WSG-CBM: WEAKLY SUPERVISED GROUNDING FOR CONCEPT BOTTLENECK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept Bottleneck Models (CBMs) improve interpretability by introducing human-understandable concepts as an intermediate representation between inputs and predictions. However, accurate concept prediction does not necessarily imply faithful concept reasoning. A model may correctly predict a concept while relying on visual regions that are semantically unrelated to that concept, causing a gap between concept correctness and concept grounding. We argue that concept grounding is a missing ingredient in concept bottleneck learning. To address this challenge, we formulate concept grounding as a weakly supervised concept-specific spatial evidence recovery problem. We propose the Weakly Supervised Grounded Concept Bottleneck Model (WSG-CBM), which learns concept-specific spatial evidence using only image-level concept supervision. To recover meaningful concept evidence under weak supervision, we further introduce grounding objectives based on evidence concentration, disentanglement, and reliance. Unlike existing grounded concept learning approaches, WSG-CBM requires neither concept localization annotations, segmentation masks, bounding boxes, nor external vision-language supervision. Experiments demonstrate that WSG-CBM consistently provides localized and faithful spatial evidence while maintaining competitive concept and classification performance. Our results show that concept grounding can emerge from concept supervision alone and highlight the importance of grounding as a first-class objective for faithful concept-based reasoning.

1 INTRODUCTION

Concept Bottleneck Models (CBMs) (Koh et al., 2020) introduce human-understandable concepts as an intermediate representation between inputs and predictions, enabling concept-level inspection and intervention. By exposing concepts rather than latent features, CBMs are widely viewed as an interpretable alternative to deep neural networks. However, the interpretability of CBMs implicitly relies on a stronger assumption: a predicted concept should be supported by the spatial evidence it is intended to represent. This assumption is often *violated*. While concept labels indicate whether a concept is present, they do not specify where the corresponding spatial evidence occurs. As illustrated in Figure 1, a model may correctly predict a concept while relying on semantically irrelevant regions (Aysel et al., 2025), spurious contextual cues (Margeloiu et al., 2021), or correlated concepts (Havasi et al., 2022). As a result, CBMs provide concept-level explanations for class predictions, yet the input-to-concept mapping remains an uninterpretable black box. We therefore argue that concept grounding should be treated as a first-class learning objective.

Recent works have explored locality-aware concept learning by introducing external concept representations or pretrained vision-language models to guide concept localization (Huang et al., 2024; Raman et al., 2025; Benou & Raviv, 2025; Xie et al., 2025; Zhang et al., 2026b). While effective, these approaches rely on additional sources of supervision beyond image-level concept annotations. We instead ask a more fundamental question:

> *Can concept-specific spatial evidence be recovered from concept supervision alone?*

We formulate this challenge as a weakly supervised concept grounding problem. Specifically, for each concept, there exists a ground-truth spatial support that contains the spatial evidence responsi-

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.

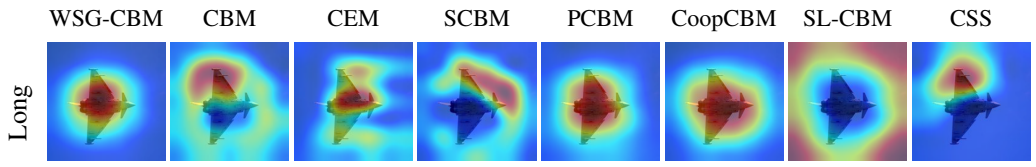

Figure 1: Qualitative comparison of concept grounding on RIVAL10. WSG-CBM directly learns concept-specific spatial evidence maps, whereas concept spatial evidence for other CBMs is obtained using Grad-CAM when explicit grounding maps are unavailable.

ble for that concept, and the objective is to recover such supports without localization annotations or external priors.

To address this, we propose *Weakly Supervised Grounded Concept Bottleneck Models (WSG-CBM)* for concept-specific spatial evidence recovery. The model first generates concept-specific spatial evidence maps and extracts a concept-agnostic evidence prior from the visual representation, then couples the two through evidence-guided aggregation. To recover meaningful concept-specific spatial evidence under weak supervision, we investigate three grounding objectives: (1) *concentration*, encouraging concepts to be supported by compact evidence regions; (2) *disentanglement*, encouraging different concepts to rely on distinct evidence; (3) *reliance*, requiring concept predictions to depend on their identified evidence. To instantiate the proposed grounding objectives, we introduce a set of grounding losses and systematically study their contribution to concept grounding. The results show that WSG-CBM consistently learns localized and faithful concept-specific spatial evidence while preserving competitive concept and classification performance.

Our contributions are three-fold:

- We identify concept grounding as a missing learning objective in concept bottleneck learning and formulate it as concept-specific spatial evidence recovery under weak supervision.
- We propose WSG-CBM, a weakly supervised concept grounding framework that learns concept-specific spatial evidence using only image-level concept annotations, without localization supervision, segmentation masks, bounding boxes, or vision-language priors.
- Experiments on CUB, AWA2, and RIVAL10 demonstrate that WSG-CBM consistently provides localized and faithful spatial evidence while maintaining competitive concept and classification performance.

