BEYOND CONCEPT GENERATION: INTERPRETABLE CONCEPT SCORING FOR DEBIASING CLIP-BASED CBMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept Bottleneck Models (CBMs) explain predictions through human-readable concepts. However, label-free CBMs built on CLIP image-text similarities may inherit biases from CLIP representations, allowing visually or contextually associated concepts to serve as semantically unreliable shortcuts. Addressing such shortcuts is non-trivial: a concept that provides valid evidence for one class may form a misleading association with another. Therefore, globally removing potentially problematic concepts can be harmful, as it may discard useful information for other classes. To address this challenge, we introduce ScoreCBM, an interpretable concept-scoring framework for debiasing CLIP-based CBMs. Moving large language models (LLMs) beyond their conventional role of concept generation, we use them to explicitly score the semantic relevance of class-concept pairs. The resulting scores provide a class-specific semantic prior that guides two complementary interventions: score-guided activation attenuation suppresses low-relevance concept activations during projection learning, while adaptive ℓ_1 regularization penalizes low-relevance concept-class connections in the final sparse classifier. Our framework requires neither group labels nor modifications to CLIP, while providing a traceable and editable interface for model intervention. To scale to large datasets, we further introduce CLIP-based top-r pre-filtering, reducing LLM scoring complexity from quadratic to linear in the number of classes. Experiments across biased and standard benchmarks demonstrate consistent improvements in predictive performance. On BAR, our method improves LF-CBM accuracy from 84.25% to 89.39%, surpassing the 88.91% non-interpretable ERM baseline. The framework further scales to the 1,000-class ImageNet, improving LF-CBM from 71.87% to 75.28%.

1 INTRODUCTION

Deep neural networks (DNNs) have become the dominant approach for visual recognition, achieving high accuracy in decision-critical domains such as medical image analysis and autonomous driving perception. However, their opaque predictions complicate failure analysis and prevent meaningful human oversight. Concept Bottleneck Models (CBMs) (Koh et al., 2020) address this limitation by introducing an intermediate layer of human-interpretable concepts: the model first predicts concepts from the input and then predicts the class from those concepts. Early CBMs relied on costly human annotations to supervise these concepts (Chen et al., 2020), while recent label-free CBMs reduce or remove this burden by leveraging vision-language models (VLMs) such as CLIP to align concepts with images (Oikarinen et al., 2023; Yuksekgonul et al., 2023). While greatly improving scalability, this reliance on VLMs also exposes CBMs to biases inherited from the underlying VLM representations.

In CLIP-based CBMs, mitigating such shortcuts is particularly challenging because concept relevance is class-dependent: the same concept may provide valid evidence for one class while forming a misleading association with another. As illustrated in Fig. 1, when a flower and a butterfly co-occur in an image, concepts describing both objects would be activated and may spuriously contribute to predictions of the other class. For example, *white petals* may become a cue for predicting *butterfly*. While concept removal may seem like a straightforward solution, naively removing potentially

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.

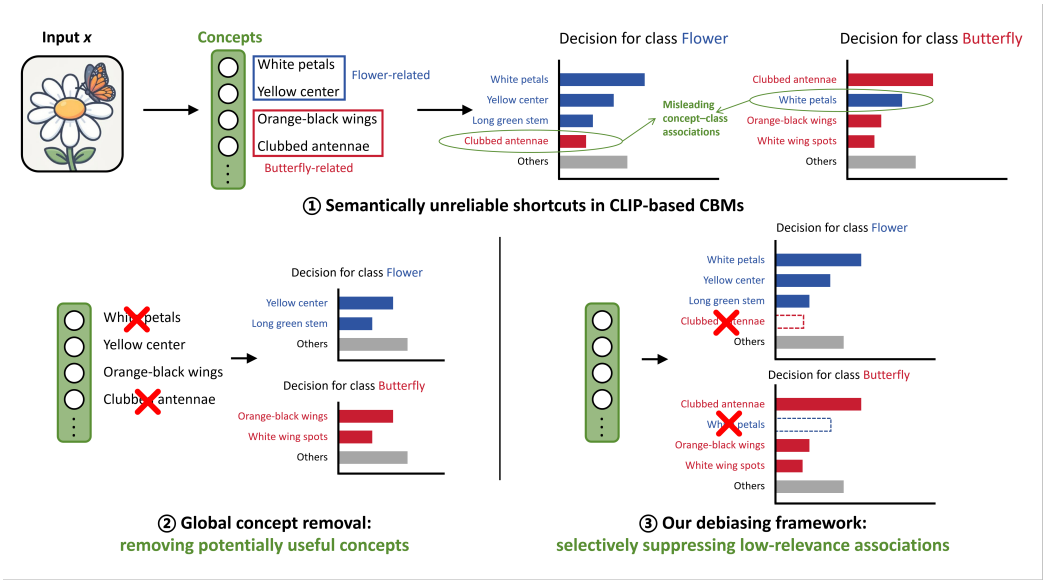


Figure 1: (1) Visually or contextually associated concepts can serve as semantically unreliable shortcuts in CLIP-based CBMs. (2) Globally removing potentially problematic concepts may discard useful information for classes to which they are relevant. (3) ScoreCBM instead uses class-specific relevance to selectively suppress low-relevance associations while preserving the same concepts as valid evidence for other classes.

problematic concepts can also discard useful information for other classes. Removing *white petals*, for instance, would eliminate relevant evidence for predicting *flower*. This suggests that what should be evaluated is not a concept in isolation, but *its relevance to individual classes.*

Prior work has addressed biases in CLIP representations and CBMs from multiple perspectives. Early CLIP debiasing methods rely on human- or LLM-annotated textual descriptions of spurious attributes and mitigate such biases by projecting CLIP representations into a space invariant to these attributes (Chuang et al., 2023; Dehdashtian et al., 2024; Molahasani et al., 2025). Beyond directly debiasing CLIP representations, prior studies have mitigated biases in concept-based models through causal adjustment of confounded concept representations (Bahadori & Heckerman, 2020), concept-level model editing (Yuksekgonul et al., 2023), and the identification and removal of spurious concepts (Kim et al., 2024). However, how to explicitly leverage class-concept semantic relevance for debiasing remains underexplored.

To address this limitation, we propose ScoreCBM, an interpretable concept-scoring framework for debiasing CLIP-based CBMs. While LLMs have been widely used in LF-CBMs for concept generation, we move beyond this conventional role and further leverage their semantic knowledge to assess class-concept relevance. Specifically, for each class-concept pair, we query an LLM to score the semantic relevance of the concept to the corresponding class. As illustrated in Fig. 1, rather than globally removing potentially problematic concepts, our framework selectively suppresses low-relevance associations with individual classes. For instance, the model is discouraged from relying on *white petals* when predicting *butterfly*, while preserving the same concept as relevant evidence for predicting *flower*.

The resulting relevance scores provide a class-specific semantic prior for intervening at two stages of a LF-CBM-style training pipeline. First, CLIP similarities are commonly used as supervision for concept learning, but may contain misleading signals arising from CLIP's biases. We therefore introduce score-guided activation attenuation to suppress such signals according to their class-specific relevance. Second, downstream classifiers may still rely on concepts that are semantically irrelevant to individual classes. We address this through adaptive ℓ_1 regularization, which discourages such low-relevance concept-class associations during classifier learning. Together, these two interventions use the same semantic prior to regulate both concept learning and classifier learning, providing complementary mechanisms for suppressing low-relevance associations. To scale ScoreCBM to large

datasets, we further introduce CLIP-based top-r pre-filtering, which reduces the number of LLM evaluations from quadratic to linear in the number of classes.

We evaluate ScoreCBM across both biased and standard image classification benchmarks. ScoreCBM consistently improves over the underlying LF-CBM baselines, with particularly strong gains on biased datasets, while scaling effectively to large-scale benchmarks such as ImageNet.

Our main contributions are summarized as follows:

- **We identify the class-dependent nature of concept relevance as an important challenge for debiasing CLIP-based CBMs.** Rather than treating potentially problematic concepts as globally unreliable, we explicitly assess the semantic relevance of class-concept pairs, enabling low-relevance associations to be selectively suppressed without globally removing useful concepts.
- **We propose ScoreCBM, an interpretable concept-scoring framework that uses LLM-derived class-concept relevance as a semantic prior for debiasing.** The same relevance signal guides score-based activation attenuation during concept learning and adaptive ℓ_1 regularization during classifier learning, providing complementary interventions at two stages of the training pipeline.
- **We demonstrate the effectiveness, scalability, traceability, and editability of ScoreCBM through quantitative and qualitative evaluations.** Across biased and standard benchmarks, ScoreCBM consistently improves over LF-CBM-style baselines.

2 Related Work

Concept bottleneck models and label-free concept learning. Concept Bottleneck Models (CBMs) introduce human-interpretable concepts as an intermediate representation between inputs and predictions, enabling concept-level inspection and intervention (Koh et al., 2020). Their reliance on concept annotations, however, limits their scalability. Recent advances in pretrained vision–language models (VLMs) have substantially reduced this dependence by providing a shared representation space for visual inputs and textual concepts. PCBM demonstrated that concept bottlenecks can be constructed post hoc from pretrained representations and natural-language concepts (Yuksekgonul et al., 2023), while LaBo and LF-CBM further combined language-model-generated concept sets with CLIP-based image-concept alignment to enable scalable concept learning without concept annotations (Yang et al., 2023; Oikarinen et al., 2023).

