BEYOND CONCEPT ACCURACY: SPATIAL GROUNDING IN CONCEPT BOTTLENECK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept Bottleneck Models (CBMs) aim to make predictions interpretable by decomposing reasoning into human-interpretable concepts and downstream decisions. While existing CBM research has extensively studied whether task predictions faithfully follow concept representations, it has paid less attention to how concept predictions depend on spatially localized information. Consequently, concept accuracy alone provides limited insight into the spatial evidence supporting a concept. To this end, we introduce a spatial grounding framework for concept representations that characterizes concept support through two complementary properties: spatial sufficiency and spatial necessity. Spatial sufficiency measures whether a region contains enough information to preserve a concept prediction, while spatial necessity measures whether removing that region disrupts the prediction. Together, these measures characterize how concept predictions depend on spatially localized information within the concept-generating representation. Across CUB-200-2011 and RIVAL-10, we find that concept and task accuracy do not determine spatial grounding; comparable predictive performance yields substantially different grounding behavior. Grounding depends on both the concept bottleneck and backbone representation, while joint concept-task optimization does not substantially change grounding. Finally, we show that attribution-based localization and spatial grounding capture distinct properties of concept representations. These findings establish spatial grounding as a complementary dimension for evaluating the interpretability and faithfulness of visual CBMs.

1 INTRODUCTION

Concept Bottleneck Models (CBMs) (Koh et al., 2020; Knab et al., 2026) decompose prediction into two steps. The first maps an input to a set of human-understandable concepts, and the second maps these concepts to a task label. This structure allows users to inspect, modify, and intervene on concepts to understand or correct a model's prediction. The goal is therefore to explain decisions through human-understandable concepts rather than latent features.

Most CBM research focuses on the second step, from concepts to task predictions. Prior work studies this relationship through concept interventions (Espinosa Zarlenga et al., 2023; Shin et al., 2023), causal analyses (Vandenhirtz et al., 2024), counterfactual reasoning (Dominici et al., 2025), and leakage measures (Havasi et al., 2022; Parisini et al., 2026). These studies ask: *when the concepts are known, does the task prediction follow from them?* In contrast, the first step, from input to concepts, remains less studied. Here, the reasoning is performed by a deep network and is largely hidden. Concept accuracy is often used as evidence that a model has learned a concept correctly, but accuracy alone does not reveal the evidence used to make the prediction. However, a model may correctly predict a concept from context, correlated attributes, or spurious visual evidence (Margeloiu et al., 2021).

Recent work has begun to study concept localization (Huang et al., 2024; Raman et al., 2024; Aysel et al., 2025; Zhang et al., 2026). These approaches typically use attribution maps or input perturbations to identify regions associated with a concept. However, highlighting a region does not establish that it contains the information used to produce the concept prediction. In particular, a highly attributed region need not be sufficient to produce the concept, nor necessary for it. Localization therefore does not by itself characterize the spatial evidence underlying a concept prediction.

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: Illustrative examples of the three observed concept regimes. Yellow contours denote ground-truth concept segmentations. Green and red frames indicate correct and incorrect concept predictions by the CBM, respectively. From left to right: grounded (high sufficiency, high necessity), redundant (high sufficiency, low necessity), and weak (low sufficiency, low necessity) concepts. No concepts exhibiting high necessity and low sufficiency were observed in the dataset.

In this work, we ask a different question: *where in the learned representation is the information that supports a concept?* For this we introduce *spatial support*, which characterizes the regions of the concept-generating representation that support a concept prediction. We measure support through two complementary properties: *spatial sufficiency*, which asks whether a region alone preserves the concept prediction, and *spatial necessity*, which asks whether removing the region eliminates the prediction. We further combine them into a single *grounding* score. Figure 1 illustrates the different types of spatial support. A concept is *grounded* when its region is both sufficient and necessary, *redundant* when its region is sufficient but the same evidence is also available elsewhere, and *weak* when neither property is high.

We perform these interventions on the concept-generating representation rather than on the input image. Since this is the representation consumed by the concept predictor, the intervention directly probes the spatial information available to the concept prediction. It also avoids the distribution shifts introduced by pixel-space masking. We further introduce the *Spatial Concept Influence Matrix (SCIM)*, which measures how intervening on the region associated with one concept affects the predictions of other concepts. We use SCIM to quantify *concept purity*, measuring the specificity of a spatial intervention to its corresponding concept.

