COVERAGE-AWARE SEMANTIC REPRESENTATION LEARNING FOR UNDERREPRESENTED VISUAL CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern vision models increasingly rely on rich semantic representations that extend beyond class labels to include descriptive concepts, attributes, and contextual cues. However, semantic concepts are not uniformly represented across classes: a concept may be frequent globally yet remain underrepresented within a class, resulting in low class–concept coverage. We formalize this class–concept phenomenon as *Semantic Coverage Imbalance (SCI)* and study its relationship with learned semantic representations. We find that lower class–concept coverage is consistently associated with weaker semantic representations. To address this, we introduce *Coverage-aware Semantic Representation Learning*, which shares more semantic structure for low-coverage or uncertain relations while preserving more pair-specific structure for high-coverage relations. We instantiate this principle in SEMCOVNET with *Coverage-Calibrated Semantic Sharing (CCSS)*, a coverage- and uncertainty-aware partial-pooling mechanism. Using a range of vision datasets spanning facial attributes and fine-grained recognition, as well as real-world medical datasets, we show that SCI is widespread, degrades representation quality, and can be mitigated through coverage-aware semantic representation learning.

1 INTRODUCTION

Semantic concepts provide an interpretable link between visual observations and task-level predictions, capturing class-relevant structure such as attributes, object parts, contextual cues, and clinically meaningful findings in medical imaging (Koh et al., 2020; Zhang et al., 2025). Concept-based models make such structure explicit and can support representations that are both discriminative and semantically meaningful (Gao et al., 2026). However, semantic concepts are not uniformly represented across classes (Zeng et al., 2024; Liu & He, 2025; Choi et al., 2025). We use **semantic evidence** to denote the concept annotations or scores observed in training examples of a given class. A concept may be common overall yet receive minimal coverage within a particular class, resulting in low coverage, as illustrated in Figure 1(a). This raises a fundamental question: *how does the amount of semantic evidence for class and concept relationships affect their learned representation, and can semantic sharing be adapted to that evidence?*

We study this at the class–concept pair level, since the same concept can have very different coverage across classes. Prior work has demonstrated within-class attribute imbalance and class-dependent concept distributions (Tang et al., 2022; Zhou et al., 2025; Choi et al., 2025), but it does not formalize the empirical coverage of each class–concept pair or study how this coverage affects the learned semantic representation. For example, an attribute may be common across bird species yet rare within one species. As Figure 1(a) illustrates, class and concept marginals can therefore overlook a long-tailed distribution of class–concept coverage. We refer to this phenomenon as **Semantic Coverage Imbalance (SCI)**, uneven empirical coverage of class–concept pairs, characterized by a long-tailed coverage distribution within a dataset.

SCI matters because low class–concept coverage can make semantic representations less reliable. As illustrated in Figure 1(b), high-coverage pairs are supported by many observations and result in compact representations, whereas low coverage pairs have sparse observations and can yield noisier

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.

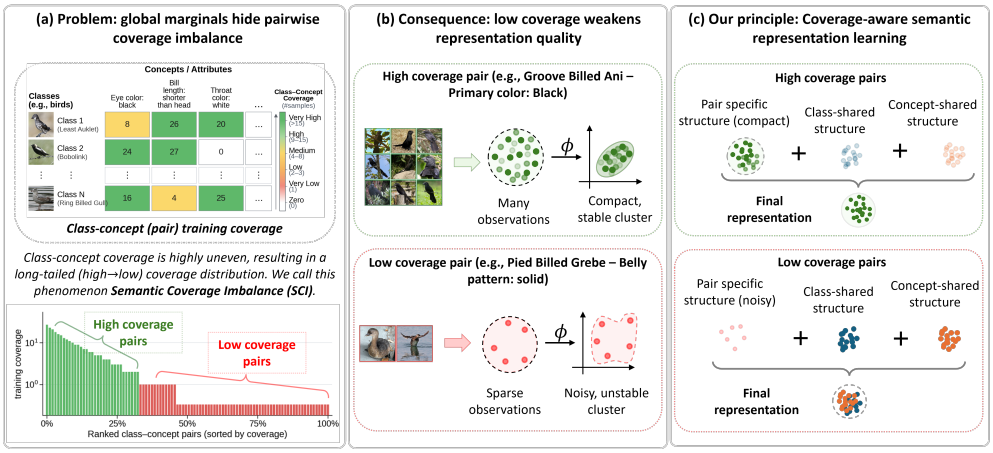

Figure 1: **Motivation and principle of coverage-aware semantic representation learning.** (a) Semantic Coverage Imbalance (SCI) occurs when class–concept pair coverage is highly uneven, even when class and concept marginals appear adequate. (b) We show that low coverage weakens semantic representations. (c) Coverage-aware semantic representation learning adapts more pair-specific structure for high-coverage pairs and shares more structure for low coverage or uncertain pairs.

representations. We examine this relationship in Section 4.2 and find that lower class–concept coverage is associated with weaker semantic representations. This is especially important in medical imaging, where diagnosis depends on clinically meaningful visual findings (Yang et al., 2024; Gao et al., 2024; Huy et al., 2025). In Derm7pt, rare diagnoses and rare findings create clinically important low-coverage class–concept pairs despite expert-defined diagnostic criteria for the 7-point melanoma checklist (Kawahara et al., 2019).

Long-tailed methods mainly rebalance classes (Kang et al., 2020; Menon et al., 2021; Luo et al., 2025), while generalized long-tail learning considers within-class attribute imbalance (Tang et al., 2022; Zhou et al., 2025). Group-robust methods target underperforming subpopulations (Sagawa et al., 2020; Liu et al., 2021; Park et al., 2026), and concept-based methods improve grounding, class-specific structure, or semantic representations (Zhang et al., 2025; Gao et al., 2026; Choi et al., 2025). However, existing methods generally optimize class frequencies, predefined groups, or attribute distributions rather than using class–concept relations to show how they learn semantic representations. We discuss related work in more detail in Appendix B.

Contributions. To address this limitation, we identify and formalize **Semantic Coverage Imbalance (SCI)**. We introduce **Coverage-aware Semantic Representation Learning**, illustrated in Figure 1**(c)**, which shares more semantic structure for low coverage or uncertain relations while preserving pair-specific structure for high coverage relations. We instantiate this principle in **SemCovNet** with **Coverage-Calibrated Semantic Sharing (CCSS)**, which adapts sharing using pair support and representation uncertainty. Across CelebA, CUB, Waterbirds, Derm7pt, and MILK10k, spanning binary, categorical, expert-defined, and soft semantic evidence, we find that SCI is widespread and that lower class–concept coverage is associated with weaker semantic representations. Controlled coverage interventions further show that reducing pair-specific coverage degrades representation quality, while coverage-aware learning mitigates this effect. Overall, our results suggest that semantic supervision depends not only on which concepts are annotated, but also on *how* their coverage is distributed across class–concept relations. We next formalize SCI and its representation-level consequence.

2 Formalizing Semantic Coverage Imbalance in Visual Learning

We now formalize **Semantic Coverage Imbalance (SCI)**. Figure 2 illustrates the intuitive view of SCI: how latent semantic relations are represented in observed training evidence and the resulting learned representations.

2.1 Problem Setup

Let $\mathcal{X}$ denote the observation space, $\mathbb{Y} := \{1, \ldots, T\}$ the set of class labels, and $\mathbb{K} := \{1, \ldots, K\}$ the set of semantic concepts (e.g., attributes, parts, or clinical findings). The training set is $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, \boldsymbol{a}_i)\}_{i=1}^{n}$, $\boldsymbol{a}_i \in [0,1]^K$, where $a_{i,k}$ denotes the evidence for concept k in image $\boldsymbol{x}_i$. This definition is used for binary attributes, categorical attributes expanded into atomic concepts, expert annotations, and soft semantic evidence.

Each **class–concept pair** (y, k) defines a relation in a semantic relation space $\mathcal{S}$, which the training data grounds in a set of observed examples by mapping $\mathcal{G} : (y, k) \mapsto \mathcal{D}_{y,k}$, where $\mathcal{D}_{y,k}$ contains the observations that provide evidence for (y, k). For binary, multi-level, or categorical annotations, the observations that support pair (y, k) are $\mathcal{D}_{y,k} := \{\boldsymbol{x}_i : y_i = y,\ a_{i,k} = 1\}$. For soft semantic annotations, pair support is represented by the corresponding evidence mass.

2.2 Class–Concept Coverage

We measure the available semantic evidence through class-conditioned concept coverage.

Definition 2.1 (class–concept coverage)**.** *Let* $n_y := \sum_{i=1}^{n} \mathbf{1}\{y_i = y\}$. *The empirical coverage of class* y *and concept* k *is*

$$C_{y,k} := \mathbb{E}[\mathrm{a}_k \mid \mathrm{y} = y] = \frac{1}{n_y} \sum_{i:y_i=y} a_{i,k}. \qquad (1)$$

For binary concepts, $C_{y,k}$ is the within-class prevalence of concept k; for soft evidence, it is the mean evidence mass within class y. We use *coverage* to denote the amount of observed support available for learning relation (y, k). We also define the effective support as the cardinality of $\mathcal{D}_{y,k}$ as

$$n_{y,k}^{\mathrm{eff}} := |\mathcal{D}_{y,k}| = \sum_{i:y_i=y} a_{i,k}, \qquad (2)$$

the total observed evidence contributing to pair (y, k). We refer to a pair with large $C_{y,k}$ (or $n_{y,k}^{\mathrm{eff}}$) as *high-coverage* (well-supported), and a pair with small $C_{y,k}$ as *low-coverage* (undercovered).

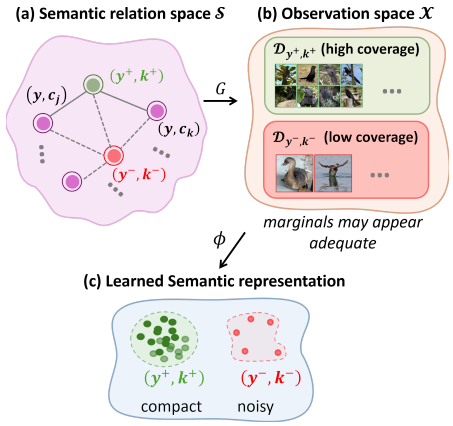

Figure 2: **Intuitive view of Semantic Coverage Imbalance (SCI).** (a) Each class–concept pair (y, k) defines a relation in the semantic space $\mathcal{S}$. (b) In the observation space $\mathcal{X}$, the training data grounds each semantic relation as $\mathcal{G} : (y, k) \mapsto \mathcal{D}_{y,k}$, producing high- and low-coverage pairs. (c) We hypothesize that low coverage leads to weaker learned semantic representations. We refer to the phenomenon of uneven class–concept pair coverage as SCI.

Property. class–concept coverage is pair-specific and cannot be determined from class or concept marginals alone. Since $\mathbb{E}[\mathrm{a}_k] = \sum_{y=1}^{T} P(\mathrm{y} = y)\, C_{y,k}$, the global prevalence of concept k is only a class-frequency-weighted average of its class-conditioned coverage. This pairwise view is what separates semantic coverage from standard class or concept frequency (Figure 2).

Consequently, coverage must be considered over the full set of class–concept pairs. When ranked by $C_{y,k}$, these pairs can exhibit a substantial low-coverage tail, as illustrated in Figure 1(a). We formalize this uneven class–concept pair coverage distribution as SCI in Section 2.3.