2 Related Work

Faithfulness of Concept Bottleneck Models Concept Bottleneck Models (CBMs) (Koh et al., 2020) introduce human-interpretable concepts as an intermediate representation between inputs and predictions, enabling concept-level inspection and intervention. While early work primarily focused on concept prediction and downstream accuracy, subsequent studies have shown that accurate concepts do not necessarily imply faithful reasoning (Knab et al., 2026). In particular, concept predictors may exploit spurious visual cues or side-channel information, leading to concept leakage (Havasi et al., 2022; Parisini et al., 2026) and unreliable explanations. To address this issue, prior work has explored concept intervention (Koh et al., 2020; Espinosa Zarlenga et al., 2023; Shin et al., 2023), bottleneck regularization (Selvaraj et al., 2024), and causal concept modeling (Vandenhirtz et al., 2024) to improve robustness, intervention effectiveness, and explanation faithfulness. However, these works mainly focus on the faithfulness of the mapping from concept to prediction without asking whether a predicted concept is supported by the visual evidence it is intended to represent.

Spatially Grounded Concept Learning Recent work has increasingly questioned whether concept predictions are grounded in the visual evidence they are intended to represent. Margeloiu et al. (2021) first showed that concept explanations often fail to align with human-expected visual regions, a finding later confirmed by Aysel et al. (2025). Building on this observation, Huang et al. (2024) studied concept trustworthiness through spatial evidence, while Raman et al. (2025) analyzed the locality properties of concept predictors. More recent approaches explicitly incorporate spatial struc-

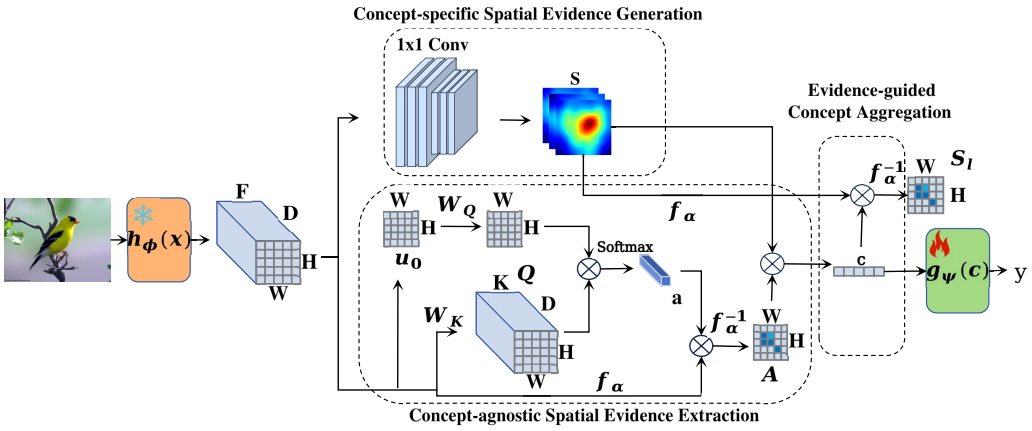

Figure 2: WSG-CBM overview. Concept-specific evidence maps $\mathcal{S}$ and a concept-agnostic evidence prior A are extracted from backbone features $\mathbf{F} = h_\phi(\boldsymbol{x})$ and combined through evidence-guided aggregation, resulting in grounded concept representations $\boldsymbol{c}$ for downstream prediction.

ture into concept learning, including DOT-CBM (Xie et al., 2025), which models concept-region interactions via optimal transport, SALF-CBM (Benou & Raviv, 2025), which leverages pretrained vision-language models, and SL-CBM (Zhang et al., 2026b), which introduces semantic locality constraints. However, these methods typically rely on external concept representations, additional supervision, or pretrained vision-language priors. In contrast, we study concept grounding in the standard supervised CBM setting and aim to recover concept-specific visual evidence using only image-level concept annotations, without localization supervision or external knowledge.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{(\boldsymbol{x}_n, \mathcal{C}_n, l_n)\}_{n=1}^N$ be a dataset, where $\boldsymbol{x}_n \in \mathcal{X}$ is an image, $\mathcal{C}_n \in \{0,1\}^C$ denotes C predefined concepts, and l_n is the class label. A Concept Bottleneck Model (CBM) consists of a concept predictor $f_\theta : \mathcal{X} \to \mathbb{R}^C$ and a class predictor $g_\psi : \mathbb{R}^C \to \mathbb{R}^K$. Given an image $\boldsymbol{x}$, the model predicts concept logits $\boldsymbol{c} = f_\theta(\boldsymbol{x})$ and class logits $\boldsymbol{y} = g_\psi(\boldsymbol{c})$.

