SL-CBM: Enhancing Concept Bottleneck Models with Semantic Locality for Better Interpretability

Hanwei Zhang [1], Luo Cheng [2,5], Rui Wen [3], Yang Zhang [4], Lijun Zhang [5], Holger Hermanns [1]

[1] Saarland University
[2] University of Chinese Academy of Sciences
[3] Institute of Science Tokyo
[4] CISPA Helmholtz Center for Information Security
[5] Key Laboratory of System Software (Chinese Academy of Sciences) and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
{zhang,hermanns}@depend.uni-saarland.de, {chengluo,zhanglj}@ios.ac.cn, rui.w.b9ce@m.isct.ac.jp, zhang@cispa.de

Abstract

Explainable AI (XAI) is crucial for building transparent and trustworthy machine learning systems, especially in high-stakes domains. Concept Bottleneck Models (CBMs) have emerged as a promising ante-hoc approach that provides interpretable, concept-level explanations by explicitly modeling human-understandable concepts. However, existing CBMs often suffer from poor locality faithfulness, failing to spatially align concepts with meaningful image regions, which limits their interpretability and reliability. In this work, we propose SL-CBM (CBM with Semantic Locality), a novel extension that enforces locality faithfulness by generating spatially coherent saliency maps at both concept and class levels. SL-CBM integrates a 1×1 convolutional layer with a cross-attention mechanism to enhance alignment between concepts, image regions, and final predictions. Unlike prior methods, SL-CBM produces faithful saliency maps inherently tied to the model's internal reasoning, facilitating more effective debugging and intervention. Extensive experiments on image datasets demonstrate that SL-CBM substantially improves locality faithfulness, explanation quality, and intervention efficacy while maintaining competitive classification accuracy. Our ablation studies highlight the importance of contrastive and entropy-based regularization for balancing accuracy, sparsity, and faithfulness. Overall, SL-CBM bridges the gap between concept-based reasoning and spatial explainability, setting a new standard for interpretable and trustworthy concept-based models.

Code — https://github.com/Uzukidd/sl-cbm
RIVAL10 — https://mmoayeri.github.io/RIVAL10/index.html
CUB — https://www.vision.caltech.edu/datasets/cub_200_2011/
PCBM — https://github.com/mertyg/post-hoc-cbm
CCS — https://github.com/NMS05/Improving-Concept-Alignment-in-Vision-Language-Concept-Bottleneck-Models
CLIP-ViT — https://huggingface.co/laion/CLIP-ViT-B-16-laion2B-s34B-b88K

Introduction

The widespread adoption of AI has heightened concerns around AI alignment, with explainable AI (XAI) increasingly recognized as central to ensuring aligned, transparent, and trustworthy systems. Among existing XAI techniques, saliency methods have gained popularity due to their intuitive, low-cost, and post-hoc nature. However, studies have shown that these methods often lack faithfulness and fail to deliver reliable explanations in high-stakes settings (Zhang, Figueroa, and Hermanns 2024; Kares et al. 2025). As a more faithful alternative, Concept Bottleneck Models (CBMs) (Koh et al. 2020) incorporate a layer of human-interpretable concepts between inputs and predictions, providing concept-level insights into model behavior. Unlike post-hoc methods, CBMs operate in an ante-hoc manner, ensuring that explanations directly reflect the model's decision-making process. They also support human intervention through concept correction (Chauhan et al. 2023). Due to their practicality, low conversion cost (Yuksekgonul, Wang, and Zou 2022), and recent extensions to multimodal settings (Selvaraj et al. 2024), CBMs are gaining increasing attention as a viable framework for explainable and aligned AI.

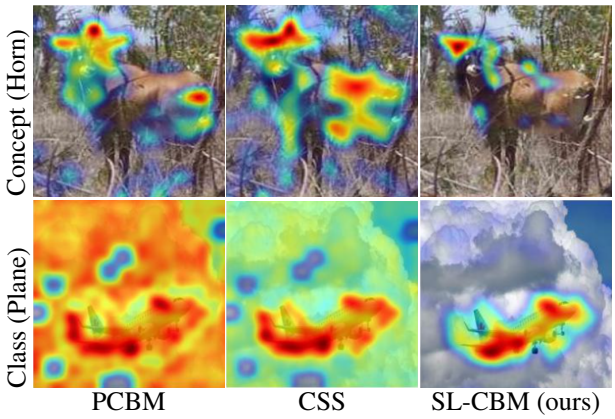

Figure 1: Saliency maps of state-of-the-art CBMs and SL-CBM at both concept and class levels. Saliency maps of PCBM (Yuksekgonul, Wang, and Zou 2022) and CSS (Selvaraj et al. 2024) are generated using GradCAM (Selvaraju et al. 2017), while SL-CBM produces its own saliency maps.

However, existing CBMs struggle with locality faithfulness, particularly in image-based tasks where concepts often fail to align with relevant image regions or contribute meaningfully to the final prediction (Margeloiu et al. 2021; Raman et al. 2023; Furby et al. 2023). Improving locality faithfulness is key to making CBM explanations more reliable and understandable, especially in high-risk settings, where it enables better human-guided interventions, though current CBM interventions mainly rely on automation, limiting human control. Efforts to improve concept trustworthiness have explored aligning concepts with classes and enforcing cross-layer/image alignment (Huang et al. 2024; Selvaraj et al. 2024), yet existing approaches remain limited in capturing meaningful alignment between concepts and images. To the best of our knowledge, our work is the first to directly address these limitations by explicitly enhancing the locality faithfulness of CBMs. As illustrated in Figure 1, existing methods such as PCBM (Yuksekgonul, Wang, and Zou 2022) and CSS (Selvaraj et al. 2024) often highlight irrelevant regions, *e.g.*, in generating saliency maps for the concept *Horn*, or even more drastically, for the class *Plane*.