2.3 Semantic Coverage Imbalance

Definition 2.2 (Semantic Coverage Imbalance)**.** *Let* $\mathcal{P} := \{(y, k) \in \mathbb{Y} \times \mathbb{K} : |\mathcal{D}_y| > 0\}$ *denote the valid class–concept pairs in the training set, with empirical coverage* $C_{y,k}$ *defined in Equation 1. We define* **Semantic Coverage Imbalance (SCI)** *as the uneven distribution of empirical coverage* $\{C_{y,k}\}_{(y,k)\in\mathcal{P}}$ *across class–concept pairs within a dataset.*

SCI *is a pair-level property and is not determined by class or concept marginals alone. In particular, a class–concept pair may have low coverage even when its class is common, and its concept is*

frequently observed overall. Thus, class and concept marginals can appear well represented overall while their specific class–concept relation has low coverage.

Remark 2.1 (Long-tailed structure of class–concept coverage). Ranking the coverage values as $C_{(1)} \geq C_{(2)} \geq \cdots \geq C_{(N)}$, where $N = |\mathcal{P}|$, provides the class–concept pair analogue of the ranked class-frequency curve commonly used to characterize long-tailed label distributions (Cui et al., 2019). In practice, SCI appears as a long low-coverage tail containing many class–concept pairs, rather than as a single rare relation.

Representation hypothesis. Figure 2(c) illustrates the representation hypothesis underlying our study. Let $\phi : \mathcal{X} \to \mathbb{R}^D$ denote a learned visual representation, and let $q_{y,k}(\phi)$ measure the representation quality of class–concept pair (y, k). We hypothesize that

$$C_{y,k} \downarrow \implies \mathbb{E}[q_{y,k}(\phi)] \downarrow, \tag{3}$$

i.e., lower class–concept coverage corresponds to lower expected representation quality. Importantly, SCI is a property of the training distribution and is defined independently of the learned model. Whether lower coverage is associated with weaker semantic representations is therefore an empirical question, which we examine in Section 4.2 and further test through controlled coverage interventions in Section 4.4.

3 Method: Coverage-Aware Semantic Representation Learning

Section 2 formalizes SCI as uneven class–concept pair coverage and motivates the hypothesis that low coverage can weaken semantic representations. We therefore instantiate our **Coverage-aware Semantic Representation Learning** principle in **SemCovNet**: high coverage pairs retain more pair-specific structure, while low coverage or uncertain pairs share more semantic structure. As illustrated in Figure 3, a Concept Query Module maps dense image features to visual concept tokens for each semantic concept. Coverage-Calibrated Semantic Sharing (CCSS) then determines how strongly each class–concept pair relies on shared semantic structure versus its observed pair-specific representation.

3.1 Learning Visual Concepts

Given image $\boldsymbol{x}_i$, a visual encoder $f_{\boldsymbol{\theta}}$ produces dense patch features $\boldsymbol{F}_i^{\text{patch}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \{\boldsymbol{f}_{i,p}\}_{p=1}^{P}$. In our implementation, we initialize $f_{\boldsymbol{\theta}}$ from DINOv3[1]. For each semantic concept k, we use a learnable concept query embedding $\boldsymbol{q}_k \in \mathbb{R}^D$.

A Concept Query Module uses cross-attention to extract the visual realization of each concept:

$$\boldsymbol{z}_{i,k} = \Phi\left(\boldsymbol{q}_k, \boldsymbol{F}_i^{\text{patch}}\right). \tag{4}$$

Here, Φ denotes the cross-attention operator and $\boldsymbol{z}_{i,k} \in \mathbb{R}^D$ is the resulting visual concept token.

The semantic evidence $a_{i,k}$ is *not* an input to Equation 4. The concept token $\boldsymbol{z}_{i,k}$ is inferred from the image alone, while $a_{i,k}$ is used only as training supervision: $\hat{a}_{i,k} = h_k(\boldsymbol{z}_{i,k})$, giving the semantic supervision loss $\mathcal{L}_{\text{concept}}$ in Equation 12.

3.2 Pair-Specific and Shared Semantic Structure

We estimate the observed class–concept pair representation by aggregating the concept tokens from Equation 4 using the available semantic evidence:

$$\bar{\boldsymbol{z}}_{y,k} = \frac{\sum_{i:y_i=y} a_{i,k}\boldsymbol{z}_{i,k}}{n_{y,k}^{\text{eff}} + \epsilon}. \tag{5}$$

[1]We use DINOv3 as the visual encoder: its self-supervised pretraining on large-scale unlabeled data yields structured representations that transfer across tasks (classification, segmentation, depth, tracking) and domains (natural, medical, satellite) without text supervision (Siméoni et al., 2025)

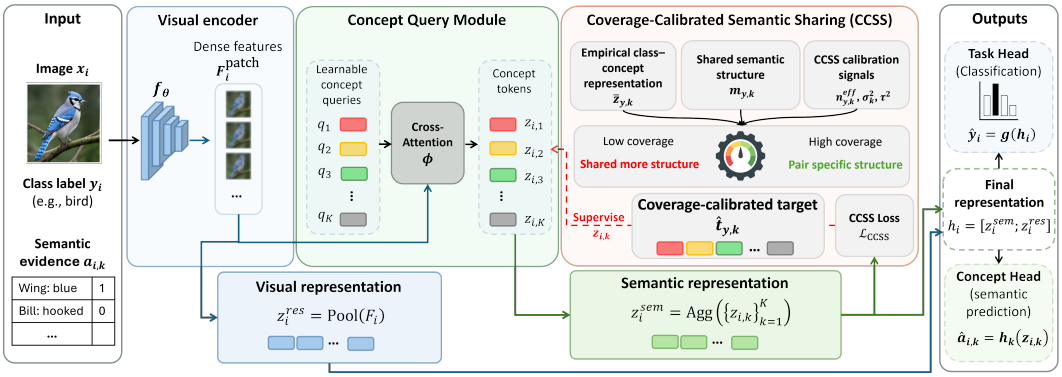

Figure 3: **SEMCOVNET for coverage-aware semantic representation learning.** A Concept Query Module maps dense features to image-driven concept tokens. For each class–concept pair, CCSS combines the empirical class–concept representation, a shared semantic structure, and calibration signals. Low-coverage or uncertain pairs shrink toward the shared structure, while high-coverage pairs retain more pair-specific structure. The resulting target supervises the concept tokens, and a residual pathway preserves complementary visual information for task prediction.

The prototype $\bar{\boldsymbol{z}}_{y,k}$ is a stop-gradient statistic computed only from the training split for each class–concept pair. We define a shared semantic structure as:

$$\boldsymbol{m}_{y,k} = \underbrace{\boldsymbol{\mu}_0}_{\text{global structure}} + \underbrace{\tilde{\boldsymbol{\mu}}_y}_{\text{class-shared structure}} + \underbrace{\tilde{\boldsymbol{\mu}}_k}_{\text{concept-shared structure}}, \tag{6}$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^D$ is a global learnable vector shared by classes and concepts. $\boldsymbol{\mu}_y$ is a learnable vector for class y, shared by every concept in that class; $\boldsymbol{\mu}_k$ is a learnable vector for concept k, shared by every class it appears in. To make this decomposition identifiable, we use $\tilde{\boldsymbol{\mu}}_y = \boldsymbol{\mu}_y - \frac{1}{C}\sum_{y'=1}^{C} \boldsymbol{\mu}_{y'}$, and $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k - \frac{1}{K}\sum_{k'=1}^{K} \boldsymbol{\mu}_{k'}$, so that each term in Equation 6 captures a distinct source of structure, with no overlap between them (See Appendix D.1 for gradient routing).

3.3 COVERAGE-CALIBRATED SEMANTIC SHARING (CCSS)

The reliability of the aggregated concept token $\bar{\boldsymbol{z}}_{y,k}$ depends on how much evidence supports the class–concept pair and how visually variable the concept is. Therefore, we estimate the uncertainty of each concept k as the variance of each concept token:

$$\sigma_k^2 = \frac{\sum_{y=1}^{C}\sum_{i:y_i=y} a_{i,k}\,\|\boldsymbol{z}_{i,k} - \bar{\boldsymbol{z}}_{y,k}\|_2^2}{D\sum_{y=1}^{C} n_{y,k}^{\text{eff}} + \epsilon}. \tag{7}$$

We also estimate the uncertainty of the shared semantic structure defined in Equation 6 as a global between-pair variance of how much pairs vary around their shared prior, net of sampling noise. Let $\mathbb{V}$ be the set of valid class–concept pairs with $n_{y,k}^{\text{eff}} > 0$:

$$\tau^2 = \left[\frac{1}{|\mathbb{V}|}\sum_{(y,k)\in\mathbb{V}}\left(\frac{\|\bar{\boldsymbol{z}}_{y,k} - \boldsymbol{m}_{y,k}\|_2^2}{D} - \frac{\sigma_k^2}{n_{y,k}^{\text{eff}}}\right)\right]_+. \tag{8}$$

The subtracted term removes the sampling noise expected purely from finite support; $[\cdot]_+$ floors the result at zero, since a variance cannot be negative.

CCSS combines σ_k^2, τ^2, and effective support $n_{y,k}^{\text{eff}}$ (Equation 2) into a *pair-specificity weight*:

$$\alpha_{y,k} = \frac{n_{y,k}^{\text{eff}}\tau^2}{n_{y,k}^{\text{eff}}\tau^2 + \sigma_k^2} \in [0,1], \tag{9}$$

where $\alpha_{y,k} = 0$ when $n_{y,k}^{\text{eff}} = 0$: a pair with no observed evidence relies entirely on the shared structure. The resulting coverage-calibrated target is

$$\hat{\boldsymbol{t}}_{y,k} = (1 - \alpha_{y,k})\boldsymbol{m}_{y,k} + \alpha_{y,k}\bar{\boldsymbol{z}}_{y,k}. \tag{10}$$

CCSS therefore depends on shared structure when evidence is low coverage or uncertain, and increasingly allows class-concept specialization as stable evidence increases.

3.4 Coverage-Aware Semantic Representation Learning Objective

We align each visual concept token with its coverage-calibrated target using semantic evidence weighted cosine loss:

$$\mathcal{L}_{\text{CCSS}} = \frac{\sum_{i=1}^{B}\sum_{k=1}^{K} a_{i,k}\left[1 - \cos\left(\boldsymbol{z}_{i,k}, \hat{\boldsymbol{t}}_{y_i,k}\right)\right]}{\sum_{i=1}^{B}\sum_{k=1}^{K} a_{i,k} + \epsilon}. \tag{11}$$

The target $\hat{\boldsymbol{t}}_{y,k}$ in Equation 10 is formed in the unnormalized latent space; we apply normalization only within the cosine similarity. Thus, low-support (low coverage) or uncertain pairs are regularized more strongly toward shared semantic structure, while well-supported (high coverage) pairs retain more pair-specific structure.

The annotated concept vocabulary may not capture all visual information. We therefore retain a residual visual pathway, $\boldsymbol{z}_i^{\text{res}} = \text{Pool}(\boldsymbol{F}_i)$, and aggregate the concept tokens as $\boldsymbol{z}_i^{\text{sem}} = \text{Agg}(\{\boldsymbol{z}_{i,k}\}_{k=1}^{K})$. The final task representation is $\boldsymbol{h}_i = [\boldsymbol{z}_i^{\text{sem}}; \boldsymbol{z}_i^{\text{res}}]$.

We optimize the task and semantic supervision objectives as

$$\mathcal{L}_{\text{task}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(g(\boldsymbol{h}_i),\, y_i)\,, \qquad \mathcal{L}_{\text{concept}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K} \ell_{\text{sem}}(\hat{a}_{i,k}, a_{i,k})\,, \tag{12}$$

where g is the task classifier, $\hat{a}_{i,k} = h_k(\boldsymbol{z}_{i,k})$ is the semantic prediction from Section 3.1, and ℓ_{sem} denotes the dataset-specific semantic supervision loss (Appendix C).