Weakly Supervised Concept Grounding. Accurate concept prediction does not guarantee that a concept is grounded in the spatial evidence it is intended to represent. We therefore distinguish between *concept correctness* and *concept grounding*. Let $\Omega = \{1, \dots, H\} \times \{1, \dots, W\}$ denote the image lattice. For each concept r, we assume the existence of a ground-truth evidence region $Z_r \subseteq \Omega$, which contains the spatial evidence supporting concept r. Although concept labels C_r are available during training, the evidence regions $\{Z_r\}_{r=1}^C$ are never explicitly required. We formulate this concept grounding issue as a weakly supervised localization problem. The goal is to learn

$$f_\theta : \boldsymbol{x} \mapsto (\boldsymbol{c}, \mathcal{S}),$$

where

$$\mathcal{S} = \{S_r\}_{r=1}^C, \qquad S_r \in \mathbb{R}^{H \times W},$$

is a set of concept-specific spatial evidence maps. The learned representation should satisfy

$$\sigma(c_r) \approx P(C_r = 1 \mid \boldsymbol{x}), \qquad \text{(concept correctness)} \tag{1}$$

$$S_r(i,j) \propto P((i,j) \in Z_r \mid \boldsymbol{x}, C_r)\,, \qquad \text{(concept grounding)} \tag{2}$$

ensuring accurate concept prediction, while requiring predictions to be supported by concept-specific spatial evidence. Since the ground-truth evidence region Z_r is unavailable during training, concept grounding must be learned solely from image-level concept supervision.

3.2 Grounding by Evidence Aggregation

Recovering the concept-specific spatial evidence supports $\{Z_r\}_{r=1}^{C}$ from image-level concept supervision alone is highly under-constrained. A concept predictor may achieve high concept accuracy while relying on semantically irrelevant regions. To learn grounded concepts, concept predictions must be explicitly coupled with the spatial evidence that supports them. To achieve this, we propose the *Weakly Supervised Grounded Concept Bottleneck Model (WSG-CBM)*, shown in Figure 2. WSG-CBM consists of three stages: (1) generating concept-specific spatial evidence maps, (2) extracting a concept-agnostic evidence prior, and (3) aggregating concept predictions according to the inferred spatial evidence.

Concept-specific Spatial Evidence Generation. Given an input image $\boldsymbol{x}$, a pretrained visual backbone h_ϕ produces a spatial feature map

$$\mathbf{F} = h_\phi(\boldsymbol{x}) \in \mathbb{R}^{D\times H\times W}.$$

To preserve spatial information throughout concept prediction, we project the backbone representation into the concept space using a 1×1 convolution,

$$\mathcal{S} = \mathrm{Conv}_{1\times 1}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W},$$

where each map $S_r \in \mathbb{R}^{H\times W}$ corresponds to the spatial evidence of concept r.

Concept-agnostic Spatial Evidence Extraction Concept-specific spatial evidence maps alone do not indicate which image regions are emphasized by the visual representation. We therefore extract a concept-agnostic evidence prior from the same backbone features using SimPool (Psomas et al., 2023). Specifically, SimPool produces a spatial object evidence map $A \in \mathbb{R}^{H\times W}$, which captures where the backbone concentrates its discriminative evidence for the class prediction. Importantly, A is shared across concepts and does *not* localize any individual concept. Instead, it estimates object-level spatial evidence and provides a concept-agnostic prior for recovering concept-specific spatial evidence under weak supervision.

Evidence-Guided Concept Aggregation We obtain image-level concept predictions by aggregating concept spatial evidence according to the concept-agnostic spatial evidence prior:

$$c_r = \sum_{i=1}^{H}\sum_{j=1}^{W} A_{i,j} S_{r,i,j}, \tag{3}$$

namely, concepts receive high scores only when their spatial evidence aligns with the object-level spatial evidence prior, encouraging spatially grounded concept representations. The refined object spatial evidence is obtained by $S_l = f_\alpha^{-1}(\mathcal{S} \otimes \boldsymbol{c})$, where f_α is pooling operation defined in Psomas et al. (2023).

3.3 Grounding Objective

Although WSG-CBM encourages evidence-grounded concept prediction, recovering concept-specific spatial evidence maps from image-level supervision alone is inherently ambiguous. Similar to weakly supervised localization, many potential spatial evidence maps may explain the same concept label. Therefore, recognition supervision alone is insufficient to guarantee meaningful spatially grounded concepts. In addition to the binary cross-entropy concept loss and cross-entropy classification loss, we introduce three grounding objectives: **(1) Concentration**, encouraging compact concept evidence; **(2) Disentanglement**, encouraging distinct evidence across concepts; and **(3) Reliance**, requiring concept predictions to causally depend on their supporting evidence.

3.3.1 Concentration

Concept Compactness Loss. To encourage spatially compact evidence, we penalize activations that are far from the dominant evidence. Let $(i_r^*, j_r^*) = \arg\max_{i,j} S_{r,i,j}$ denote the peak activation of concept r. We define

$$\mathcal{L}_{\mathrm{CPT}} = \frac{1}{C}\sum_{r=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} \sigma(S_{r,i,j})\Big[(i-i_r^*)^2 + (j-j_r^*)^2\Big]. \tag{4}$$

Minimizing $\mathcal{L}_{\mathrm{CPT}}$ encourages concept evidence to concentrate around its dominant activation.