Across diverse CBM architectures, backbones, and training schemes, we show that concept and task accuracy alone do not determine spatial grounding. Grounding varies with the learned concept representation and bottleneck design, while attribution-based localization can capture properties different from spatial grounding. These findings establish spatial grounding as a complementary dimension of CBM interpretability.

Our contributions are as follows:

Input-to-concept faithfulness. We identify the spatial source of concept evidence as an important underexplored dimension of CBM interpretability, beyond concept prediction accuracy.

A framework for spatial support. We introduce spatial sufficiency, spatial necessity, and their joint grounding score on the concept-generating representation. We further introduce SCIM and concept purity to quantify cross-concept effects.

A controlled empirical study. We systematically vary the bottleneck, backbone, training scheme, and representation layer to characterize which factors affect spatial grounding.

2 Related Work

Concept-to-task faithfulness. CBMs are typically evaluated using concept prediction and downstream task accuracy, measuring whether concepts are correctly predicted and whether they support the final task decision (Koh et al., 2020). Consequently, much of the CBM literature has focused on the *concept-to-task* relationship: whether task predictions faithfully depend on the human-understandable concepts. This has been studied through concept interventions (Koh et al., 2020; Espinosa Zarlenga et al., 2023; Shin et al., 2023), causal analyses (Vandenhirtz et al., 2024), and intervention-aware learning strategies. A related line of work examines whether concept represen-

tations contain information beyond their intended semantics. Concept leakage (Havasi et al., 2022) and its extensions, including concept-task and inter-concept leakage (Parisini et al., 2026), assess whether concept representations encode unintended information. Other measures, such as the Oracle Impurity Score (OIS) (Zarlenga et al., 2023) and information-theoretic leakage metrics (Makonnen et al.), further characterize the purity and interpretability of learned concepts. Together, these approaches characterize *what information is available in concept representations* and *how concepts influence task predictions*, but provide limited evidence about the spatial origin of the information used to predict a concept.

Input-to-concept faithfulness and concept grounding. In contrast, relatively little work has examined the *input-to-concept* relationship: whether a concept prediction is based on the intended visual evidence. Margeloiu et al. (2021) highlighted this issue by showing that CBMs can achieve accurate concept predictions without necessarily learning the intended concept evidence. Subsequent work has primarily studied this problem through concept localization and explanation alignment. The Concept Trustworthiness Score (CTS) (Huang et al., 2024), Concept Location Metric (CLM) (Aysel et al., 2025), and locality analyses of CBMs (Raman et al., 2024) evaluate whether model explanations align with semantically meaningful regions, while spatially-aware architectures such as SL-CBM (Zhang et al., 2026) explicitly incorporate spatial information into concept learning. These approaches primarily ask *where an explanation points*. Our objective is different: we ask *where the information supporting a concept prediction is encoded*. We therefore introduce a representation-level spatial grounding framework based on *spatial sufficiency* and *spatial necessity*, which directly characterizes the spatial support of concept evidence.

3 Spatial Support Theory of Concept Representations

3.1 Preliminaries

A CBM consists of a feature extractor h, a concept predictor f, and a task predictor g. Given an input x, the feature extractor produces a spatial representation $F = h(x) \in \mathbb{R}^{H\times W\times D}$, from which the concept vector and task prediction are obtained as $c = f(F), \quad y = g(c)$, where $c = (c_1, \ldots, c_C)$. We study the spatial evidence supporting a concept prediction and ask two complementary questions: *is a region sufficient to produce the concept?* and *is the region necessary for producing the concept?*

3.2 Spatial Intervention and Support

Concept values alone do not reveal where the evidence supporting a concept is encoded. We therefore introduce *spatial support*, which characterizes the spatial locations whose information contributes to a concept prediction. Let $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$ denote the spatial domain of F. For a region $R \subseteq \Omega$, let M_R be its binary spatial mask and define the masked representation as $F_R = M_R \odot F$, where $\odot$ denotes element-wise multiplication. We denote the fully masked representation by $F_\emptyset$.

For concept c_i, the effect of intervening on region R is measured by the change in its prediction,

$$\Delta_i(R) = |f_i(F) - f_i(F_R)| . \tag{1}$$

The corresponding spatial support set at threshold τ is

$$\mathcal{S}_i(\tau) = \{R \subseteq \Omega : \Delta_i(R) > \tau\} . \tag{2}$$

Thus, $\mathcal{S}_i(\tau)$ contains regions whose intervention substantially changes the prediction of concept c_i. All interventions are performed on the representation F consumed by the concept predictor. This directly probes spatial information available to the concept prediction, rather than attributing concept support to intermediate features that may not directly determine the final concept.