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{concept}}\mathcal{L}_{\text{concept}} + \lambda_{\text{CCSS}}\mathcal{L}_{\text{CCSS}}, \tag{13}$$

where λ_{concept} and λ_{CCSS} are fixed scalar weights for the concept supervision and coverage-aware alignment losses. We first warm up the Concept Query Module and prediction heads with $\mathcal{L}_{\text{task}} + \lambda_{\text{concept}}\mathcal{L}_{\text{concept}}$, then compute the training-set CCSS statistics and activate $\mathcal{L}_{\text{CCSS}}$ while adapting the selected DINOv3 blocks. $\bar{\boldsymbol{z}}_{y,k}$, σ_k^2, and τ^2 remain stop-gradient quantities throughout.

Semantic evidence and CCSS statistics are used only during training. Therefore, at inference, only the image is required:

$$\boldsymbol{x}_i \xrightarrow{f_{\boldsymbol{\theta}}} \boldsymbol{F}_i \longrightarrow \{\boldsymbol{z}_{i,k}\}_{k=1}^{K}, \quad \begin{cases} \boldsymbol{z}_i^{\text{sem}} = \text{Agg}(\{\boldsymbol{z}_{i,k}\}_{k=1}^{K}), \\ \boldsymbol{z}_i^{\text{res}} = \text{Pool}(\boldsymbol{F}_i), \end{cases} \longrightarrow \hat{y}_i = g(\boldsymbol{h}_i), \quad \hat{a}_{i,k} = h_k(\boldsymbol{z}_{i,k}). \tag{14}$$

4 Results and Discussion

Experimental setup. We evaluate the proposed coverage-aware semantic representation learning on CelebA, CUB, Waterbirds, Derm7pt, and MILK10k. We use the available semantic evidence in each dataset, including binary attributes, expert-defined clinical concepts, and continuous soft semantic scores. We use DINOv3 as the visual encoder and conduct observational and controlled analyses, as well as comparisons with concept-learning, imbalance-aware, and robustness baselines under identical training protocols. For analysis, we rank class–concept pairs by training coverage $C_{y,k}$ and use the 25th and 75th percentiles to define Tail ($Q_{0\text{–}25\%}$), Middle ($Q_{25\text{–}75\%}$), and Head ($Q_{75\text{–}100\%}$) groups. **Tail/Middle/Head pairs are used only for reporting and visualization; SemCovNet *does not* use them as training thresholds.** Appendix C provides full details of the dataset, model, baselines, training, and evaluation protocol.

4.1 SCI is Widespread Across Semantic Visual Datasets

We first investigate whether SCI occurs in real training data. For each class–concept pair (y, k), we compute training coverage $C_{y,k}$. Figure 4 shows strongly uneven coverage across CelebA, CUB, and Derm7pt. The fraction of class–concept pairs with $C_{y,k} < 0.05$ is 26.9%, 67.2%, and 33.9%, respectively; in CUB, 53.9% of pairs have zero training coverage. These results show that low-coverage class–concept pairs are common across distinct visual domains.

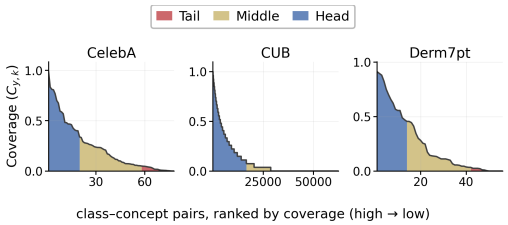

Figure 4: **SCI across visual domains.** Ranked training-set class–concept coverage $C_{y,k}$ for CelebA, CUB, and Derm7pt. Coverage is highly uneven: 26.9%, 67.2%, and 33.9% of class–concept pairs have $C_{y,k} < 0.05$ in CelebA, CUB, and Derm7pt, respectively.

Appendix E.1 provides a detailed characterization of SCI across all five datasets, including Waterbirds and MILK10k. Appendix E.1 further shows that class frequency alone does not explain this structure: class-balanced resampling largely preserves the coverage ordering, and permutation tests reveal substantial across-class variation in concept coverage. We next examine whether lower coverage is associated with weaker semantic representations.

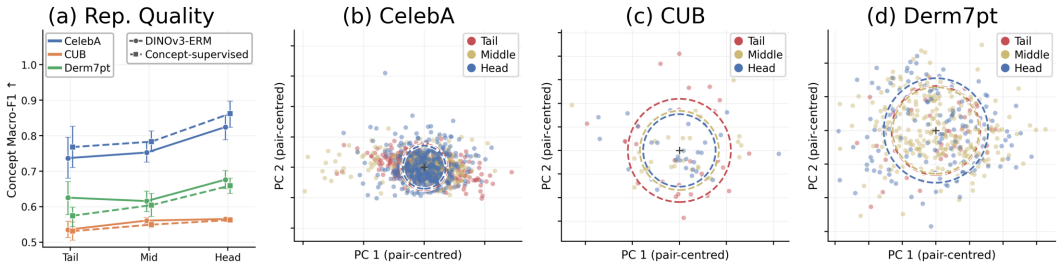

Figure 5: **Lower coverage is associated with weaker semantic representations. (a)** Frozen linear-probe class-conditioned concept Macro-F1 from the learned concept tokens across Tail, Middle, and Head class–concept pairs for DINOv3–ERM and concept supervised learning. Error bars show 95% CIs over class–concept pairs. **(b–d)** Concept supervised representations projected using a common PCA basis fitted separately for each dataset. Dispersion varies across coverage groups and datasets; $+$ denotes class–concept centroids.

4.2 Lower Coverage Is Associated with Weaker Semantic Representations

Figure 5(a) shows a clear coverage-dependent representation gap. Using the same frozen linear probe, DINOv3–ERM exhibits Head–Tail class-conditioned concept Macro-F1 gaps of 0.087, 0.030, and 0.051 on CelebA, CUB, and Derm7pt, respectively. Concept supervision does not remove this gap: the corresponding gaps remain 0.094, 0.031, and 0.084. Thus, semantic supervision alone does not eliminate the representation disparity between low- and high-coverage class–concept pairs. These results support our representation hypothesis in Section 2.3.

The representation geometry provides complementary evidence. Figure 5(b–d) shows greater Tail dispersion than Head dispersion on CelebA and CUB, decreasing from 7.12 to 6.45 and from 10.60 to 8.77, respectively. Derm7pt retains the frozen-probe coverage gap, but its geometric trend is less consistent; we therefore treat PCA geometry as qualitative supporting evidence. Overall, lower class–concept coverage is associated with weaker semantic representations, motivating coverage-aware learning that shares more semantic structure for low-coverage or uncertain pairs while preserving pair-specific structure for high-coverage pairs.

4.3 CCSS Improves Low-Coverage Representations

We next test whether the proposed coverage-aware learning principle mitigates this gap. We test this in two steps. First, we isolate the effect of our proposed Coverage-Calibrated Semantic Sharing (CCSS) using a controlled model ladder in which matched variants differ only by adding CCSS.

Table 1: **Controlled CCSS model ladder on CUB and CelebA.** The ladder isolates the effect of CCSS under frozen and LoRA-adapted DINOv3 encoders (last three layers). Concept denotes standard semantic supervision without CCSS. Tail reports class-conditioned concept Macro-F1 and $\Delta_{\mathrm{H-T}}$ the Head–Tail representation gap. Highlighted rows show the paired Tail gain from adding CCSS under the same adaptation setting, with 95% CIs.

Variant	Concept supervised	CCSS	Adaptation	CUB			CelebA		
				Task↑	Tail↑	$\Delta_{\mathrm{H-T}}$↓	Task↑	Tail↑	$\Delta_{\mathrm{H-T}}$↓
ERM	×	×	Frozen	0.886±0.006	N/A	N/A	0.953±0.001	N/A	N/A
Concept	✓	×	Frozen	0.881±0.007	0.003±0.002	0.171±0.032	0.954±0.004	0.441±0.102	0.426±0.091
Concept+CCSS (Ours)	✓	✓	Frozen	0.888±0.002	0.005±0.001	0.145±0.011	0.956±0.002	0.463±0.006	0.406±0.008
Δ_{Tail} **(Concept+CCSS − Concept), Frozen [95% CI]**				**0.002** [0.001, 0.002]			**0.022** [0.015, 0.060]		
Concept	✓	×	LoRA	0.886±0.003	0.004±0.002	0.188±0.026	0.949±0.006	0.411±0.072	0.448±0.049
Concept+CCSS (Ours)	✓	✓	LoRA	0.890±0.002	0.005±0.001	0.166±0.011	0.955±0.001	0.433±0.010	0.430±0.014
Δ_{Tail} **(Concept+CCSS − Concept), LoRA [95% CI]**				**0.001** [0.000, 0.002]			**0.023** [0.009, 0.060]		

Second, we compare SEMCOVNET with existing representation-, imbalance-, and concept-learning methods; we provide full comparisons in Appendix E.4.

Does CCSS itself matter? Within each adaptation setting, Concept and Concept+CCSS use the same backbone, semantic supervision, classifier, training schedule, and seeds. Thus, their difference isolates coverage-aware semantic sharing. As shown in Table 1, adding CCSS consistently improves Tail representation quality and reduces the Head–Tail gap under both frozen and LoRA-adapted DINOv3 encoders. Tail Macro-F1 improves by $+0.002$ and $+0.001$ on CUB, and by $+0.022$ and $+0.023$ on CelebA, under frozen and adapted encoders, respectively. Task performance is preserved or slightly improved.

These controlled comparisons separate the effect of CCSS from concept supervision and backbone adaptation. Concept supervision alone does not remove the coverage-dependent representation gap; explicitly adapting semantic sharing to class and concept support and uncertainty provides additional gains for low-coverage pairs. This supports the proposed principle: low coverage or uncertain class–concept pairs benefit from shared semantic structure, while high coverage pairs retain more pair-specific structure.

Comparison with existing methods. SEMCOVNET remains competitive in task and semantic representation performance. Appendix E.4 reports full SOTA comparisons.

4.4 CONTROLLED COVERAGE INTERVENTIONS

We next test whether reducing class–concept coverage degrades semantic representation quality, and whether coverage-aware learning mitigates this effect. We select CUB because it exhibits the strongest coverage sparsity (Figure 4(b)) and a clear coverage-dependent difference in representation dispersion (Figure 5(c)). We intervene on selected CUB class–concept pairs by retaining 100%, 75%, 50%, 25%, one example, or 0% of their pair-specific training evidence. We intervene only on the target pair: its class and concept remain observed through other training pairs, and the test set is unchanged. We measure representation quality using the same frozen-probe protocol as in Section 4.2, with all comparisons paired over the same target pairs. Appendix E.2 provides full target-pair selection, intervention details, and results.

Figure 6(a) shows small change in SEMCOVNET representation quality as pair-specific evidence is removed, from 0.679 at full coverage to 0.682

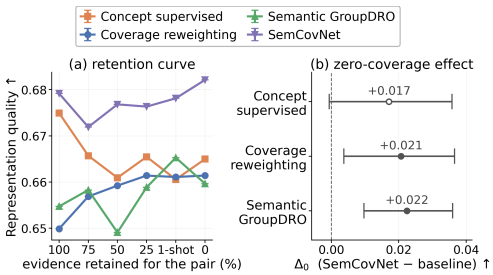

Figure 6: **Controlled class–concept coverage intervention on CUB. (a)** Frozen-probe representation quality as training evidence for target class–concept pairs is reduced from 100% to 0%; "one" retains one positive example. **(b)** Paired zero-coverage effect Δ_0 positive values indicate higher representation quality for SEMCOVNET. Error bars show paired 95% CIs over target pairs. At zero coverage, SEMCOVNET retains higher representation quality than Coverage reweight and Semantic GroupDRO.

at zero coverage. The zero-coverage setting provides the strongest test: the target class–concept pair is absent from training, while its constituent class and concept remain observed elsewhere. At zero coverage, SEMCOVNET improves q_{repr} over Coverage Reweight by $+0.021$ (95% CI $[+0.004, +0.037]$) and over Semantic GroupDRO by $+0.022$ ($[+0.010, +0.036]$). The point estimate over standard concept supervision is also positive ($+0.017$), although its 95% CI includes zero ($[-0.001, +0.036]$).