Inspired by CAM (Zhou et al. 2016) and CBMs (Koh et al. 2020), we propose *CBM with Semantic Locality (SL-CBM)*, a simple yet effective structure to enforce locality faithfulness by aligning concept saliency maps with images' concept projection. SL-CBM generates both concept- and class-level saliency maps alongside predictions, with class saliency maps derived by linearly combining concept maps using class-specific weights. As shown in Figure 1, SL-CBM better localizes concepts like *Horn* and focuses on objects like *Plane* without highlighting irrelevant areas. Quantitative evaluation with XAI and localization metrics confirms SL-CBM's improved locality faithfulness, explainability, and intervention. In summary, our contributions are as follows:

- We propose SL-CBM, a model that generates saliency maps and concept-based explanations to enhance human understanding of the decision-making process.
- By enforcing locality faithfulness, SL-CBM enhances the alignment between image space, concept space, and class predictions, thereby improving the model's interpretability and reliability.
- We systematically evaluate SL-CBM's concept- and class-level accuracy, locality faithfulness, and intervention effectiveness, demonstrating improved explanation accuracy and faithfulness with potential to enhance human understanding and model refinement.

Related Work

Concept Bottleneck Models. The use of concept bottlenecks in deep neural networks for task-specific solutions or explainability is well-established (Yi et al. 2018; Chen, Bei, and Rudin 2020; Losch, Fritz, and Schiele 2019; Kim et al. 2018; De Santis et al. 2024). However, Koh et al. (2020) first formally defined CBMs as a concept project backbone network paired with a classifier. CBMs address three goals: *Interpretability* (identifying important concepts), *Predictability* (predicting targets from concepts), and *Intervenability* (improving predictions by replacing concept values with ground truth). Due to interpretability, intervenability, and adaptability (Dominici et al. 2024), CBMs have gained prominence in XAI.

Recent research focuses on enhancing CBMs through improved concept quality and broader applicability. One direction refines concept annotations using language models: *e.g.*, GPT-3-generated concept sets (Oikarinen et al. 2023) and LaBo's submodular selection of discriminative, CLIP-aligned concepts (Yang et al. 2023). Another approach replaces rigid concept definitions with *soft concepts*: PCBMs enable data-efficient conversion of pretrained models into CBMs (Yuksekgonul, Wang, and Zou 2022), while ProbCBM introduces probabilistic embeddings to handle data ambiguity (Kim et al. 2023). Further innovations include autoregressive concept predictors (Havasi, Parbhoo, and Doshi-Velez 2022) and ChatGPT-guided concept augmentation (Tan et al. 2024). Our research does not focus on concept quality; therefore, we use a predefined concept set across all CBM models for fair comparison. While performance may be improved with a refined concept set, this lies beyond the scope of our study.

Locality Faithfulness of CBMs. Despite progress, CBMs face persistent issues in *locality faithfulness*—ensuring concepts align with spatially or semantically localized input features. Studies reveal that CBMs often fail to learn localized concept representations: Margeloiu et al. (2021) find limited concept interpretability using Integrated Gradients (Sundararajan, Taly, and Yan 2017), while Raman et al. (2023) demonstrate poor spatial and semantic locality. Furby et al. (2023) corroborate this via LRP (Bach et al. 2015), showing concepts rarely map to distinct input regions. To address this, Huang et al. (2024) propose GradCAM-based (Selvaraju et al. 2017) evaluation of concept trustworthiness, and Selvaraj et al. (2024) advocate for locality alignment between concepts and classes. Interactive approaches, such as human-in-the-loop concept labeling (Chauhan et al. 2023), aim to improve faithfulness by grounding concepts in human oversight. These efforts primarily aim to enhance concept quality for better explanations. However, the issue of locality faithfulness has been identified but remains unaddressed. Our work aims to bridge this gap in the field.

Method

Problem Formulation

Consider Concept Bottleneck Models (CBMs) as a pair (f, g) consisting of a concept project network $f : \mathcal{X} \to \mathbb{R}^C$, which maps an input image $\mathbf{x} \in \mathcal{X}$ to a concept space $\mathbb{R}^C$ containing C predefined concepts, and a classifier $g : \mathbb{R}^C \to \mathbb{R}^K$, which maps the predicted concept embedding to one of K target classes. Let l_{gt} represent the ground truth class label of $\mathbf{x}$ and let $\mathcal{C}_{gt}$ denote the set of ground truth concept labels associated with this input. Let $k := |\mathcal{C}_{gt}|$ be its cardinality. Define the predicted set $\mathcal{P}$ as the concept indices of the top-k values of $f(\mathbf{x})$:

$$\mathcal{P} := \{i \in \mathcal{I} | f(\mathbf{x})_i \text{ is in the top-}k \text{ of the } \{f(\mathbf{x})_j\}_{j \in \mathcal{I}}\},$$

where $\mathcal{I}$ is the full set of candidate concept indices and $f(\mathbf{x})_j$ denotes the value of j^{th} concept c^j. Concept accu-