In particular, LF-CBM helped establish a scalable training paradigm in which VLM-derived image–concept similarities provide supervision for learning an interpretable concept bottleneck. Subsequent work has built on or reworked different components of this broader label-free paradigm, including concept selection (Yang et al., 2023), concept completeness (Shang et al., 2024; Liu et al., 2025), automated concept discovery (Rao et al., 2024; Prasse et al., 2025), sparse and open-vocabulary concept representations (Semenov et al., 2024; Tan et al., 2024), and improved visual grounding (Srivastava et al., 2024; Jeon et al., 2025). Despite their methodological differences, VLM-derived associations between images and textual concepts have become a fundamental building block for a broad family of label-free CBMs.

Debiasing CLIP and CLIP-based CBMs. As a widely used vision–language foundation model, CLIP is known to encode spurious correlations and dataset biases inherited from large-scale pretraining (Agarwal et al., 2021; Ruggeri & Nozza, 2023). A growing body of work has therefore proposed debiasing methods that intervene on prompts, features, or representation spaces (Chuang et al., 2023; Dehdashtian et al., 2024; Molahasani et al., 2025; Zhang et al., 2025a). As CLIP is increasingly used to provide supervision for concept learning, recent studies have shown that its biases and spurious associations can propagate into downstream concept bottlenecks (Wang et al., 2024; Sammani et al., 2026). To address this issue, existing approaches have explored concept-level model editing (Yuksekgonul et al., 2023), identifying and filtering spurious concepts (Kim et al., 2024), improving concept grounding (Xie et al., 2025; Jeon et al., 2025), and examining the reliability of VLM-derived concept supervision (Yalavarthi et al., 2026). Most of these approaches treat undesirable concepts or associations primarily at a global level. Some recent work instead considers class-dependent structure: Discover and Cure (DISC) identifies class-level spurious concepts through

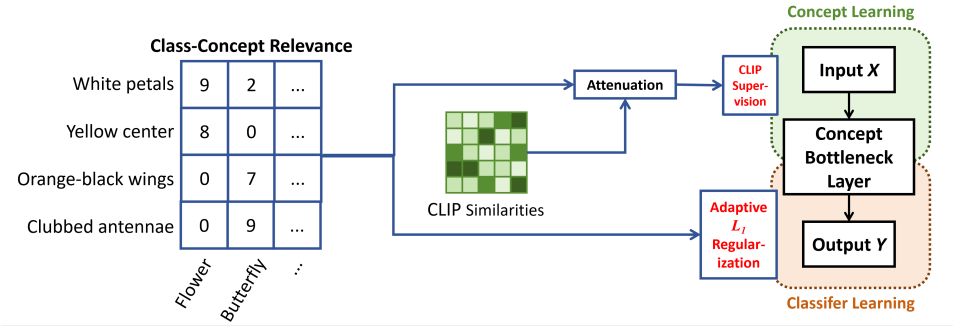


Figure 2: The framework of ScoreCBM.

their instability across environments (Wu et al., 2023), while ALBM constructs class-specific concept spaces to reduce spurious cue inference in language bottleneck models (Zhang et al., 2025b).

While these studies introduce class-dependent treatments of undesirable concepts or cues, they primarily identify or restructure the concepts associated with individual classes, without explicitly representing class-concept semantic relevance as a debiasing signal. ScoreCBM instead models this relevance directly and uses it to regulate both concept learning and classifier learning.

3 Methodology of ScoreCBM

In this section, we present ScoreCBM in detail. We first review the LF-CBM pipeline and introduce the necessary notation in Sec. 3.1, followed by our formulation of class-concept relevance in Sec. 3.2. We then describe how the resulting relevance prior is incorporated into concept learning and classifier learning through score-guided attenuation and adaptive ℓ_1 regularization in Secs. 3.3 and 3.4, respectively. Fig. 2 provides an overview of ScoreCBM.

3.1 Problem Setup and LF-CBM Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled image classification dataset with classes $\mathcal{Y} = \{y_1, \ldots, y_K\}$, and let $\mathcal{C} = \{c_1, \ldots, c_M\}$ denote a vocabulary of human-readable concepts. Following the LF-CBM setting, this vocabulary can be generated from class names using an LLM and does not require image-level concept annotations. Let E_I and E_T be the frozen image and text encoders of CLIP. The CLIP activation of concept c_j for image x_i is

$$A_{ij} = \cos(E_I(x_i), E_T(c_j)), \qquad \mathbf{A} \in \mathbb{R}^{N \times M}. \tag{1}$$

These activations serve as concept supervision in LF-CBM (Oikarinen et al., 2023). Specifically, given a target visual representation $\mathbf{z}_i = f(x_i)$, a linear projection W_c is learned such that $\widehat{\mathbf{a}}_i = W_c \mathbf{z}_i$ matches the corresponding CLIP concept activations. The projected activations are standardized and passed to a sparse linear classifier,

$$\widehat{\mathbf{y}}_i = W_g \widehat{\mathbf{a}}_i + \mathbf{b}, \tag{2}$$

where each entry $W_{g,kj}$ represents the contribution of concept c_j to class y_k. This pipeline exposes two possible sources of problematic concept use: unreliable CLIP activations can be learned by W_c, and the final classifier can assign a large weight to a concept that is semantically irrelevant to a particular class. ScoreCBM addresses both sources using a shared class–concept relevance prior, as illustrated in Fig. 2.

3.2 Class–Concept Relevance

Existing label-free CBMs primarily employ LLMs to generate candidate concept vocabularies. We additionally use an LLM as a *relevance critic* after the vocabulary has been constructed. For every concept–class pair, we ask the LLM to assess the semantic relevance of the concept to the

class according to a predefined scoring rubric, with higher scores indicating stronger and more discriminative class relevance. This produces a score matrix

$$S_{jk} = \mathrm{LLMScore}(c_j, y_k), \qquad \mathbf{S} \in \{0, \dots, 10\}^{M \times K}. \tag{3}$$

Importantly, S_{jk} does not measure whether c_j is a generally valid or visually observable concept. Instead, it measures the LLM-estimated class-conditional relevance of c_j to y_k, with larger values indicating stronger relevance. The same concept can therefore receive different scores for different classes. For example, *white petals* can receive a high score for *flower* but a low score for *butterfly*, even though it may appear in images of both classes.

We treat these scores as an external relevance prior rather than ground-truth concept annotations. Scoring is performed once offline, and the cached matrix is reused throughout training, requiring no LLM calls at inference time. The complete scoring rubric and prompt are provided in Appendix A.2.

Scalability. For datasets with large label spaces, exhaustively evaluating all MK class–concept pairs can be costly. We therefore use CLIP text similarity to pre-filter candidate pairs, retaining only the top-r concepts for each class for LLM scoring and assigning the minimum relevance score to the remaining pairs. CLIP serves only as a candidate retriever, while the retained pairs are evaluated by the LLM using the same relevance rubric. This reduces the number of LLM evaluations from MK to rK. Because the concept vocabulary contains a fixed number of candidates per class, M scales linearly with K. Thus, for fixed r, the number of LLM evaluations scales from $O(K^2)$ to $O(K)$, making relevance scoring practical for large-scale datasets.

3.3 Score-Guided Activation Attenuation

The activation matrix in Eq. 1 may assign high values to concepts that are visually associated with an image but are not reliable evidence for its class. Directly fitting these values encourages the concept projection to reproduce the same associations. We therefore condition the projection target on the relevance of each concept to the training class. For an example (x_i, y_i), we define

$$R_{ij} = \begin{cases} \alpha, & S_{j,y_i} \le \tau, \\ 1, & S_{j,y_i} > \tau, \end{cases} \qquad 0 \le \alpha \le 1, \tag{4}$$

where τ is a relevance threshold and α is an attenuation factor. The corrected concept supervision is

$$\widetilde{A}_{ij} = R_{ij} A_{ij}. \tag{5}$$

Thus, a low-relevance association is weakened rather than requiring the concept to be globally removed. High-relevance uses of the same concept for other classes remain unchanged.

Following LF-CBM, W_c is optimized using the cubed cosine similarity objective, denoted by $d(\mathbf{u}, \mathbf{v}) = \cos\big((\mathbf{u} - \bar{u})^{\odot 3}, (\mathbf{v} - \bar{v})^{\odot 3}\big)$, where $\odot 3$ indicates element-wise cubing. For a minibatch $\mathcal{B}$, let $\mathbf{Z}_{\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}| \times d}$ stack the backbone features and define $\widehat{\mathbf{A}}_{\mathcal{B}} = \mathbf{Z}_{\mathcal{B}} W_c^\top$. Our projection objective is

$$W_c^* = \arg\min_{W_c} -\frac{1}{M} \sum_{j=1}^{M} d\Big(\widehat{\mathbf{A}}_{\mathcal{B},:,j}, \widetilde{\mathbf{A}}_{\mathcal{B},:,j}\Big). \tag{6}$$

Attenuation is applied only when constructing targets for projection learning. At inference time, concept activations are produced directly by the learned W_c; neither ground-truth class labels nor additional LLM queries are required.

