C2MF: Consistent and Concept-Unified Matrix Factorization for Interpretable and Robust Concept Discovery

Anonymous authors
Paper under double-blind review

Abstract

Deep neural networks have achieved remarkable performance in various domains, but their opacity remains a significant challenge, particularly in high-risk applications. Traditional attribution methods highlight important input regions but fail to reveal the underlying semantic concepts driving model decisions. Recent methods like TCAV and CRAFT attempt to address this gap by extracting interpretable concepts, but they suffer from limitations such as distribution mismatch between training and inference, reliance on non-negative activation constraints, and the lack of a shared concept dictionary across categories. In this paper, we introduce Consistent and Concept-Unified Matrix Factorization (C2MF) method, a novel approach that overcomes these issues. By leveraging full-image representations instead of cropped sub-regions, C2MF ensures consistency between training and inference distributions, improving robustness and confidence calibration. We also relax the non-negativity constraint, allowing both positive and negative concept activations, which enhances the flexibility and fidelity of learned concepts. Furthermore, we propose a shared global concept dictionary across all categories, enabling concept reuse and improving interpretability. Through extensive experiments on ImageNet and CUB datasets, we demonstrate that C2MF outperforms state-of-the-art methods in terms of concept faithfulness, category reconstruction accuracy, and generalization across categories. Our code is available at: https://anonymous.4open.science/r/C2MF-E760/.

1 Introduction

Deep neural networks have achieved remarkable performance in various domains such as vision, language, and more (Khan et al., 2020). However, their opacity presents challenges in safety-critical applications (Wang & Chung, 2022). In high-risk environments like healthcare (Corfmat et al., 2025) and autonomous driving (Ajenaghughrure et al., 2020), understanding the reasons behind a model's specific decisions is crucial, rather than simply identifying the areas of focus. Many explanation techniques (Selvaraju et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017) generate saliency maps or heatmaps to highlight important input pixels, but these attribution-based methods only show which parts of the input influence the model's output, without revealing the semantic concepts behind the model's reasoning.

Concept-based interpretability methods aim to bridge this gap by constructing explanations with human-understandable concepts. A notable example is the Concept Activation Vector Test (TCAV) (Kim et al., 2018), which starts with a set of user-defined concept examples (e.g., images with stripes or medical scans showing specific lesions) to define a concept vector in the network's activation space. TCAV then measures the directional derivative of the model output with respect to the concept vector, quantifying the sensitivity of predictions to that concept. By leveraging user-defined concept examples, TCAV provides semantic insights into what the model has learned. However, TCAV requires manually curated concept datasets and cannot automatically discover new concepts, which limits its ability to capture the full range of concepts learned by the model based on human prior knowledge.

To address this issue, recent research has focused on automating concept discovery and attribution. The Concept Recursive Activation FacTorization (CRAFT) (Fel et al., 2023) method extracts concepts and their associated image locations simultaneously. Specifically, CRAFT collects a large set of activations for target categories from intermediate layers of a pre-trained network (often using random cropping) and applies Recursive Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) to these activations. The