These interventions strengthen the observational result in Section 4.2: reducing pair-specific coverage degrades the evidence available to learn that relation, while SEMCOVNET remains more robust when direct pair evidence is absent. Reweighting or group-robust objectives can only act on observed training evidence, whereas CCSS can rely more strongly on semantic structure shared across the class and concept when pair-specific support is limited. The zero-coverage comparison therefore supports the proposed coverage-aware learning principle.

4.5 COVERAGE-AWARE LEARNING ACROSS CLINICAL SEMANTIC SUPERVISION

We next evaluate coverage-aware learning under clinical semantic supervision. On Derm7pt, which contains only 413 training samples with expert-defined clinical concepts under the 7-point melanoma gold-standard setting, Figure 7(a) shows that representation quality remains coverage-dependent. SEMCOVNET improves Tail class-conditioned concept Macro-F1 from 0.562 to 0.573 and reduces the Head–Tail gap from 0.090 to 0.055, with the main gain concentrated on low-coverage pairs.

We further consider continuous soft semantic evidence on MILK10k. We perturb the continuous training evidence to make the semantic target less informative. As Figure 7(b) shows, Tail representation quality and task AUROC remain stable over $\eta \in [0, 0.4]$. The concept supervised reference is similarly stable, so this experiment tests the applicability of coverage-aware learning to continuous and uncertain semantic evidence rather than uniform robustness. Together, these results show that the same coverage-aware principle applies across expert-defined and soft clinical semantic supervision. Full results and perturbation details are provided in Appendix E.3.

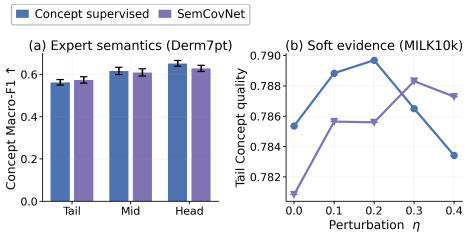

Figure 7: **Coverage-aware learning under clinical semantic supervision. (a)** On Derm7pt expert-defined clinical concepts, SEMCOVNET improves Tail representation quality and reduces the Head–Tail gap over Concept supervision. **(b)** On MILK10k continuous soft semantic supervision, Tail representation quality remains stable as semantic evidence becomes less informative. Error bars show 95% CIs.

5 CONCLUSION

We study semantic supervision at the class–concept relation level and formalize uneven empirical class–concept pair coverage as **Semantic Coverage Imbalance (SCI)**. Across diverse vision datasets, we find that lower class–concept coverage is associated with weaker semantic representations, motivating coverage-aware sharing of semantic structure for low-coverage or uncertain pairs. Our proposed principle, **Coverage-aware Semantic Representation Learning**, mitigates this effect, *even* when training evidence for a target class–concept pair is absent. These findings also show that strong task performance can hide weak representations for low-coverage semantic relations.

The broader implication is that semantic supervision should not be evaluated only by which concepts are annotated or how frequent they are globally. What matters is also how semantic evidence is distributed across class–concept pairs. This perspective is especially relevant in settings such as medical imaging, where insufficient but clinically meaningful relations may receive minimal training evidence. By explicitly considering class–concept pair coverage, SCI provides a new lens for analyzing semantic supervision, while coverage-aware learning shows a principled way to reduce the resulting representation gaps. This work shows that SCI and coverage-aware semantic representation learning provide a valuable perspective for learning more reliable semantic representations under uneven coverage.

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; SEMCOVNET is not a clinically validated diagnostic system and should not be used for clinical decision-making without appropriate external validation.

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

Itay Benou and Tammy Riklin Raviv. Show and tell: Visually explainable deep neural nets via spatially-aware concept bottleneck models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30063–30072, 2025.

Maty Bohacek, Thomas Fel, Maneesh Agrawala, and Ekdeep Singh Lubana. Uncovering conceptual blindspots in generative image models using sparse autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 25715–25766, 2026.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

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.

Shujian Gao, Yuan Wang, Chenglong Ma, Xin Gao, Jiangtao Yan, Junzhi Ning, Cheng Tang, Changkai Ji, Huihui Xu, Wei Li, et al. Escaping low-rank traps: Interpretable visual concept learning via implicit vector quantization. In *International Conference on Learning Representations*, volume 2026, pp. 95771–95805, 2026.

Yunhe Gao, Difei Gu, Mu Zhou, and Dimitris Metaxas. Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15010. Springer Nature Switzerland, October 2024.

Lijie Hu, Tianhao Huang, Huanyi Xie, Xilin Gong, Chenyang Ren, Zhengyu Hu, Lu Yu, Ping Ma, and Di Wang. Semi-supervised concept bottleneck models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2110–2119. IEEE, 2025.

Ta Duc Huy, Sen Kim Tran, Phan Nguyen, Nguyen Hoang Tran, Tran Bao Sam, Anton Van Den Hengel, Zhibin Liao, Johan W Verjans, Minh-Son To, and Vu Minh Hieu Phan. Interactive medical image analysis with concept-based similarity reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30797–30806, 2025.

Jinyong Jeong, Hyungu Kahng, and Seoung Bum Kim. Multi-expert distributionally robust optimization for out-of-distribution generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Lz5BUjArK4.

Sungho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In *International Conference on Learning Representations*, volume 2026, pp. 121167–121191, 2026.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, mar 2019. ISSN 2168-2194. doi: 10.1109/JBHI.2018.2824327.

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.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Yang Liu, Tianwei Zhang, and Shi Gu. Hybrid concept bottleneck models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20179–20189, 2025.

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.

Jiaan Luo, Feng Hong, Qiang Hu, Xiaofeng Cao, Feng Liu, and Jiangchao Yao. Long-tailed recognition with model rebalancing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jDKhljBQb8.

Aditya Krishna Menon, Sadeep Jayasumana, Ameya S Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

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.

Seonghwan Park, Jueun Mun, Donghyun Oh, and Namhoon Lee. An analysis of concept bottleneck models: Measuring, understanding, and mitigating the impact of noisy annotations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=s3TvewYDKk.

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6877–6886. IEEE, 2022.

Subeen Park, Joowang Kim, Hakyung Lee, Kyungwoo Song, et al. Spurious correlation-aware embedding regularization for worst-group robustness. In *International Conference on Learning Representations*, volume 2026, pp. 156420–156445, 2026.

Emiliano Penaloza, Tianyue H. Zhang, Laurent Charlin, and Mateo Espinosa Zarlenga. Addressing concept mislabeling in concept bottleneck models through preference optimization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48762–48788. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/penaloza25a.html.

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.

Katharina Prasse, Patrick Knab, Sascha Marton, Christian Bartelt, and Margret Keuper. DCBM: Data-efficient visual concept bottleneck models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49752–49782. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/prasse25a.html.

Max Ruiz Luyten and Mihaela van der Schaar. A theoretical design of concept sets: improving the predictability of concept bottleneck models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100160–100195. Curran Associates, Inc., 2024. doi: 10.52202/079017-3178. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b5a412531110b92961fa13c90938806a-Paper-Conference.pdf.

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.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Jm2aK3sDJD.

Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision*, pp. 123–138. Springer, 2024.

Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision*, pp. 709–726. Springer, 2022.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

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

Yan Xie, Zequn Zeng, Hao Zhang, Yucheng Ding, Yi Wang, Zhengjue Wang, Bo Chen, and Hongwei Liu. Discovering fine-grained visual-concept relations by disentangled optimal transport concept bottleneck models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30199–30209. IEEE, 2025.

Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S Yao, Chris Callison-Burch, James C Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. *Advances in neural information processing systems*, 37:90683–90713, 2024.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

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

Jianyang Zhang, Qianli Luo, Guowu Yang, Wenjing Yang, Weide Liu, Guosheng Lin, and Fengmao Lv. Attribute-formed class-specific concept space: Endowing language bottleneck model with better interpretability and scalability. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30291–30300. IEEE, 2025.

Rui Zhang, Xingbo Du, Junchi Yan, and Shihua Zhang. Decoupling concept bottleneck model, 2023. URL https://openreview.net/forum?id=vVbUB9oWUup.

Zhe Zhao, Zhiheng Gong, Pengkun Wang, HaiBin Wen, Cankun Guo, Bo Xue, Xi Lin, Zhenkun Wang, Qingfu Zhang, and Yang Wang. TS-MOF: Two-stage multi-objective fine-tuning for long-tailed recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=S82Afyfbj3.

Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19550–19559. IEEE, 2023.

Xiaoling Zhou, Ou Wu, and Nan Yang. Class and attribute-aware logit adjustment for generalized long-tail learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22991–22999, 2025.

A APPENDIX

B Extended Related Work **15**

C Experimental Setup **16**

C.1 Datasets and Semantic Evidence 16

C.2 Baseline Methods 16

C.3 Evaluation Protocol 17

C.4 Implementation Details 18

D Coverage-Calibrated Semantic Sharing: Derivation and Diagnostics **18**

D.1 How CCSS Updates the Shared Semantic Structure 19

E Extended Results and Discussion **20**

E.1 Detailed Characterization of Semantic Coverage Imbalance 20

E.2 Full Results for Controlled Coverage Interventions 21

E.2.1 Setup and Target-Pair Selection 21

E.2.2 Additional Results and Discussion 21

E.3 Semantic Coverage under Expert and Soft Clinical Supervision 22

E.4 Comparison with Existing Methods 23

E.5 Spurious-Correlation and Shortcut Robustness on WaterBirds 23

E.6 Ablation Study 25

E.6.1 What Component of Coverage-Aware Learning Matters? 25

E.6.2 Why CCSS Instead of a Simpler α Rule? 25

E.6.3 Shared Evidence for Low Coverage pairs 25

E.7 Sensitivity to the CCSS Weight 26

B Extended Related Work

Concept-based visual representation learning. Concept Bottleneck Models (CBMs) make predictions through human-interpretable concepts (Koh et al., 2020). Recent work improves concept discovery, grounding, and scalability using vision–language models and richer visual representations (Oikarinen et al., 2023; Yuksekgonul et al., 2023; Srivastava et al., 2024; Tan et al., 2024; Liu et al., 2025; Prasse et al., 2025). Fine-grained methods model local or class-specific concept structure: DOT-CBM aligns image patches with concepts (Xie et al., 2025), ALBM categorizes concepts in class-specific attribute spaces (Zhang et al., 2025), SALF-CBM spatially grounds concept evidence (Benou & Raviv, 2025), and IVQ-CBM studies many-to-many visual–concept alignment and representation collapse (Gao et al., 2026). Torralba & Efros (2011) show that visual semantics are far from uniform: modern datasets remain distinguishable through transferable semantic signals (Zeng et al., 2024; Liu & He, 2025), and ConceptScope shows uneven class-wise distributions of target, context, and bias concepts (Choi et al., 2025). These works treat concepts as useful semantic signals for representation and dataset analysis, but do not show how the *amount of training evidence* supporting each class–concept relation affects the learned representation.