3.3 Spatial Sufficiency and Necessity

Spatial sufficiency asks whether a region contains enough information to reproduce the concept prediction when all other spatial locations are removed. For a region R, we define its spatial sufficiency

for concept c_i as

$$\text{sufficiency}(R, c_i) = \frac{|f_i(F_R) - f_i(F_\emptyset)|}{|f_i(F) - f_i(F_\emptyset)| + \epsilon}. \tag{3}$$

The normalization compares the evidence retained by R with the total concept evidence relative to the fully masked representation. We clip the result to $[0, 1]$. High sufficiency indicates that the concept can largely be reproduced from R alone. However, sufficiency does not imply that the evidence is localized: the same information may also be encoded elsewhere in the representation. We therefore introduce spatial necessity, which measures how strongly the concept prediction depends on information contained in R. Specifically, we remove R while retaining its complement $\bar{R} = \Omega \backslash R$ and define

$$\text{necessity}(R, c_i) = 1 - \frac{|f_i(F_{\bar{R}}) - f_i(F_\emptyset)|}{|f_i(F) - f_i(F_\emptyset)| + \epsilon}. \tag{4}$$

We again clip the result to $[0, 1]$.

High necessity indicates that removing R eliminates a substantial amount of the concept evidence, whereas low necessity indicates that the concept can still be predicted from the remaining representation. Thus, sufficiency captures the *presence* of concept evidence in a region, while necessity captures the model's *dependence* on that region.

3.4 Spatial Grounding

A region is spatially grounded for a concept when it is both sufficient to produce the concept and necessary for maintaining its predictions. We therefore define the spatial grounding score as

$$\text{grounding}(R, c_i) = \text{sufficiency}(R, c_i) \cdot \text{necessity}(R, c_i). \tag{5}$$

The product requires both properties to be high and yields a score in $[0, 1]$.

The two components distinguish different forms of spatial organization. A region can be sufficient but redundant when the same concept evidence is also encoded elsewhere. Conversely, a region can be necessary but insufficient when the concept depends on complementary information distributed across multiple regions. We therefore distinguish four qualitative regimes:

sufficiency	necessity	**Spatial support regime**
High	High	Grounded
High	Low	Redundant
Low	High	Distributed / complementary
Low	Low	Weak / unsupported

Table 1: Spatial support regimes defined by sufficiency and necessity.

Region Size The proposed measures characterize the information contained in a specified region and are therefore not invariant to region size. Larger regions retain more of the representation and may consequently achieve higher sufficiency simply by containing more information.

We quantify the spatial coverage of a region R as $\rho(R) = \frac{|R|}{|\Omega|}$. Spatial coverage provides important context for interpreting sufficiency and grounding: comparisons between regions with substantially different coverage may otherwise conflate spatial localization with the amount of representation retained.

3.5 Spatial Concept Influence

A spatial intervention associated with one concept may also affect other concept predictions. Given a region R_i associated with concept c_i, we define the *Spatial Concept Influence Matrix (SCIM)* as

$$A_{ij} = |f_j(F) - f_j(F_{R_i})|. \tag{6}$$

The diagonal entries A_{ii} measure the self-influence of the region on its associated concept, while the off-diagonal entries A_{ij} measure cross-concept influence.

Concept purity. To quantify how specifically a concept-associated region affects its own concept relative to the remaining concepts, we define

$$\text{purity}(c_i) = \frac{A_{ii}}{\sum_{j=1}^{C} A_{ij} + \epsilon}. \tag{7}$$

High purity indicates that intervening on R_i predominantly affects c_i, whereas low purity indicates stronger effects on other concepts and therefore greater spatial entanglement.

Purity depends on the size of the concept vocabulary. If an intervention affects all concepts equally, for example, its expected purity is approximately $1/C$. Purity is therefore most naturally interpreted within a fixed concept vocabulary or relative to an appropriate reference level.

Relationship Between Grounding and Cross-Concept Influence Grounding and purity capture complementary properties. Spatial sufficiency and necessity characterize the relationship between a concept and its associated region, $R_i \longleftrightarrow c_i$. whereas SCIM measures the effect of that region on the complete concept set, $R_i \longrightarrow \{c_1, \ldots, c_C\}$.