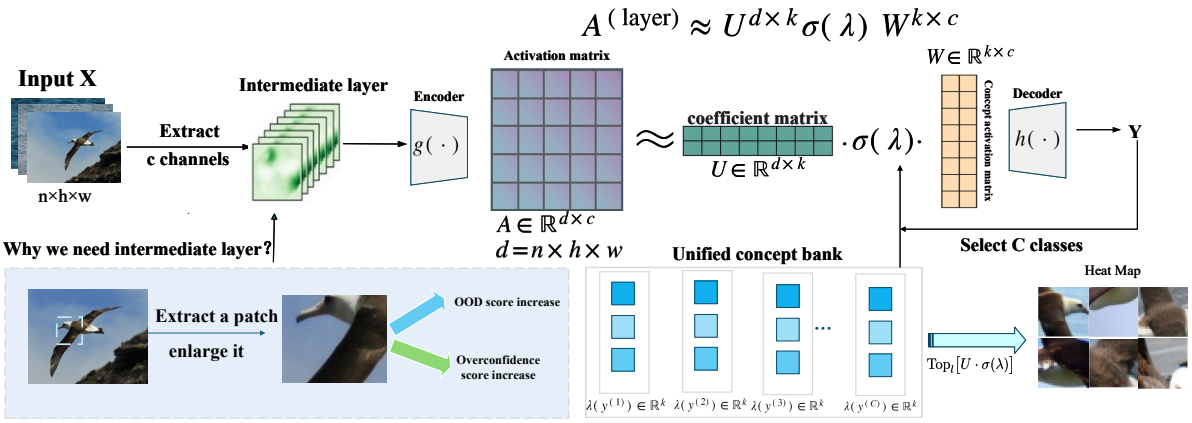


Figure 1: Structure of Consistent and Concept-Unified Matrix Factorization (C2MF).

decomposition generates a set of basis vectors, each representing a distinct high-level concept, along with coefficients indicating the strength of each concept in each image. CRAFT then ranks these concepts using global sensitivity metrics such as the Sobol index (Fel et al., 2023), generating concept attribution maps that highlight the locations where each concept appears in the given input image. In this way, CRAFT explains the concepts the model uses and their locations in the image, without the need for manually labeled concepts.

However, CRAFT has several limitations: it learns concepts from random image patches but applies them to full images, causing a distribution gap that leads to unreliable activations and overconfident explanations; its non-negativity constraint forbids negative features, which may carry important contrasting information, and removing them increases correlations between matrices and mixes background features, reducing decomposition fidelity (see Section 2.2); moreover, it extracts concepts independently for each category without enforcing cross-category sparsity, resulting in redundant and less interpretable concepts.

To address these challenges, we propose Consistent and Concept-Unified Matrix Factorization (C2MF). C2MF retains the concept decomposition framework but introduces three key innovations. First, it performs concept learning and inference on intermediate layer feature representations of the model, rather than on image patches. By aligning the training and testing distributions, this consistency significantly improves the robustness and calibrability of concept activations. Second, C2MF removes the non-negativity constraint and allows concept activation vectors to have signed values (positive or negative), leading to more reliable concept decomposition. Third, C2MF introduces a unified global concept dictionary shared across all categories, where each category adjusts the dictionary via sparse activation weights. We constrain each weight between 0 and 1 using a *Sigmoid* activation function and encourage sparsity through an L_1 penalty term, ensuring that each category uses only a small subset of concepts. These innovations collectively produce a set of concepts that can be shared and reused across categories, leading to more compact and interpretable explanations. Figure 1 shows the structure of our C2MF. The main contributions of this paper are as follows:

- We propose a consistent training and inference framework for concept decomposition, which operates on the complete image and eliminates the distribution mismatch problem present in previous methods, thereby enhancing the confidence and robustness of concept-based explanations.
- By removing the non-negativity constraint on concept activation functions and design a unified global concept dictionary, we construct a more comprehensive matrix factorization model.
- Extensive experiments on the CUB and ImageNet datasets demonstrate that our C2MF method achieves optimal performance in terms of concept fidelity, concept sparsity, concept reconstruction accuracy, consistency in concept decomposition, and concept clustering ablation.

2 METHOD

In this section, we first introduce the problem definition and then present our C2MF approach. The Related work section can be found in the App. A of this paper.

2.1 PROBLEM DEFINITION

In this section, we provide a formal definition of the problem. Let $f(x) = h(g(x))$ denote a pre-trained deep network, where $g(\cdot)$ is the encoder that generates intermediate features, and $h(\cdot)$ is a specific classifier that produces the predictions of the model. Given a dataset $\mathcal{D} = \{\langle x^{(i)}, y^{(i)} \rangle\}_{i=1}^{n}$ containing n samples, where $x^{(i)} \in \mathbb{R}^{wh\times 3}$ (where h and w indicate the image height and width, 3 corresponds to the number of channels, and Y denotes the number of classes). For each image $x^{(i)}$ in the dataset, the intermediate feature representation extracted by the encoder $g(\cdot)$ is denoted as $x_i \in \mathbb{R}^{wh\times c}$. The collection of all such feature activations across the dataset forms a matrix $A \in \mathbb{R}^{nwh\times c}$.where n is the number of feature representations extracted. Our goal is to extract k disentangled concept prototypes that represent the underlying semantic factors within the activation space.