Learning under class and attribute imbalance. Long-tailed learning addresses unequal class frequencies through resampling, reweighting, margin or logit adjustment, decoupled training, contrastive learning, and model adaptation (Cui et al., 2019; Cao et al., 2019; Kang et al., 2020; Menon et al., 2021; Park et al., 2022; Luo et al., 2025). Generalized long-tail learning further shows that attributes can be imbalanced *within* a class (Tang et al., 2022; Zhou et al., 2025), and Zhong et al. (2023) show that imbalance can distort feature geometry and reduce minority-class discrimination. Group-robust methods instead target poorly performing subpopulations through worst-group optimization, error-based reweighting, or classifier retraining (Sagawa et al., 2020; Nam et al., 2020; Liu et al., 2021; Kirichenko et al., 2023; Jeong et al., 2025). These approaches work primarily over classes, attributes, samples, or predefined groups, rather than the empirical coverage of individual class–concept pairs. In contrast, SemCovNet uses the effective training support of each class–concept pair, together with semantic uncertainty, to determine how strongly its representation shares structure across classes and concepts.

From semantic imbalance to coverage-aware learning. Concept-level failures also appear in learned models: generative models can systematically suppress or exaggerate concepts relative to natural data (Bohacek et al., 2026), and concept-based classifiers can suffer from incomplete supervision, noisy concepts, or collapsed representations (Penaloza et al., 2025; Park et al., 2025; Hu et al., 2025; Gao et al., 2026). We study a more inherent source of failure: uneven empirical coverage across class–concept pairs in the training data. Under SCI, class and concept marginals may appear adequate while particular class–concept pairs have limited effective support, yielding noisier semantic representations. This motivates *coverage-aware semantic representation learning*: low coverage or uncertain pairs should share more semantic structure, while high coverage pairs should retain more pair-specific structure. SemCovNet instantiates this principle through CCSS and combines the resulting concept representations with a residual visual pathway.

Limits of concept supervision. Beyond CBMs, recent work studies whether a concept vocabulary is complete or well-specified. Ruiz Luyten & van der Schaar (2024) formalize what makes a concept set good, its expressiveness and its inductive bias for the downstream model, and show that a well-chosen set improves sample efficiency and out-of-distribution robustness, motivating a method for selecting informative, non-redundant concepts. Zhang et al. (2023) show that an insufficient concept set forces a bottleneck to blend explicit and implicit information, distorting both the learned concepts and the final label; their Decoupling CBM separates this mixture into explicit and implicit branches to recover accuracy without abandoning interpretability. SCI studies a different question, also distinct from prior dataset-bias work (Torralba & Efros, 2011; Zeng et al., 2024; Liu & He, 2025; Bohacek et al., 2026; Choi et al., 2025): even with a complete, well-specified vocabulary, a class–concept pair can still receive abundant evidence in one class and almost none in another, and we study whether this unequal distribution affects the representation learned for that pair.

Subgroup robustness and semantic coverage. Group-robust methods correct for imbalance at a specific unit of analysis, and each implicitly assumes that unit has enough evidence to be estimated reliably. GroupDRO (Sagawa et al., 2020) minimizes a minimax objective over predefined groups, so it needs the group label in advance and enough examples per group to estimate each group's

loss reliably. JTT (Liu et al., 2021) avoids predefined groups by first training a standard model and upweighting the examples it misclassifies, using early mistakes as a proxy for minority-group membership; this proxy is only informative if enough misclassified examples exist to upweight. DFR (Kirichenko et al., 2023) instead freezes the features from standard training and retrains only the last layer on a small group-balanced sample, which shows the features already carry the signal but still requires a held-out set large enough to balance every group. In each case, imbalance is corrected at the level of a group, a misclassified example, or a balanced held-out set but not at the level of class–concept relation. A class–concept pair with very little support can therefore fall inside a group, or a balanced sample, that looks adequate on average, while the pair itself remains of minimum support and hidden from these objectives.

C Experimental Setup

C.1 Datasets and Semantic Evidence

We evaluate across five visual datasets spanning natural images, fine-grained recognition, spurious correlations, and medical imaging. The datasets provide different forms of semantic evidence, allowing us to test whether SCI extends beyond a single annotation type. Table 3 reports dataset-level coverage statistics. In this study, we compute coverage $C_{y,k}$ only from the training split.

CelebA (Liu et al., 2015) provides binary facial attributes and is used as a general visual-recognition benchmark. CelebA 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. We use the 39 available attributes as atomic semantic concepts. Therefore, $a_{i,k} \in 0, 1$ indicates whether concept k is present in image $\boldsymbol{x}_i$. This setting evaluates SCI under dense binary attribute supervision.

CUB (Wah et al., 2011) contains 200 bird species with 312 annotated visual attributes. We treat the 312 attributes as atomic concepts and use the 28 attribute families only as semantic metadata. All coverage computation, representation learning, and pair-level analysis operate on atomic class–concept pairs. CUB is particularly useful for studying sparse and zero-coverage relations because many species–attribute combinations obtain small or no direct training evidence.

WaterBirds (Sagawa et al., 2020) is included as a complementary spurious-correlation benchmark. We use its standard class and group annotations to evaluate robustness under background correlation. Its role differs from the primary SCI characterization: it tests whether coverage-aware semantic learning remains competitive in a setting commonly studied through group robustness.

Derm7pt (Kawahara et al., 2019) associates skin-lesion diagnoses with gold-standard expert-defined clinical descriptors for the 7-point melanoma checklist. We expand categorical clinical criteria into atomic semantic concepts, yielding 28 concepts organized into seven clinical descriptors. The descriptors are used only for organization; class–concept pairs remain the primary unit of coverage and representation analysis. Derm7pt therefore evaluates SCI with expert-defined clinical semantic supervision.

MILK10k (Philipp et al., 2025) provides continuous semantic evidence rather than binary concept labels. MILK10k officially provides seven MONET-driven descriptor soft scores and uses them as soft supervision during training. This setting tests whether coverage-aware learning extends to uncertain and continuous semantic supervision.

C.2 Baseline Methods

We compare SemCovNet with established and recent methods from foundation-model representations, imbalance-aware learning, group robustness, and concept-based representation learning. We use each baseline to test a different explanation for weakly under-covered semantics.

Representation and last-layer baselines. DINOv3–ERM uses DINOv3 (Siméoni et al., 2025) features with a linear prediction head and provides the main no-coverage-awareness control. DFR (Kirichenko et al., 2023) retrains the final classifier on fixed features and tests whether a strong representation already contains the information needed for robust prediction. These baselines separate representation quality from specialized semantic learning.

Reweighting and robust learning. JTT (Liu et al., 2021) emphasizes examples that are misclassified by an initial model and tests whether error-driven reweighting is sufficient. GroupDRO (Sagawa et al., 2020) directly optimizes worst-group performance. We additionally consider recent representation- and distribution-level robustness methods, including SCER (Park et al., 2026), which regularizes embeddings against spurious correlations, and Hierarchical DRO (Jo et al., 2026), which models both inter-group and intra-group distributional uncertainty.

Long-tail learning. MORE (Luo et al., 2025) addresses long-tailed recognition by reallocating model capacity toward underrepresented classes, while TS-MOF (Zhao et al., 2025) performs multi-objective classifier adaptation over frozen representations. These comparisons test whether methods designed for frequency imbalance also address the class-conditioned semantic coverage problem studied here.

Concept-based learning. We compare with recent concept-learning approaches where applicable. IVQ-CBM (Gao et al., 2026) learns high-rank visual concept representations through implicit vector quantization. These methods test whether stronger concept representations, concept robustness, or structured concept learning can resolve SCI.

Controlled baselines and ablations. *Concept Supervision* uses the same semantic supervision without CCSS. *Coverage Reweighting* adjusts the contribution of low coverage semantic relations but does not borrow shared semantic structure. *Semantic GroupDRO* applies group-robust optimization (Sagawa et al., 2020) to semantic groups or pairs, testing whether worst-group optimization alone is sufficient. Finally, we compare *Frozen* and *LoRA* variants while keeping the DINOv3 encoder fixed; the latter applies parameter-efficient adaptation to the final three transformer blocks. Within each controlled comparison, we match the backbone, supervision, classifier, training schedule, and seeds; only the tested mechanism changes.

C.3 Evaluation Protocol

Task performance. We use the standard task metric for each benchmark. We report balanced accuracy (BAcc) on CelebA, Top-1 accuracy on CUB, average accuracy on WaterBirds, and AUROC on Derm7pt and MILK10k. BAcc accounts for class imbalance in CelebA, Top-1 accuracy is the standard metric for CUB's multi-class recognition task, and AUROC provides a threshold-independent measure for the medical classification tasks. Using the task-specific metric for each dataset also allows comparison with prior work.

Tail/Middle/Head Coverage pairs. For semantic analysis, we rank class–concept pairs by their training coverage $C_{y,k}$ and use the 25th and 75th percentiles to define

$$\text{Tail} : Q_{0\text{–}25\%}, \qquad \text{Middle} : Q_{25\text{–}75\%}, \qquad \text{Head} : Q_{75\text{–}100\%}.$$

We compute the Tail/Middle/Head pairs once from the training split and fix them for evaluation. We use them only for reporting and visualization; our learning method does not use them as thresholds. We use the official test set only to evaluate the representation learned for each training-defined class–concept pair.

Semantic representation quality. Class–concept pair-level semantic analyses use the learned concept representation $\boldsymbol{z}^{\text{sem}}$. We report class-conditioned concept Macro-F1 for Tail, Middle, and Head pair representation quality and the Head–Tail gap $\Delta_{\text{H}-\text{T}} = \text{Head} - \text{Tail}$. Frozen linear probes measure how well each concept is represented without updating the encoder. Unless stated otherwise, we aggregate pair-level probe results over class–concept pairs within each training coverage pair. We use visual concept tokens $z_{i,k}$ for semantic representation analyses.

Task and diagnostic representations. We compute task performance from the model's saved predictions or logits rather than from concept probes. When a task-level linear probe is reported, we use the joint semantic–visual representation $\boldsymbol{h}_i$. DINOv3-ERM baseline is evaluated on its global feature for task performance and its class token `CLS` representation. This separation keeps task performance distinct from how well individual semantic representations perform.

95% CIs. We report mean performance across training seeds and use 95% bootstrap confidence intervals over class–concept pairs for pair-level analyses. For controlled pair interventions, the class–concept pair is the experimental unit: we first aggregate seed results within each pair, then compute confidence intervals across pairs.

Table 2: **Implementation details.** Methods use the same backbone, data splits, optimization settings, and random seeds.

Setting	Configuration
Visual encoder	DINOv3 ViT-B/16
Input resolution	224×224
Encoder adaptation	Frozen / LoRA on final 3 transformer blocks
LoRA rank r	8
LoRA scaling α	16
LoRA dropout	0.05
Optimizer	AdamW
Learning rate	3e−4
Weight decay	0.05
Batch size	64
Training epochs	20
Concept-loss weight	1.0
CCSS-loss weight	0.25
Coverage source	Training split only
Semantic unit	class–concept pair (y, k)

C.4 Implementation Details

We use DINOv3 ViT-B/16 as the primary visual encoder. In the frozen setting, we keep the pretrained encoder fixed, while in the parameter-efficient setting, we apply LoRA to the final three transformer blocks. The semantic branch produces concept-specific representations $z_{i,k}$, and the model retains a residual visual representation so that prediction is not restricted to the predefined concept vocabulary.

We instantiate ℓ_{sem} (Equation 12) based on the specific dataset's concept annotations: binary cross-entropy for binary concepts, categorical supervision for grouped categorical attributes where applicable, and a soft-target loss for continuous semantic evidence.

For controlled comparisons, we use identical data splits, intervention manifests, initialization seeds, and evaluation protocol across methods. Concept and Concept+CCSS variants also share the same backbone, semantic supervision, classifier, and optimization schedule, so their difference isolates CCSS.