3.4 Score-Adaptive Sparse Classification

Activation attenuation reduces unreliable supervision for concept learning, but does not directly prevent the downstream classifier from exploiting low-relevance concept–class connections. We therefore construct a penalty matrix $\mathbf{\Lambda} \in \mathbb{R}_+^{K \times M}$ from the same class–concept relevance prior to regularize such connections more strongly. Among possible relevance-to-penalty mappings, we use a simple thresholded variant by default, which we find to perform best empirically among the mappings considered (see Appendix B.1):

$$\Lambda_{kj}^{\mathrm{raw}} = \begin{cases} \mu, & S_{jk} < \tau_{\mathrm{cls}}, \\ 1, & S_{jk} \geq \tau_{\mathrm{cls}}, \end{cases} \qquad \mu > 1. \tag{7}$$

Low-relevance connections therefore receive stronger regularization. To separate this redistribution from simply increasing the overall regularization strength, we normalize the matrix to have unit mean:

$$\Lambda_{kj} = \frac{\Lambda_{kj}^{\mathrm{raw}}}{\frac{1}{KM}\sum_{k'=1}^{K}\sum_{j'=1}^{M}\Lambda_{k'j'}^{\mathrm{raw}}}. \tag{8}$$

We incorporate the normalized penalty matrix into the ℓ_1 regularization of the SAGA classifier. The resulting classification objective is

$$\mathcal{L}_{\mathrm{cls}} = \mathcal{L}_{\mathrm{CE}} + \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{M}\Lambda_{kj}|W_{g,kj}| + \frac{\lambda_2}{2}\|W_g\|_F^2. \tag{9}$$

We retain the uniform ℓ_2 component of the original SAGA formulation and apply the relevance-adaptive penalty only to its ℓ_1 component. This objective discourages concepts from strongly influencing classes for which they have low semantic relevance, while retaining their ability to support classes for which they are relevant.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate ScoreCBM on four standard classification benchmarks, CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), Stanford40 (Yao et al., 2011), and CUB-200 (Wah et al., 2011), as well as two debiasing benchmarks, 200-class Waterbirds-Species (Sagawa et al., 2020) and BAR (Nam et al., 2020). Waterbirds-Species preserves the 200-way species labels of CUB-200 while introducing spurious background correlations, rather than using the standard binary Waterbirds task. We further evaluate scalability on Places365 (Zhou et al., 2018) and ImageNet-1K (Deng et al., 2009). We report top-1 classification accuracy with mean and standard deviation over multiple runs.

Implementation. ScoreCBM uses frozen CLIP-RN50 visual representations (Radford et al., 2021) for most datasets, while CUB-200 and Waterbirds-Species follow the CUB-trained ResNet-18 protocol (He et al., 2016). For ImageNet-1K, we use a ResNet-50 backbone with CLIP ViT-B/16 for image–concept supervision, and additionally report a matched CLIP-RN50 setting. Concept vocabularies are generated following LF-CBM, and class–concept relevance scores are computed offline, with CLIP candidate pre-filtering used for Places365 and ImageNet-1K. We use Gemini as the default relevance scorer and evaluate alternative relevance estimators in Appendix B.2. No LLM calls are required at inference time.

Baselines. We compare with ERM (Gulrajani & Lopez-Paz, 2021), LF-CBM (Oikarinen et al., 2023), U-F^2-CBM (Sammani et al., 2026), DN-CBM (Rao et al., 2024), DCBM (Prasse et al., 2025), LaBo (Yang et al., 2023), and UCBM (Schrodi et al., 2025). Further implementation details, hyperparameters, prompts, and dataset descriptions are provided in Appendix A.

4.2 Main Results

Table 1 reports the main classification results. ScoreCBM consistently improves upon LF-CBM across all seven benchmarks, with particularly pronounced gains on the two debiasing benchmarks. On Waterbirds-Species, ScoreCBM improves overall accuracy by 4.23 points and worst-group accuracy by 3.43 points over LF-CBM. On BAR, it improves accuracy by 5.14 points. These results indicate that explicitly modeling class–concept relevance is particularly beneficial under background and contextual biases, while also consistently improving performance on standard classification benchmarks. Notably, the improvements substantially close the performance gap between LF-CBM and the non-concept ERM baseline on the debiasing benchmarks: ScoreCBM achieves comparable performance to ERM on Waterbirds-Species (62.05% vs. 62.09% ACC; 43.20% vs. 43.71% WGA) and slightly higher accuracy on BAR (89.39% vs. 88.91%).

Table 1: Comparative performance on standard and debiasing benchmarks. Columns marked † use the CUB-trained ResNet-18 protocol, and all other entries use frozen CLIP-RN50. We report top-1 accuracy (ACC, %) for all datasets and additionally report worst-group accuracy (WGA, %) for Waterbirds-Species. Bold and underlined values denote the best and second-best results among CBMs, respectively.

Method	Standard Benchmarks					Debiasing Benchmarks		
	CIFAR-10	CIFAR-100	Stanford40	Places365	CUB-200†	Waterbirds-Species†		BAR
						Acc.	WGA	
ERM	87.83±0.02	68.84±0.10	90.44±0.12	49.08±0.004	76.34±0.05	62.09±0.28	43.71±0.90	88.91±0.77
LF-CBM	86.51±0.10	67.78±0.80	89.14±0.17	49.20±0.24	74.82±0.13	57.82±0.18	39.77±0.32	84.25±0.93
LaBo	87.68±0.01	65.16±0.07	**90.45±0.01**	51.93±0.01	–	–	–	88.58±0.09
DN-CBM	86.55±0.15	67.93±0.07	88.06±0.04	49.65±0.07	–	–	–	74.57±0.09
UCBM	82.52±0.03	68.31±0.07	59.73±0.21	50.65±0.06	74.85±0.06	56.72±0.28	32.66±0.24	44.85±1.15
DCBM (SAM2)	84.81±0.06	62.70±0.48	87.06±0.19	47.76±0.11	**76.09±0.04**	57.90±0.13	38.94±0.00	82.26±0.15
U-F^2-CBM	85.36±0.61	61.20±0.29	87.50±0.20	39.56±0.18	59.95±0.07	43.50±0.39	29.65±0.89	85.37±3.21
ScoreCBM	**88.55±0.09**	**69.21±0.10**	89.55±0.17	**51.98±0.06**	75.42±0.05	**62.05±0.08**	**43.20±0.70**	**89.39±0.40**

Table 2: ImageNet-1K top-1 accuracy (%).

Method	Backbone	Accuracy
Supervised ResNet-50		
Standard ERM	ResNet-50	76.13
LF-CBM	ResNet-50	71.87±0.02
UCBM	ResNet-50	69.64±0.02
U-F^2-CBM	ResNet-50	73.90
ScoreCBM	ResNet-50	**75.28±0.03**
CLIP-RN50		
LF-CBM	CLIP-RN50	67.45±0.03
ScoreCBM	CLIP-RN50	**71.41±0.09**

Table 3: Effect of CLIP candidate pre-filtering on BAR. All settings use Gemini relevance scores and identical downstream configurations.

Setting	Pairs	Reduction	Accuracy
Full	1,320	—	**89.39±0.40**
Top-100	600	54.5%	88.94±0.46
Top-50	300	77.3%	88.99±0.26
Top-30	180	86.4%	87.92±0.38
Top-20	120	90.9%	87.40±0.36

4.3 Large-Scale Evaluation

We further evaluate ScoreCBM on ImageNet-1K to examine its scalability. As shown in Table 2, ScoreCBM achieves 75.28% top-1 accuracy with supervised ResNet-50, improving LF-CBM by 3.41 points and U-F^2-CBM by 1.38 points, while approaching the corresponding ERM accuracy of 76.13%. The improvement over LF-CBM also persists under a matched CLIP-RN50 setting (67.45% → 71.41%). We next examine the cost–accuracy trade-off of CLIP candidate pre-filtering. As shown in Table 3, substantial filtering on BAR reduces LLM evaluations with only modest performance degradation; similar behavior is observed on CIFAR-10, where top-50 filtering performs comparably to full scoring (88.50% to 88.55%). This reduction becomes particularly important on ImageNet-1K: top-200 filtering reduces the 10.553M possible class–concept pairs to 200K LLM evaluations (98.1% reduction). These results show that ScoreCBM retains its improvements over LF-CBM at ImageNet scale while candidate pre-filtering makes relevance scoring practical for large concept vocabularies and label spaces.

4.4 Ablation Studies

We ablate the two relevance-guided interventions in ScoreCBM while keeping the concept vocabulary and all other training settings fixed. As shown in Table 4, activation attenuation substantially improves the matched baseline from 85.06% to 88.24%, whereas adaptive ℓ_1 regularization alone does not yield an improvement. Combining both interventions achieves the best accuracy of 89.39%, further improving upon attenuation alone by 1.15 percentage points. Note that the matched baseline differs from the LF-CBM result in Table 1 because this ablation uses the same Gemini concept vocabulary across all configurations, thereby isolating the effects of the two interventions. These results indicate that attenuation provides the primary individual gain, while adaptive ℓ_1 regularization is most effective when applied jointly with relevance-guided attenuation. This interaction is consistent with their complementary roles: attenuation acts upstream by reshaping the concept bottleneck, whereas

Table 4: Core component ablation on BAR. All rows use identical settings otherwise.

Attenuation	Adaptive ℓ_1	Accuracy
×	×	85.06±0.39
✓	×	88.24±0.33
×	✓	84.63±0.32
✓	✓	**89.39±0.40**

Table 5: Distribution of LLM relevance scores for the top-10 concepts with the largest positive contributions to the predicted class on BAR. ScoreCBM substantially reduces reliance on low-relevance associations compared with the matched LF-CBM baseline.

Method	Score 0	Scores 1–4	Scores 5–8	Scores 9–10
LF-CBM	10.05%	8.37%	38.43%	43.15%
ScoreCBM	**0.01%**	**0.13%**	49.19%	50.67%

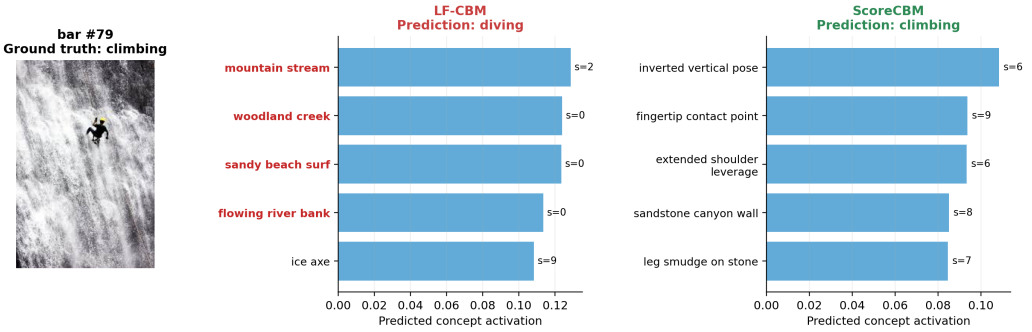


Figure 3: Concept predictions for a BAR example corrected by ScoreCBM. We show the five largest raw concept activations. s denotes each concept's relevance score to the ground-truth class, *climbing*.

adaptive regularization acts downstream by discouraging the classifier from exploiting remaining low-relevance associations.