2.2 CONSISTENT & CONCEPT-UNIFIED MATRIX FACTORIZATION (C2MF)

To address the issues of training and inference inconsistency, non-negativity constraint distortion, and the lack of a unified concept dictionary in the CRAFT method mentioned in the introduction, we propose a new concept discovery approach—Consistency-Training and Concept-Unified Matrix Factorization (C2MF).

To perform non-negative matrix factorization, the CRAFT method constructs sub-regions $x'^{(i)} \in \mathbb{R}^{(s\times s)\times 3}$ by randomly cropping the input image $x^{(i)}$, where s denotes the width and height of each sub-region. These sub-regions are then resized to match the input image's dimensions (i.e., $w \times h$) for concept extraction. It is worth noting that this approach causes a significant mismatch between the input distribution during training, which is based on cropped sub-regions, and the whole-image distribution during inference. This mismatch can lead to out-of-distribution (OOD) risks and miscalibrated confidence. The sub-region input magnifies local textures while losing contextual information, whereas the whole-image input retains global semantics. As a result, the features learned during training are often not aligned with the inference process, thus exacerbating the out-of-distribution issue.

As illustrated in Fig. 5 and discussed in App. C, the confidence distributions of complete images and randomly cropped *image patches* are nearly separable across datasets and backbones: complete images concentrate in a high-confidence regime, whereas patches shift to markedly lower confidence with minimal overlap. Consistently, our OOD evaluation in App. B (Tab. 4) shows elevated operational errors (e.g., FPR@95TPR) when the scoring threshold must preserve high recall, indicating that scores induced by patch-based training intrude less into the high-confidence region where whole images dominate. Together, these observations substantiate that patch-based training induces a distributional shift that both degrades calibration and increases OOD risk.[1]

In contrast, our C2MF method decomposes the intermediate features with spatial invariance during training by directly using the whole image $x^{(i)} \in \mathbb{R}^{(w\times h)\times c}$ as input. This ensures that the training distribution aligns with the deployment distribution, thereby significantly reducing errors and overconfidence caused by the inconsistency in cropping.

Additionally, C2MF is not limited to purely non-negative forms for concept representations. As a traditional concept extraction method based on NMF, CRAFT requires the activation values $A \in \mathbb{R}^{nwh\times c}$ to be non-negative. It decomposes the positive activation values through NMF into a concept activation matrix $W \in \mathbb{R}^{k\times c}$, containing several non-negative bases, and a coefficient matrix $U \in \mathbb{R}^{nwh\times k}$. Here, k is the predefined number of concepts. This decomposition process can be expressed as follows:

$$A \approx UW, \quad \text{s.t. } U \geq 0 \text{ and } W \geq 0 \tag{1}$$

Although this approach aids in obtaining interpretable concept representations, it also introduces the *common bias* effect: to ensure the non-negativity of the concept activation matrix W and the coefficient matrix

[1] Moreover, modern deep networks tend to be overconfident on unseen inputs (Zhang et al., 2025), further weakening the link between confidence and accuracy.

U, the same positive component is added to both matrices during the matrix factorization process. This makes the two matrices align more closely in the high-dimensional space, thereby exaggerating their correlation. To address this, C2MF allows the concept activation vectors in W to take negative values, thus avoiding the uniform positive bias inherent in non-negative bases. In other words, we relax the non-negativity constraint and allow the concept activation matrix to have both positive and negative components during the factorization process. This enables different concepts to cancel or distinguish each other, leading to more flexible and mutually exclusive semantic representations. This is particularly beneficial in networks that do not use ReLU or other non-negative activation functions, as relying excessively on NMF may struggle to capture the complex structure of the original activation space. Below we give the mathematical proof of artificially increasing the correlation between matrices by adding the same positive component during the matrix factorization process.

