfulness)**.** The post-hoc concept scores $\mathbf{z}(x) = g(f_\theta(x))$ are locally faithful to classifier-relevant representation directions. Specifically, changes in concept score z_k correspond to bounded perturbations along representation directions that affect the classifier output.

Let L_k denote the expected local sensitivity of the classifier to perturbations associated with concept k. We assume that this sensitivity is captured empirically by the semantic reliance score SR_k up to a constant factor, i.e., $L_k \leq C_0\mathrm{SR}_k$, for some constant $C_0 > 0$.

Assumption 2.4 (Concept instability)**.** Some concepts exhibit distribution-dependent variation:

$$\Delta_k := \left|\mathbb{E}[z_k \mid \text{source}] - \mathbb{E}[z_k \mid \text{target}]\right| > 0.$$

Proposition 2.1 (SI and semantic separability)**.** *Under Assumptions 2.1–2.2, larger pairwise separation $\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2$ reduces the semantic confusion risk between classes c and c'. Consequently, larger $\mathcal{I}_{\mathrm{SI}}$ corresponds to larger average semantic separation and can improve class separability when semantic directions are task-relevant.*

Remark. SI reflects how well classes are separated in semantic space. However, SI alone does not guarantee low worst-class risk, since an average separation measure does not ensure that every class pair is equally well separated.

Proposition 2.2 (SR and post-hoc shift vulnerability)**.** *Let $\Delta\ell(x) := \ell(h_\theta(f_\theta(x_{\mathrm{target}})), y) - \ell(h_\theta(f_\theta(x_{\mathrm{source}})), y)$ denote the change in prediction loss induced by a source-to-target distribution*

shift. Under Assumptions 2.3 and 2.4, the expected absolute loss perturbation satisfies $\mathbb{E}[|\Delta\ell|] \leq C_0 \sum_{k=1}^{K} \mathrm{SR}_k \Delta_k$. *Therefore, the Unified Reliance Score*

$$\mathrm{URS} := \sum_{k=1}^{K} \mathrm{SR}_k \Delta_k \tag{7}$$

serves as a post-hoc diagnostic of shift vulnerability.

Remark. URS captures the interaction between reliance and instability: concepts that strongly influence predictions and also shift across environments become liabilities under distribution shift.

Proposition 2.3 (SI and SR capture complementary model-level properties)**.** *There can exist trained models with identical* $\mathcal{I}_{\mathrm{SI}}$ *but different SR patterns. Under shifts that affect differently relied-upon concepts, such models can exhibit different shift behavior. Thus, SI alone does not determine semantic reliance or shift vulnerability.*

Remark. SI characterizes variation in semantic representation structure across classes, whereas SR characterizes how strongly semantic evidence is associated with prediction outcomes. Together, they provide complementary views of how uneven semantic coverage is reflected in a trained model; neither alone captures both structure and reliance.

SI and SR are complementary. Semantic Imbalance (SI) captures variation in semantic representation structure across classes, while Semantic Reliance (SR) captures how strongly semantic evidence is associated with prediction outcomes. Models with similar SI can exhibit different SR and, consequently, different behavior under distribution shift.

Proposition 2.1 is empirically supported in §3.2, where $\mathcal{I}_{\mathrm{SI}}$ is associated with worst-class performance when semantic evidence is task-aligned. Proposition 2.2 is supported in §3.4, where SR and URS are associated with shortcut reliance and degradation under distribution shift. Proposition 2.3 is illustrated empirically in §3.1, where models with similar SI can exhibit different SR.

3 Experiments

We evaluate our post-hoc framework through five analyses. First, we examine how SI and SR vary across models and datasets (§3.1). Next, we study whether these model-level properties are associated with worst-class performance and generalization gaps (§3.2). We then test whether these associations remain beyond class imbalance and dataset-level bias (§3.3). Next, we examine whether Semantic Reliance (SR) is associated with failure under distribution shift (§3.4). Finally, we investigate whether the framework can identify unstable semantic reliance patterns that inform robustness interventions (Appendix F.1).

Setup. We evaluate ResNet-50, ConvNeXt, ViT-B/16, and CLIP-ViT-B/32 across WaterBirds (Sagawa et al., 2020), CelebA (Liu et al., 2015), CIFAR-100 (Krizhevsky et al., 2009), NICO++ (Zhang et al., 2023), ISIC-DICM-17K (Ahammed et al., 2025), and MILK10k (Philipp et al., 2025). For each model–dataset pair, we report the task-specific performance metric together with $\mathcal{I}_{\mathrm{SI}}$ and $\mathcal{I}_{\mathrm{SR}}$ (Eqs. 4 and 5). Results are averaged over three seeds and reported in Table 4 (Appendix F). Full experimental settings are provided in Appendix G, and concept sources are described in Appendix G.2.

3.1 SI–SR Across Learned Representations

In this experiment, we examine how the model-level effects of uneven semantic coverage vary across visual models and datasets. We characterize each model–dataset pair in the joint SI–SR space (Fig. 2). SI captures variation in semantic representation structure across classes, while SR captures how strongly semantic evidence is associated with prediction outcomes.

SI–SR Space. Figure 2 visualizes the evaluated model–dataset pairs in the SI–SR space. We use the median SI and SR values across the evaluated settings to divide the two axes, yielding four descriptive regimes (Table 2, Appendix E): *(i) high SI, high SR*, where semantic representation structure varies substantially across classes and semantic evidence is strongly associated with prediction outcomes;

(ii) high SI, low SR, where class-level semantic structure varies but semantic reliance is relatively weak; *(iii) low SI, high SR*, where class-level semantic structure is more uniform but semantic reliance remains strong; and *(iv) low SI, low SR*, where both class-level variation and semantic reliance are relatively weak. The SI–SR space therefore separates variation in semantic representation structure from semantic reliance, revealing model-level differences that a single scalar summary cannot capture.

Findings. *(i) SI–SR patterns appear across all evaluated settings* (Prop. 2.3). All model–dataset pairs show non-trivial SI and/or SR, indicating that variation in semantic representation structure and semantic reliance is common across the evaluated models and domains. *(ii) Datasets occupy different SI–SR regimes.* WaterBirds shows moderate SI and high SR, consistent with strong reliance on background-related semantics. CelebA shows higher SI but lower SR, indicating substantial class-level semantic variation with weaker reliance on that evidence. CIFAR-100 also lies in a relatively low-SR regime, whereas NICO++ shows lower SI and higher SR, consistent with stronger reliance on contextual semantics. The medical datasets show relatively low SI but different levels of SR, reflecting shared clinical descriptors with varying semantic reliance. *(iii) Architecture affects SI–SR position.* The same dataset can occupy different regions of the SI–SR space across architectures, showing that the model-level effects of uneven semantic coverage depend on the learned representation. *(iv) SI and SR are complementary* (Prop. 2.3). High SI does not necessarily imply high SR, and low SI does not imply weak semantic reliance. This supports analyzing semantic representation structure and semantic reliance jointly rather than with a single scalar measure.

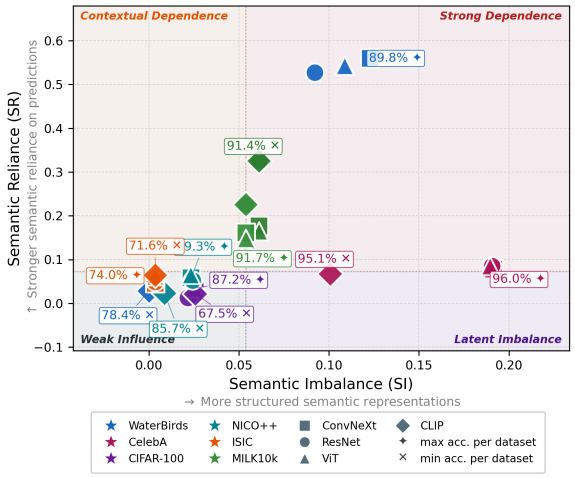

Figure 2: **SI–SR space across learned representations.** Each point represents a model–dataset pair positioned by Semantic Imbalance (SI) and Semantic Reliance (SR). Markers indicate the best- and worst-performing models within each dataset according to the corresponding task metric. The four regions show different combinations of semantic representation structure and semantic reliance across models and visual learning settings.