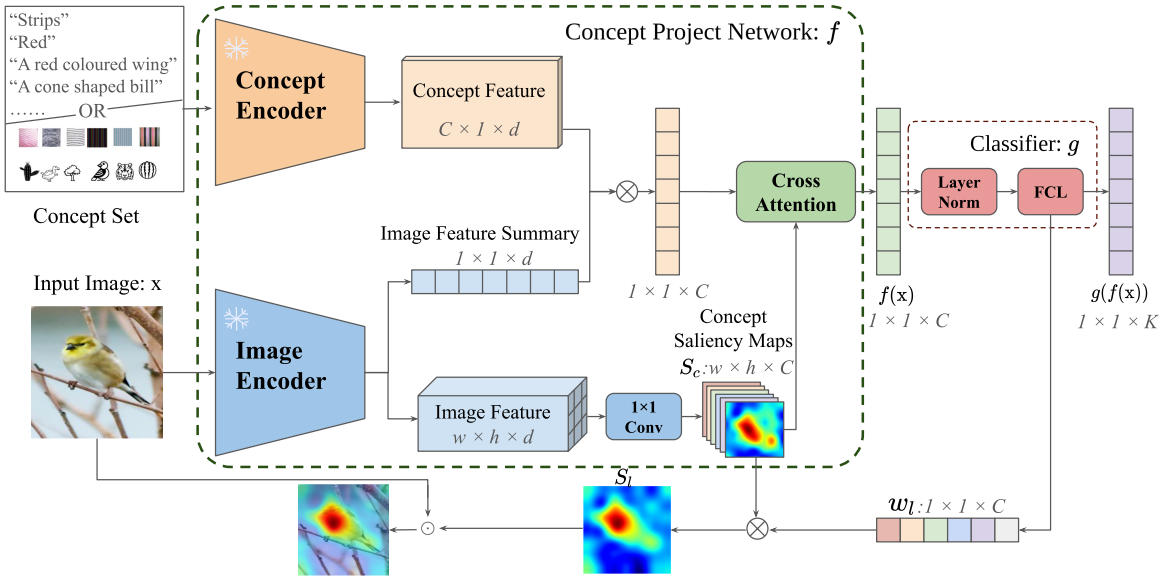

Figure 2: SL-CBM Overview: Given an input image $\mathbf{x}$, a concept set, the fixed concept and image encoders extract concept and image features, and an image feature summary. Projecting the image summary onto concept features yields a similarity vector. A 1×1 convolution generates concept saliency maps $S_\mathbf{c}$, which, with the similarity vector, are refined via cross-attention into $f(\mathbf{x})$. preserving locality and concept relevance. A classifier then produces logit $g(f(\mathbf{x}))$, and class saliency map S_l is computed by weighting $S_\mathbf{c}$ with the class-specific FCL weight $\mathbf{w}_l$.

racy is evaluated based on the overlap between the predicted set $\mathcal{P}$ and the ground truth set $\mathcal{C}_{gt}$, *i.e.*

$$\frac{|\mathcal{P} \cap \mathcal{C}_{gt}|}{k}.$$

The class predicted label is given by

$$l := \arg\max_i g(f(\mathbf{x}))_i,$$

where $g(\cdot)_i$ denotes the predicted logit value of the class i^{th}. The CBM prediction is considered correct when $l = l_{gt}$.

Let S_{c^i} denote the saliency map for the i^{th} concept c^i and let S_l denote the saliency map for the class l. The *locality faithfulness* of CBMs can be defined at two levels: concept-level and class-level. At the concept level, the locality faithfulness ensures that the saliency map of the i^{th} concept highlights the most relevant information for concept c^i. Thus when the input image is masked by the saliency map of concept i, it should emphasize the corresponding concept, formally expressed as

$$\arg\max_j f(\mathbf{x} \odot S_{c^i})_j = i.$$

At the class level, locality faithfulness aligns with traditional saliency maps, meaning that the saliency map should preserve the information relevant to the class prediction. This can be formulated as:

$$\arg\max_i g(f(\mathbf{x} \odot S_{l_{gt}}))_i = \arg\max_i g(f(\mathbf{x}))_i = l_{gt}.$$

Our objective is to maximize accuracy while enhancing locality faithfulness at both the concept and class levels.

CBM with Semantic Locality

To improve CBM interpretability, we propose *CBM with Semantic Locality (SL-CBM)*, which generates semantic saliency maps at both the concept and class levels. As Figure 2 shows, SL-CBM extends the concept projection network by adding a branch to learn concept saliency, enabling localized explanations at the concept level. The concept feature is derived from a dedicated concept encoder spanning the concept subspace. While preserving the conventional projection of input embeddings, *i.e.* image feature summary, onto the concept subspace, SL-CBM also derives a separate image feature for saliency generation from a shared image encoder. The concept-based image representation $f(x)$, used for classification, is generated by combining concept saliency maps and concept subspace projections via a cross-attention module, encouraging local, interpretable features.

Concept Project Network with Pre-trained Backbone. Existing CBMs operate in two main settings: (1) using a vision backbone with concept activation vectors (CAVs) (Kim et al. 2018) to learn concept vectors as in PCBM (Yuksekgonul, Wang, and Zou 2022), or (2) leveraging vision-language models for both image and concept embeddings. SL-CBM is compatible with any pretrained backbone, ensuring flexibility. With CNNs, it extracts image features

from the last convolutional layer and uses average pooling for the image summary. With transformers, it uses spatial tokens as image features and the CLS token as the summary.