Proof. First, we prove that shifting the data to be non-negative (i.e., $A' = A+\delta \cdot \mathbf{1}_{nwh}\mathbf{1}_c^T, \delta > 0$),$d = min(A)$ makes it easier for NMF to introduce a uniform positive component when learning U and W. Let $A \in \mathbb{R}^{nwh \times c}$, NMF uses the following loss function:

$$\min_{U\geq 0, W\geq 0} \|A' - UW\|_F^2, \quad A' = A + \delta \cdot \mathbf{1}_{nwh}\mathbf{1}_c^T, \quad \delta > 0 \tag{2}$$

Where $\mathbf{1}_{nwh}(\mathbf{1}_c^T)$ are vectors of all 1. Due to the fact that NMF can only "non-negative" (i.e., the optimization is constrained to non-negative values), to minimize the loss, the best approach is to have all vectors pointing in the same direction (along $\mathbf{1}$), i.e., learning the global alignment. Let's define the $\langle, \rangle$ the sum of all elements. For any decomposition $UW = \sum_{r=1}^{k} u_r w_r$ ($u_r \in \mathbb{R}^n wh_{\geq 0}, w_r \in \mathbb{R}^c_{\geq 0}$), we have:

$$\langle u_r w_r, \mathbf{1}_{nwh}\rangle = (\mathbf{1}^T u_r)(\mathbf{1}^T w_r) \quad \Rightarrow \quad \langle UW, \mathbf{1}\mathbf{1}^T\rangle = \sum_{r=1}^{k}(\mathbf{1}^T u_r)(\mathbf{1}^T w_r) \tag{3}$$

If UW is close to $A' = A + \delta\mathbf{1}\mathbf{1}^T$, then the total sum must satisfy:

$$\sum_{r=1}^{k}(\mathbf{1}^T u_r)(\mathbf{1}^T w_r) \approx \mathbf{1}_{nwh}^T A \mathbf{1}_c + \delta cnwh \tag{4}$$

When comparing with A, the right-side term increases by $dcnwh$. Due to the fact that $(\mathbf{1}^T u_r), (\mathbf{1}^T w_r) \geq 0$, the direct form of this "external factor" is to increase the impact of all u_r, w_r in the direction of $\mathbf{1}$ (i.e., aligning all the vectors along the same direction, a "unified basis"). In the case of non-negative constraints, this addition of 1 can only increase the similarity in the same direction (because negative values cannot be used in this case), so this "regularization" term is inevitable. In addition, the objective function $f(U, W) = \|A' - UW\|_F^2$ has the following gradients:

$$\nabla_U f = (UW - A')W = (UW - A)W - \delta(\mathbf{1}^T)W \tag{5}$$

$$\nabla_W f = (UW - A')^T U = (UW - A)^T U - \delta(\mathbf{1}^T)U \tag{6}$$

Note that $(\mathbf{1}\mathbf{1}_c^T)W = \mathbf{1}_{nwh}(\mathbf{1}\mathbf{1}^T W)$ is completely aligned along each row. Therefore, $-\delta(\mathbf{1}\mathbf{1}^T)W$ represents a consistent bias term (≤ 0) for $\nabla_U f$, which drives U and W to move towards increasing the positive values, i.e., learning to share the positive base line (unified correct term). The same applies to $\nabla_W f$.