Consequently, high grounding does not imply high purity: a region may be both sufficient and necessary for its concept while also carrying information shared with other concepts. Conversely, high purity does not imply high grounding, since an intervention may primarily affect the intended concept even when the region is insufficient to reproduce it. Thus, *grounding* measures joint spatial support, whereas *purity* measures the specificity of that support to the associated concept.

3.6 Interpretation and Scope

Our framework characterizes functional dependence within the learned representation. The interventions should therefore not be interpreted as causal interventions in the data-generating process: high necessity means that the model depends on the corresponding representation region under the specified intervention, not that the underlying visual attribute is causally necessary.

Similarly, spatial grounding characterizes *functional spatial support* in the learned representation rather than semantic localization in the input image. The framework therefore allows concept evidence to be localized, redundant, or distributed across multiple regions.

4 Experiments

We begin by describing the experimental setup, then investigate what shapes and distinguishes concept grounding. Finally, we show that grounding captures information beyond attribution-based localization. We defer additional implementation details, full experimental results, significance analyses, and visualizations to the Appendix.

4.1 Experimental Setup

Dataset. We evaluate all methods on **CUB-200-2011** (Wah et al., 2011), which provides 312 human-annotated bird attributes and fine-grained species labels; following Koh et al. (2020), we use the selected set of 112 concepts. To evaluate spatial grounding, we use the CUB-70 test-set part annotations (Behzadi-Khormouji & Oramas, 2023), comprising 1,888 images and 36,447 positive (present) concept-image pairs. For each positive (present) concept, we associate the concept with its corresponding anatomical part mask. When a concept does not have a dedicated part annotation, we use the whole-bird segmentation mask. Negative (absent) concepts are assigned an empty mask and are not used as intervention targets. We reproduce key experiments on the RIVAL-10 dataset (Moayeri et al., 2022), which maps CIFAR-10 (Krizhevsky et al., 2009) classes to ImageNet (Krizhevsky et al., 2012) and contains around 26,000 images with 18 visual attributes and segmentations.

Models. We evaluate a diverse set of concept bottleneck architectures, including CBM (Koh et al., 2020), CEM (Espinosa Zarlenga et al., 2022), SCBM (Vandenhirtz et al., 2024), SL-CBM (Zhang et al., 2026), CT (Rigotti et al., 2021), VLG-CBM (Srivastava et al., 2024), and FVLC (Lai et al., 2024). Backbone experiments include ResNet-18/34/50 (He et al., 2016), Inception-v3 (Szegedy

et al., 2016), ViT-B/16, ViT-L/16 (Dosovitskiy et al., 2021), DeiT-B/16 (Touvron et al., 2021), and Swin-Tiny (Liu et al., 2021). All backbones are initialized from ImageNet-pretrained weights, fine-tuned on the target dataset, and then frozen unless explicitly stated otherwise.

Protocol. For each concept, we resize the ground-truth segmentation mask from image space to the spatial resolution of the representation under analysis, yielding a concept-specific region R and its complement $\bar{R}$. We apply the same procedure across all models, layers, and datasets. We evaluate spatial support on the concept-generating representation, corresponding to the final convolutional feature map for CNNs and the patch-token grid for transformers (justified in Appendix B). Task accuracy (Acc_y) is reported as top-1 classification accuracy. Concept accuracy (Acc_c) is computed by thresholding concept predictions at 0.5 and averaging accuracy across concepts.

4.2 What Determines and Distinguishes Spatial Grounding?

Having established where spatial support should be measured, we next ask what determines concept grounding. We consider three possible sources: the concept bottleneck $g \circ f : \mathbb{R}^{H\times W\times D} \to \mathbb{R}^C \to \mathcal{Y}$, the backbone representation $h : \mathcal{X} \to \mathbb{R}^{H\times W\times D}$, and the optimization procedure used during training. To disentangle their effects, we run controlled experiments in which we vary only one factor at a time while holding the others fixed. This allows us to identify whether grounding primarily arises from concept formation, feature representation, or optimization. Due to space constraints, we present only the evidence most directly supporting our core findings and defer the full experimental analysis to Appendix C.

(1) Accuracy does not imply grounding. (2) Grounding is shaped by the bottleneck and spatial structure of the backbone, but not much by the training strategy. (3) Grounding is different from attribution localization and purity.