Concept-wise Spatial Entropy. Compact evidence should occupy only a small subset of spatial locations. For each concept map S_r, we construct a spatial probability distribution $p_r^{(i,j)} = \frac{\exp(S_{r,i,j}/\tau)}{\sum_{i',j'}\exp(S_{r,i',j'}/\tau)}$ and compute its entropy $H_r = -\sum_{i,j} p_r^{(i,j)} \log p_r^{(i,j)}$. We apply entropy minimization only to presented concepts:

$$\mathcal{L}_{\mathrm{ECW}} = \frac{\sum_r C_r H_r}{\sum_r C_r + \epsilon} - \frac{\sum_r (1 - C_r) H_r}{\sum_r (1 - C_r) + \epsilon}. \tag{5}$$

Minimizing $\mathcal{L}_{\mathrm{ECW}}$ encourages presented concepts to be localized to compact evidence regions.

3.3.2 Disentanglement

Channel-wise Entropy. To encourage different concepts to compete for spatial evidence, we normalize concept-specific spatial evidence across channels $q_r^{(i,j)} = \frac{\exp(S_{r,i,j}/\tau)}{\sum_{r'=1}^{C}\exp(S_{r',i,j}/\tau)}$, and minimize the resulting channel entropy

$$\mathcal{L}_{\mathrm{PEnt}} = -\frac{1}{C}\sum_{i}^{W}\sum_{j}^{H}\sum_{r=1}^{C} q_r^{(i,j)} \log q_r^{(i,j)}. \tag{6}$$

$\mathcal{L}_{PEnt}$ loss encourages each spatial location to be dominated by a small number of concepts.

Orthogonality loss. To encourage different concepts to rely on distinct evidence, we enforce orthogonality between concept-specific spatial evidence maps. Let $\hat{S} \in \mathbb{R}^{C \times HW}$ denote the flattened concept maps and $\bar{S} = \hat{S}/\|\hat{S}\|$ their normalized form. The orthogonality loss is

$$\mathcal{L}_{\mathrm{Orth}} = \left\| \bar{S}\bar{S}^{\top} - I \right\|_F^2, \tag{7}$$

where I is the identity matrix. $\mathcal{L}_{Orth}$ loss discourages overlap between concept-specific evidence and promotes evidence separation.

3.3.3 Reliance

Erase loss. To enforce evidence reliance, we require concept predictions to depend on their recovered evidence. For a positive concept r, we construct an erase mask $\bar{M}_r = \mathbb{1}[S_r < \tau]$, remove the corresponding evidence region, and recompute concept logits $\boldsymbol{c}'$. The erase objective is

$$\mathcal{L}_{\mathrm{Erase}} = -(c_r - c'_r) + \sum_{k \neq r}(c_k - c'_k)^2. \tag{8}$$

Minimizing $\mathcal{L}_{\mathrm{Erase}}$ encourages the target concept to be sensitive to the removal of its supporting spatial evidence while keeping unrelated concepts invariant, thereby promoting faithful spatial concept-evidence dependence.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate WSG-CBM on three widely used CBM benchmarks: CUB-200-2011 (Wah et al., 2011), AWA2 (Xian et al., 2019), and RIVAL10 (Moayeri et al., 2022). CUB-200-2011 contains nearly 12K bird images annotated with 112 semantic attributes, AWA2 contains over 37K animal images annotated with 85 attributes, and RIVAL10 contains approximately 26K images with 18 human-defined concepts and class-level segmentation masks. We evaluate the localization of CUB-200-2011 with part segmentation annotations from the CUB70 test set (Behzadi-Khormouji & Oramas, 2023), comprising 1,888 images.

Models We compare our method to CBM (Koh et al., 2020), CEM (Espinosa Zarlenga et al., 2022), SCBM (Vandenhirtz et al., 2024), CoopCBM (Sheth & Ebrahimi Kahou, 2023), PCBM Yuksekgonul et al. (2022), CSS Selvaraj et al. (2024), SL-CBM (Zhang et al., 2026a) with their official implementation. For WSG-CBM, we use a pretrained ResNet-18 (He et al., 2016) as the default frozen backbone.

Evaluation Protocol We evaluate prediction correctness, localization, faithfulness, and intervention effectiveness. Concept correctness is measured by class (Acc_y) and concept (Acc_c) accuracy and interpretable prediction (NEC-5 and ANEC metrics Srivastava et al. (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 RIVAL-10, and $[5, 10, 15, 20, 25, 30]$ for AWA2 and CUB. Localization is evaluated through the spatial alignment between concept activation maps and external localization annotations using IoU-based metrics (IoU, Dice, C-IoU, *cf*. subsection A.3). Faithfulness is assessed using perturbation-based measures, including Average Drop (AD), Average Increase (AI) (Selvaraju et al., 2017), Average Gain (AG) (Zhang et al., 2023), and ROAD (Rong et al., 2022). Finally, intervention effectiveness (Koh et al., 2020) is measured by replacing predicted concepts with ground-truth concepts during inference and reporting the resulting classification accuracy. For methods that do not provide internal concept grounding maps, we employ Grad-CAM (Selvaraju et al., 2017) to generate concept- ($S_{C_{gt}}$) and class-level ($S_{l_{gt}}$) spatial evidence, which is subsequently used for localization evaluation.