3.2 SI, SR, and Performance Disparity

We examine whether Semantic Imbalance (SI) and Semantic Reliance (SR) are associated with class-level performance, with particular focus on worst-class behavior. For each model–dataset combination, we compute SI, SR, macro-F1, and worst-class F1. We analyze their relationships using OLS regression (Fig. 3) and report Pearson (r) and Spearman (ρ) correlations in Table 6 (Appendix).

Findings. *(i) Higher SI is not necessarily harmful.* SI is positively associated with macro-F1 and worst-class F1 in several datasets, including WaterBirds ($r{=}{+}0.984$) and CelebA ($r{=}{+}0.878$). This indicates that greater variation in semantic representation structure can coexist with better class-level performance when the semantic evidence is task-aligned, consistent with Proposition 2.1. *(ii) SR shows strong but domain-dependent associations.* SR exhibits strong correlations with performance in several datasets (e.g., NICO++: $\rho{=}{+}0.930$; MILK10k: $\rho{=}{-}0.868$), but the direction varies across domains. This suggests that semantic reliance is beneficial when the relied-upon evidence is task-aligned, but can be harmful when that evidence is unstable or spurious. *(iii) SI and SR provide complementary information.* Models with similar SI can exhibit different SR and different performance, and high SI does not necessarily imply poor worst-class behavior. These results support analyzing semantic representation structure and semantic reliance jointly rather than using either quantity alone (Proposition 2.3).

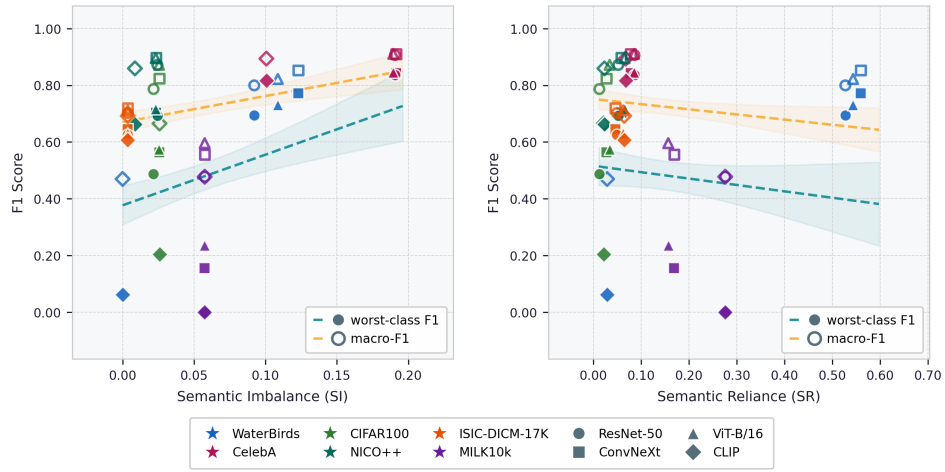

Figure 3: **SI and SR are associated with class-level performance.** OLS regression of SI (left) and SR (right) against macro-F1 (open markers) and worst-class F1 (filled markers) across datasets. Each point represents a model–dataset combination averaged over three seeds. SI shows generally positive associations with performance, while SR exhibits stronger dataset-dependent behavior.

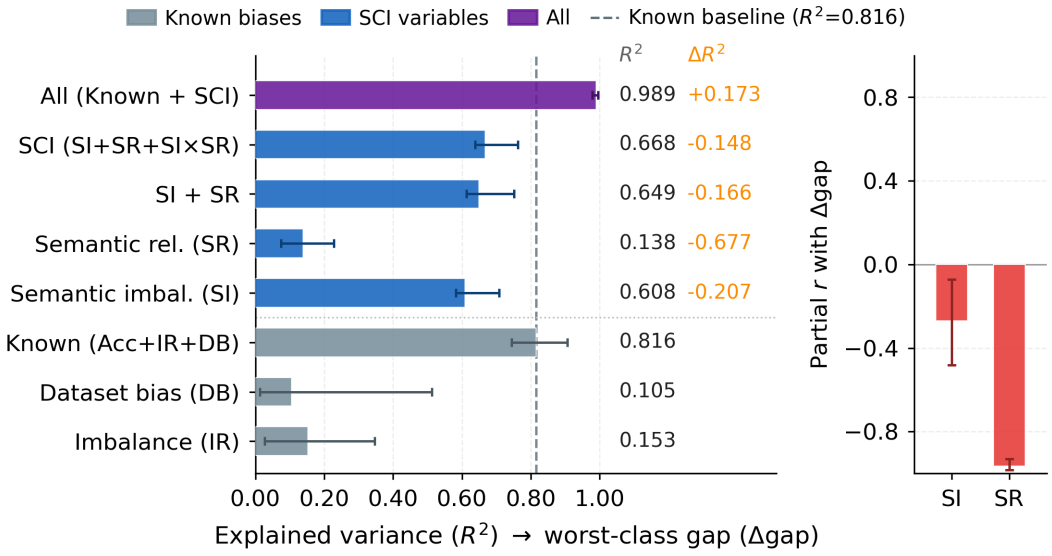

Figure 4: **SI and SR provide information beyond known factors.** *Left:* Incremental R^2 for predicting the worst-class gap (Δ_{gap}) as SI and SR are added to a baseline containing overall performance, class imbalance, and dataset-bias proxies. *Right:* Partial correlations with Δ_{gap} after controlling for class imbalance and dataset bias. Adding SI and SR increases explained variance, with SR showing the strongest conditional association.

3.3 SI and SR Beyond Class Imbalance and Dataset Bias

We test whether SI and SR provide information about class-level performance disparity beyond class imbalance and dataset bias. We define Δ_{gap} = macro-F1 − worst-class F1, where larger values indicate greater disparity. The baseline includes overall performance, training-set class imbalance, and a dataset-bias proxy based on linear prediction of available nuisance labels from learned representations. We then add SI, SR, and their interaction. All predictors are z-score standardized. Figure 4 reports incremental R^2 and partial correlations.

Findings. *(i) Known factors explain substantial but incomplete variation.* The baseline explains $R^2 = 0.816$ of the variation in Δ_{gap}. *(ii) SI and SR add explanatory information.* Adding SI, SR, and their interaction increases R^2 to 0.989 ($\Delta R^2 = +0.173$). Thus, the model-level semantic diagnostics capture variation in worst-class disparity that is not fully explained by the baseline factors. *(iii) SR shows the strongest conditional association.* After controlling for class imbalance and dataset bias, SR has a strong negative partial correlation with Δ_{gap} ($r = -0.963$, $p < 0.001$), while SI shows a weaker association. This indicates that semantic reliance provides information about class-level performance disparity beyond how semantic representation structure varies across classes.

3.4 Semantic Reliance and Shift Vulnerability

We examine whether reliance on shift-unstable semantic concepts is associated with model vulnerability under distribution shift (Proposition 2.2). We consider three settings: *(i) WaterBirds*, comparing shortcut-aligned and conflict groups; *(ii) NICO++*, using leave-one-domain-out evaluation; and *(iii) MILK10k*, using cross-modality transfer (Derm→Clinic and Clinic→Derm). SI and SR are computed from source-domain representations, and URS (Eq. 7) summarizes reliance on concepts that vary across domains. Shift degradation is measured on the corresponding target split.