4.5 Analysis of Class–Concept Relevance

We analyze how relevance guidance affects the associations learned and used by ScoreCBM. For each BAR test image, we retain the ten concepts with the largest positive contributions to the predicted class and group them by their LLM relevance scores. As shown in Table 5, low-relevance associations (scores 0–4) account for 18.42% of the matched LF-CBM baseline's top contributions, but only 0.14% of ScoreCBM's, while high-relevance associations (scores 5–10) increase from 81.58% to 99.86%. This consistent shift across ten seeds shows that relevance guidance substantially changes the class–concept associations used for prediction. Fig. 3 further shows that relevance guidance can affect the concept bottleneck. For a climbing image featuring a prominent waterfall, LF-CBM primarily activates water-related concepts and incorrectly predicts diving. ScoreCBM instead places more emphasis on a broader range of evidence, including body pose, contact points, and rock structure, leading to the correct prediction. We also illustrate why relevance should be modeled at the class–concept level. In Fig. 4, *sandy beach* forms a misleading association with *palm tree* but remains valid evidence for *sea*. Similarly, visual resemblance causes *roll cage* to support an incorrect *throwing* prediction while remaining useful for *racing*. ScoreCBM suppresses these low-relevance uses without globally removing the concepts.

Finally, the explicit relevance scores provide a direct interface for model editing. As shown in Fig. 5, correcting a single class–concept relevance score from 0 to 6 changes both the corresponding concept activation and classifier weight, increasing its contribution to the reading class and correcting the prediction from *riding a bike* to *reading*. The resulting changes are directly traceable to the edited class–concept association.

5 Conclusion

We introduced ScoreCBM, an interpretable concept-scoring framework for debiasing CLIP-based concept bottleneck models. By explicitly modeling semantic relevance at the class–concept level, ScoreCBM selectively suppresses low-relevance associations without globally removing concepts

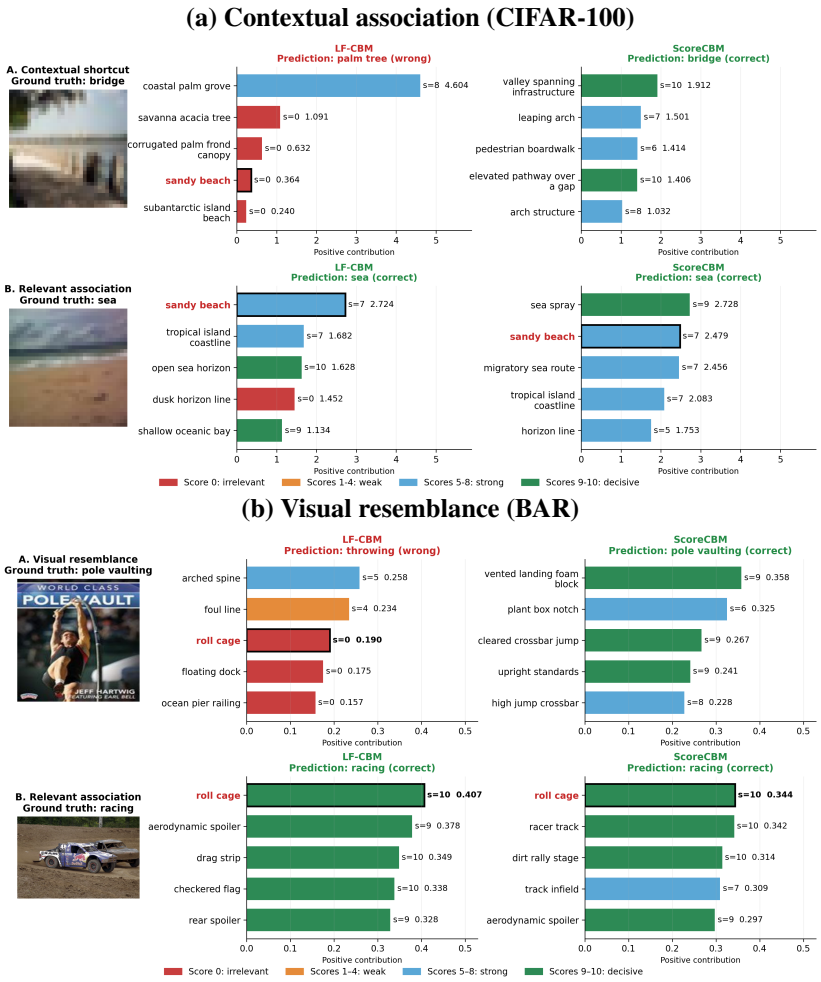


Figure 4: Suppression and preservation of class-concept associations. ScoreCBM suppresses low-relevance associations while preserving the same concepts as evidence for relevant classes.

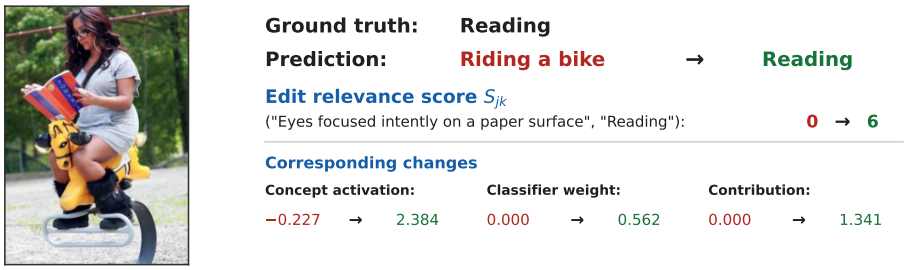


Figure 5: Editing a class–concept relevance score in ScoreCBM.

that remain informative for other classes. LLM-derived relevance scores provide a shared semantic prior for intervening in both concept learning and classifier learning. Across biased and standard benchmarks, ScoreCBM consistently improves over LF-CBM-style baselines and scales effectively to ImageNet-1K. Our analyses further show that relevance guidance substantially reduces reliance on low-relevance associations and promotes more class-relevant concept bottlenecks. A limitation of ScoreCBM is its reliance on LLM-derived relevance judgments, which can be imperfect. Systematically evaluating and improving the reliability of LLMs or other potential scorers is beyond the scope of this work and remains an important direction for future research.

AI DISCLOSURE

In this work, we used generative AI tools for assisting in designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, and generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proof are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, drafting parts of a research paper, transcribing recordings of research material, summarizing or analyzing existing literature, searching for information, editing a research paper to improving readability, identifying relevant literature, formatting references, proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. We manually ensure the citations are correct. LLM-generated code was verified and tested for correctness by the authors. The core research idea and scientific contributions were developed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code is available in the supplementary materials.

REFERENCES

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications, 2021. URL https://arxiv.org/abs/2108.02818.

Mohammad Taha Bahadori and David E Heckerman. Debiasing concept-based explanations with causal analysis. *arXiv preprint arXiv:2007.11500*, 2020.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:772–782, 2020. doi: 10.1038/s42256-020-00265-z. URL https://doi.org/10.1038/s42256-020-00265-z.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts, 2023. URL https://arxiv.org/abs/2302.00070.

Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6fb3465e95fbfbde193d3df17769f59c-Abstract-Conference.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://www.image-net.org/about.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Sujin Jeon, Hyundo Lee, Eungseo Kim, Sanghack Lee, Byoung-Tak Zhang, and Inwoo Hwang. Locality-aware concept bottleneck model, 2025. URL https://arxiv.org/abs/2508.14562.

Jeeyung Kim, Ze Wang, and Qiang Qiu. Constructing concept-based models to mitigate spurious correlations with minimal human effort. In *European Conference on Computer Vision*, pp. 137–153. Springer, 2024. URL https://arxiv.org/abs/2407.08947.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yang Liu, Tianwei Zhang, and Shi Gu. Hybrid concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Hybrid_Concept_Bottleneck_Models_CVPR_2025_paper.html.

Mahdiyar Molahasani, Azadeh Motamedi, Michael Greenspan, Il-Min Kim, and Ali Etemad. Prism: Reducing spurious implicit biases in vision-language models with llm-guided embedding projection, 2025. URL https://arxiv.org/abs/2507.08979.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html.

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

Katharina Prasse, Patrick Knab, Sascha Marton, Christian Bartelt, and Margret Keuper. DCBM: Data-efficient visual concept bottleneck models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49752–49782. PMLR, 2025. URL https://proceedings.mlr.press/v267/prasse25a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9973_ECCV_2024_paper.php.

Gabriele Ruggeri and Debora Nozza. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6445–6455. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.403. URL https://aclanthology.org/2023.findings-acl.403/.

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*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Fawaz Sammani, Jonas Fischer, and Nikos Deligiannis. CLIP-free, label-free, unsupervised concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 3262–3272, 2026. URL https://arxiv.org/abs/2503.10981.

Simon Schrodi, Julian Schur, Max Argus, and Thomas Brox. Selective concept bottleneck models without predefined concepts. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=PMO30TLI4l.