Grounding $\neq$ Accuracy. Comparing accuracy (Acc_c,Acc_y) with spatial grounding ($\mathrm{sufficiency}$, $\mathrm{necessity}$, $\mathrm{grounding}$) across Table 2, grounding exhibits only a weak relationship with accuracy. High concept accuracy does not necessarily imply strong grounding. For instance, on CUB, CBM, CEM, SCBM, and VLG-CBM achieve similar $\mathrm{grounding}$ $(0.14 - 0.16)$ despite concept accuracies ranging from 91.5% to 95.6%. Conversely, SL-CBM matches VLG-CBM's concept accuracy while showing markedly different grounding behavior. The correlation analysis in Appendix Table 10 further shows weak and inconsistent relationships between predictive accuracy and grounding across datasets and model families. These results indicate that concept correctness and spatial support capture distinct properties of a concept representation.

Bottleneck. As shown in Table 2 (A), most CBM variants learn similar support structures despite differing predictive performance, consistent with the fact that they do not explicitly enforce spatial locality and grounding. In contrast, SL-CBM, which directly promotes concept locality, exhibits a qualitatively different support pattern. SL-CBM markedly increases spatial sufficiency by enforcing concept locality. On RIVAL-10, this results in higher grounding, whereas on CUB it produces a high-sufficiency, low-necessity regime, revealing that concept evidence remains redundantly distributed outside the target region. This demonstrates that bottleneck design can substantially reshape grounding behavior.

Backbone. As shown in Table 2 (B), convolutional backbones consistently achieve higher grounding than transformers, and grounding increases further with CNN depth. Differences in feature-map resolution can bias grounding comparisons between CNNs and transformer backbones. To control for this effect, we project transformer concept regions onto a common 7×7 grid before intervention. This resolution-matched evaluation substantially improves transformer grounding, closing roughly one-third to one-half of the CNN-transformer gap, leaving CNNs consistently ahead. Swin-Tiny achieves stronger grounding than vanilla vision transformers. These findings suggest that locality-preserving inductive biases promote grounded concept representations.

Training strategy. The optimization strategy has little effect on spatial grounding Table 2 (C). Sequential and joint training produce very similar grounding despite small differences in predictive

(A) Bottleneck comparison

Bottleneck	Acc_c (%)	Acc_y (%)	sufficiency ↑	necessity ↑	grounding ↑	% grounded	purity (%) ↑
CUB, frozen ResNet-18							
CBM	94.0	67.4	0.35 ± 0.28	0.30 ± 0.25	0.15 ± 0.21	27.0	0.57 ± 0.44
CEM	95.6	74.3	0.34 ± 0.27	0.31 ± 0.26	0.16 ± 0.21	28.5	0.63 ± 0.41
SCBM	91.5	70.6	0.33 ± 0.27	0.28 ± 0.24	0.14 ± 0.20	24.0	0.47 ± 0.40
SL-CBM	93.8	67.4	0.92 ± 0.19	0.08 ± 0.16	0.08 ± 0.16	12.4	1.13 ± 0.85
VLG-CBM	93.8	66.2	0.35 ± 0.28	0.29 ± 0.25	0.15 ± 0.21	26.5	0.53 ± 0.40
FVLC	56.8	65.6	0.31 ± 0.26	0.27 ± 0.24	0.13 ± 0.19	22.6	0.96 ± 0.63
RIVAL-10, frozen CLIP ViT-B/16							
CBM	94.0	99.2	0.330 ± 0.301	0.297 ± 0.291	0.174 ± 0.248	30.8	3.01 ± 1.92
CEM	93.4	99.1	0.362 ± 0.320	0.296 ± 0.300	0.177 ± 0.259	29.5	1.92 ± 1.53
SCBM	94.0	98.4	0.336 ± 0.296	0.315 ± 0.287	0.183 ± 0.247	32.4	3.57 ± 2.15
SL-CBM	89.7	99.2	0.946 ± 0.153	0.270 ± 0.353	0.270 ± 0.353	38.2	2.96 ± 4.65
VLG-CBM	93.8	99.1	0.337 ± 0.305	0.297 ± 0.292	0.176 ± 0.251	30.7	3.00 ± 1.98
FVLC	62.9	94.3	0.359 ± 0.343	0.274 ± 0.304	0.174 ± 0.269	28.3	6.07 ± 4.98