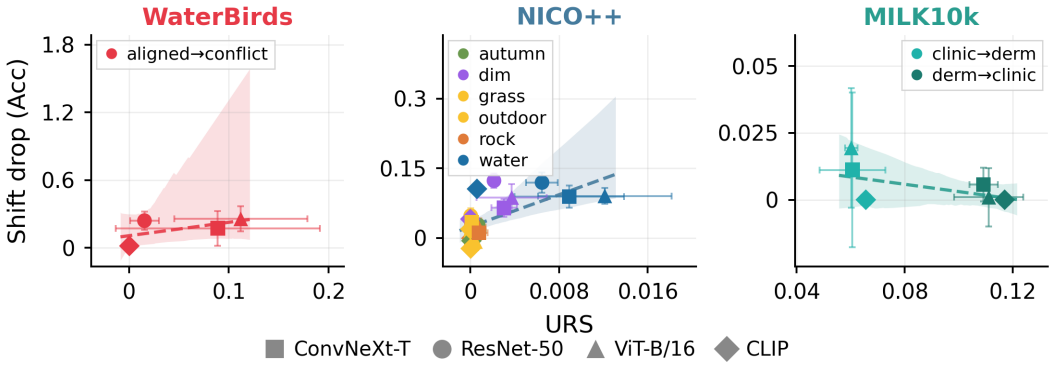

Figure 5: **Semantic reliance is associated with shift degradation.** Unified Reliance Score (URS) versus source-to-target performance degradation across WaterBirds, NICO++, and MILK10k. Higher URS is associated with larger degradation under shortcut and context shift, while this relationship is weaker under the clinical modality shift in MILK10k.

Findings. *(i) Reliance on unstable semantics is associated with shortcut failure.* On WaterBirds, higher SR and URS are associated with larger aligned-to-conflict performance drops, while SI shows a weaker relationship. This suggests that shift vulnerability depends more on which semantic evidence the model relies on than on class-level semantic variation alone. *(ii) URS extends to context shift.* On NICO++, URS is positively correlated with shift degradation (r=0.52, p<0.01), whereas SI remains weakly associated. Models relying more strongly on context-sensitive concepts show larger degradation on unseen domains. *(iii) Semantic reliance is not inherently harmful.* On MILK10k, high SR or URS does not consistently correspond to large cross-modality degradation. This suggests that reliance can remain robust when the underlying semantic concepts are clinically meaningful and stable across domains.

Overall, these results show that shift vulnerability is better understood by *what* semantic evidence a model relies on, particularly whether that evidence remains stable across domains.

3.5 URS as an Intervention-Guidance Diagnostic

We test whether URS can guide post-hoc intervention by ranking semantic directions that are both strongly relied upon and unstable under shift. URS attenuation alone gives small gains on the NICO++ shifts but does not improve WaterBirds. When combined with DFR, performance remains essentially unchanged relative to the matched DFR baseline ($\lambda = 0$; Table 1). Thus, URS is useful for *localizing candidate unstable semantic dependencies*, while effective correction still requires classifier-aware adaptation. Full controls, confidence intervals, and sensitivity analyses are provided in Appendix F.1.

Table 1: **URS-guided attenuation versus classifier adaptation.** Source-to-target performance drop (↓; mean over seeds). Full results and confidence intervals are in Appendix F.1.

Method	WaterBirds	NICO++ Dim	NICO++ Water
No mitigation	0.1742	0.0432	0.1024
URS attenuation	0.1793	0.0409	0.1015
DFR	**0.0712**	**0.0337**	**0.0981**

3.6 Discussion

Our results show that the model-level effects of uneven semantic coverage depend on how semantic information is represented and used after training. SI and SR capture these complementary properties. SI is not inherently associated with poor performance, while the effect of SR depends on the stability and relevance of the relied-upon concepts. Under shortcut and context shift, reliance on unstable semantic evidence is associated with larger degradation, whereas stable, task-relevant semantics can remain useful across domains.

This distinction also clarifies the role of post-hoc intervention. Our analysis can localize unstable semantic dependencies as candidate intervention targets, but attenuating these directions alone does not consistently improve robustness beyond classifier-aware adaptation (§3.5; Appendix F.1). Thus, diagnosis and correction are complementary: the framework identifies *what* semantic evidence may be problematic, while effective mitigation may still require adapting the decision boundary. The medical experiments further show that semantic reliance should not be reduced uniformly, since clinically meaningful concepts can remain stable and useful across domains.

More broadly, our framework extends SCI analysis from data-level coverage to post-hoc diagnosis of trained models. By separating Semantic Imbalance (SI) from Semantic Reliance (SR), it enables models under similar coverage conditions to be distinguished by how they represent and use semantic evidence. This provides a practical basis for semantic auditing, failure localization, and domain-aware robustness analysis without modifying the learned representation.

4 Conclusion

We introduced a post-hoc, model-agnostic framework for analyzing how uneven semantic coverage is reflected in the representations and behavior of trained visual models. The framework separates Semantic Imbalance (SI), which captures variation in semantic representation structure across classes, from Semantic Reliance (SR), which captures how strongly semantic evidence is associated with prediction outcomes.

Across diverse datasets and architectures, SI and SR reveal model-level differences that are not fully explained by class imbalance or dataset-level bias. Our results further show that robustness depends not simply on the amount of semantic reliance, but on *which* concepts a model relies on and whether those concepts remain stable under distribution shift. The framework can therefore help localize unstable semantic dependencies and guide subsequent intervention, while classifier-aware adaptation remains important for effective correction. Overall, this work advances SCI toward post-hoc semantic auditing of already-trained visual models and provides a foundation for analyzing representation behavior and robustness under uneven semantic coverage.

Limitation. Our analysis depends on the semantic coverage of the chosen concept vocabulary. If important latent features are not represented by the concept function g, SI and SR may not capture their contribution to model behavior. This limitation is common to concept-based analysis and motivates future work on adaptive concept discovery and domain-specific semantic probes.

AI use statement

We used Grammarly to refine language, including grammar, phrasing, clarity, and conciseness. Claude was used for limited assistance with code implementation and debugging. The authors reviewed all AI-assisted text and code. The authors produced and verified experimental results and

reported measurements; no synthetic datasets were generated using AI tools. The authors take full responsibility for the methods, experiments, claims, results, references, and final content of this work.

Ethics statement

This work uses existing visual benchmark datasets, including medical imaging datasets, and follows their stated access and usage conditions. This study collected no new human subjects or patient data. Our analysis explicitly examines uneven semantic coverage, which may reflect or amplify underrepresentation in real-world data. The medical experiments are intended for research evaluation only.

Reproducibility statement

We provide the implementation code as supplementary material. The appendix documents the datasets and preprocessing, model and training configurations, coverage computation, evaluation protocols, controlled coverage interventions, and implementation details required to reproduce the experiments. It also reports additional analyses, ablations, and coverage-stratified results. These materials support independent reproduction and verification of the reported findings.

References

Sakib Ahammed, Xia Cui, Wenqi Lu, and Moi Hoon Yap. Skin lesion classification using dermoscopic images and clinical metadata: Insights from multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 222–230, June 2025.

Sakib Ahammed, Xia Cui, Xinqi Fan, Wenqi Lu, and Moi Hoon Yap. Semcovnet: Towards fair and semantic coverage-aware learning for underrepresented visual concepts. *arXiv preprint arXiv:2602.16917*, 2026.

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *ICML*, 2024.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Jinho Choi, Hyesu Lim, Steffen Schneider, and Jaegul Choo. Conceptscope: Characterizing dataset bias via disentangled visual concepts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=lkmlNHuzY4.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

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

Sangwoo Hong, Youngseok Yoon, Hyungjun Joo, and Jungwoo Lee. Gbmix: Enhancing fairness by group-balanced mixup. *IEEE Access*, 12:18877–18887, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature medicine*, 30(4):1154–1165, 2024.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

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

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

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SctfBCLmWo.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Tschandl Philipp, Akay Nisa Bengü, Rosendahl Cliff, Rotemberg Veronica, Todorovska Verche, Weber Jochen, Wolber Anna Katharina, Müller Christoph, Kurtansky Nicholas, Halpern Allan, et al. Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

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.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 10962–10971, 2023.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, pp. 37765–37786. PMLR, 2023.

Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. *Advances in Neural Information Processing Systems*, 37:61839–61871, 2024.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16036–16047, 2023.

A Appendix

B Extended Related Work **13**

C Proofs of Theoretical Propositions **13**

C.1 Proof of Proposition 2.1 13

C.2 Proof of Proposition 2.2 14

C.3 Proof of Proposition 2.3 15

D Metric Properties and Estimation Details **15**

E Interpretation of SI–SR space **16**

F Additional Experiment Results **16**

F.1 URS-Guided Intervention Analysis 16

F.2 Performance and semantic statistics across datasets and models 17

F.3 Correlation between semantic quantities and model performance across datasets . . 17

G Experimental Settings **17**

G.1 Datasets. 17

G.2 Semantic Evidence Extraction 18

G.3 Metrics 21

G.4 Implementation Details 21

G.5 Probe Protocol 21

G.6 Training Details for All Probes 22

B Extended Related Work

Dataset and Distribution-Level Bias A large number of recent studies have shown that visual models encode dataset-specific signals, leading to persistent bias across benchmarks. Liu and He (Liu & He, 2025) demonstrated that modern neural networks can reliably predict the dataset of origin for an image across large-scale collections such as YFCC (Thomee et al., 2016), CC (Changpinyo et al., 2021), and DataComp (Gadre et al., 2023). Zeng et al. (Zeng et al., 2024) further decompose dataset biases into semantic, structural, and frequency components, showing that semantic discrepancy is the dominant component of such bias. These findings establish that models capture dataset-specific features, and these biases persist even under standard training pipelines. However, these approaches focus on differences across datasets, characterizing bias at the distribution level. They do not explain *why* models trained on a single dataset still exhibit large performance disparities across classes. In contrast, our framework characterizes the SCI as a representation-level bias at the class level, measuring how unevenly meaningful concepts are relied upon *within* a single dataset rather than across datasets.

Concept-Based Analysis and Probing Concept-based methods provide tools to analyze how semantic concepts are encoded in learned representations. TCAV measures concept influence by representing concepts as directions in activation space and estimating their effect on predictions using directional derivatives (Kim et al., 2018). Concept Bottleneck Models (CBMs) incorporate human-specified concepts as intermediate predictions within model architectures (Wang et al., 2023; Koh et al., 2020; Kulkarni et al., 2025). More recently, vision-language models such as CLIP (Radford et al., 2021) enable scalable concept probing via image–text similarity, enabling semantic analysis without additional supervision (Oikarinen et al., 2023). While these methods quantify which concepts are encoded or influential, they primarily serve as interpretability tools. They do not characterize *how* concept usage varies across classes within a model, nor whether such variation leads to systematic failures. Our framework measures *how unevenly* concepts are relied upon across classes, providing a systematic *analysis* of how semantic reliance relates to robustness and failure behavior.

Concept-Class Correlations and Bias Analysis Concept-based bias analysis has been studied to identify which concepts are present to understand how they drive model failures. ConceptScope (Choi et al., 2025) proposes an annotation-free framework that discovers and categorizes human-interpretable visual concepts into *target*, *context*, and *bias* types based on their correlation with class labels to audit dataset bias. Unsupervised concept discovery can identify shared semantic structures across samples, enabling shortcut reliance to be reduced without requiring group annotations (Arefin et al., 2024). At the class level, the degree of spurious correlation between concepts and classes can be quantified and used directly to guide debiasing (Wu et al., 2023). Together, these lines of work study bias *in* data and *between* concepts, but not how models allocate semantic reliance unevenly across classes within a representation. SCI addresses this gap, presenting a quantitative, representation-level characterization that moves from describing what concepts exist to explaining how uneven semantic reliance across classes is associated with model failures.

C Proofs of Theoretical Propositions

C.1 Proof of Proposition 2.1

Proof. Let $\mathbf{z}(x) = g(f_\theta(x))$ denote the post-hoc concept representation extracted from the learned features. By Assumption 2.1, for each class c, $\mathbf{z}(x) \mid y = c$ is sub-Gaussian with mean $\boldsymbol{\mu}_c$ and shared scale σ^2.

Consider two classes c and c'. Define the normalized semantic direction

$$\mathbf{v}_{c,c'} = \frac{\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2}.$$

The projected class means differ by

$$\mathbf{v}_{c,c'}^\top \boldsymbol{\mu}_c - \mathbf{v}_{c,c'}^\top \boldsymbol{\mu}_{c'} = \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2.$$

Since the projected concept scores are sub-Gaussian, a standard tail bound gives

$$\Pr(\text{semantic confusion between } c \text{ and } c') \leq \exp\left(-\frac{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^2}{8\sigma^2}\right).$$

Thus, increasing the pairwise distance $\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2$ reduces pairwise semantic confusion risk.

Using the variance identity

$$\frac{1}{C}\sum_{c=1}^{C}\|\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}\|_2^2 = \frac{1}{2C^2}\sum_{c,c'}\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^2,$$

larger $\mathcal{I}_{\mathrm{SI}}$ corresponds to larger average pairwise semantic separation. Under Assumption 2.2, these semantic directions are task-relevant, so improved semantic separation can improve class separability.

Importantly, this establishes improved semantic separability in the post-hoc concept space, rather than a direct guarantee on classifier error. □

C.2 Proof of Proposition 2.2

Proof. The framework is post-hoc: the classifier is $h_\theta(f_\theta(x))$, while the concept vector $\mathbf{z}(x) = g(f_\theta(x))$ is extracted from the learned representation only for analysis. We therefore do not assume that the classifier directly operates on $\mathbf{z}(x)$.

Let

$$\Delta\ell(x) := \ell(h_\theta(f_\theta(x_{\mathrm{target}})), y) - \ell(h_\theta(f_\theta(x_{\mathrm{source}})), y)$$

denote the change in prediction loss induced by a source-to-target shift.

By Assumption 2.3, a change in concept score z_k corresponds to a bounded perturbation along a classifier-relevant representation direction $\mathbf{u}_k$. Let $\delta_k(x)$ denote the induced perturbation magnitude associated with concept k. A first-order expansion of the loss in representation space gives

$$\Delta\ell(x) \approx \sum_{k=1}^{K} \nabla_h \ell(h_\theta(f_\theta(x)), y)^\top \mathbf{u}_k \, \delta_k(x).$$

Taking absolute values and applying the triangle inequality,

$$|\Delta\ell(x)| \le \sum_{k=1}^{K} \left|\nabla_h \ell(h_\theta(f_\theta(x)), y)^\top \mathbf{u}_k\right| \, |\delta_k(x)|.$$

Taking expectations gives

$$\mathbb{E}[|\Delta\ell|] \le \sum_{k=1}^{K} \mathbb{E}\left[\left|\nabla_h \ell(h_\theta(f_\theta(x)), y)^\top \mathbf{u}_k\right| \, |\delta_k(x)|\right].$$

Let

$$L_k := \mathbb{E}\left[\left|\nabla_h \ell(h_\theta(f_\theta(x)), y)^\top \mathbf{u}_k\right|\right]$$

denote the expected local sensitivity of the prediction loss to the representation direction associated with concept k. Under Assumption 2.4, the expected magnitude of the concept shift is

$$\Delta_k = \left|\mathbb{E}[z_k \mid \mathrm{source}] - \mathbb{E}[z_k \mid \mathrm{target}]\right|.$$

Assuming the induced representation perturbation is bounded in expectation by the corresponding concept shift magnitude, i.e.,

$$\mathbb{E}[|\delta_k(x)|] \le \Delta_k,$$

we obtain

$$\mathbb{E}[|\Delta\ell|] \le \sum_{k=1}^{K} L_k \Delta_k.$$

By Assumption 2.3, the empirical semantic reliance score SR_k captures this local sensitivity up to a constant factor:

$$L_k \le C_0 \mathrm{SR}_k,$$

for some constant $C_0 > 0$. Therefore,

$$\mathbb{E}[|\Delta\ell|] \leq C_0 \sum_{k=1}^{K} \mathrm{SR}_k \Delta_k = C_0 \mathrm{URS}.$$

Thus, URS is not a causal equality between concept shift and error. Rather, it is a post-hoc diagnostic upper bound for shift vulnerability: models that rely more strongly on unstable concepts have larger potential loss perturbation under distribution shift. □

C.3 Proof of Proposition 2.3

Proof. We construct two post-hoc semantic representations with the same class mean geometry but different reliance patterns.