Grounding Objective			Accuracy		NEC-5	IoU ↑		AG ↑	
Concentration	Disentanglement	Reliance	$\mathrm{Acc_c}$	$\mathrm{Acc_y}$		$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$
×	×	×	89.46	75.06	33.34	7.60	25.61	19.82	21.05
$\mathcal{L}_{\mathrm{ECW}}$	×	×	89.40	75.01	36.47	10.74	30.75	20.14	21.36
$\mathcal{L}_{\mathrm{CPT}}$	×	×	88.91	74.09	37.11	8.90	27.62	17.07	18.28
×	$\mathcal{L}_{\mathrm{PEnt}}$	×	88.93	73.94	37.81	7.55	2575	18.47	19.71
×	$\mathcal{L}_{\mathrm{Orth}}$	×	88.83	73.83	38.49	7.50	26.02	18.76	19.69
×	×	$\mathcal{L}_{\mathrm{Erase}}$	87.93	74.02	36.45	7.24	24.22	13.72	14.28
$\mathcal{L}_{\mathrm{ECW}}$	$\mathcal{L}_{\mathrm{PEnt}}$	×	87.03	73.33	40.25	9.14	33.68	19.15	19.92
$\mathcal{L}_{\mathrm{CPT}}$	$\mathcal{L}_{\mathrm{PEnt}}$	×	88.92	73.87	38.11	7.72	26.07	18.35	19.58
$\mathcal{L}_{\mathrm{ECW}} + \mathcal{L}_{\mathrm{CPT}}$	×	×	87.03	73.35	40.01	9.14	33.72	19.22	19.98
$\mathcal{L}_{\mathrm{ECW}} + \mathcal{L}_{\mathrm{CPT}}$	$\mathcal{L}_{\mathrm{PEnt}} + \mathcal{L}_{\mathrm{Orth}}$	$\mathcal{L}_{\mathrm{Erase}}$	78.69	69.90	22.56	9.38	39.54	9.61	9.27
$\mathcal{L}_{\mathrm{ECW}} + \mathcal{L}_{\mathrm{CPT}}$	$\mathcal{L}_{\mathrm{PEnt}} + \mathcal{L}_{\mathrm{Orth}}$	×	86.98	73.20	39.75	9.16	33.91	19.18	19.89
$\mathcal{L}_{\mathrm{ECW}} + \mathcal{L}_{\mathrm{CPT}}$	$\mathcal{L}_{\mathrm{PEnt}}$	×	88.79	74.80	36.73	11.09	35.74	22.13	22.94

Table 1: Ablation study of grounding objectives on CUB.

4.2 Ablation on Grounding Objectives

We instantiate the proposed grounding objectives using several grounding losses and evaluate their individual contributions through ablation studies. Unless otherwise specified, Table 1 uses $\lambda_{\mathrm{ECW}} = 10$, $\lambda_{\mathrm{CPT}} = 10^{-6}$, $\lambda_{\mathrm{PEnt}} = 0.005$, $\lambda_{\mathrm{Orth}} = 1$, and $\lambda_{\mathrm{Erase}} = 1$ as weight for corresponding loss. Additional ablations, including loss-weight sensitivity analyses, are provided in subsection B.1.

As shown in Table 1, we observe three notable trends: (1) losses for the concentration provide the largest gains in concept grounding ($\mathcal{L}_{\mathrm{ECW}}$ provides the largest gain), indicating that recovering compact evidence is fundamental under weak supervision; (2) losses for disentanglement do not improve grounding on their own; they become effective after concept spatial evidence has been concentrated, suggesting that disentanglement is most beneficial once compact concept supports have been established; (3) the loss for reliance considerably degrades predictive performance, suggesting that enforcing causal dependence through evidence erasure is challenging in the weakly supervised setting. Overall, the combination of concentration (both $\mathcal{L}_{\mathrm{ECW}}$ and $\mathcal{L}_{\mathrm{CPT}}$) and disentanglement ($\mathcal{L}_{\mathrm{PEnt}}$) yields the best trade-off between prediction performance, grounding quality, and faithfulness, and is therefore adopted in the final model.

Method	RIVAL10				AWA2				CUB			
	Acc_c	Acc_y	NEC-5	ANEC	Acc_c	Acc_y	NEC-5	ANEC	Acc_c	Acc_y	NEC-5	ANEC
CBM	**89.59**	97.05	93.62	95.67	96.25	<u>93.43</u>	<u>57.82</u>	76.01	86.22	68.62	20.75	37.59
CEM	88.02	<u>98.41</u>	51.55	56.43	<u>96.70</u>	90.70	**78.58**	**87.21**	87.86	72.07	17.86	52.87
SCBM	87.74	95.19	65.21	85.37	**97.86**	91.99	44.25	<u>79.80</u>	<u>88.19</u>	67.97	4.02	22.83
CoopCBM	84.78	**98.90**	86.04	94.56	87.92	**93.73**	45.43	68.67	42.65	**75.26**	26.65	53.00
PCBM	83.65	94.82	51.23	60.09	83.39	82.28	33.17	53.43	68.81	57.49	18.16	38.36
CSS	85.57	96.35	94.15	95.42	87.20	86.93	52.10	70.23	57.34	60.43	25.65	25.65
SL-CBM	<u>88.79</u>	97.77	<u>96.24</u>	<u>97.02</u>	93.19	85.20	54.87	75.86	81.80	58.96	<u>30.93</u>	<u>45.36</u>
WSG-CBM	87.77	97.75	**96.29**	**97.21**	87.49	92.47	56.05	77.66	**88.61**	<u>74.63</u>	**35.23**	**63.46**