(B) Backbone ablation

Backbone	Grid	Acc_c (%)	Acc_y (%)	sufficiency ↑	necessity ↑	grounding ↑	% grounded	purity (%)↑
CUB, CBM								
ResNet-18	7×7	94.0	67.4	0.35 ± 0.28	0.30 ± 0.25	0.15 ± 0.21	27.0	0.57 ± 0.44
ResNet-34	7×7	93.9	71.8	0.42 ± 0.32	0.34 ± 0.29	0.21 ± 0.26	34.7	0.49 ± 0.42
ResNet-50	7×7	95.4	73.3	0.45 ± 0.33	0.37 ± 0.30	0.24 ± 0.28	39.8	0.51 ± 0.42
Inception-v3	8×8	96.8	76.9	0.43 ± 0.31	0.40 ± 0.30	0.25 ± 0.28	43.4	0.61 ± 0.35
ViT-B/16	14×14	95.7	74.5	0.18 ± 0.20	0.15 ± 0.17	0.05 ± 0.12	6.9	0.59 ± 0.40
DeiT-Base/16	14×14	95.9	75.5	0.19 ± 0.21	0.17 ± 0.18	0.06 ± 0.13	8.3	0.62 ± 0.38
ViT-Large/16	14×14	97.0	78.9	0.19 ± 0.20	0.18 ± 0.18	0.06 ± 0.13	8.9	0.63 ± 0.35
Swin-Tiny	7×7	96.4	76.1	0.24 ± 0.21	0.22 ± 0.19	0.09 ± 0.14	13.1	0.70 ± 0.38
Control regions resolved on a common 7×7 grid								
ViT-B/16	14×14	95.7	74.5	0.24 ± 0.22	0.21 ± 0.20	0.09 ± 0.15	12.7	0.59 ± 0.40
DeiT-Base/16	14×14	95.9	75.5	0.26 ± 0.23	0.23 ± 0.20	0.10 ± 0.16	15.1	0.62 ± 0.38
ViT-Large/16	14×14	97.0	78.9	0.26 ± 0.23	0.24 ± 0.21	0.10 ± 0.16	15.9	0.62 ± 0.35

(C) Training scheme ablation

Scheme	Acc_c (%)	Acc_y (%)	sufficiency ↑	necessity ↑	grounding ↑	% grounded	purity (%) ↑
CUB, CBM (Inception-v3)							
Sequential	97.2	72.2	0.36 ± 0.27	0.34 ± 0.26	0.19 ± 0.23	33.5	0.62 ± 0.22
Joint	97.4	71.8	0.34 ± 0.26	0.33 ± 0.25	0.17 ± 0.22	31.0	0.69 ± 0.23

Table 2: Bottleneck comparison, backbone ablation and training scheme ablation. % Grounded denotes the proportion of examples with a grounding score greater than 0.2.

performance, suggesting the spatial structure of concept representations is more strongly determined by architectural design than by the choice between training schemes.

Purity ≠ Grounding. Table 2 shows that grounding does not imply purity. Learning a concept from the intended spatial region does not automatically produce disentangled concept representations. Grounding and spatial purity capture complementary properties of concept representations. Grounding measures whether a concept is supported by its associated spatial region, whereas purity measures whether the effect of that region is specific to the corresponding concept.

Visualization of SCIM. As shown in Figure 2, concept interactions are strongly influenced by bottleneck design. While CBM, CEM, SCBM, and VLG-CBM learn similar SCIM structures, SL-CBM and FVLC exhibit qualitatively different interaction patterns. Despite these differences, several concepts (e.g., *hairy*, *metallic*, *rectangular*, *wet*, *long*, *tall*, and *patterned*) consistently show weak interactions across architectures, suggesting they are less spatially entangled than other concepts.

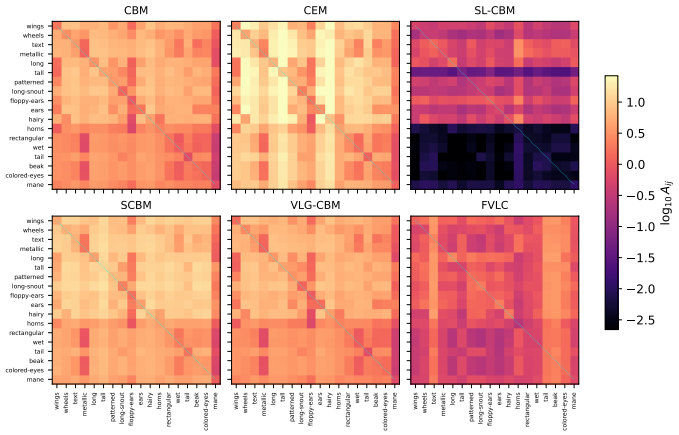


Figure 2: SCIM visualization for six bottleneck architectures on RIVAL-10 (log color scale).