Andrei Semenov, Vladimir Ivanov, Aleksandr Beznosikov, and Alexander Gasnikov. Sparse concept bottleneck models: Gumbel tricks in contrastive learning, 2024. URL https://arxiv.org/abs/2404.03323.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11040, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Shang_Incremental_Residual_Concept_Bottleneck_Models_CVPR_2024_paper.html.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79057–79094, 2024. doi: 10.52202/079017-2510. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90043ebd68500f9efe84fedf860a64f3-Abstract-Conference.html.

Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision*, volume 15144 of *Lecture Notes in Computer Science*, pp. 123–138. Springer, 2024. doi: 10.1007/978-3-031-73016-0_8. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/11588.pdf.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of CLIPs to spurious features. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3893. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dd59fad18638714e6c447a3b7b9c4160-Abstract-Conference.html.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37765–37786. PMLR, 2023. URL https://proceedings.mlr.press/v202/wu23w.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30199–30209, 2025. URL https://arxiv.org/abs/2505.07209.

Bharat Chandra Yalavarthi, Nalini Ratha, and Venu Govindaraju. Label-free mitigation of spurious correlations in vlms using sparse autoencoders. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 31820–31838, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/359242d225d2fba218a85e147803f8b7-Paper-Conference.pdf.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yang_Language_in_a_Bottle_Language_Model_Guided_Concept_Bottlenecks_for_CVPR_2023_paper.html.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE*

International Conference on Computer Vision, 2011. doi: 10.1109/ICCV.2011.6126386. URL https://geometry.stanford.edu/paper.php?id=yjklgl-harlbaap-11.

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

Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. Joint vision-language social bias removal for CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4246–4255, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Joint_Vision-Language_Social_Bias_Removal_for_CLIP_CVPR_2025_paper.html.

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, 2025b.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://places2.csail.mit.edu/.

A Implementation and Experimental Details

A.1 Overall ScoreCBM Pipeline

Our implementation follows LF-CBM for constructing CLIP-based concept supervision, learning a linear concept projection, standardizing projected activations, and training a sparse classifier (Oikarinen et al., 2023). The overall workflow is

$$\begin{aligned}\text{Concept Generation} \rightarrow \text{Class–concept Relevance Scoring} \rightarrow \text{Score-guided Activation Attenuation} \\ \rightarrow \text{Concept Projection Learning} \rightarrow \text{Sparse Classification with Adaptive} \ell_1 \text{ Regularization.}\end{aligned} \tag{10}$$

Concept generation produces a human-readable concept vocabulary without image-level concept annotations. Class–concept relevance scoring uses an LLM to estimate the relevance of each concept to each class, producing the relevance scores used as semantic priors during training. These scores are used in two stages: they guide activation attenuation when constructing concept supervision, and determine the adaptive ℓ_1 penalty applied during sparse classifier learning.

The image encoder and CLIP text encoder remain frozen throughout training. Relevance scores are computed once offline and cached for subsequent experiments. No LLM calls are required during training or inference.

A.2 Concept Vocabulary and Class–Concept Relevance Scoring

We generate concept candidates from class names without image-level concept annotations following LF-CBM (Oikarinen et al., 2023). We use a slightly adapted prompt that favors static, photographically observable evidence, including objects or parts, scene and environment, human appearance or pose, and fine-grained visual attributes, rather than abstract or non-visual descriptions. Generated phrases are normalized and exactly deduplicated. For larger vocabularies, we additionally remove phrases longer than 50 characters, phrases overly similar to class names, and near-duplicates with cosine similarity above 0.85. These preprocessing steps are applied before relevance scoring.

Given the resulting vocabulary, we estimate class–concept relevance using the scoring rubric below. The rubric evaluates whether a concept provides discriminative semantic evidence for the target class, rather than whether it can merely co-occur with that class.

```
You are a Computer Vision expert specializing in feature-discriminative analysis for image
↪  classification. Your task is to evaluate the relevance of specific visual features for a
↪  target category.

TARGET CATEGORY: {target_category}

SCORING GUIDELINES:
- Score 0: Irrelevant, generic, or highly confusing (appears frequently in other categories).
- Score 1-4: Weak association (common posture/background, lacks discriminative power).
- Score 5-8: Strong association (frequently present in this category, but has some overlap
↪  with similar activities).
- Score 9-10: Decisive feature (the defining criterion, unique and essential to this
↪  category).

IMPORTANT RULES:
1. MANDATORY: You must provide a score and reasoning for EVERY SINGLE feature phrase in the
↪  input list.
2. DO NOT skip, summarize, or group any features together.
3. If a feature is unclear, score it 0 and explain why.
4. If the list is long, maintain high output density. Do not stop until every input phrase has
↪  been processed.

INPUT FEATURES:
{feature_list_str}

REQUIREMENTS:
Return the result strictly as a valid JSON object matching the requested schema.
```

Gemini is used for the small- and medium-scale experiments, with gemini-3.1-flash-lite, temperature 0.1, JSON-constrained output, and batches of 15 concept phrases. Places365 and ImageNet-1K use the same scoring rubric with a local Qwen2.5-14B model. Each retained pair is scored once unless

Table 6: Concept vocabularies and offline relevance-scoring protocols. "Full" scores every pair; top-r queries only the r concepts with highest CLIP text–text similarity for each class for LLM relevance scoring.

Dataset	Classes	Concepts	Scorer	Candidate policy
CIFAR-10	10	429	Gemini	Full
CIFAR-100	100	5,411	Gemini	top-100
Stanford40	40	1,640	Gemini	top-200
CUB-200	200	3,302	Gemini	top-100
Waterbirds-Species	200	3,302	Gemini	top-100
BAR	6	220	Gemini	Full
Places365	365	5,897	Qwen2.5-14B	top-200
ImageNet-1K	1,000	10,553	Qwen2.5-14B	top-200

Table 7: ScoreCBM configurations.

Setting	Backbone	τ	α	τ_{cls}	μ	λ	Iter.
CIFAR-10	CLIP-RN50	4	0.5	4	20	7×10^{-4}	300
CIFAR-100	CLIP-RN50	3	0.3	3	30	7×10^{-4}	300
Stanford40	CLIP-RN50	3	0.3	3	30	7×10^{-4}	300
CUB-200	ResNet-18-CUB	2	0.5	3	30	7×10^{-4}	300
Waterbirds-Species	ResNet-18-CUB	2	0.5	3	30	7×10^{-4}	300
BAR	CLIP-RN50	4	0.5	4	20	7×10^{-4}	300
Places365	CLIP-RN50	3	0.3	3	30	7×10^{-4}	300
ImageNet-1K	ResNet-50	3	0.3	3	30	1×10^{-4}	80

otherwise specified in an ablation. Scores are parsed as integers in $\{0, \ldots, 10\}$ and cached. The scorer also returns a textual explanation for each concept, but only the phrase–score mapping is used by ScoreCBM.

For top-r scoring, each concept and the prompt *a photo of* $\{class\}$ is encoded with the CLIP text encoder and ranked by cosine similarity independently for each class. Only the top-r pairs are submitted to the relevance scorer, while all remaining pairs receive a score of zero. CLIP therefore serves only as a candidate retriever rather than as the relevance estimator. Importantly, pre-filtering does not remove concepts from the bottleneck, and the complete M-concept vocabulary is retained. Table 6 summarizes the resulting vocabulary sizes and scoring configurations used in all experiments.

A.3 Training and Hyperparameters

Unless stated otherwise, image and CLIP activations are precomputed and reused across seeds. Projection learning uses Adam with learning rate 10^{-3} for 1,000 steps and the LF-CBM cubed-cosine objective. The final classifier uses SAGA with learning rate 0.1. We use the default elastic-net mixing coefficient of 0.99, corresponding to $\lambda_1 = 0.99\lambda$ and $\lambda_2 = 0.01\lambda$ in Eq. 9. CLIP-activation-extraction, projection, and SAGA batch sizes are 512, 50,000, and 256, respectively. Table 7 lists the ScoreCBM configurations used in our experiments.

A.4 Dataset Descriptions

CIFAR-10 and CIFAR-100. CIFAR-10 and CIFAR-100 are standard object classification benchmarks containing 60,000 natural images. CIFAR-10 includes 10 classes, while CIFAR-100 includes 100 classes grouped into 20 superclasses. We use these datasets to evaluate ScoreCBM on small-scale object recognition tasks.

Stanford40. Stanford40 is an action recognition dataset containing 9,532 images from 40 human action categories. Compared with object-centric benchmarks, action categories often involve contextual and pose-related visual cues.

Table 8: BAR attenuation-only accuracy (%, mean ± standard deviation over ten seeds) on the V3/Gemini concept set, with 300 SAGA iterations and no adaptive ℓ_1.

$\tau\backslash\alpha$	0.05	0.1	0.2	0.3	0.5
1	85.55±0.32	85.58±0.30	85.83±0.29	86.01±0.22	86.41±0.56
2	85.87±0.32	85.89±0.23	86.07±0.27	86.30±0.31	86.79±0.58
3	85.81±0.29	85.86±0.35	86.07±0.29	86.36±0.35	87.66±0.76
4	86.07±0.68	86.18±0.73	86.80±0.63	87.11±0.69	**88.24±0.35**
5	87.09±0.32	87.22±0.32	87.25±0.38	87.34±0.43	87.92±0.37

CUB-200. CUB-200-2011 is a fine-grained bird classification dataset containing 11,788 images from 200 bird species. The dataset provides detailed species-level categories and is commonly used to evaluate fine-grained visual recognition.

Waterbirds-Species. Waterbirds-Species uses the same images and background construction as the standard Waterbirds benchmark. The original Waterbirds data are derived from the 200 bird species in CUB-200, with the original backgrounds replaced by land- or water-associated scenes from Places365. The standard benchmark collapses these species labels into the binary landbird-versus-waterbird classification task. In Waterbirds-Species, we instead retain the original 200-way species labels, while preserving the same background correlations and distribution shifts. This yields a fine-grained classification setting in which background remains a potential spurious cue, but the prediction target is the bird species rather than its landbird/waterbird superclass. We use the CUB-200 concept vocabulary and representation, and no habitat or group annotations are used for training or inference.