Table 2: Acc_c prediction and intervention performance. Best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

Method		RIVAL10				AWA2				CUB			
		AD↓	AI↑	AG↑	ROAD↓	AD↓	AI↑	AG↑	ROAD↓	AD↓	AI↑	AG↑	ROAD↓
CBM	$S_{C_{gt}}$	8.99	26.75	7.32	72.96	**4.69**	32.43	11.07	60.67	15.53	35.25	10.66	40.88
	$S_{l_{gt}}$	9.35	26.09	7.24	63.68	**4.32**	33.75	11.91	**84.86**	10.95	41.71	13.32	75.42
CEM	$S_{C_{gt}}$	40.14	26.88	19.51	91.70	11.00	32.67	20.40	74.42	15.39	29.33	9.31	39.50
	$S_{l_{gt}}$	36.62	26.44	19.28	51.87	6.21	28.05	17.51	90.99	7.94	37.59	12.37	<u>62.84</u>
SCBM	$S_{C_{gt}}$	30.70	26.21	17.24	76.30	24.31	16.22	10.99	59.56	16.86	36.34	<u>11.61</u>	39.25
	$S_{l_{gt}}$	11.68	31.20	21.55	64.62	11.75	20.42	14.05	92.94	10.75	44.19	<u>14.17</u>	83.58
CoopCBM	$S_{C_{gt}}$	8.99	44.62	13.62	82.99	6.93	38.64	10.80	61.33	**5.83**	**56.94**	7.77	37.73
	$S_{l_{gt}}$	10.70	41.04	12.54	<u>61.04</u>	6.61	39.82	11.27	96.15	**2.64**	**55.33**	2.99	83.59
PCBM	$S_{C_{gt}}$	7.90	36.36	9.98	**68.84**	8.27	38.75	10.11	**51.94**	10.23	26.72	2.64	34.41
	$S_{l_{gt}}$	8.04	35.38	9.55	65.69	8.32	39.32	10.53	<u>84.99</u>	8.06	28.46	2.50	78.25
CSS	$S_{C_{gt}}$	**7.17**	**50.84**	<u>21.92</u>	76.06	9.49	40.49	**21.70**	<u>52.90</u>	45.70	24.34	9.45	<u>29.63</u>
	$S_{l_{gt}}$	**6.04**	**51.74**	<u>21.48</u>	72.77	8.29	41.91	**22.61**	94.17	30.63	26.15	6.63	81.56
SL-CBM	$S_{C_{gt}}$	9.04	47.22	18.37	89.26	9.98	<u>43.73</u>	19.35	92.21	6.88	44.57	11.10	29.96
	$S_{l_{gt}}$	9.00	46.86	17.70	91.07	9.82	<u>43.82</u>	19.21	99.11	<u>4.79</u>	46.90	10.28	90.73
WSG-CBM	$S_{C_{gt}}$	<u>7.25</u>	<u>48.48</u>	**23.19**	<u>70.62</u>	<u>6.55</u>	**45.10**	<u>21.38</u>	58.83	<u>6.31</u>	<u>48.75</u>	**22.51**	**20.67**
	$S_{l_{gt}}$	<u>7.16</u>	<u>48.56</u>	**23.24**	**54.16**	<u>6.15</u>	**46.23**	<u>21.86</u>	87.31	5.94	<u>49.91</u>	**23.34**	**40.63**

Table 3: Comparison of faithfulness of spatial evidence (*i.e.* saliency maps). Best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

4.3 Comparison

Class and Concept Recognition. Table 2 highlights the recognition performance of WSG-CBM in comparison to current state-of-the-art methods. We observe that our approach sustains competitive results in most datasets, retaining a position among the top-3 performing models. In particular, we draw attention to the performance in NEC-5 and ANEC as they measure the quality of the concept representations produced by the bottleneck. On CUB and RIVAL10 we observe that WSG-CBM attains the best performance, while on AWA2, CEM outperforms all other methods. These results indicate that grounded concepts are more useful for downstream reasoning than concepts that are merely predictive. The discrepancy between concept accuracy and NEC metrics suggests that concept correctness alone is insufficient for faithful concept reasoning; grounded concepts are substantially more actionable for downstream decision-making.