All Tail/Middle/Head coverage pairs are computed from the training split only. Dataset validation and test splits are never used to estimate coverage or select intervention pairs. We select hyperparameters using validation data and use the test split only for final evaluation. Table 2 presents the optimizer, learning rates, batch sizes, number of epochs, LoRA configuration, loss weights, input resolution, and dataset-specific preprocessing settings.

D Coverage-Calibrated Semantic Sharing: Derivation and Diagnostics

Hierarchical model and posterior derivation. CCSS follows from a hierarchical Gaussian model for the latent coverage-calibrated representation $\hat{\boldsymbol{t}}_{y,k}$ of each class–concept pair:

$$\hat{\boldsymbol{t}}_{y,k} \sim \mathcal{N}(\boldsymbol{m}_{y,k},\, \tau^2 \boldsymbol{I}), \qquad \bar{\boldsymbol{z}}_{y,k} \mid \hat{\boldsymbol{t}}_{y,k} \sim \mathcal{N}\left(\hat{\boldsymbol{t}}_{y,k},\, \frac{\sigma_k^2}{n_{y,k}^{\text{eff}}} \boldsymbol{I}\right), \tag{15}$$

where $\boldsymbol{m}_{y,k}$ is the shared prior mean (Equation 6); τ^2 is how far a true pair representation can drift from that prior; $\sigma_k^2 / n_{y,k}^{\text{eff}}$ is the sampling variance of $\bar{\boldsymbol{z}}_{y,k}$, built from $n_{y,k}^{\text{eff}}$ effective observations. Under this Gaussian prior and likelihood, the posterior mean of $\hat{\boldsymbol{t}}_{y,k}$ given $\bar{\boldsymbol{z}}_{y,k}$ is exactly Equation 10, with $\alpha_{y,k}$ in Equation 9 the corresponding shrinkage weight. Equation 10 is therefore the Bayes-optimal estimate of $\hat{\boldsymbol{t}}_{y,k}$ under this model, for any $\tau^2, \sigma_k^2 > 0$, not a heuristic interpolation.

The τ^2 estimator. Averaging $\frac{1}{D}\|\bar{\boldsymbol{z}}_{y,k} - \boldsymbol{m}_{y,k}\|_2^2$ over pairs conflates two sources of spread: true variation of $\hat{\boldsymbol{t}}_{y,k}$ around $\boldsymbol{m}_{y,k}$, and sampling noise in $\bar{\boldsymbol{z}}_{y,k}$ from finite $n_{y,k}^{\text{eff}}$. Under Equation 15, $\mathbb{E}\big[\|\bar{\boldsymbol{z}}_{y,k} - \boldsymbol{m}_{y,k}\|_2^2\big] = D\tau^2 + D\sigma_k^2/n_{y,k}^{\text{eff}}$, so subtracting the second term and averaging over $\mathbb{V}$ gives an unbiased estimate of τ^2, as in Equation 8. Any single pair's corrected term can be negative from noise alone; $[\cdot]_+$ enforces the model's non-negativity constraint on the final estimate.

Zero-support pairs. When $n_{y,k}^{\text{eff}} = 0$, Equation 9 gives $\alpha_{y,k} = 0$ regardless of τ^2, σ_k^2, so Equation 10 reduces to $\hat{\boldsymbol{t}}_{y,k} = \boldsymbol{m}_{y,k}$: a pair with no observed evidence is represented entirely by the shared prior.

We add ϵ to the denominators of Equation 7 and $\bar{\boldsymbol{z}}_{y,k}$, and floor τ^2 at zero via $[\cdot]_+$ in Equation 8, to avoid division by zero and negative variances from finite-sample noise.

Soft semantic evidence. For soft annotations, $a_{i,k} \in [0,1]$ is a confidence score for the presence of concept k in image $\boldsymbol{x}_i$. The class–concept coverage becomes $C_{y,k} = \frac{1}{|\mathcal{D}_y|}\sum_{i:y_i=y} a_{i,k}$, the mean confidence for concept k within class y. The effective support becomes $n_{y,k}^{\text{eff}} = \sum_{i:y_i=y} a_{i,k}$, the total evidence mass for pair (y,k). Each observation contributes in proportion to its confidence score, not as a hard binary count.

D.1 How CCSS Updates the Shared Semantic Structure

The shared semantic structure Equation 6 is

$$\boldsymbol{m}_{y,k} = \boldsymbol{\mu}_0 + \tilde{\boldsymbol{\mu}}_y + \tilde{\boldsymbol{\mu}}_k$$

contains learnable global, class-shared, and concept-shared parameters. These parameters do not receive separate global-, class-, or concept-level supervision. Instead, the instance-level CCSS objective learns them through the class–concept pair associated with each visual concept token.

For image i with class label y_i, the concept token $\boldsymbol{z}_{i,k}$ is aligned with the coverage-calibrated target Equation 10 is

$$\hat{\boldsymbol{t}}_{y_i,k} = (1 - \alpha_{y_i,k})\boldsymbol{m}_{y_i,k} + \alpha_{y_i,k}\bar{\boldsymbol{z}}_{y_i,k}.$$

Thus, the minibatch does not require an additional pair-level label: the pair is determined directly by the observed class y_i and concept index k, while $a_{i,k}$ determines the contribution of that concept to the loss. Specifically, Equation 11 is

$$\mathcal{L}_{\text{CCSS}} = \frac{\sum_{i=1}^{B}\sum_{k=1}^{K} a_{i,k}\left[1 - \cos\big(\boldsymbol{z}_{i,k}, \hat{\boldsymbol{t}}_{y_i,k}\big)\right]}{\sum_{i=1}^{B}\sum_{k=1}^{K} a_{i,k} + \epsilon}.$$

During CCSS optimization, $\bar{\boldsymbol{z}}_{y,k}$ and the statistics used to compute $\alpha_{y,k}$ are treated as stop-gradient quantities, whereas $\boldsymbol{m}_{y,k}$ remains differentiable. Hence, for a class–concept pair (y,k),

$$\boldsymbol{g}_{y,k} := \frac{\partial \mathcal{L}_{\text{CCSS}}}{\partial \boldsymbol{m}_{y,k}} = (1 - \alpha_{y,k})\frac{\partial \mathcal{L}_{\text{CCSS}}}{\partial \hat{\boldsymbol{t}}_{y,k}}. \tag{16}$$

All instances in the minibatch with the same (y,k) contribute to the same pairwise gradient $\boldsymbol{g}_{y,k}$. Because the same parameters are reused across multiple pairs, these pairwise gradients are then aggregated according to the sharing structure:

$$\nabla_{\boldsymbol{\mu}_0}\mathcal{L}_{\text{CCSS}} = \sum_{y,k}\boldsymbol{g}_{y,k}, \tag{17}$$

while $\boldsymbol{\mu}_y$ receives the aggregate signal from pairs sharing class y, and $\boldsymbol{\mu}_k$ from pairs sharing concept k. With the centered effects $\tilde{\boldsymbol{\mu}}_y = \boldsymbol{\mu}_y - \frac{1}{C}\sum_c \boldsymbol{\mu}_c$ and $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k - \frac{1}{K}\sum_j \boldsymbol{\mu}_j$, their gradients are

$$\nabla_{\boldsymbol{\mu}_y}\mathcal{L}_{\text{CCSS}} = \sum_k \boldsymbol{g}_{y,k} - \frac{1}{C}\sum_{y'}\sum_k \boldsymbol{g}_{y',k}, \tag{18}$$

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{L}_{\text{CCSS}} = \sum_y \boldsymbol{g}_{y,k} - \frac{1}{K}\sum_{k'}\sum_y \boldsymbol{g}_{y,k'}. \tag{19}$$

Table 3: **Dataset-level characterization of Semantic Coverage Imbalance.** Coverage $C_{y,k}$ is computed per class–concept pair on the training split only. Near-zero and Zero denote the fraction of pairs with $C_{y,k} < 0.05$ and $C_{y,k} = 0$, respectively.

Dataset	#train	Class	Concept	Pairs	Q_{25}	Median	Q_{75}	Near-zero (%)	Zero (%)
CelebA	162770	2	39	78	0.048	0.143	0.338	26.9	0.0
CUB	5394	200	312	62,400	0.000	0.000	0.111	67.2	53.9
Derm7pt	413	2	28	56	0.031	0.117	0.455	33.9	12.5
Waterbirds	4795	2	2	4	0.050	0.500	0.950	25.0	0.0
MILK10k	3668	11	7	77	0.253	0.368	0.433	0.0	0.0

Table 4: **Statistical controls for SCI.** Significant heterogeneity is the share of concepts whose across-class coverage variation exceeds a prevalence-preserving permutation null after BH–FDR correction ($q < 0.05$). ρ_{bal} is the Spearman correlation between full-data and class-balanced coverage estimates; "—" denotes not applicable.

Dataset	Significant heterogeneity (%)	ρ_{bal}
CelebA	97.4	1.000
CUB	96.5	0.996
Derm7pt	64.3	0.993
Waterbirds	100.0	0.800
MILK10k	85.7	—

Therefore, $\boldsymbol{\mu}_0$ aggregates information across all observed class–concept pairs, $\tilde{\boldsymbol{\mu}}_y$ captures deviations shared across concepts of class y, and $\tilde{\boldsymbol{\mu}}_k$ captures deviations shared across classes for concept k. Importantly, this structure is induced by parameter sharing rather than by additional supervision. Moreover, the factor $(1 - \alpha_{y,k})$ makes sharing adaptive: low coverage (low-support) or uncertain pairs are pulled more strongly toward shared semantic structure, whereas high coverage (well-supported) pairs place greater weight on their empirical pair-specific representation $\bar{\boldsymbol{z}}_{y,k}$.

E Extended Results and Discussion

E.1 Detailed Characterization of Semantic Coverage Imbalance

Tables 3 and 4 report the full training-set coverage statistics underlying Fig. 4, including two datasets: Waterbirds and MILK10k. CelebA, CUB, and Derm7pt exhibit substantial variation in class-conditioned concept coverage despite large differences in dataset size, number of classes, and semantic vocabulary. The permutation analysis further shows that the observed across-class heterogeneity is substantially greater than expected from global concept prevalence and finite sampling alone. Class-balanced resampling preserves the original coverage ordering almost perfectly, confirming that the observed semantic structure is not an artifact of target-class counts. Waterbirds and MILK10k also follow the same pattern: substantial across-class coverage heterogeneity (100.0% and 85.7% of concepts significant) that class-balanced resampling does not explain ($\rho_{\text{bal}} = 0.800$ for Waterbirds; not computed for MILK10k due to its small sample size).

CUB exhibits particularly sparse coverage because its 312 fine-grained attributes form highly structured class–concept compositions; some zero-coverage pairs may therefore represent genuine semantic incompatibility rather than harmful low coverage. We retain all pairs when characterizing the dataset-level coverage structure, but apply predefined support criteria in subsequent representation and performance analyses. Thus, Section 4.1 establishes the existence and structure of SCI without assuming that every naturally rare or absent class–concept pair constitutes a learning failure.

Table 5: **Full coverage-response intervention on CUB.** We report frozen-probe representation quality q_{repr} for each method as target-pair training evidence decreases from 100% to zero. "One" retains a single positive example and therefore has no nominal retention fraction. Only the target pair's own evidence is intervened on; the corresponding class and concept remain observable through other training pairs. We report concept Macro-F1 as the mean over the target pairs with 95% CIs. We average 5 seeds within each pair before bootstrapping. The final column reports the $\Delta_{100\%\to 0\%}$, difference between 100% and 0% change. Smaller degradation indicates greater robustness to loss of pair-specific evidence; positive values indicate degradation, while values near zero indicate stability.