BAR. BAR is an action recognition benchmark containing six human action categories with correlated background information in the training set. The test set removes these background correlations, allowing evaluation of whether models rely on contextual biases learned during training. We use BAR to evaluate concept learning under spurious contextual associations.

Places365. Places365 is a large-scale scene recognition dataset containing 365 scene categories and approximately 1.8 million training images. We include Places365 to evaluate ScoreCBM on large-scale scene recognition and diverse visual concept spaces.

ImageNet-1K. ImageNet-1K is a large-scale object classification benchmark containing 1,000 categories and approximately 1.28 million training images. It covers a broad range of object categories and remains one of the standard benchmarks for evaluating large-scale visual recognition models. We use ImageNet-1K to evaluate ScoreCBM under a large number of classes and concept candidates. Unlike smaller benchmarks, the large class and concept space requires efficient relevance score computation.

B Analysis and Robustness Studies

B.1 Relevance Score Mapping Analysis

The class–concept relevance scores in ScoreCBM can be incorporated through different mappings: attenuation during concept learning, and adaptive ℓ_1 regularization during classifier learning. We first study the attenuation mapping on BAR by evaluating different threshold-factor pairs (τ, α). Table 8 summarizes the results. We further compare different score mappings for adaptive ℓ_1 regularization. Table 9 compares linear, exponential, and hard mappings under the same attenuation setting. All evaluated adaptive mappings improve over the uniform ℓ_1 baseline. The hard mapping achieves the highest accuracy among the tested choices, with $(\tau_{\text{cls}}, \mu) = (4, 15)$ and $(4, 20)$ reaching $89.39\pm0.40\%$. These results suggest that the relevance prior can be used with different relevance-to-penalty mappings, while the hard mapping provides an effective choice for the reported ScoreCBM configuration.

Table 9: BAR adaptive-ℓ_1 mapping comparison (top-1 accuracy in %, mean ± standard deviation over ten seeds). All rows use V3/Gemini concept scores, attenuation $(\tau, \alpha) = (4, 0.5)$, and 300 SAGA iterations.

Mapping	Setting	Hyperparameters	Accuracy (%)
No adaptive ℓ_1	Uniform	–	88.24±0.35
Linear	$g(s) = 1 - s/\mu$	$\mu = 10$	88.87±0.49
		$\mu = 20$	88.69±0.42
		$\mu = 30$	88.47±0.46
		$\mu = 40$	88.38±0.39
		$\mu = 50$	88.38±0.39
Exponential	$g(s) = e^{-\mu s}$	$\mu = 0.1$	88.87±0.53
		$\mu = 0.2$	88.99±0.42
		$\mu = 0.3$	88.98±0.38
		$\mu = 0.5$	88.81±0.32
		$\mu = 1.0$	88.49±0.20
Hard	$g(s) = \mu$ if $s < \tau_{\text{cls}}$; 1 otherwise	$(\tau_{\text{cls}}, \mu) = (4, 5)$	89.28±0.44
		$(4, 10)$	89.36±0.41
		$(4, 15)$	89.39±0.40
		$(4, 20)$	**89.39±0.40**
		$(4, 25)$	89.36±0.42
		$(4, 30)$	89.36±0.39
		$(4, 40)$	89.37±0.40
		$(4, 50)$	89.37±0.42
		$(5, 5)$	89.13±0.45
		$(5, 10)$	89.13±0.40
		$(5, 15)$	89.13±0.42
		$(5, 20)$	89.13±0.42
		$(5, 25)$	89.13±0.42
		$(5, 30)$	89.14±0.40
		$(5, 40)$	89.16±0.41
		$(5, 50)$	89.20±0.44

Table 10: ScoreCBM test accuracy (%, mean ± std) under different relevance-score sources and controls. Random Score independently replaces every class–concept score with an integer sampled uniformly from 0 to 10. No Scorer denotes the matched LF-CBM baseline without relevance scores.

Dataset	Gemini	Qwen2.5-14B	CLIP text–text	Random Score	No Scorer
BAR	89.39±0.40	87.65±0.56	87.05±0.59	85.57±0.61	85.06±0.39

B.2 Alternative Scorer Analysis

We conduct experiments to evaluate ScoreCBM under different relevance scorers. Table 10 compares ScoreCBM with different relevance-score sources, including Gemini, Qwen2.5-14B, and ViT-B/16 CLIP text–text similarity. Gemini and Qwen scores are obtained from LLM-based relevance scoring, while the CLIP-based alternative derives relevance scores from text similarity between each concept phrase and the class-name prompt *a photo of* $\{class\}$. The resulting similarities are calibrated to the 0–10 relevance range before being used by ScoreCBM. On BAR, both LLM-based relevance priors outperform the CLIP-derived prior, with Gemini and Qwen2.5-14B achieving 89.39% and 87.65% accuracy, respectively, compared with 87.05% for the CLIP-based alternative. As a control, independently replacing all 1,320 BAR class–concept scores with uniform random integers from 0 to 10 yields $(85.57 \pm 0.61)\%$ across ten model seeds. The performance variation across scorers indicates that the quality of the estimated relevance prior affects downstream performance, while the overall framework remains applicable to different relevance estimators. In BAR, the raw CLIP text similarity shows moderate agreement with the Gemini scores (Pearson $r \approx 0.34$), suggesting that direct text similarity captures a related but different relevance signal.

Table 11: BAR concept–class disagreements between CLIP text association and LLM relevance, in both directions. CLIP scores are text–text similarities between concept phrases and class prompts, min–max calibrated to 0–10 separately for each class.

Concept	Target class	CLIP	LLM
High CLIP association, low LLM relevance			
high jump crossbar	throwing	8.02	0
cleared crossbar jump	throwing	9.25	1
splashing water surface	throwing	8.37	0
fish on hook	throwing	7.99	0
water surface	climbing	8.56	0
fish on hook	climbing	8.95	0
casting motion	climbing	8.30	0
baseball	racing	8.04	0
Low CLIP association, high LLM relevance			
straightened pointed toes	diving	2.06	9
blue tiled pool floor	diving	0.27	7
protective foam mats	pole vaulting	2.81	9
flexible carbon fiber pole	fishing	2.92	8

Table 12: Robustness to relevance-score corruption on BAR. The corruption ratio ρ denotes the percentage of CLIP-retrieved class–concept pairs whose relevance scores are replaced by random values. Zero scores introduced by pre-filtering are kept unchanged. Results are top-1 accuracy (%, mean $\pm$ std).

Candidate set	$\rho = 0\%$	$\rho = 10\%$	$\rho = 20\%$	$\rho = 50\%$	$\rho = 80\%$
Full	89.39 ± 0.40	88.88 ± 0.46	87.55 ± 0.25	86.97 ± 0.20	86.04 ± 0.43
Top-100	88.94 ± 0.46	88.82 ± 0.36	88.29 ± 0.44	87.25 ± 0.25	86.76 ± 0.42
Top-50	88.99 ± 0.26	88.55 ± 0.17	88.61 ± 0.24	87.94 ± 0.42	86.99 ± 0.23
Top-30	87.92 ± 0.38	87.63 ± 0.41	87.45 ± 0.39	87.57 ± 0.46	87.17 ± 0.53
Top-20	87.40 ± 0.36	87.51 ± 0.47	87.39 ± 0.48	87.71 ± 0.46	87.26 ± 0.48

B.3 CLIP–LLM Disagreement Analysis

We inspect disagreement between CLIP text association and LLM-based relevance scores. The goal is not to quantitatively measure image-level concept activation, but to illustrate why CLIP association alone may be insufficient as a relevance prior. Table 11 shows examples from BAR. The CLIP column denotes ViT-B/16 concept-text versus class-name-text similarity, min–max calibrated to 0–10 separately for each class. The CLIP column represents text-level similarity rather than image-specific concept activation.

High-CLIP but low-LLM examples, such as *high jump crossbar* for throwing and *water surface* for climbing, demonstrate cases where CLIP captures problematic associations but fails to distinguish them from class-discriminative evidence. Conversely, low-CLIP but high-LLM examples, such as *straightened pointed toes* for diving and *protective foam mats* for pole vaulting, show that LLM-based relevance scoring can provide additional fine-grained or domain-specific semantic cues that are weakly reflected in CLIP text similarity.

B.4 Robustness to Relevance Score Corruption

ScoreCBM relies on class–concept relevance scores to guide both concept supervision and adaptive sparse classification. To evaluate the robustness of this relevance prior, we randomly corrupt a proportion of the retained relevance scores after CLIP-based pre-filtering. Specifically, for a given corruption ratio ρ, ρ percent of the top-r retrieved class–concept pairs are randomly selected with their relevance scores replaced by random values. Concepts removed by pre-filtering (with score zero) remain unchanged. We evaluate different pre-filtering sizes together with increasing corruption ratios.

Table 13: Group-wise top-1 accuracy (%) on Waterbirds-Species. The prediction task remains 200-way species classification, and bird superclass and background annotations are used only to define the four evaluation groups. Landbird–Land and Waterbird–Water are aligned groups, while Landbird–Water and Waterbird–Land are conflict groups. WGA denotes worst-group accuracy. Results are mean ± standard deviation; ERM follows its original five-seed population-standard-deviation convention, while the three-seed CBM results use sample standard deviation. ERM is excluded from best-CBM highlighting.