Saliency Maps at Concept and Class levels. To enable the concept projection network f to generate saliency maps at both concept and class levels, we use a 1×1 *convolution* to learn weights over spatial features, producing concept saliency maps $S_{\mathbf{c}}$ that preserve semantic locality. To reinforce the faithfulness of concept saliency maps, we use a *cross-attention module* to promote alignment between the concept saliency maps $S_{\mathbf{c}}$ and image feature summary projected onto the concept subspace. For class-level explanations, we extract the classifier weight $\mathbf{w}_l$ from the Fully Connected Layer (FCL) for class l and linearly combine the concept saliency maps, *i.e.* $S_l := \sum \mathbf{w}_l S_{\mathbf{c}}$.

Loss Function. To ensure both concept- and class-level accuracy while maintaining the locality faithfulness of the generated saliency maps, SL-CBM is trained using the following loss functions: 1) **Class Accuracy**: We use standard *Cross-Entropy Loss*

$$\mathcal{L}_{ce} := -g(f(\mathbf{x}))_{l_{gt}} + \log \sum_j \exp(g(f(\mathbf{x})))_j)$$

to enforce the class precisions. 2) **Concept Accuracy**: *Concept Accuracy Loss* $\mathcal{L}_{ca}$ is defined as

$$\mathcal{L}_{ca} := \mathcal{L}_1(\gamma(s(f(\mathbf{x})) - s(\mathbb{1}(\mathcal{C}_{gt})))),$$

where $s(\cdot)$ is softmax function, $\mathcal{L}_1(\cdot)$ is standard mean absolute error, γ scales the loss, and $\mathbb{1}(\mathcal{C}_{gt})$ is the indicator vector of ground-truth concepts. 3) **Saliency Sparisity**: To encourage concise and meaningful explanations, *Entropy Loss*

$$\mathcal{L}_e := \sum_{i,j} H(s(S_{\mathbf{c}}^{(i,j)})),$$

is applied over spatial positions (i, j) of the concept saliency map $S_{\mathbf{c}}$, where $H(\cdot)$ demotes the entropy. 4) **Optional: Concept Consistency**: To further refine SL-CBM, we optionally include *Contrastive Loss*

$$\mathcal{L}_c := \frac{1}{n} \sum_{i,j} -\log \frac{e^{sim(f_i, f_j)/\tau}}{\sum_{m \neq i} e^{sim(f_i, f_m)/\tau}},$$

to promote intra-class similarity and inter-class distinction of concept embeddings. $sim(f_i, f_j)$ represents the similarity between the concpet embeddings $f(\mathbf{x})_i$ and $f(\mathbf{x})_j$, n is the number of examples (mini-batch), and τ is the temperature parameter. The total loss combines these terms,

$$\mathcal{L} := \lambda_{ce}\mathcal{L}_{ce} + \lambda_{ca}\mathcal{L}_{ca} + \lambda_e\mathcal{L}_e + \lambda_c\mathcal{L}_c,$$

balancing classification accuracy, concept fidelity, and saliency map interpretability.

Evaluating Locality Faithfulness

To fairly assess locality faithfulness, we use two setups: 1) with ground truth segmentation annotations at both the concept and class levels, and 2) without such annotations.

In the first scenario, saliency maps at both levels are denoted as S, and corresponding binary ground truth masks as M, where spatial locations contributing to a concept or class are marked 1, otherwise 0. Saliency maps S are binarized as $B := \begin{cases} 1 & \text{if } S^{(i,j)} > 0.5, \\ 0 & \text{otherwise.} \end{cases}$ Evaluation uses *Intersection over Union* (IoU) and *Dice coefficient*, measuring overlap between B and M. To address inflated scores caused by large saliency regions, we introduce *Compact IoU* (C-IoU), replacing the denominator with the area of B alone. Additionally, all metrics are weighted by image classification accuracy to integrate predictive correctness.

In the second scenario without annotations, saliency maps undergo max-min normalization per standard practice (Zhang, Figueroa, and Hermanns 2024). Faithfulness is assessed via Average Drop (AD), Average Increase (AI)(Chattopadhay et al. 2018), and Average Gain (AG)(Zhang et al. 2024). AD measures the decrease in class probability when masking the image, AI measures the fraction of cases where masking increases probability, and AG quantifies the overall gain in predictive power. AG is preferred as it reliably detects adversarial cases like FakeCAM (Poppi et al. 2021), unlike AD and AI, making it a more robust metric for saliency evaluation in explainable AI.

Experiments

Dataset. We use the RIVAL-10 dataset (Moayeri et al. 2022), which maps CIFAR-10 (Krizhevsky, Hinton et al. 2009) classes to ImageNet (Krizhevsky, Sutskever, and Hinton 2012) and contains around 26,000 images with 18 visual attributes and segmentations. We train on the training dataset (21,098 images) and evaluate on the test set (5,286 images), reporting all results on the test set. Additionally, we use the CUB-200-2011 dataset (Wah et al. 2011), a fine-grained categorization benchmark with 11,788 images across 200 bird subcategories, annotated with part locations, attributes, and bounding boxes. We train on 4,796 images and test on 5,794, following PCBM (Yuksekgonul, Wang, and Zou 2022) by using an imbalanced dataset sampler.