Method	100%	75%	50%	25%	One	0%	$\Delta_{100\%\to 0\%}$
Concept supervised	0.675 [0.630, 0.717]	0.666 [0.619, 0.708]	0.661 [0.617, 0.703]	0.665 [0.619, 0.709]	0.661 [0.617, 0.704]	0.665 [0.619, 0.708]	0.010
Coverage reweighting	0.650 [0.608, 0.691]	0.657 [0.613, 0.698]	0.659 [0.618, 0.699]	0.661 [0.617, 0.702]	0.661 [0.620, 0.701]	0.661 [0.623, 0.697]	-0.012
Semantic GroupDRO	0.655 [0.614, 0.696]	0.658 [0.617, 0.698]	0.649 [0.605, 0.689]	0.659 [0.618, 0.698]	0.665 [0.621, 0.707]	0.660 [0.617, 0.698]	-0.005
LoRA Concept	0.664 [0.623, 0.704]	0.658 [0.618, 0.697]	0.665 [0.628, 0.702]	0.669 [0.628, 0.709]	0.661 [0.620, 0.703]	0.670 [0.629, 0.710]	-0.006
SEMCOVNET (Ours)	**0.679 [0.631, 0.724]**	**0.672 [0.623, 0.717]**	**0.677 [0.629, 0.722]**	**0.676 [0.627, 0.721]**	**0.678 [0.631, 0.721]**	**0.682 [0.634, 0.726]**	-0.003

Table 6: **Representation-sharing ablations under extreme semantic under-coverage.** We evaluate structural variants in the one-example and zero-evidence conditions, where borrowing shared semantic structure matters most. We designed these variants as low-coverage mechanism ablations and therefore did not evaluate them over the full retention trajectory. We report results as means over the target class–concept pairs with 95% CIs. We average 5 seeds within each pair before bootstrapping.

Method	One example	Zero evidence
No class sharing (μ_k only)	0.678 [0.627,0.724]	0.678 [0.628,0.722]
Shared prior only	0.667 [0.625,0.708]	0.667 [0.624,0.709]
Fixed specialization	0.677 [0.627,0.723]	**0.684 [0.635,0.728]**
CCSS (Ours)	**0.678 [0.631,0.721]**	0.682 [0.634,0.726]

E.2 FULL RESULTS FOR CONTROLLED COVERAGE INTERVENTIONS

E.2.1 SETUP AND TARGET-PAIR SELECTION

We conduct the controlled intervention on CUB, where the same semantic concept can occur across multiple bird classes. We first enumerate all class–concept pairs (y, k) using the training split only. A pair is eligible if it has sufficient positive training support for controlled retention, sufficient test support for stable evaluation, and the corresponding concept k remains observed in other classes. We also require class y to remain represented by other semantic concepts. These constraints confirm that the intervention removes evidence for the *target combination*, rather than removing the class or concept itself. Applying these criteria yields 24 target class–concept pairs, which are fixed before any intervention or model comparison in this experiment.

For each target pair, we retain 100%, 75%, 50%, 25%, one positive example, or 0% of its original training evidence. All non-target samples and the validation/test sets remain unchanged. Therefore, we manipulate only $C_{y,k}$, while shared evidence about class y and concept k remains available elsewhere in the training data. We evaluate all methods on the same pairs and intervention manifests.

We measure representation quality using the same frozen-probe protocol as in Section 4.2. After training each encoder, we freeze its representation and evaluate a linear probe on the unchanged test examples of the target pair; class-conditional concept Macro-F1 is the primary metric. We first average results across seeds within each pair, then obtain confidence intervals by bootstrapping the 24 target pairs (2,000 resamples). The class–concept pair is therefore the experimental unit.

E.2.2 ADDITIONAL RESULTS AND DISCUSSION

Table 5 reports the complete Section 4.4 coverage-response results. Each intervention changes only the training evidence associated with a predefined target class–concept pair. The corresponding class and concept remain represented elsewhere in the training set, isolating the ability to compose previously learned class and concept information from the simpler problem of recognizing an entirely

Table 7: **Expert clinical semantic supervision on Derm7pt.** Task metrics evaluate diagnosis, while concept metrics evaluate the learned semantic representation. Tail AUROC and the $\Delta_{\mathrm{H-T}}$ Head–Tail gap use coverage strata defined from the training split; a smaller positive gap and smaller coverage slope β_1 indicate less dependence of representation quality on semantic coverage. Results are reported as mean ± standard deviation over five seeds.

Method	Task			Concept			Coverage Slope $\beta_1 \downarrow$
	AUROC ↑	M-F1 ↑	BAcc ↑	F1 ↑	Tail AUROC ↑	$\Delta_{\mathrm{H-T}} \downarrow$	
Concept supervised	0.848±0.007	0.738±0.007	0.717±0.011	**0.485±0.016**	0.647±0.049	0.085±0.049	0.115±0.030
LoRA Concept	0.858±0.005	0.751±0.012	0.729±0.014	0.419±0.196	0.615±0.021	0.110±0.021	0.071±0.020
Semantic only	**0.867±0.006**	0.761±0.008	0.745±0.005	0.404±0.137	0.683±0.055	0.038±0.053	0.101±0.027
Semantic + residual	0.855±0.008	0.762±0.007	0.743±0.008	0.490±0.010	0.655±0.057	0.063±0.046	0.092±0.024
Frozen SEMCOVNET	0.852±0.007	0.756±0.008	0.736±0.007	0.365±0.130	0.673±0.061	0.054±0.062	0.093±0.038
SEMCOVNET (Ours)	0.856±0.007	**0.767±0.011**	**0.759±0.018**	0.447±0.023	**0.686±0.020**	**0.025±0.015**	**0.068±0.010**

unseen class or concept. We use the same target pairs for every method and retention level. We first average seeds within each pair, then estimate uncertainty by paired bootstrap over pairs, making the class–concept pair the experimental unit.

SEMCOVNET is stable across the full intervention range: representation quality 0.679 at 100% retention and 0.682 at zero retention. Concept supervision also transfers substantial information to the unseen combination, explaining why its zero-coverage difference from SEMCOVNET is positive but not statistically significant ($\Delta_0 = +0.017$, 95% CI $[-0.001, +0.036]$). In contrast, SEMCOVNET significantly exceeds coverage reweighting and Semantic GroupDRO when there is zero evidence. This distinction demonstrates that both methods can emphasize poorly represented observed pairs, but neither creates new evidence when the exact pair is absent.

Table 6 results show that removing class sharing (μ_k only) changes the zero-coverage result by only $+0.004$ in favor of SemCovNet, with a confidence interval spanning zero. The coverage-controlled intervention alone does not statistically isolate the class-sharing representation's contribution at zero coverage. Similarly, several specialization variants remain competitive in this extreme condition. We therefore use this analysis to establish *compositional recovery under missing pair evidence*, while separate component ablations assess which parts of the coverage-calibrated parameterization drive the overall gains.

E.3 SEMANTIC COVERAGE UNDER EXPERT AND SOFT CLINICAL SUPERVISION

This section provides the complete results underlying Figure 7 in Section 4.5. We examine whether the coverage-aware semantic representation learning principle remains applicable across heterogeneous clinical semantic supervision. Derm7pt provides expert-defined categorical clinical concepts, whereas MILK10k provides continuous soft semantic evidence. We experiment with both datasets under identical settings for our proposed method; only the form of semantic supervision differs.

Derm7pt. Figure 7(a) reports frozen-probe class-conditioned concept Macro-F1 across Tail/Middle/Head coverage pairs. Table 7 provides the corresponding task and concept representation quality metrics across seeds. SEMCOVNET improves Tail representation quality and substantially reduces the Head–Tail gap compared to standard concept supervision. Middle and Head performance are slightly lower, indicating that the gain is concentrated on low-coverage clinical semantics rather than reflecting a uniform improvement across all pairs. The task-level results also improve in Macro-F1 and Balanced Accuracy, although SEMCOVNET remains competitive with standard concept supervision on every concept metric.

Derm7pt further supports the association between coverage and representation quality through the frozen-probe analysis. However, its smaller sample size and diverse clinical appearances make the geometric dispersion trend less consistent across coverage groups, as shown in Figure 5(c). We therefore interpret the Derm7pt result primarily through the coverage-stratified representation metrics rather than geometric compactness.

Table 8: **Soft semantic supervision on MILK10k.** We perturb the continuous training evidence to make the semantic target less informative. Coverage strata are fixed from the unperturbed training evidence. Tail Concept F1 measures representation quality for under-covered semantics, Task AUROC measures clinical discrimination, and Soft Spearman measures agreement with the continuous semantic teacher.

η	Method	Task AUROC ↑	Tail Concept F1 ↑	Soft Spearman ↑
0.0	Concept supervised	0.877±0.017	0.785±0.004	0.655±0.114
	SEMCOVNET	0.887±0.018	0.781±0.021	0.637±0.024
0.1	Concept supervised	0.879±0.022	0.789±0.001	0.643±0.115
	SEMCOVNET	0.887±0.019	0.786±0.027	0.624±0.022
0.2	Concept supervised	0.879±0.021	0.790±0.006	0.626±0.111
	SEMCOVNET	0.888±0.019	0.786±0.025	0.604±0.016
0.3	Concept supervised	0.879±0.020	0.787±0.006	0.608±0.104
	SEMCOVNET	0.888±0.018	0.788±0.012	0.586±0.014
0.4	Concept supervised	0.880±0.021	0.783±0.008	0.588±0.098
	SEMCOVNET	0.888±0.019	0.787±0.013	0.572±0.020

MILK10k. For MILK10k, we retain the original continuous semantic scores and perturb only the training evidence as

$$a' = (1 - \eta)a + 0.5\eta.$$

Increasing η makes the semantic evidence progressively less informative while preserving the continuous supervision interface. Tail/Middel/Head coverage pairs are defined once from the unperturbed training evidence and kept fixed across perturbation levels.

Table 8 shows that SEMCOVNET preserves both Tail semantic quality and task discrimination across this perturbation range. Soft-target reconstruction itself remains more competitive than concept supervision; this is expected because the objective is to learn coverage-aware visual representations rather than to reproduce the external soft concept score extractor as closely as possible. Accordingly, this experiment is not intended to establish uniform robustness to semantic corruption. Instead, it shows that coverage-aware semantic representation learning remains applicable when semantic evidence is continuous and uncertain.

E.4 COMPARISON WITH EXISTING METHODS

We compare SEMCOVNET with strong representation-, imbalance-, and concept-learning baselines. We use the standard task metric for each because the datasets define different prediction tasks and evaluation protocols.

Table 9 shows that coverage-aware learning remains competitive at the task level while targeting a different failure mode: under-represented class–concept semantics. SEMCOVNET achieves the strongest performance on CelebA and MILK10k and is competitive on CUB, indicating that addressing SCI need not sacrifice overall recognition performance. It does not outperform on Derm7pt. Derm7pt is small (413 training samples only) and clinically heterogeneous, making pair-specific specialization harder to learn and increasing variance across runs. In this high-stakes medical setting, however, aggregate task performance alone can hide weak representations of rare clinical concepts; a model can perform well overall while still learning weak representations for low-coverage class–concept pairs. This motivates our separate Tail and Head analyses, particularly in clinical settings where rare semantic evidence can be important despite limited effect on aggregate AUROC.

E.5 SPURIOUS-CORRELATION AND SHORTCUT ROBUSTNESS ON WATERBIRDS

WaterBirds provides a complementary test of whether coverage-aware semantic representation learning stays competitive under canonical spurious-correlation evaluation. WaterBirds defines four fixed groups from the target label (landbird/waterbird) and background (land/water). These groups, considered as pairs, are highly uneven in training coverage (Tail $= 0.0502$, Middle $= 0.50$, Head $= 0.9498$): two pairs, waterbirds on land and landbirds on water, have coverage below 0.05, although none has zero coverage. We therefore report standard group-robustness metrics.