4.3 Spatial Grounding is Distinct from Attribution Localization

Attribution-based localization is widely used as a proxy for concept grounding in CBMs. Yet localization quality and concept grounding do not need to coincide. We compare grounding against Grad-CAM (Selvaraju et al., 2017), Score-CAM (Wang et al., 2020), Integrated Gradients (IG) (Sundararajan et al., 2017), a random baseline (Random), and SL-CBM's learned saliency maps (Learned). We evaluate localization using IoU and the Pointing Game (PG) (Zhang et al., 2018b).

Table 3 (A) shows substantial variation in localization metrics across attribution methods, while the corresponding grounding behavior can remain different. For example, on CUB, SL-CBM exhibits low grounding despite non-trivial localization scores from its learned saliency maps and post-hoc attribution methods.

We therefore use attribution maps to construct support regions with matched spatial coverage and directly evaluate their grounding. As shown in Table 3 (B), the regions identified by different attribution methods produce substantially different necessity and grounding scores. Notably, the learned saliency regions increase grounding from 0.076 for the ground-truth concept mask to 0.161, while object-level regions also improve over the ground-truth mask. These results indicate that the region that best aligns with a human-defined concept mask need not be the region that functionally supports the model's concept prediction.

4.4 Purity and Existing Measures of Concept Faithfulness

Spatial purity measures the specificity of concept-supporting regions and captures a notion of spatial modularity complementary to grounding. We compare it against established measures of impurity, leakage, and intervention utility, including RI (a concept-wise decomposition of OIS) (Zarlenga et al., 2023), CTL, ICL Parisini et al. (2026), and intervention gain (ΔI_c) Koh et al. (2020). We conduct experiments on CUB across six bottleneck architectures with a shared ResNet-18 backbone. More details and experiments can be found in Appendix F.

According to Table 4, spatial purity exhibits no consistent alignment with existing measures of impurity, leakage, or intervention utility. While highly pure models such as SL-CBM often exhibit reduced concept interactions, this relationship does not hold universally across architectures. Conversely, models with strong utility metrics do not necessarily achieve high purity. These results suggest that spatial purity captures a distinct aspect of concept representations, complementary to representational impurity, information leakage, and intervention-based measures of faithfulness.

5 Conclusion

Our results show that concept accuracy alone does not characterize how a concept is represented spatially. Models with similar predictive performance can exhibit substantially different spatial ground-

(A) Spatial grounding *vs*. attribution-based locality

Model	Spatial Grounding	Attribution	IoU(R)↑	IoU($\bar{R}$)↓	PG ↑
CUB					
CBM	sufficiency : 0.359±0.269 necessity : 0.342±0.261 grounding : 0.187±0.229	Grad-CAM	0.124±0.142	0.374±0.149	0.238±0.426
		Score-CAM	0.123±0.139	0.459±0.132	0.258±0.438
		IG	0.025±0.030	0.006±0.005	0.277±0.447
		Random	0.073±0.090	0.820±0.086	0.067±0.250
SL-CBM	sufficiency : 0.921±0.189 necessity : 0.076±0.162 grounding : 0.076±0.162	Grad-CAM	0.068±0.093	0.338±0.209	0.114±0.317
		Score-CAM	0.086±0.103	0.676±0.138	0.284±0.451
		IG	0.029±0.035	0.010±0.008	0.267±0.442
		Learned	0.135±0.158	0.060±0.040	0.274±0.446
		Random	0.067±0.083	0.825±0.083	0.059±0.235
RIVAL-10					
CBM	sufficiency : 0.330±0.301 necessity : 0.297±0.291 grounding : 0.174±0.248	Grad-CAM	0.173±0.199	0.427±0.198	0.260±0.439
		Score-CAM	0.226±0.238	0.774±0.238	0.124±0.330
		IG	0.010±0.015	0.010±0.009	0.178±0.382
		Random	0.215±0.221	0.696±0.206	0.176±0.381
SL-CBM	sufficiency : 0.946±0.153 necessity : 0.270±0.353 grounding : 0.270±0.353	Grad-CAM	0.120±0.166	0.287±0.311	0.261±0.439
		Score-CAM	0.226±0.238	0.774±0.238	0.119±0.324
		IG	0.013±0.024	0.002±0.003	0.571±0.495
		Learned	0.037±0.058	0.005±0.005	0.561±0.496
		Random	0.215±0.221	0.696±0.206	0.176±0.381