Method	Landbird-Land	Landbird-Water	Waterbird-Land	Waterbird-Water	WGA
ERM	66.96±0.17	65.47±0.28	43.71±0.90	51.50±0.73	43.71±0.90
LF-CBM	64.41±0.40	59.38±0.18	39.77±0.32	47.98±0.68	39.77±0.32
UCBM	63.46±0.42	60.53±0.16	32.66±0.24	43.72±0.48	32.66±0.24
DCBM	64.36±0.31	59.73±0.19	38.94±0.00	47.77±0.09	38.94±0.00
U-F^2-CBM	48.80±0.41	45.28±0.89	29.65±0.89	32.55±1.35	29.65±0.89
ScoreCBM	**67.69±0.07**	**64.73±0.14**	**43.20±0.70**	**51.71±0.41**	**43.20±0.70**

Table 12 reports the results on BAR. ScoreCBM remains stable under moderate score corruption: across different candidate sizes, 10%–20% randomized scores only cause limited performance degradation. Increasing corruption gradually reduces performance, indicating that the model benefits from meaningful relevance ordering rather than depending on exact score values. Interestingly, stronger pre-filtering provides additional robustness under heavy corruption, as the candidate concept set becomes less affected by noisy relevance estimates. Still, even with 80% corruption, ScoreCBM remains consistently above the original LF-CBM baseline, which does not use class–concept relevance scores. This indicates that the relevance prior remains beneficial even when a substantial proportion of its values are corrupted.

C Additional Experimental Results

C.1 Group-wise Results on Waterbirds-Species

We further report group-wise performance on Waterbirds-Species in Table 13. Importantly, Waterbirds-Species remains a 200-way bird-species classification task, and the landbird/waterbird superclass and land/water background annotations are used only to define evaluation groups.

Let $y_i \in \{1, \dots, 200\}$ denote the species label of image x_i, $b(y_i) \in \{\text{Landbird}, \text{Waterbird}\}$ its bird superclass, and $h_i \in \{\text{Land}, \text{Water}\}$ its background type. For each superclass–background pair (b, h), we define

$$\mathcal{G}_{b,h} = \{i \mid b(y_i) = b,\ h_i = h\}.$$

The group-wise accuracy is computed using the original 200-way species prediction:

$$\text{Acc}_{b,h} = \frac{1}{|\mathcal{G}_{b,h}|} \sum_{i \in \mathcal{G}_{b,h}} \mathbf{1}[\hat{y}_i = y_i],$$

and worst-group accuracy is

$$\text{WGA} = \min_{(b,h)} \text{Acc}_{b,h}.$$

Thus, WGA measures the lowest species-classification accuracy across the four subgroups. Landbird–Land and Waterbird–Water are aligned groups, whereas Landbird–Water and Waterbird–Land are conflict groups.

ScoreCBM improves over LF-CBM across all four groups. In particular, the WGA increases from 39.77% for LF-CBM to 43.20% for ScoreCBM, showing improved robustness under background distribution shifts.

C.2 Additional Results on Candidate Pre-filtering

Table 14 reports additional comparisons between full class–concept scoring and CLIP-based top-r candidate pre-filtering, with Qwen used as the relevance scorer. We also evaluate an end-to-end Qwen

Table 14: Additional results on CLIP-based candidate pre-filtering. "Concept source" denotes the model used for concept generation, while "Scorer" denotes the relevance scorer. "Scored pairs" denotes the number of class–concept pairs evaluated by the relevance scorer. Results are top-1 accuracy (%, mean ± standard deviation).

Dataset	Concept source	Scorer	Policy	# Scored pairs	Accuracy
BAR	Gemini	Qwen	Full	1,470	87.65±0.56
BAR	Gemini	Qwen	top-50	300	88.29±0.49
BAR	Qwen	Qwen	Full	1,866	87.84±0.67
BAR	Qwen	Qwen	top-50	300	88.56±0.37
CIFAR-10	Gemini	Qwen	Full	4,340	87.93±0.12
CIFAR-10	Gemini	Qwen	top-50	500	87.36±0.07

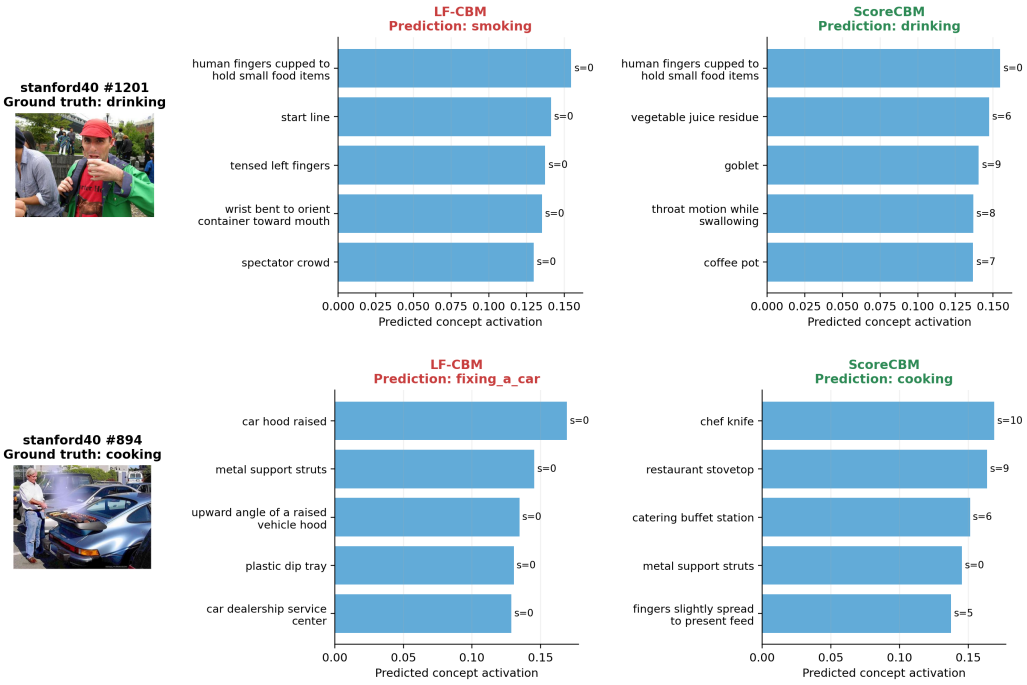


Figure 6: Additional concept-prediction examples on Stanford40.

configuration in which Qwen is used for both concept generation and relevance scoring. Across these settings, top-r pre-filtering substantially reduces the number of scored class–concept pairs while maintaining comparable downstream accuracy, and can even improve performance in some cases.

D Additional Qualitative Results

This appendix provides additional qualitative examples complementing the analyses in Sec. 4.5. Fig 6 shows two Stanford40 cases at the concept-activation level. In both examples, LF-CBM emphasizes concepts that are only weakly relevant to the ground-truth class and makes an incorrect prediction, whereas ScoreCBM shifts the concept bottleneck toward more class-relevant evidence and recovers the correct prediction.

Fig 7 further examines the concepts that contribute positively to the predicted class across Stanford40, ImageNet, and Places365. The examples cover several types of misleading evidence, including contextual associations, pose-related ambiguity, visual resemblance, and scene-level cues. Across these cases, ScoreCBM reduces the influence of low-relevance concepts while assigning larger

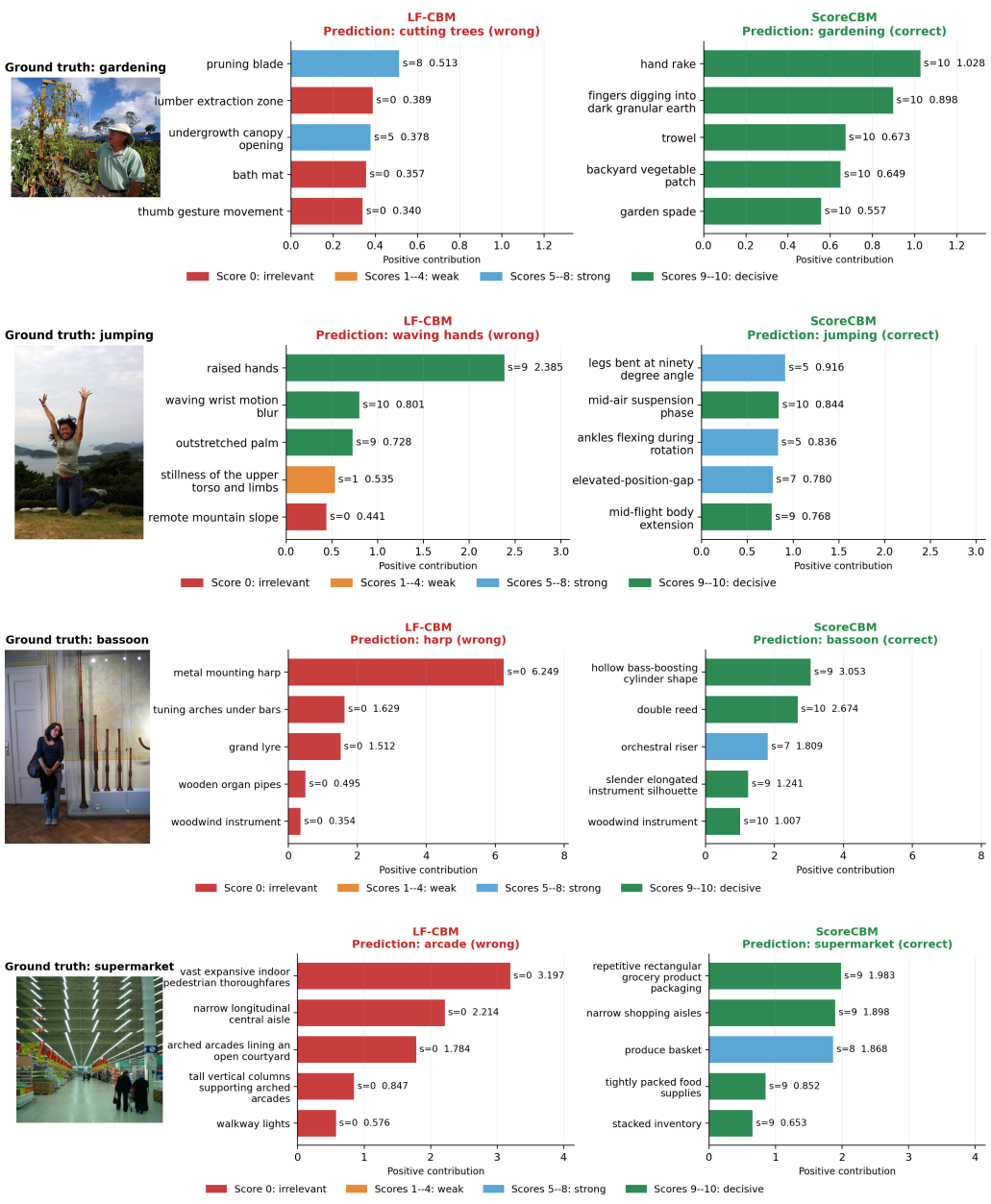


Figure 7: Additional positive concept-contribution examples. From top to bottom: Stanford40 *gardening*, Stanford40 *jumping*, ImageNet *bassoon*, and Places365 *supermarket*.