Table 9: **Task-level comparison with existing methods.** We report the standard task metric for each dataset: CelebA BAcc, CUB Top-1 accuracy, and AUROC for Derm7pt and MILK10k. Higher is better.

Method	CelebA	CUB	Derm7pt	MILK10k
JTT	0.887±0.004	0.890±0.004	0.852±0.008	0.879±0.019
DFR	0.933±0.001	0.837±0.002	**0.866±0.011**	0.859±0.017
MORE	0.925±0.005	**0.892±0.001**	0.854±0.009	0.879±0.013
TS-MOF	0.907±0.013	0.885±0.001	0.858±0.003	0.878±0.014
IVQ-CBM	0.932±0.002	0.881±0.008	0.870±0.005	0.861±0.015
SEMCOVNET (Ours)	**0.955±0.001**	0.891±0.002	0.851±0.018	**0.887±0.009**

Table 10: **Group robustness on WaterBirds.** WGA is the minimum accuracy across the four target×background groups, and Avg–WGA measures the gap between average and worst-group accuracy.

Method	**Avg. Acc. ↑**	**WGA ↑**	**Avg–WGA ↓**
Concept supervised	0.920±0.016	0.816±0.026	0.103±0.017
JTT	0.917±0.010	0.822±0.020	0.095±0.015
MORE	0.930±0.003	0.819±0.036	0.112±0.039
IVQ-CBM	0.910±0.025	0.776±0.021	0.134±0.026
GroupDRO	0.939±0.003	0.882±0.013	0.057±0.013
Hierarchical-DRO	0.944±0.005	0.891±0.011	0.052±0.010
SCER	**0.950±0.011**	**0.923±0.004**	**0.027±0.011**
SEMCOVNET (Ours)	0.926±0.008	0.830±0.034	0.095±0.034

Group robustness. Table 10 reports average accuracy (Avg. Acc.), worst-group accuracy (WGA), and the Average–WGA gap. Dedicated group-robustness methods perform best on these metrics. SCER achieves the highest WGA (0.923) and the smallest Average–WGA gap (0.027), followed by Hierarchical-DRO and GroupDRO. This behavior is expected: these methods are designed to improve performance over the same fixed target×background groups used by the WaterBirds evaluation.

SEMCOVNET achieves 0.926 average accuracy and 0.830 WGA. It improves WGA over concept supervision (0.816) while reducing the Average–WGA gap from 0.103 to 0.095. However, it does not match methods that explicitly optimize the known spurious groups. We view this as complementary rather than contradictory to our main results: SEMCOVNET is designed for uneven class–concept coverage, not for direct worst-group optimization along a predefined nuisance axis.

Per-group behavior. The same pattern shows in the individual groups (Table 11). Concept supervised methods remain strong on aligned bird–background combinations but degrade on the conflicting combinations. SEMCOVNET improves robustness relative to its matched non-group-aware controls, particularly without explicitly optimizing the WaterBirds group labels. In contrast, SCER, GroupDRO, and Hierarchical-DRO substantially improve the conflicting groups because their objectives directly target group-level robustness.

Relation to SCI. WaterBirds and SCI probe different forms of training imbalance. WGA evaluates minimum performance along one predefined group axis (bird class×background) while the Average–WGA gap provides the disparity induced by that same partition. SCI instead analyzes how semantic coverage varies across the full set of class–concept relations. Therefore, a method can perform strongly on WaterBirds by explicitly correcting the known background groups without addressing uneven coverage across a broader semantic vocabulary.

This distinction also explains why group-robustness methods perform better in this experiment. Their training objective and the WaterBirds evaluation use closely aligned group structure, whereas SEMCOVNET does not optimize WGA directly. Nevertheless, SEMCOVNET remains competitive in average accuracy and improves worst-group performance over concept supervised settings. These results suggest that coverage-aware semantic representation learning is compatible with canonical

Table 11: **Per-group accuracy on WaterBirds.** Groups correspond to the four combinations of bird class and background. Dedicated group-robustness methods particularly improve the conflicting landbird-on-water and waterbird-on-land groups.

Method	Land/Land	Land/Water	Water/Land	Water/Water
Concept supervised	0.997±0.001	0.858±0.036	0.816±0.026	0.968±0.005
JTT	0.996±0.001	0.861±0.024	0.824±0.026	0.973±0.007
MORE	0.997±0.002	0.882±0.021	0.823±0.044	0.975±0.009
IVQ-CBM	**0.998±0.001**	0.871±0.022	0.783±0.020	0.956±0.003
GroupDRO	0.994±0.002	0.894±0.010	0.883±0.014	0.955±0.005
Hierarchical-DRO	0.994±0.001	0.905±0.012	0.891±0.011	0.956±0.004
SCER	0.977±0.016	**0.944±0.010**	**0.926±0.003**	0.929±0.014
SEMCOVNET (Ours)	0.996±0.001	0.868±0.023	0.830±0.034	**0.975±0.006**

shortcut robustness while addressing a different, more fine-grained source of representation weakness.

E.6 ABLATION STUDY

E.6.1 WHAT COMPONENT OF COVERAGE-AWARE LEARNING MATTERS?

We first decompose SEMCOVNET into concept supervision, shared semantic structure, pair-specific specialization, coverage calibration, and DINOv3 adaptation (Table 12). Each row changes one component compared to the initial configuration.

Concept supervision improves Tail Concept F1 from 0.528 to 0.550. Adding shared class and concept structure further increases it to 0.577, while Mid and Head performance remains nearly unchanged. In contrast, fixed specialization provides no additional Tail gain, showing that simply adding pair-specific capacity is insufficient. Coverage-dependent specialization maintains or improves Tail performance across the calibrated variants. Finally, combining the full statistical CCSS mechanism with DINOv3 adaptation gives the strongest Tail result (0.619 ± 0.054) and the largest reduction in the Head–Tail gap (-0.069 ± 0.055).

Overall, the ablation suggests that the main benefit does not come from pair-specific specialization alone. Low coverage pairs benefit from shared semantic structure, while coverage calibration determines how much pair-specific evidence to use.

E.6.2 WHY CCSS INSTEAD OF A SIMPLER α RULE?

We test whether we can replace the specialization coefficient $\alpha_{y,k}$ with a simpler coverage-based rule (Table 13). We compare raw frequency, coverage alone, effective support, support with variance information, and the full CCSS estimator under the same training setting.

Using effective support already improves Tail Concept F1 over raw frequency (0.625 vs. 0.608), showing that the amount of semantic evidence is more informative than a raw count alone. However, no particular statistic corresponds to the full CCSS rule. Full CCSS achieves the highest mean Tail performance (0.634 ± 0.057), while Mid and Head performance remains stable. It also gives the largest Head–Tail gap reduction (-0.087 ± 0.060).

These results support using a statistically calibrated specialization rule rather than a fixed coverage-based gate. CCSS jointly accounts for the available effective evidence and its reliability when deciding how strongly a class–concept pair should specialize.

E.6.3 SHARED EVIDENCE FOR LOW COVERAGE PAIRS

Table 14 isolates the shared structure used by coverage-aware learning. Tail representation quality improves progressively from concept-level evidence to class-shared structure to combining global, class, and concept evidence sharing ($0.579 \rightarrow 0.598 \rightarrow 0.608$, respectively); Mid and Head performance remains consistent. These results show that low coverage class–concept pairs should not be learned independently from limited paired evidence. They benefit from progressively more gen-

Table 12: **Component decomposition of coverage-aware learning on CUB.** The ladder adds concept supervision, shared semantic structure, pair-specific specialization, coverage calibration, and DINOv3 adaptation.

Variant	Concept	Shared	Pair spec.	Calibration / adaptation	Tail F1↑	$\Delta_{\mathrm{H-T}}\downarrow$
ERM	–	–	–	–	0.528±0.000	0.019±0.000
Concept	✓	–	–	–	0.550±0.039	-0.005±0.041
Shared	✓	✓	–	–	0.577±0.037	-0.032±0.037
Fixed-spec	✓	✓	✓	Fixed	0.571±0.029	-0.020±0.030
CCSS frozen	✓	✓	✓	Statistical / frozen	0.577±0.039	-0.028±0.040
SEMCOVNET (Ours)	✓	✓	✓	Statistical / LoRA-last3	**0.619**±0.054	**-0.069**±0.055

Table 13: **Component ablation of coverage-aware learning on CUB.** Each row introduces one additional component, from concept supervision and shared semantic structure to pair-specific specialization, coverage calibration, and DINOv3 adaptation. We report results as mean ± standard deviation across five seeds.

α **rule**	**Tail F1**↑	$\Delta_{\mathrm{H-T}}\downarrow$
Frequency	0.608±0.057	-0.062±0.058
Coverage	0.617±0.042	-0.072±0.043
Effective support	0.625±0.028	-0.079±0.029
$n+\sigma^2$	0.617±0.042	-0.069±0.044
CCSS (Ours)	**0.634**±0.057	**-0.087**±0.060

Table 14: **Shared-prior decomposition on CUB.** Effect of shared semantic structure on Tail, Middle, and Head representation quality.

Shared prior	Method	Tail F1↑	$\Delta_{\mathrm{H-T}}\downarrow$
Concept	$\tilde{\boldsymbol{\mu}}_k$	0.579±0.008	-0.032±0.002
Class + concept	$\tilde{\boldsymbol{\mu}}_y+\tilde{\boldsymbol{\mu}}_k$	0.598±0.004	-0.047±0.003
Global + class + concept (Ours)	$\boldsymbol{\mu}_0+\tilde{\boldsymbol{\mu}}_y+\tilde{\boldsymbol{\mu}}_k$	**0.608**±0.007	**-0.057**±0.005

Table 15: **Sensitivity to λ_{CCSS} on CUB.** We vary the CCSS weight around the main setting $\lambda^*_{\mathrm{CCSS}}=0.25$. Δ reports the paired change in Tail Macro-F1 relative to $\lambda^*_{\mathrm{CCSS}}$ with 95% CIs.

λ/λ^*	λ_{CCSS}	Tail F1 ↑	H–T Gap ↓	Δ vs. λ^* [95% CI]
0.25×	0.0625	0.580±0.056	-0.033±0.056	-0.006 [-0.020, +0.006]
0.5×	0.1250	0.606±0.056	-0.058±0.055	-0.002 [-0.011, +0.007]
1×	0.2500	0.599±0.071	-0.047±0.072	–
2×	0.5000	0.617±0.068	-0.065±0.069	+0.008 [-0.001, +0.015]
4×	1.0000	0.608±0.057	-0.056±0.056	+0.008 [-0.007, +0.021]

eral semantic structure (concept, class, and global) that provides a more stable sharing prior until sufficient pair-specific evidence supports specialization. This supports structured semantic sharing across class and concept as the prior for coverage-aware specialization.

E.7 SENSITIVITY TO THE CCSS WEIGHT

We investigate sensitivity to the CCSS loss weight λ_{CCSS} on CUB by varying it from $0.25\times$ to $4\times$ the value used in the main experiments ($\lambda^*_{\mathrm{CCSS}}=0.25$). Across this range, Tail concept performance remains stable, with Macro-F1 between 0.580 and 0.617, while Head performance changes little (Table 15). The corresponding mean specialization coefficient α also remains stable.

Although $2 \times \lambda^*_{\mathrm{CCSS}}$ gives the highest numerical Tail score, none of the tested settings differs significantly from the chosen $\lambda^*_{\mathrm{CCSS}}$ in the paired analysis over the 24 evaluation pairs. This shows that the improvements in coverage-aware representation are consistent and do not rely on precise tuning of λ_{CCSS}.