Faithfulness of Spatial Evidence. Table 3 reveals that WSG-CBM consistently achieves competitive performance, ranking among the top methods across most faithfulness metrics. The gains are

Method	RIVAL10						CUB					
	IoU ↑		Dice ↑		C-IoU ↑		IoU ↑		Dice ↑		C-IoU ↑	
	$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$	$S_{C_{gt}}$	$S_{l_{gt}}$
CBM	<u>24.73</u>	31.34	<u>34.32</u>	43.68	36.96	66.20	5.34	17.64	9.27	26.76	5.92	46.07
CEM	9.64	17.07	14.62	25.20	<u>50.87</u>	67.67	<u>11.08</u>	25.52	<u>17.42</u>	36.50	14.84	71.15
SCBM	9.20	28.79	14.00	42.08	48.77	54.88	6.09	19.88	10.37	30.28	6.92	38.59
CoopCBM	21.47	20.07	30.13	27.73	32.07	<u>80.36</u>	2.25	22.36	3.87	32.96	2.83	45.78
PCBM	19.98	<u>42.19</u>	28.23	<u>57.07</u>	28.97	61.50	7.83	29.33	12.58	41.48	10.62	56.06
CSS	10.06	25.09	16.42	38.23	25.36	64.91	4.94	<u>31.82</u>	8.03	<u>46.70</u>	6.98	65.42
SL-CBM	0.03	0.16	0.06	0.18	**69.96**	**89.62**	8.88	20.40	14.34	29.93	<u>11.85</u>	<u>72.28</u>
WSG-CBM	**25.76**	**43.51**	**35.44**	**58.85**	40.31	78.83	**11.11**	**34.22**	**17.65**	**48.18**	**13.12**	**74.22**

Table 4: Comparison of localization of spatial evidence (*i.e.* saliency maps). Best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

most pronounced on CUB, where AG increases from 11.10 to 22.51, and ROAD decreases from 29.96 to 20.67 compared with SL-CBM. The combination of low AD and high AI/AG suggests that the recovered concept-specific spatial evidence captures regions that are both predictive and relevant to the model's decision process. Moreover, the strong ROAD performance demonstrates that the identified evidence remains informative under progressive perturbations, providing further evidence that WSG-CBM learns concept representations supported by meaningful spatial evidence rather than spurious correlations.

Localization of Spatial Evidence. Table 4 reveals that WSG-CBM consistently achieves the strongest concept localization performance, obtaining the best IoU and Dice scores on both RIVAL10 and CUB. CBMs with strong concept prediction performance (*cf*. Table 2) can still exhibit poor grounding quality. In particular, SL-CBM achieves near-zero concept IoU on RIVAL10 despite competitive NEC performance, highlighting the distinction between concept correctness and concept grounding. WSG-CBM improves both concept-level and class-level localization; together with the faithfulness results (*cf*. Table 3), these findings support our central hypothesis that grounded concepts produce more faithful concept-based explanations.

Visualization. To assess concept grounding qualitatively, we visualize the recovered spatial evidence maps for ground-truth concepts. Figure 3 presents a representative example from CUB, illustrating how different methods identify concept-specific spatial evidence for concept prediction as well as aggregated object-level evidence. Due to space limitations, we focus our qualitative comparison on SL-CBM, which, like WSG-CBM, explicitly learns concept-specific spatial evidence. Unlike WSG-CBM, however, SL-CBM benefits from additional information beyond image-level concept supervision. Qualitative analysis suggests that recovering concept-specific spatial evidence is substantially harder than recovering object-level spatial evidence. When multiple concepts co-occur, their evidence maps often overlap and align with the same object-level support. Although the disentanglement objectives mitigate this effect, achieving fine-grained concept separation remains an open challenge for weakly supervised concept grounding.

4.4 Intervention

We evaluate the intervention effectiveness of different CBMs by replacing predicted concepts with their ground-truth annotations at test time (Koh et al., 2020). Following the setting from Shin et al. (2023), we select concepts via random intervention (RAND), uncertainty-based concept prediction (UCP), concept prediction loss (LCP), and contribution to target prediction (CCTP). Figure 4 reports classification accuracy under varying intervention budgets, providing a measure of how effectively each model utilizes corrected concept information. It shows WSG-CBM achieves the best performance. Additional intervention results for AWA2 and RIVAL10 are provided in subsection B.2.

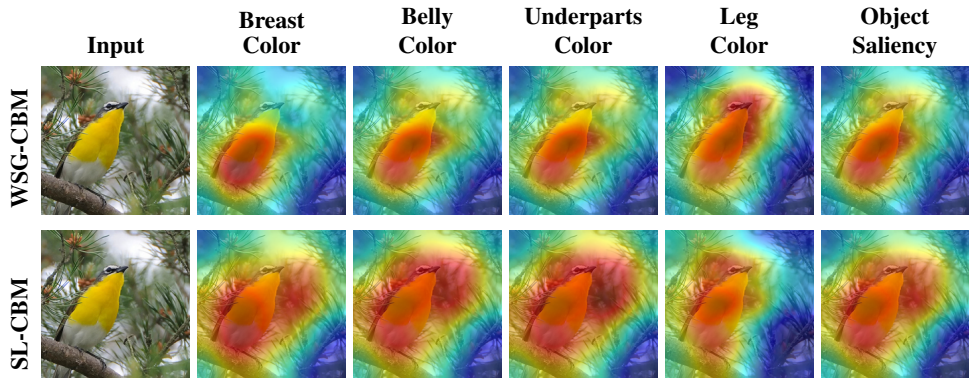

Figure 3: Qualitative comparison of concept and object localization on CUB. For each method, we show the input image, four concept-specific spatial evidence maps, and the corresponding object-level evidence map. More visualizations across CBMs and datasets are provided in Appendix C.