Let two representations share identical class-conditional concept means $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$. Since $\mathcal{I}_{\mathrm{SI}}$ depends only on the dispersion of these means, both representations have the same SI:

$$\mathcal{I}_{\mathrm{SI}}^{(A)} = \mathcal{I}_{\mathrm{SI}}^{(B)}.$$

Now suppose the concepts can be partitioned into stable concepts $\mathcal{K}_s$, for which $\Delta_k \approx 0$, and unstable concepts $\mathcal{K}_u$, for which $\Delta_k > 0$. Consider two models with the same semantic class geometry but different reliance patterns:

$$\mathrm{SR}_k^{(A)} > 0 \text{ mainly for } k \in \mathcal{K}_s, \qquad \mathrm{SR}_k^{(B)} > 0 \text{ mainly for } k \in \mathcal{K}_u.$$

Then their Unified Reliance Scores differ:

$$\mathrm{URS}^{(A)} = \sum_{k \in \mathcal{K}_s} \mathrm{SR}_k^{(A)} \Delta_k \approx 0,$$

whereas

$$\mathrm{URS}^{(B)} = \sum_{k \in \mathcal{K}_u} \mathrm{SR}_k^{(B)} \Delta_k > 0.$$

By Proposition 2.2, the second model has larger potential loss perturbation under distribution shift despite having the same SI. Thus, SI alone cannot determine robustness. SR is required to distinguish whether the model relies on stable or unstable semantic evidence.

Therefore, SCI is inherently two-dimensional: SI captures semantic separation or imbalance in the representation, while SR captures the model's reliance on that semantic evidence. □

D Metric Properties and Estimation Details

Alternative norms for SI. We use the ℓ_2 norm in Eq. equation 4 for simplicity and its connection to variance decomposition. Alternative norms (ℓ_1) can be substituted without changing the formulation; we find ℓ_2 most stable across datasets in practice.

Empirical estimation. Population expectations in Eq. equation 4–equation 5 are approximated using empirical sample means. Let $\hat{\mathcal{D}}_c$ denote the set of samples belonging to class c. We estimate

$$\hat{\mu}_{c,k} := \frac{1}{|\hat{\mathcal{D}}_c|} \sum_{x \in \hat{\mathcal{D}}_c} z_k(x), \tag{8}$$

with $\hat{\mu}_{c,k}^{+}$ and $\hat{\mu}_{c,k}^{-}$ computed analogously over correctly and incorrectly predicted sample subsets, respectively.

Threshold ϵ. The threshold ϵ in Definition **??** is a qualitative device to distinguish non-trivial imbalance from noise. In all experiments we use $\mathcal{I}_{\mathrm{SI}}$ as continuous quantitative measures, making ϵ unnecessary in practice.

Table 2: Interpretation of the SI–SR space. SI (Semantic Imbalance) captures variation of semantic support across classes, while SR (Semantic Reliance) measures the dependence of model predictions on semantic evidence. The four regimes correspond to different modes of representation and behavior.

SI	SR	Regime Name	Interpretation
High	High	Strong Semantic Dependence	Uneven semantic structure that is strongly used by the model; can reflect shortcut or task-aligned reliance depending on data
High	Low	Latent Semantic Imbalance	Semantic structure exists but is weakly utilized; imbalance is present but does not drive predictions
Low	High	Contextual Dependence	Balanced semantic structure, but the model relies on shared or contextual semantics
Low	Low	Weak Semantic Influence	Semantic signals have limited influence on predictions; model decisions are weakly explained by measured concepts

E Interpretation of SI–SR space

The SI–SR space provides a unified view of representation and behavior. SI captures the structural distribution of semantic support across classes, while SR captures whether and how this structure influences predictions. Importantly, high SI does not necessarily imply harmful bias, and high SR does not necessarily imply failure. Instead, different regions correspond to distinct semantic regimes, which vary across datasets and models. This decomposition reveals that semantic imbalance (SI) and semantic reliance (SR) play fundamentally different roles: SI reflects representation geometry, while SR reflects decision behavior. Their interaction determines whether semantic structure improves performance, is ignored, or leads to undesirable reliance. This motivates analyzing SCI as a two-dimensional phenomenon rather than a single scalar.

F Additional Experiment Results

F.1 URS-Guided Intervention Analysis

Goal. We test whether URS can identify semantic directions that are useful targets for post-hoc intervention. This experiment evaluates the *specificity of the diagnostic*, rather than proposing URS attenuation as a standalone robustness method.

URS ranking. For concept k, we combine Semantic Reliance with its observed instability:

$$\alpha_k = \mathrm{SR}_k \, \Delta_k, \tag{9}$$

where larger α_k indicates stronger reliance on a concept whose semantic evidence changes across the specified source–target shift. URS is therefore conditional on an observed shift; it is not intended as a universal predictor of unseen future domains.

Feature attenuation. Let $\mathbf{h} = f_\theta(x) \in \mathbb{R}^d$ and let $\mathbf{w}_k$ denote the linear-probe direction for concept k. To reduce removal of class-discriminative information, we project each direction away from the class-mean subspace,

$$\tilde{\mathbf{w}}_k = (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{w}_k, \tag{10}$$

where $\mathbf{U}$ spans class-mean differences estimated from group-balanced training data. For the selected high-URS concepts $\mathcal{S}$, we apply

$$\mathbf{h}' = \mathbf{h} - \lambda \sum_{k \in \mathcal{S}} \tilde{\alpha}_k \, \mathrm{proj}_{\tilde{\mathbf{w}}_k}(\mathbf{h}), \tag{11}$$

where λ is selected using source validation only.

Comparison with DFR. URS attenuation and DFR (Kirichenko et al., 2023) address different stages: URS ranks potentially unstable semantic directions in the representation, whereas DFR adapts the classifier decision boundary. We therefore evaluate URS attenuation alone, DFR, and URS attenuation followed by DFR. For the latter, $\lambda = 0$ exactly recovers the matched DFR re-run.

Table 3: **Post-hoc intervention results.** Shift drop is source minus target performance (↓). ΔDrop is the reduction relative to no mitigation (↑). Intervals are paired seed-bootstrap 95% CIs.

Shift	Method	Shift Drop ↓	ΔDrop [95% CI] ↑
WaterBirds	No mitigation	0.1742 ± 0.0153	–
	URS attenuation	0.1793 ± 0.0128	-0.0050 $[-0.0148, 0.0131]$
	DFR	$\mathbf{0.0712 \pm 0.0102}$	$+0.1030$ $[0.0898, 0.1229]$
NICO++ Dim	No mitigation	0.0432 ± 0.0026	–
	URS attenuation	0.0409 ± 0.0043	$+0.0024$ $[0.0000, 0.0042]$
	DFR	$\mathbf{0.0337 \pm 0.0019}$	$+0.0095$ $[0.0088, 0.0103]$
NICO++ Water	No mitigation	0.1024 ± 0.0049	–
	URS attenuation	0.1015 ± 0.0045	$+0.0009$ $[0.0006, 0.0014]$
	DFR	$\mathbf{0.0981 \pm 0.0106}$	$+0.0043$ $[-0.0028, 0.0082]$

Does URS add to DFR? The matched DFR+URS experiment shows little additional effect once the classifier has been adapted. On WaterBirds, shift drop changes from 0.0789 for the matched DFR re-run to 0.0792 with URS attenuation. For NICO++ Dim and Water, the corresponding changes are $0.0301 \rightarrow 0.0306$ and $0.0995 \rightarrow 0.0996$. Thus, the current evidence does not support an additive robustness gain from URS attenuation beyond DFR.