contributions to concepts that are more semantically aligned with the target class. These examples complement the aggregate analysis in Table 5 by illustrating how the shift toward higher-relevance concepts manifests in individual predictions.

Finally, Fig 8 provides additional examples of relevance-score editing. Editing the relevance of *hand gripping a cylindrical container* to *Drinking* corrects two distinct Stanford40 images, while editing *railroad track alignment* to *Streetcar* corrects a CIFAR-100 example. In each case, the single class–concept edit produces traceable changes in the learned concept activation, classifier weight, and resulting class contribution, further illustrating the direct editability of the relevance prior.

Fig. 9 provides a direct examination of how score-guided activation attenuation affects concept learning. Fig 9(a) shows training examples from BAR, Stanford40, and ImageNet where CLIP



Ground truth: Drinking
Prediction: Brushing Teeth → Drinking
Edit relevance score S_{jk}
("hand gripping a cylindrical container", "Drinking"): 0 → 10
Corresponding changes

Concept activation:			Classifier weight:			Contribution:		
0.294	→	2.583	0.000	→	0.357	0.000	→	0.921



Ground truth: Drinking
Prediction: Pouring Liquid → Drinking
Edit relevance score S_{jk}
("hand gripping a cylindrical container", "Drinking"): 0 → 10
Corresponding changes

Concept activation:			Classifier weight:			Contribution:		
1.535	→	2.401	0.000	→	0.357	0.000	→	0.856



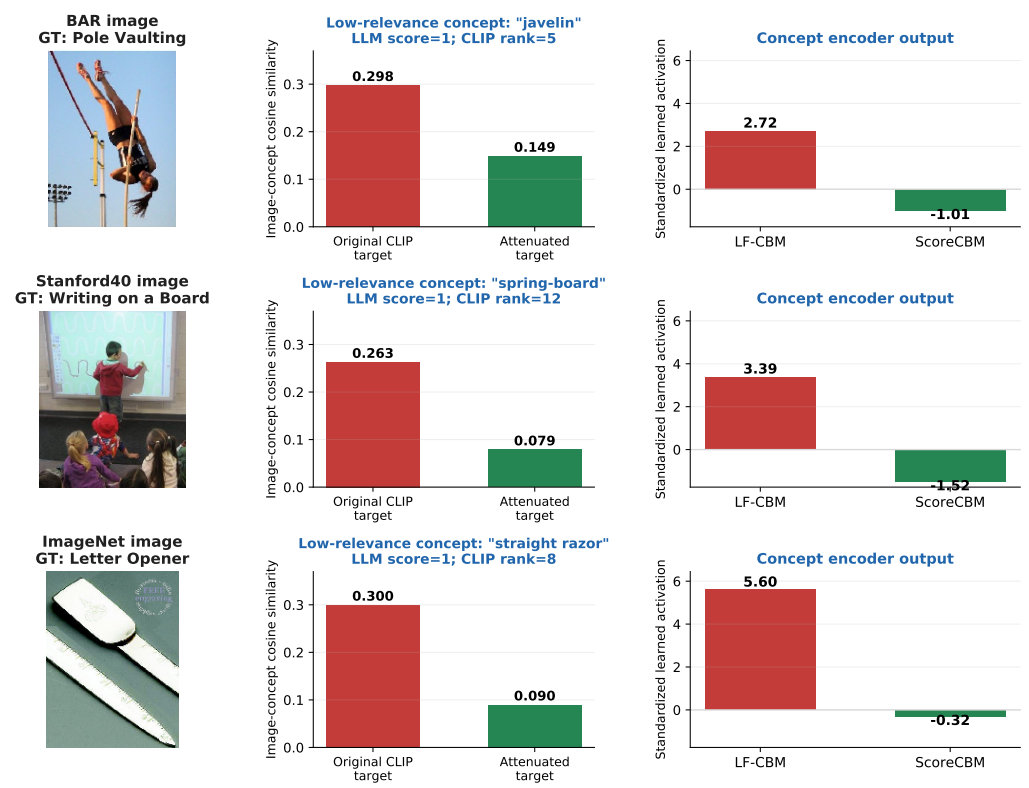
Ground truth: Streetcar
Prediction: Bus → Streetcar
Edit relevance score S_{jk}
("railroad track alignment", "Streetcar"): 0 → 10
Corresponding changes

Concept activation:			Classifier weight:			Contribution:		
1.780	→	4.060	0.000	→	0.484	0.000	→	1.966

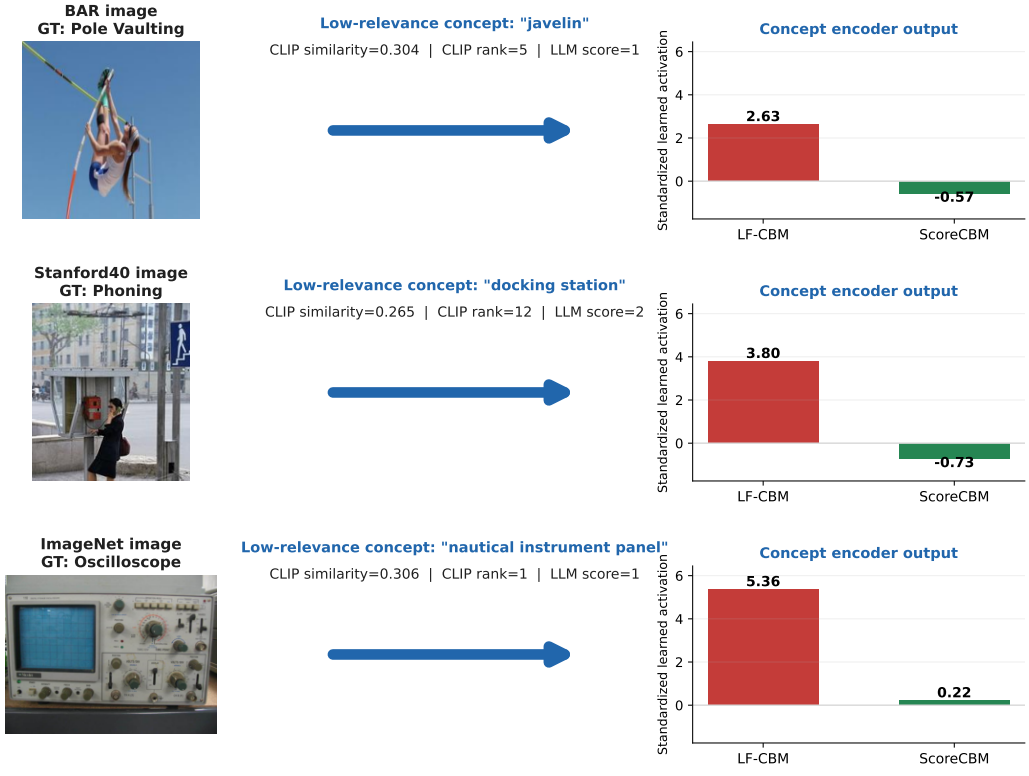
Figure 8: Additional examples of editing individual class–concept relevance scores. Top and middle: two Stanford40 *Drinking* images corrected by editing the same class–concept score. Bottom: a CIFAR-100 *Streetcar* image corrected by editing the relevance of *railroad track alignment*. Red and green denote values before and after the edit, respectively.

assigns relatively high similarity to concepts that receive low relevance scores for the ground-truth class. ScoreCBM attenuates these supervision targets before projection learning, and the resulting concept encoder assigns substantially lower activations to the corresponding concepts than LF-CBM. This shows that attenuation not only modifies the training targets, but is also reflected in the concept representations learned by the projection layer.

Fig. 9(b) further illustrates that this effect is not confined to the individual training examples whose targets are attenuated. Similar misleading associations can recur in held-out images, and the learned ScoreCBM concept bottleneck remains less responsive to them. For example, the same low-relevance association between *javelin* and *pole vaulting* appears in both the training and held-out BAR examples. After attenuating this relation during training, ScoreCBM also substantially reduces the *javelin* activation on the held-out image. This recurrence suggests that relevance-guided attenuation reshapes the learned concept bottleneck more broadly, reducing recurring low-relevance associations beyond the training examples used to construct the attenuated targets.



(a) Training-set examples of score-guided activation attenuation.



(b) Corresponding behavior on held-out images.

Figure 9: Qualitative analysis of score-guided activation attenuation across BAR, Stanford40, and ImageNet.