Models. We compare SL-CBM with two state-of-the-art CBMs: PCBM (Yuksekgonul, Wang, and Zou 2022) and CCS (Selvaraj et al. 2024). We evaluate PCBM with ResNet50 and ViT-B16 backbones, while CSS uses a ViT-B16-based CLIP model with its original training parameters. SL-CBM employs a ViT-B16-based CLIP model and is trained with a learning rate of 0.0003 using the Adam optimizer. By default, we set $\lambda_{ce} = 1$, $\lambda_{ca} = 10^4$, $\lambda_e = 5$, and $\lambda_c = 0$. To test CBM adaptability, we also use ResNet18[1] as a backbone, adjusting SL-CBM's λ_e to 1.0 while keeping other parameters unchanged.

Environment Settings. Experiments ran on Ubuntu 22.04.5 with a Xeon 8336C CPU, 125GB RAM, 127GB swap, and four RTX 4090 GPUs (24GB VRAM each), using Python 3.12.11, CUDA 12.4, and PyTorch 2.x with GPU acceleration. Unless noted, a single RTX 4090 was used.

[1]Pre-trained weights from torchcv (You et al. 2019)

Method		Accuracy		Interpretable Prediction		Locality Faithfulness											
						With Annotation						Without Annotation					
						IoU ↑		Dice ↑		C-IoU ↑		AD ↓		AI ↑		AG ↑	
		Concept	Class	NEC-5	ANEC	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$
PCBM	ResNet50	49.75	89.05	51.49	71.06	6.16	0.09	9.90	0.14	21.61	0.45	0.62	0.63	46.73	46.17	0.67	0.65
	ViT-B16	39.97	90.18	18.12	54.65	4.65	0.00	7.34	0.00	13.87	0.15	**0.55**	**0.57**	**65.38**	**63.64**	9.69	8.94
CSS		85.69	98.68	95.08	97.45	17.13	0.00	25.76	0.00	39.32	0.27	7.30	7.29	34.28	34.42	10.53	10.87
SL-CBM		**90.16**	**99.11**	**98.83**	**99.03**	**16.21**	**0.14**	**24.44**	**0.21**	**44.57**	**0.47**	3.65	3.68	47.87	48.07	**13.36**	**13.58**

Table 1: Comparison of SL-CBM with state-of-the-art CBMs on RIVAL-10 in terms of accuracy, as well as locality faithfulness metrics with annotation (IoU, Dice, C-IoU) and without annotation (AD, AI, AG) at both concept-level and class-level. ↑ signifies that a higher value is preferable for the metric, while ↓ indicates that a lower value is better. All values are presented as percentages. The best results are highlighted in bold.