Interpretation. The results separate *failure localization* from *failure correction*. URS identifies candidate semantic dependencies by jointly considering reliance and observed instability. Attenuating those directions alone provides small gains on the NICO++ shifts, but not on WaterBirds, while DFR produces substantially larger improvements by adapting the decision boundary. We therefore interpret URS as an intervention-guidance diagnostic rather than a standalone mitigation method.

F.2 Performance and semantic statistics across datasets and models

F.3 Correlation between semantic quantities and model performance across datasets

G Experimental Settings

G.1 Datasets.

We evaluate SCI across diverse visual domains to capture different types of semantic structure and bias. Summary of dataset configurations and concept sources are in Table 7.

WaterBirds (Sagawa et al., 2020) is a spurious-correlation benchmark for waterbird vs. landbird classification with background bias. Training is strongly skewed (95% of waterbirds on water, 95% of landbirds on land; 4,056 images), while the test set is background-balanced (5,794 images), enabling controlled analysis of shortcut reliance.

CelebA (Liu et al., 2015) (attribute-rich) contains 162,770 training and 19,962 test images of celebrity faces annotated with 40 binary attributes. We treat `blond_hair` as the binary classification target and use the remaining 39 attributes as concept sources.

CIFAR-100 (Krizhevsky et al., 2009) is a general object recognition benchmark with 100 fine-grained classes, 50,000 training and 10,000 test images. It contains no explicit attribute annotations; concept activations are obtained via vision–language probing using CLIP with class-specific textual descriptors from (Oikarinen et al., 2023).

NICO++ (Zhang et al., 2023) is a domain generalisation benchmark containing objects photographed across six contexts (e.g., *water*, *grass*, *dim*, *outdoor*). We adopt a leave-one-domain-out protocol to evaluate sensitivity to unseen contexts, enabling analysis of context-dependent semantic reliance under realistic distribution shift.

MILK10k (Philipp et al., 2025) is a cross-modality medical imaging benchmark containing 10,480 skin-lesion images spanning dermoscopic and clinical (non-dermoscopic) modalities. Each image is

Table 4: Core performance and semantic statistics across datasets and models (mean ± std across 5 seeds). Accuracy is reported in percent; the remaining quantities retain the scales used in the updated summary. We report Top-1 Accuracy, Semantic Imbalance (SI), Semantic Reliance (SR), macro-F1, and worst-class F1 (Worst-F1). These results provide the primary baseline reference for subsequent SCI analyses (§3.1–§3.4).

Dataset	Model	Accuracy	$\mathcal{I}_{\mathrm{SI}}$	$\mathcal{I}_{\mathrm{SR}}$	macro-F1	Worst-F1
WaterBirds	ResNet-50	85.57±1.365	0.0921±0.0044	0.5271±0.0176	0.7996±0.0152	0.6936±0.0207
	ConvNeXt	89.76±1.819	0.1230±0.0182	0.5596±0.0267	0.8526±0.0227	0.7712±0.0339
	ViT-B/16	86.91±2.120	0.1087±0.0230	0.5434±0.0161	0.8221±0.0247	0.7309±0.0354
	CLIP	78.38±0.044	0.0000±0.0000	0.0286±0.0024	0.4698±0.0019	0.0617±0.0036
CelebA	ResNet-50	95.84±0.105	0.1906±0.0030	0.0858±0.0031	0.9065±0.0032	0.8369±0.0058
	ConvNeXt	95.96±0.077	0.1914±0.0041	0.0779±0.0041	0.9100±0.0030	0.8431±0.0056
	ViT-B/16	95.88±0.131	0.1895±0.0013	0.0854±0.0100	0.9106±0.0031	0.8449±0.0060
	CLIP	95.05±0.014	0.1006±0.0002	0.0675±0.0001	0.8939±0.0002	0.8164±0.0004
CIFAR-100	ResNet-50	78.72±0.256	0.0217±0.0251	0.0124±0.0011	0.7867±0.0027	0.4877±0.0021
	ConvNeXt	82.34±0.509	0.0257±0.0272	0.0271±0.0047	0.8236±0.0049	0.5632±0.0186
	ViT-B/16	87.16±0.504	0.0255±0.0268	0.0340±0.0032	0.8718±0.0046	0.5738±0.0885
	CLIP	67.49±0.074	0.0259±0.0236	0.0216±0.0017	0.6651±0.0008	0.2049±0.0284
NICO++	ResNet-50	86.85±0.283	0.0243±0.0002	0.0520±0.0006	0.8717±0.0028	0.6932±0.0104
	ConvNeXt	89.35±0.186	0.0235±0.0009	0.0582±0.0005	0.8967±0.0021	0.7035±0.0106
	ViT-B/16	89.10±0.221	0.0232±0.0008	0.0652±0.0009	0.8952±0.0018	0.7163±0.0039
	CLIP	85.74±0.014	0.0086±0.0001	0.0230±0.0001	0.8594±0.0001	0.6629±0.0017
MILK10k (Derm)	ResNet-50	91.41±0.000	0.0611±0.0000	0.3256±0.0000	0.4776±0.0000	0.0000±0.0000
	ConvNeXt	91.68±0.222	0.0611±0.0000	0.1768±0.0772	0.5481±0.0398	0.1399±0.0787
	ViT-B/16	91.64±0.621	0.0611±0.0000	0.1653±0.0802	0.6104±0.0934	0.2652±0.1863
	CLIP	91.41±0.000	0.0611±0.0000	0.3256±0.0000	0.4776±0.0000	0.0000±0.0000
MILK10k (Clinic)	ResNet-50	91.41±0.000	0.0538±0.0000	0.2261±0.0000	0.4776±0.0000	0.0000±0.0000
	ConvNeXt	91.51±0.359	0.0538±0.0000	0.1612±0.0550	0.5635±0.0903	0.1718±0.1810
	ViT-B/16	91.72±0.238	0.0538±0.0000	0.1478±0.0414	0.5815±0.0551	0.2068±0.1101
	CLIP	91.41±0.000	0.0538±0.0000	0.2261±0.0000	0.4776±0.0000	0.0000±0.0000
ISIC-DICM-17K	ResNet-50	72.36±0.542	0.0036±0.0000	0.0491±0.0017	0.7024±0.0152	0.6252±0.0376
	ConvNeXt	73.95±0.656	0.0036±0.0000	0.0458±0.0050	0.7190±0.0162	0.6450±0.0367
	ViT-B/16	73.59±1.513	0.0036±0.0000	0.0624±0.0108	0.7115±0.0299	0.6315±0.0623
	CLIP	71.59±0.216	0.0036±0.0000	0.0644±0.0005	0.6923±0.0025	0.6069±0.0040

annotated with seven MONET-derived descriptor probability scores $\in [0, 1]$: *ulceration/crust*, *hair*, *vasculature*, *erythema*, *pigmentation*, *gel/liquid*, and *skin markings*. It enables evaluation of semantic reliance under modality shift (Derm→Clinic and Clinic→Derm), where concept distributions change across acquisition conditions while clinically valid descriptors remain partially invariant.

ISIC-DICM-17K (Ahammed et al., 2025) is a skin-lesion classification dataset comprising 17,060 dermoscopic images across multiple diagnostic categories. Each image is annotated with MONET-derived descriptor probability scores $\in [0, 1]$ for seven dermatological descriptors: *ulceration/crust*, *hair*, *vasculature*, *erythema*, *pigmentation*, *gel/liquid*, and *skin markings*. These provide expert-defined semantic evidence aligned with dermatological practice.