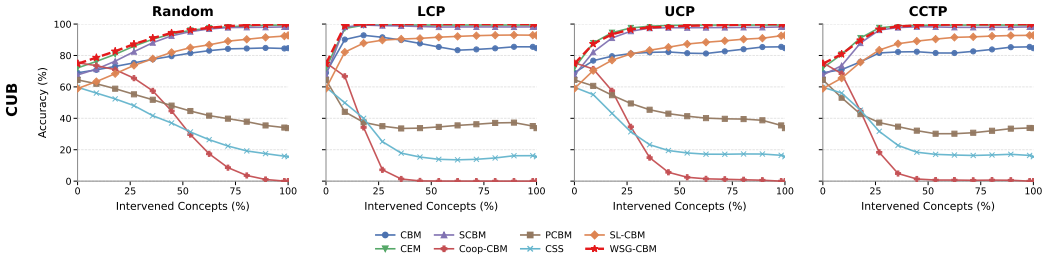

Figure 4: Concept intervention results on CUB. We report task accuracy as the proportion of intervened concepts increases under four intervention strategies: Random, LCP, UCP, and CCTP.

5 Conclusion

We argue that concept correctness alone is insufficient for faithful concept-based reasoning. A concept predictor may correctly recognize a concept while relying on spatial evidence that is inconsistent with the concept's intended semantic meaning. To address this gap, we formulate concept grounding as a weakly supervised concept-specific spatial evidence recovery problem and propose the Weakly Supervised Grounded Concept Bottleneck Model (WSG-CBM). WSG-CBM learns concept-specific spatial evidence maps using only image-level concept supervision by coupling concept-specific spatial evidence with a concept-agnostic evidence prior through evidence-guided aggregation. To recover meaningful concept evidence under weak supervision, we further introduce grounding objectives based on concentration, disentanglement, and reliance. Experiments across multiple datasets demonstrate that WSG-CBM consistently provides concept-specific evidence which achieve strong localization and faithfulness while maintaining strong concept and class prediction performance. Unlike existing approaches, our framework requires neither concept localization annotations nor external vision-language supervision, showing that concept grounding can emerge from concept supervision alone. More broadly, our findings suggest that concept grounding is a missing ingredient in concept bottleneck learning. We hope this work motivates future research toward learning concepts that are not only correct, but also grounded in the spatial evidence they are intended to represent. While WSG-CBM narrows the gap between concept correctness and concept grounding, recovering fine-grained and disentangled concept evidence remains an important direction for future work.

AI use statement. In this work, we used generative AI tools to aid polish writing, visualization, and implementation. We have not used generative AI tools for research ideation, or to draft the paper, and generating synthetic datasets, and proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated polished text, visualizations,

and code for potential plagiarism and correctness by 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Halil Ibrahim Aysel, Xiaohao Cai, and Adam Prugel-Bennett. Concept-based explainable artificial intelligence: Metrics and benchmarks. *arXiv preprint arXiv:2501.19271*, 2025.

Hamed Behzadi-Khormouji and José Oramas. A protocol for evaluating model interpretation methods from visual explanations. *Winter Conference on Applications of Computer Vision (WACV)*, 2023.

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

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 21400–21413, 2022.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

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*, pp. 770–778, 2016.

Qihan Huang, Jie Song, Jingwen Hu, Haofei Zhang, Yong Wang, and Mingli Song. On the concept trustworthiness in concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21161–21168, 2024. doi: 10.1609/aaai.v38i19.30109.

Patrick Knab, David Steinmann, Christian Bartelt, Kristian Kersting, Bernt Schiele, Thomas Seidl, Udo Schlegel, and Wolfgang Stammer. What's in the bottle? a survey and roadmap of concept bottleneck models. *Transactions on Machine Learning Research*, 2026.

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*, pp. 5338–5348, 2020.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. 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*, June 2022.

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D MacArthur, and Christopher RS Banerji. Leakage and interpretability in concept-based models. *Journal of Machine Learning Research*, 27(174):1–40, 2026.

Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it simpool: Who said supervised transformers suffer from attention deficit? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5350–5360, 2023.

Naveen Janaki Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models respect localities? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

Nithish Muthuchamy Selvaraj, Xiaobao Guo, Adams Wai-Kin Kong, and Alex Kot. Improving concept alignment in vision-language concept bottleneck models. *arXiv preprint arXiv:2405.01825*, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017.

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. *Advances in Neural Information Processing Systems*, 36:26966–26990, 2023.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International conference on machine learning*, pp. 31504–31520. PMLR, 2023.

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

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E Vogt. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:51787–51810, 2024.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

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 (CVPR)*, pp. 30199–30209, June 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *arXiv preprint arXiv:2301.07002*, 2023.

Hanwei Zhang, Luo Cheng, Rui Wen, Yang Zhang, Lijun Zhang, and Holger Hermanns. Sl-cbm: Enhancing concept bottleneck models with semantic locality for better interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38093–38101, 2026a.

Hanwei Zhang, Luo Cheng, Rui Wen, Yang Zhang, Lijun Zhang, and Holger Hermanns. Sl-cbm: Enhancing concept bottleneck models with semantic locality for better interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38093–38101, 2026b. doi: 10.1609/aaai.v40i44.41147.