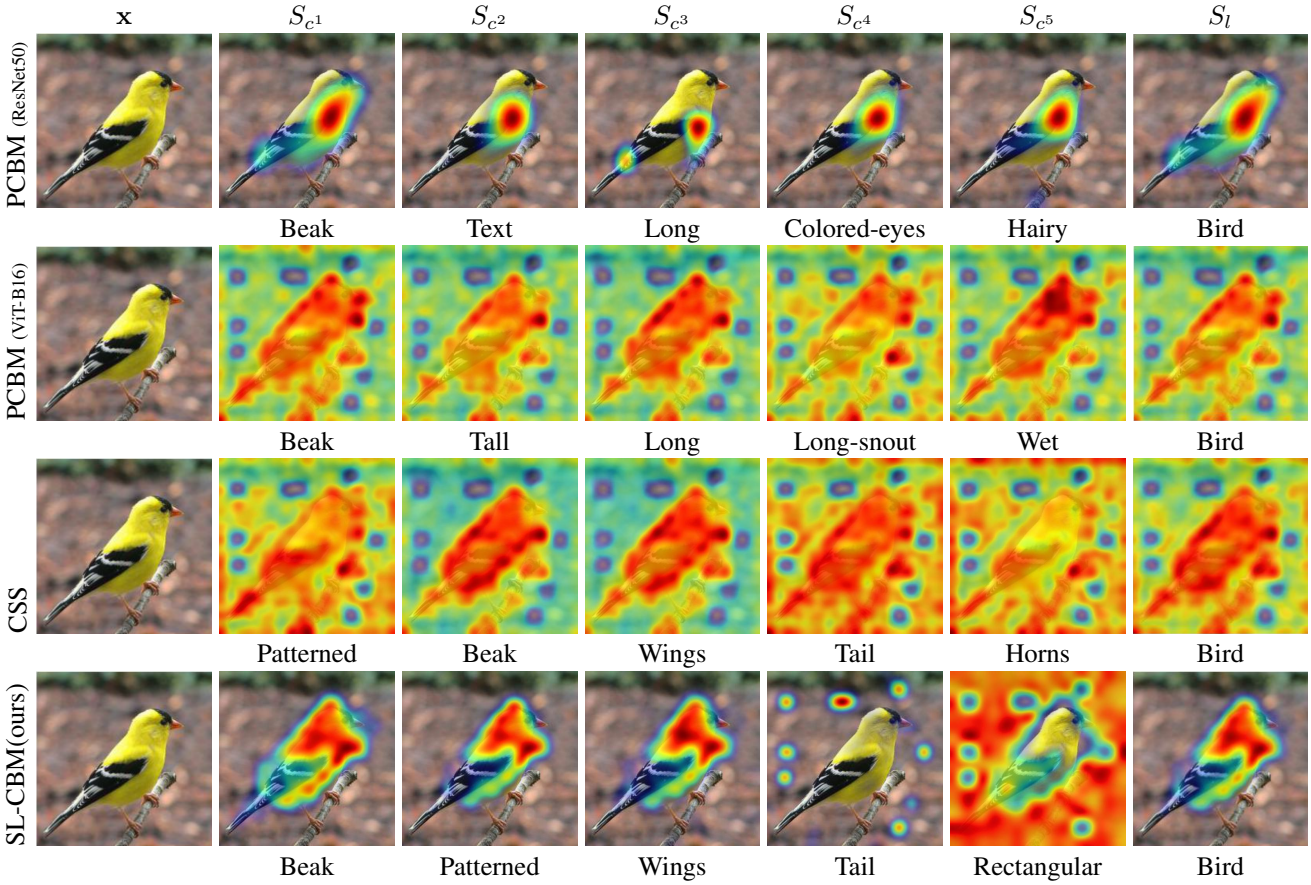

Figure 3: We present the saliency maps on RIVAL-10, showing the saliency maps for the top 5 concepts, where c^1 represents the top predicted concept, c^2 the second top, and so on, along with the predicted class l. The CBM saliency maps are generated using GradCAM (Selvaraju et al. 2017), while the SL-CBM saliency maps are produced by our proposed method.

Method	Accuracy		Interpretable Prediction		Locality Faithfulness					
					AD ↓		AI ↑		AG ↑	
	Concept	Class	NEC-5	ANEC	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$	$S_{c_{gt}}$	$S_{l_{gt}}$
PCBM	69.4	58.9	15.1	37.1	7.9	7.8	28.7	29.0	2.5	2.5
CSS	59.2	51.2	17.9	34.6	6.8	5.6	**57.5**	**58.8**	9.00	8.3
SL-CBM	**83.7**	**60.9**	**38.0**	**51.6**	**4.2**	**3.2**	49.9	51.1	**10.6**	**9.8**

Table 2: Comparison of SL-CBM with state-of-the-art CBMs on CUB in terms of accuracy, as well as locality faithfulness metrics without annotation (AD, AI, AG) at both concept-level and class-level. ↑ signifies that a higher value is preferable for the metric, while ↓ indicates that a lower value is better. All values are presented as percentages. The best results are highlighted in bold.

Evaluation Protocol. Beyond standard concept- and class-level accuracy, we assess interpretable prediction using NEC-5 and ANEC metrics (Srivastava, Yan, and Weng 2024), which measure accuracy based on decision-related concepts: NEC-5 uses the top 5, while ANEC averages over varying counts in $[5, 10, 15]$ for on RIVAL-10 and $[5, 10, 15, 20, 25, 30]$ for CUB. For locality faithfulness, we compute IoU, Dice coefficient, and C-IoU between saliency maps and segmentation masks, along with AI, AD, and AG metrics without segmentation masks. For CBMs lacking native saliency map outputs, we use Grad-CAM (Selvaraju et al. 2017) to generate them. Concept evaluation is restricted to saliency maps of ground truth concepts; class evaluation considers only the ground truth class.

Comparison

We compare SL-CBM with state-of-the-art CBMs on RIVAL-10, as it provides segmentation annotations at both the concept and class levels. This allows us to evaluate accuracy at both levels, along with locality faithfulness, using annotated and non-annotated metrics. The results are presented in Table 1. SL-CBM achieves the highest class- and concept-level accuracy as well as interpretable prediction, *i.e.* NEC-5 and ANEC. With segmentation annotations, SL-CBM surpasses other methods in locality faithfulness. In the absence of accurate annotations, we use AD, AI, and AG as surrogate metrics, where SL-CBM performs well overall, particularly excelling in AG, the most reliable measures of locality faithfulness without annotation.

SL-CBM effectively leverages pre-trained backbones. On CUB with a pre-trained ResNet18, we compare SL-CBM with other state-of-the-art CBMs using the same backbone. PCBM achieves strong and consistent performance in both accuracy and locality faithfulness in Table 2 compared to Table 1, while CSS shows degraded performance, with the lowest concept and class accuracy. SL-CBM outperforms others except in AI, which is less reliable than AG despite similar principles. Its superior locality faithfulness, especially in AG, highlights robustness across pre-trained backbones.

Visualization. Figure 3 presents saliency maps at both concept and class levels, highlighting the top five concepts per CBM. Compared to CNN and transformer-based PCBM models, CNN backbones produce more focused saliency maps. PCBM with ViT-B16, sharing the same backbone as CSS and SL-CBM, shows that GradCAM produces unreliable maps for both PCBM and CSS, lacking proper localization despite correct predictions. In contrast, SL-CBM improves both localization and prediction. For example, its saliency map for the concept *Break* identifies relevant regions, unlike other transformer-based CBMs, which often highlight the entire image. SL-CBM also produces more distinct concept maps than PCBM and CSS. While not perfect, SL-CBM inherently generates concept saliency maps, enabling failure diagnosis, *e.g.*, failing to learn *Tail* in this specific image. This diagnosis is unclear in other CBMs due to reliance on GradCAM.