Together, these six datasets provide spurious correlations, attribute supervision, open-vocabulary semantics, domain shift, and real-world medical imaging, collectively covering the full range of SCI regimes identified in our framework.

G.2 Semantic Evidence Extraction

To obtain concept activations $\mathbf{z}(x)$, we follow a unified three-stage protocol: (1) train or fine-tune the backbone model, (2) freeze the feature encoder f_θ, and (3) extract semantic evidence using dataset-specific concept sources. When explicit concept annotations are unavailable, lightweight linear probes are trained on frozen backbone features. This ensures consistent semantic evidence extraction across all models while preserving a shared semantic interpretation space.

Table 5: Complementary performance statistics for the same model–dataset pairs (mean ± std across 5 seeds). Worst-Acc is reported in percent; the remaining quantities retain the scales used in the updated summary. We report AUC, Δgap (macro-F1 − worst-class F1), best-class F1 (Best-F1), and worst-class accuracy (Worst-Acc).

Dataset	Model	AUC	Δgap	Best-F1	Worst-Acc
WaterBirds	ResNet-50	0.8987±0.0052	0.1060±0.0058	0.9056±0.0097	73.49±1.613
	ConvNeXt	0.9325±0.0188	0.0813±0.0118	0.9339±0.0126	77.65±6.053
	ViT-B/16	0.9142±0.0212	0.0912±0.0115	0.9133±0.0153	79.93±5.173
	CLIP	0.7728±0.0037	0.4080±0.0017	0.8778±0.0002	3.21±0.194
CelebA	ResNet-50	0.9846±0.0006	0.0696±0.0026	0.9762±0.0006	80.19±1.717
	ConvNeXt	0.9859±0.0004	0.0669±0.0026	0.9768±0.0004	81.53±2.004
	ViT-B/16	0.9860±0.0004	0.0656±0.0029	0.9762±0.0008	84.35±3.576
	CLIP	0.9786±0.0000	0.0775±0.0002	0.9714±0.0001	82.53±0.039
CIFAR-100	ResNet-50	0.9954±0.0001	0.2990±0.0033	0.9535±0.0065	41.00±1.897
	ConvNeXt	0.9970±0.0002	0.2604±0.0142	0.9675±0.0100	45.80±3.059
	ViT-B/16	0.9982±0.0001	0.2980±0.0845	0.9870±0.0050	46.20±10.457
	CLIP	0.9856±0.0000	0.4602±0.0282	0.9366±0.0000	14.00±2.280
NICO++	ResNet-50	0.9961±0.0002	0.1785±0.0093	0.9787±0.0020	69.74±2.103
	ConvNeXt	0.9967±0.0001	0.1933±0.0117	0.9892±0.0010	73.77±2.978
	ViT-B/16	0.9972±0.0001	0.1788±0.0028	0.9876±0.0020	71.23±2.232
	CLIP	0.9952±0.0000	0.1965±0.0017	0.9889±0.0000	62.22±0.472
MILK10k (Derm)	ResNet-50	0.8800±0.0074	0.4776±0.0000	0.9551±0.0000	0.00±0.000
	ConvNeXt	0.8910±0.0284	0.4082±0.0390	0.9563±0.0011	8.22±4.796
	ViT-B/16	0.9136±0.0080	0.3451±0.0930	0.9555±0.0035	21.33±16.306
	CLIP	0.6345±0.0108	0.4776±0.0000	0.9551±0.0000	0.00±0.000
MILK10k (Clinic)	ResNet-50	0.8519±0.0190	0.4776±0.0000	0.9551±0.0000	0.00±0.000
	ConvNeXt	0.8838±0.0144	0.3917±0.0907	0.9551±0.0020	13.78±16.651
	ViT-B/16	0.9031±0.0062	0.3748±0.0550	0.9563±0.0013	13.56±7.647
	CLIP	0.7090±0.0090	0.4776±0.0000	0.9551±0.0000	0.00±0.000
ISIC-DICM-17K	ResNet-50	0.8121±0.0026	0.0772±0.0227	0.7796±0.0090	60.68±8.583
	ConvNeXt	0.8289±0.0018	0.0741±0.0205	0.7931±0.0045	62.24±8.342
	ViT-B/16	0.8327±0.0056	0.0799±0.0337	0.7914±0.0132	59.84±11.564
	CLIP	0.7869±0.0012	0.0853±0.0017	0.7776±0.0018	56.98±0.651

Concept Sources. We use three types of concept sources to demonstrate the generality of the framework:

- *Visual attributes:* human-annotated attributes; background (*place*) for WaterBirds, 40 facial attributes for CelebA, and 6 domain contexts for NICO++.
- *Vision–language probing:* CLIP-aligned concept scores using class-specific textual descriptors using generic visual concept vocabulary (Oikarinen et al., 2023) for CIFAR-100.
- *Domain-specific descriptors:* MONET (Kim et al., 2024) dermatological descriptor scores providing clinically meaningful semantic evidence for skin-lesion datasets (ISIC-DICM-17K, MILK10k).

Concept Scoring for CIFAR-100 via CLIP. For datasets without ground-truth semantic attributes (e.g., CIFAR-100), we estimate concept-level semantic evidence using a pretrained CLIP (Radford et al., 2021) vision–language model. Let $f_{\text{img}}(\cdot)$ and $f_{\text{text}}(\cdot)$ denote the ℓ_2-normalized image and text encoders, respectively, and let $\mathcal{T} = \{t_k\}_{k=1}^{K}$ be a predefined vocabulary of K concept prompts (e.g., *"a photo of a {class}"*). All concept prompts are encoded once and cached. For image x and concept prompt t_k, the concept score is defined as the image–text cosine similarity:

$$z_k(x) \;=\; \cos\big(f_{\text{img}}(x),\, f_{\text{text}}(t_k)\big) \;=\; f_{\text{img}}(x)^{\top} f_{\text{text}}(t_k), \tag{12}$$

yielding a concept-score vector $\mathbf{z}(x) \in [0,1]^K$ per image, obtained by normalizing raw cosine similarity scores to $[0,1]$. For CIFAR-100, we use a generic visual concept vocabulary (Oikarinen

Table 6: Correlation between semantic quantities and model performance across datasets. Semantic Imbalance (SI), Semantic Reliance (SR), and Δgap (macro-F1 $-$ worst-class F1) are reported with Pearson (r) and Spearman (ρ) correlations. *Signal* indicates the dominant predictor of worst-class performance based on stronger absolute correlation. *Trend* summarizes the direction of the semantic effect: Positive indicates higher SI/SR is associated with better performance; Negative indicates higher SI/SR is associated with worse performance, and Weak indicates no clear relationship. N/A denotes cases where correlation could not be computed due to insufficient class-level variance. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Stat	SI vs F1		SR vs F1		SI, Δgap	SR, Δgap	Dominance	
		worst-class F1	macro-F1	worst-class F1	macro-F1			Signal	Trend
WaterBirds	r	+0.984***	+0.981***	+0.993***	+0.990***	-0.985***	-0.996***	SI	Positive
	ρ	+0.765**	+0.753**	+0.741**	+0.725**	-0.695*	-0.657*		
CelebA	r	+0.878***	+0.900***	+0.713**	+0.714**	-0.842***	-0.708*	SR	Positive
	ρ	+0.575	+0.568	+0.733**	+0.726**	-0.577*	-0.746**		
CIFAR100	r	+0.102	+0.014	+0.301	+0.506	-0.161	-0.086	–	Weak
	ρ	-0.014	-0.063	+0.357	+0.699*	+0.077	-0.203		
NICO++	r	+0.782**	+0.729**	+0.895***	+0.884***	-0.396	-0.391	SR	Positive
	ρ	+0.415	+0.295	+0.930***	+0.776**	-0.492	-0.435		
ISIC-DICM-17K	r	N/A	N/A	-0.773**	-0.773**	N/A	+0.734**	SR	Negative
	ρ	N/A	N/A	-0.622*	-0.720**	N/A	+0.601*		
MILK10k	r	-0.053	-0.053	-0.813***	-0.815***	+0.052	+0.810***	SR	Negative
	ρ	-0.092	-0.092	-0.868***	-0.861***	+0.092	+0.868***		

Table 7: Summary of dataset configurations and concept probe designs used in the proposed framework analysis. K denotes the number of concepts; C denotes the number of classes.