(B) Saliency map as source of R

Source of R	$\lvert R\rvert$	sufficiency ↑	necessity ↑	grounding ↑	Δ grounding
GT part mask	8.8	0.921 ± 0.189	0.076 ± 0.162	0.076 ± 0.162	–
Random	8.8	0.867 ± 0.230	0.024 ± 0.084	0.024 ± 0.084	–
Grad-CAM	7.9	0.914 ± 0.243	0.058 ± 0.119	0.058 ± 0.119	$+0.037$ [$+0.035$, $+0.038$]
Score-CAM	8.8	0.958 ± 0.158	0.136 ± 0.199	0.136 ± 0.199	$+0.112$ [$+0.109$, $+0.115$]
IG	8.8	0.938 ± 0.164	0.083 ± 0.167	0.083 ± 0.167	$+0.060$ [$+0.058$, $+0.063$]
Learned	8.0	0.904 ± 0.272	0.161 ± 0.220	0.161 ± 0.220	$+0.140$ [$+0.137$, $+0.143$]
Object	8.8	0.973 ± 0.120	0.135 ± 0.204	0.135 ± 0.204	$+0.111$ [$+0.108$, $+0.114$]

Table 3: Localization performance and comparison of support-region constructions using Grad-CAM, Score-CAM, IG, Random, and learned saliency maps (only for SL-CBM). Δ grounding is measured relative to the Random baseline, reported as mean [lower bound, upper bound].

Model	purity (%) ↑	RI ↓	CTL ↓	ICL ↓	$\Delta \mathrm{I}_c$ ↑
CBM	0.54 ± 0.25	7.69 ± 1.56	0.56 ± 0.89	1.57 ± 0.54	0.44 ± 0.53
SL-CBM	1.03 ± 0.20	7.24 ± 1.55	0.50 ± 1.20	1.15 ± 0.43	0.54 ± 0.59
VLG-CBM	0.50 ± 0.23	8.07 ± 1.86	0.81 ± 1.19	3.28 ± 1.12	0.54 ± 0.63
FVLC	0.97 ± 0.32	11.26 ± 2.53	2.28 ± 2.91	5.08 ± 2.83	-0.79 ± 0.79
CEM	0.59 ± 0.18	6.83 ± 1.39	0.31 ± 0.68	2.00 ± 0.57	0.37 ± 0.39
SCBM	0.47 ± 0.22	9.20 ± 1.82	1.86 ± 1.89	2.91 ± 0.70	1.60 ± 1.49

Table 4: Comparison between spatial purity and existing measures of concept faithfulness.

ing, indicating that concept correctness and the spatial support underlying a prediction capture distinct properties. The proposed sufficiency-necessity formulation further distinguishes different forms of spatial support. A region may contain sufficient evidence while being redundant, or it may be necessary while the concept remains distributed across multiple regions. Similarly, attribution-based localization does not necessarily imply spatial grounding: attribution identifies regions associated with a prediction, whereas grounding evaluates whether those regions functionally support the concept. Overall, we introduce spatial grounding as a complementary, representation-level perspective for evaluating visual Concept Bottleneck Models. By examining where and how concept evidence is encoded, alongside predictive accuracy and concept-to-task faithfulness, our framework provides a more complete characterization of concept representations and their interpretability.

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, 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.

Mattan S. Ben-Shachar, Daniel L"udecke, and Dominique Makowski. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815, 2020. doi: 10.21105/joss.02815. URL `https://doi.org`.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *International Conference on Learning Representations*, volume 2025, pp. 5184–5207, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

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 *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Équille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Mikael Makonnen, Moritz Vandenhirtz, Sonia Laguna, and Julia E Vogt. Measuring leakage in concept-based methods: An information theoretic approach, 2025. *URL https://arxiv.org/abs/2504.09459*.

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*, pp. 19087–19097, 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.

Naveen Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models respect localities? *arXiv preprint arXiv:2401.01259*, 2024.

Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International conference on learning representations*, 2021.

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.

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 Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

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. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 24–25, 2020.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning (ICML)*, pp. 11205–11216, 2021.

Mateo Espinosa Zarlenga, Pietro Barbiero, Zohreh Shams, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, and Mateja Jamnik. Towards robust metrics for concept representation evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11791–11799, 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, 2026.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *International Journal of Computer Vision (IJCV)*, pp. 1084–1102, 2018a.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018b.