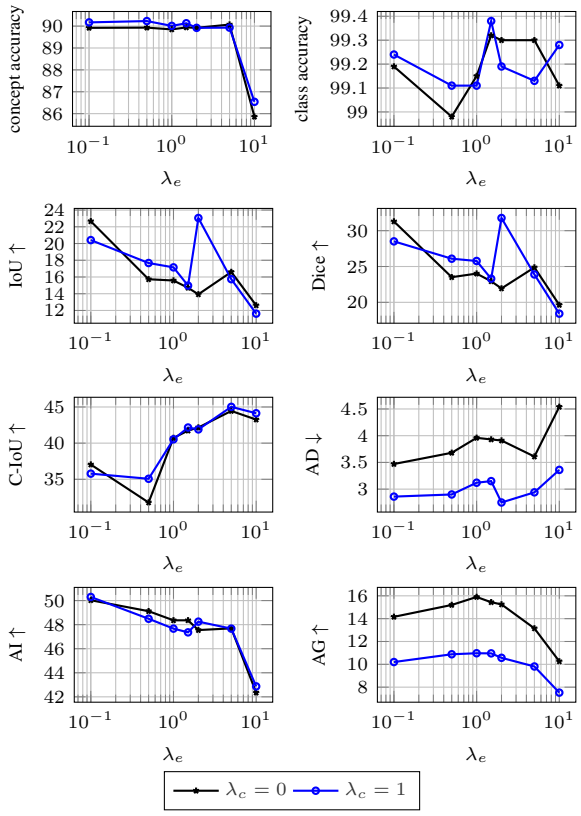

Figure 4: Ablation study on λ_e analyzing its effect on class accuracy, concept accuracy, IoU, Dice, and C-IoU with-/without $\mathcal{L}_c$ ($\lambda_c = 1$ or 0). Experiments are on RIVAL-10 with $\lambda_{ce} = 1$ and $\lambda_{ca} = 10^4$. ↑ signifies that a higher value is preferable for the metric, while ↓ indicates that a lower value is better. All values are presented as percentages.

Ablation

We perform an ablation study on λ_e and λ_c, while keeping $\lambda_{ce} = 1$ for Cross-Entropy Loss and $\lambda_{ca} = 10^4$ for Concept Accuracy Loss, following the optimal settings of

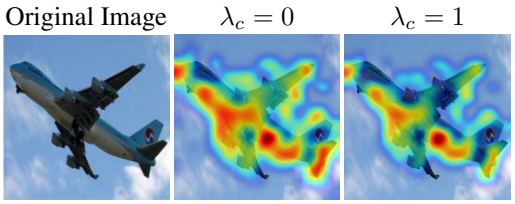

Figure 5: We present a top 1 concept saliency map on RIVAL-10 of $\lambda_c = 0$ or 1.

CCS (Selvaraj et al. 2024). Specifically, we vary λ_e for Entropy Loss over the set $\{0, 0.1, 0.5, 1.0, 1.5, 2.0, 5.0, 10.0\}$ and set λ_c for Contrastive Loss to either 1 or 0. We evaluate these configurations on RIVAL-10, focusing on concept-level metrics as concept and class locality faithfulness show similar trends. The results, depicted in Figure 4, reveal that including Contrastive Loss ($\lambda_c = 1$) leads to more unstable performance in class accuracy, IoU, and Dice scores, while maintaining comparable concept accuracy, C-IoU, and AI. Without Contrastive Loss ($\lambda_c = 0$), the model shows improved AG but reduced AD.

Visualization in Figure 5 further illustrates that when $\lambda_c = 0$, saliency maps highlight larger image regions, making AD and AG metrics more sensitive to these differences. Additionally, excessively large λ_e values cause significant performance degradation by producing overly sparse saliency maps that hinder sufficient learning. Balancing the trade-offs across all metrics,, we identify a local optimum at $\lambda_e = 5.0$. To demonstrate the adaptability of our method to new datasets and models, we conduct an experiment showing that parameters can be efficiently optimized using a small dataset and then successfully applied to the full dataset. Due to space constraints, these results are presented in the supplementary material.

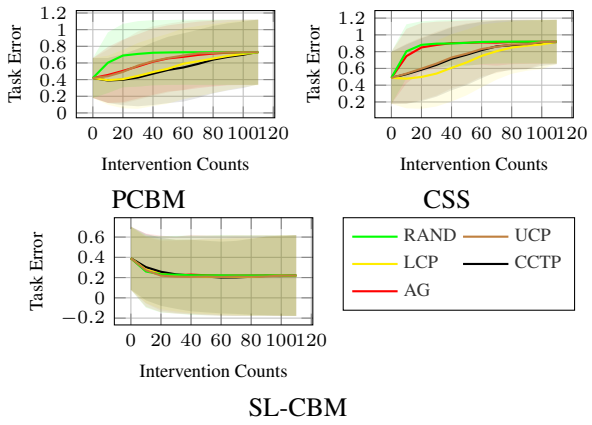

Figure 6: Intervention effectiveness on CUB: greater error reduction for the same number of concepts corrected.

Intervention

Intervention plays a critical role in correcting model outputs externally, thereby enhancing the model's applicability in real-world scenarios. Following the approach of Shin et al. (2023), we perform interventions by replacing predicted concepts with their ground truth values. We explore both random intervention (RAND) and guided interventions, where concepts are ranked using various metrics: Uncertainty in Concept Predictions (UCP), Loss on Concept Prediction (LCP), and Contribution of Concept to Target Prediction (CCTP). Additionally, we test the use of the explainability metric, *i.e*. AG, to rank and replace the top concepts with ground truth values. These experiments are conducted on the CUB dataset, which offers a sufficient number of concepts for meaningful intervention analysis. As shown in Figure 6, AG shows limited effectiveness as a metric for guiding intervention. Overall, only our proposed SL-CBM benefits from intervention, while PCBM and CSS experience degraded performance. When a model fails to learn concepts faithfully, intervention can be detrimental. In contrast, by enforcing local faithfulness, SL-CBM is better aligned with concept learning, making intervention more effective. For intervention counts, we test values in $[0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110]$ on CUB.