Dataset	Classification Task	C	Concept Source	Concept Supervision	K
WaterBirds	Bird classification	2	Background attribute (*place i.e., land/water*)	Multi-label	2
CelebA	Blond_Hair classification	2	40 binary facial attribute annotations	Multi-label	39
CIFAR-100	Object classification	100	CLIP-aligned concept vocabulary (Oikarinen et al., 2023)	Regression	50 (topk)
NICO++	Object classification	60	Context/domain attributes	Multi-label	6
ISIC-DICM-17K	Skin lesion classification	2	MONET dermatological descriptors	Regression	7
MILK10k	Skin lesion classification	2	MONET dermatological descriptors	Regression	7

MONET descriptors: *ulceration/crust*, *hair*, *vasculature*, *erythema*, *pigmentation*, *gel/liquid*, *skin markings*.
NICO++ contexts: *autumn*, *dim*, *grass*, *outdoor*, *rock*, *water*.

et al., 2023) tailored to low-resolution object categories, producing scores without any additional supervision. These scores serve as the input to our regression probe, which regresses $\mathbf{z}(x)$ onto the target label to assess concept-level predictive evidence.

Semantic probes and representation-level analysis. In this work, we characterize SCI as a representation-level phenomenon, rather than a property of any specific concept extractor. Our goal is therefore not to analyze the external probe models themselves, but to analyze how semantic evidence is distributed and utilized within the learned representations of the evaluated backbone model.

Post-hoc semantic analysis requires a semantic coordinate system that maps latent representations to interpretable concepts. For this reason, we employ external semantic probes (e.g., CLIP-aligned concept similarity or MONET descriptor predictors) only as semantic measurement functions. The evaluated representation remains the backbone representation $f_\theta(x)$ itself, while the probe outputs define the semantic evidence vector $\mathbf{z}(x)$ used for SCI analysis.

For medical datasets such as ISIC and MILK10k, we use MONET descriptor scores because they provide clinically meaningful dermatological concepts learned via expert-guided distillation on skin-lesion imagery. This enables SCI analysis in a medically relevant semantic space that is difficult to capture using generic vision-language vocabularies alone.

G.3 METRICS

We report the following metrics across all experiments:

- *Top-1 accuracy* and *macro-F1*: overall model performance.
- *Worst-class F1*: minimum per-class F1 across all classes, measuring tail performance.
- *Worst-class gap*: the difference between macro-F1 and worst-class F1, measuring the disparity between average and tail-class performance.
- *Shift Drop - Accuracy*: difference between in-distribution and out-of-distribution accuracy.
- *Semantic Imbalance* $\mathcal{I}_{\mathrm{SI}}$ (Eq. 4): class-level semantic dispersion.
- *Semantic Reliance* $\mathcal{I}_{\mathrm{SR}}$ and URS (Eqs. 5, 7): prediction dependence on semantic evidence.

All results are averaged over three random seeds.

G.4 IMPLEMENTATION DETAILS

Models and Training Protocol. We evaluate four representative architectures spanning convolutional, transformer-based, and foundation model representations: ResNet-50, ConvNeXt-T, ViT-B/16, and CLIP (ViT-B/32 visual encoder). ResNet-50, ConvNeXt-T, and ViT-B/16 are initialized from ImageNet-pretrained weights and fine-tuned on each dataset using standard cross-entropy loss. For CLIP, the visual encoder is frozen, and a linear classifier is trained on extracted features. All models are trained on identical data splits and evaluated over three random seeds to ensure reproducibility.

Hyperparameter Settings. Table 8 reports the hyperparameter settings used for all fine-tuning experiments. All models are trained with identical settings unless stated otherwise. For NICO++, the held-out domain is varied per leave-one-domain-out fold.

Table 8: Hyperparameter settings for fine-tuning training protocol.

Category	Hyperparameter	Value
Optimisation	Learning rate	3×10^{-5}
	Batch size	32
	Epochs	20
Data	Image size	224×224
	Number of workers	4
Model	Initialisation	ImageNet pretrained
	Output dimension	Dataset-specific[‡]
	CLIP encoder	Frozen (linear probe only)
Infrastructure	GPU	2x NVIDIA A100
	Random seeds	{0, 1, 2}
NICO++	Protocol	Leave-one-domain-out
	Held-out domain	Varied per fold

[‡] Output dimension equals the number of classes per dataset: WaterBirds (2), CelebA (2), CIFAR-100 (100), NICO++ (varies by split), ISIC-DICM-17K (varies), MILK10k (varies).

G.5 PROBE PROTOCOL

Let $f_\theta(x) \in \mathbb{R}^d$ denote the frozen feature representation produced by backbone f_θ for input $x \in \mathcal{X}$ (consistent with Section 2.1). For each concept $k \in \{1, \ldots, K\}$, we train a linear probe $\phi_k : \mathbb{R}^d \to \mathbb{R}^{n_k}$ on top of the frozen features and define the concept activation score as:

$$z_k(x) = \sigma\big(\mathbf{w}_k^\top f_\theta(x) + b_k\big) \in [0, 1], \tag{13}$$

where $\mathbf{w}_k \in \mathbb{R}^d$ is the probe weight vector, b_k is the bias, and σ denotes the sigmoid function for binary and multi-label concepts, or the softmax function for multi-class concepts. The resulting

scalar $z_k(x) \in [0,1]$ forms the k-th entry of the semantic evidence vector $\mathbf{z}(x) = g(f_\theta(x)) \in [0,1]^K$ (Eq. 1).

All probes share the same linear architecture with output dimension n_k determined by the concept type. Raw logits are used during training with numerically stable fused loss functions (`BCEWithLogitsLoss` for binary/multi-label, `CrossEntropyLoss` for multi-class); the activation function σ is applied only at inference time when computing $z_k(x)$.

The feature encoder f_θ is frozen prior to probe training. For TIMM-based architectures (ResNet-50, ViT-B/16, ConvNeXt-T), the classification head is removed via `reset_classifier(0)`; for CLIP, the `encode_image` pathway is used with ℓ_2-normalised outputs.

Table 9: Probe types used across datasets.

Probe type	**Output**	**Loss**	**Inference**	**Use case**
Binary	scalar	BCE w/ logits	$\sigma(\text{logit})$	One binary concept (e.g. place)
Multi-label	(B, K)	BCE w/ logits	$\sigma(\text{logits})$	K binary concepts jointly (e.g. CelebA attrs)
Multi-class	(B, C)	Cross-entropy	$\mathrm{softmax}(\text{logits})$	C-way concept (e.g. bird species)
Regression	scalar	MSE	identity	Continuous scores (e.g. CLIP, MONET)

G.6 Training Details for All Probes

All probes are trained with a shared set of hyper-parameters summarized in Table 10. Features are extracted once per split in a single forward pass and cached on CPU, so probe training does not require further GPU computation on the backbone.

Table 10: Shared concept probe hyper-parameters.

Hyper-parameter	**Value**
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	512
Epochs	50
Architecture	Single linear layer $\mathbb{R}^D \to \mathbb{R}^{n_k}$
Binary / multi-label loss	`BCEWithLogitsLoss`
Multi-class loss	`CrossEntropyLoss`
Regression loss	`MSELoss`
Feature normalization (CLIP)	ℓ_2

Constant concept columns (zero variance across the training split) are excluded from training to avoid degenerate probes. For binary and multi-label probes, any target values outside $[0,1]$ (arising from non-attribute metadata columns leaked into the concept vector) are min-max normalized to $[0,1]$ before training.