Conclusion

In this work, we present SL-CBM, a novel extension of CBMs that significantly enhances locality faithfulness by generating both concept-level and class-level saliency maps. While traditional CBMs offer concept-based explanations, they often lack meaningful spatial alignment between concepts and relevant image regions. SL-CBM overcomes this limitation through the integration of a 1×1 convolution and a cross-attention mechanism, which together improve spatial coherence and model interpretability. By providing saliency maps that faithfully reflect the model's internal reasoning, SL-CBM enables effective debugging: poor saliency quality signals concept learning failure, addressing a critical shortcoming of post-hoc explanation methods. Our experiments show that SL-CBM outperforms state-of-the-art CBMs in accuracy, locality faithfulness, and intervention performance, confirming that enforcing locality faithfulness improves concept faithfulness.

Moreover, our ablation studies reveal the essential role of contrastive loss and entropy regularization in balancing prediction accuracy, explanation faithfulness, and sparsity of saliency maps. Nevertheless, SL-CBM's effectiveness remains contingent on the quality of the predefined concept set, as it currently does not incorporate concept refinement strategies. Future work could explore integrating concept discovery and refinement to enhance explanation quality, or extend the approach toward applications in model oversight. Overall, SL-CBM represents a significant advancement in explainable AI by bridging concept-based reasoning with spatially-aware, faithful visual explanations. We believe this work lays a robust foundation for future exploration of trustworthy and interpretable concept-based models.

Acknowledgements

This work is supported by CAS Project for Young Scientists in Basic Research Grant YSBR-040, ISCAS New Cul-

tivation Project ISCAS-PYFX-202201, ISCAS Basic Research ISCAS-JCZD-202302, JST CREST JPMJCR21D3, and JSPS Grand-in-aid 23H00483. This work also received support from DFG under grant No. 389792660 as part of TRR 248[2], and funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 1010082337[3].

References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 839–847. IEEE.

Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; and Dvijotham, K. 2023. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5948–5955.

Chen, Z.; Bei, Y.; and Rudin, C. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12): 772–782.

De Santis, A.; Campi, R.; Bianchi, M.; and Brambilla, M. 2024. Visual-TCAV: Concept-based Attribution and Saliency Maps for Post-hoc Explainability in Image Classification. *arXiv preprint arXiv:2411.05698*.

Dominici, G.; Barbiero, P.; Giannini, F.; Gjoreski, M.; and Langhenirich, M. 2024. AnyCBMs: How to Turn Any Black Box into a Concept Bottleneck Model. *arXiv preprint arXiv:2405.16508*.

Furby, J.; Cunnington, D.; Braines, D.; and Preece, A. 2023. Towards a Deeper Understanding of Concept Bottleneck Models Through End-to-End Explanation. *arXiv preprint arXiv:2302.03578*.

Havasi, M.; Parbhoo, S.; and Doshi-Velez, F. 2022. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35: 23386–23397.

Huang, Q.; Song, J.; Hu, J.; Zhang, H.; Wang, Y.; and Song, M. 2024. On the Concept Trustworthiness in Concept Bottleneck Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21161–21168.

Kares, F.; Speith, T.; Zhang, H.; and Langer, M. 2025. What Makes for a Good Saliency Map? Comparing Strategies for Evaluating Saliency Maps in Explainable AI (XAI). *arXiv preprint arXiv:2504.17023*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

[2]CPEC:https://perspicuous-computing.science

[3]MISSION:https://mission-project.eu/

Kim, E.; Jung, D.; Park, S.; Kim, S.; and Yoon, S. 2023. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Losch, M.; Fritz, M.; and Schiele, B. 2019. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*.

Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; and Weller, A. 2021. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*.

Moayeri, M.; Pope, P.; Balaji, Y.; and Feizi, S. 2022. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19087–19097.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

Poppi, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2021. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2299–2304.

Raman, N.; Zarlenga, M. E.; Heo, J.; and Jamnik, M. 2023. Do Concept Bottleneck Models Obey Locality? In *XAI in Action: Past, Present, and Future Applications*.

Selvaraj, N. M.; Guo, X.; Kong, A. W.-K.; and Kot, A. 2024. Improving Concept Alignment in Vision-Language Concept Bottleneck Models. *arXiv preprint arXiv:2405.01825*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*.

Shin, S.; Jo, Y.; Ahn, S.; and Lee, N. 2023. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, 31504–31520. PMLR.

Srivastava, D.; Yan, G.; and Weng, L. 2024. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37: 79057–79094.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *ICML*.

Tan, Z.; Cheng, L.; Wang, S.; Yuan, B.; Li, J.; and Liu, H. 2024. Interpreting pretrained language models via concept bottlenecks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 56–74. Springer.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. 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*, 19187–19197.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

You, A.; Li, X.; Zhu, Z.; and Tong, Y. 2019. TorchCV: A PyTorch-Based Framework for Deep Learning in Computer Vision. https://github.com/donnyyou/torchcv.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*.

Zhang, H.; Figueroa, F. T.; and Hermanns, H. 2024. Saliency Maps Give a False Sense of Explanability to Image Classifiers: An Empirical Evaluation across Methods and Metrics. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Zhang, H.; Torres, F.; Sicre, R.; Avrithis, Y.; and Ayache, S. 2024. Opti-CAM: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding*, 248: 104101.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.