Next, we need to show that using a uniform positive component will exaggerate the correlation. Let $a, b > 0$. Consider the following formula for vectors:

$$w_r' = w_r + b \cdot \mathbf{1}_c, \quad u_r' = u_r + a \cdot \mathbf{1}_{nwh} \tag{7}$$

Where $\mathbf{1}_{nwh}$ and $\mathbf{1}_c$ are vectors of all 1 in the respective dimensions. This means that we are adding a positive constant in the direction of $\mathbf{1}$ to each concept's vector. This will influence the similarity: (i) The

two components within the matrix are more aligned (more similar); (ii) The concept similarity across the dimensions (i.e., across different concepts) is improved. For the similarity increase within the same matrix, if we consider two vectors w_p and w_q ($p \neq q$), and we have:

$$\varphi(b) = \cos\left(\angle(w_p + b\mathbf{1}, w_q + b\mathbf{1})\right) = \frac{(w_p + b\mathbf{1})^T(w_q + b\mathbf{1})}{\|w_p + b\mathbf{1}\|\|w_q + b\mathbf{1}\|} \tag{8}$$

By adding a positive constant in the direction of 1, we prove: when $b > 0$ increases, $\varphi(b)$ increases, and as $b \to \infty$, it reaches 1 (both vectors become fully aligned).

Intuitively, by adding a positive constant to each row, we are pushing them in the same direction, aligning them on the same hyperplane. The cosine between the two vectors (maximum pairwise cosine) increases, and thus the similarity increases.

For cross-matrix (between U and W) similarity increases, if we consider matrices U and W, and let u_r, w_r be the row vectors in these matrices, we have:

$$\psi(a, b) = \cos\left(\angle(u_r + a\mathbf{1}, w_s + b\mathbf{1})\right) = \frac{(u_r + a\mathbf{1})^T(w_s + b\mathbf{1})}{\|u_r + a\mathbf{1}\|\|w_s + b\mathbf{1}\|} \tag{9}$$

When using this expression, ψ increases as a and b increase, implying that the cosine similarity between U and W increases. This suggests that the "unified optimization" between the matrices U and W strengthens, leading to improved alignment between the concepts.

□

Finally, to address the issue in traditional matrix factorization methods where each class independently activates a large number of concepts and lacks sharing, C2MF introduces a shared concept dictionary and a class-specific weight mechanism. In the typical low-rank decomposition $A \approx UW$, each category or sample activates a large number of concepts, making it difficult to highlight the key concepts. In contrast, C2MF constructs a shared concept library matrix W^*, and for each category $y^{(i)} \in Y$ predicted by the classifier $h(x)$, it learns a weight vector $\lambda(y^{(i)}) \in \mathbb{R}^k$ to modulate the extent to which each concept is used. We map $\lambda(y^{(i)})$ to the interval $(0, 1)$ using the *Sigmoid* activation function, obtaining $\lambda^*(y^{(i)}) = \text{sigmoid}(\lambda(y^{(i)}))$. Then, we modulate the concept dictionary using the diagonal matrix $\text{diag}(\lambda^*(y^{(i)}))$, so that the activation representation for category $y^{(i)}$ is given by:

$$A(y^{(i)}) \approx U diag(\lambda^*(y^{(i)}))W^*, \quad \text{s.t. } U \geq 0 \tag{10}$$

Thus, the optimization objective of C2MF is:

$$\min_{U, W^*, \lambda} \sum_{y^{(i)}=1}^{Y} \left\| A(y^{(i)}) - U diag(\lambda^*(y^{(i)}))W^* \right\|_F^2 + \alpha \sum_{y^{(i)}=1}^{Y} \left\| \lambda^*(y^{(i)}) \right\|_1 \tag{11}$$

Where $\alpha > 0$ is the weight coefficient controlling the sparsity, and $\|\cdot\|_F$ denotes the Frobenius norm. CRAFT can only analyze concepts corresponding to multiple categories through clustering, but its performance is poor (as we demonstrated in Section 3.4). Therefore, even though their features lie within the same feature space, the decomposed features clearly do not belong to that space.

3 Experiment

3.1 Dataset

We evaluate the proposed C2MF framework on two benchmark datasets: **CUB-200-2011 (CUB)** (Wah et al., 2011) and **ImageNet** (Deng et al., 2009). All the details of the experiment are available in our open source code.[2]

[2]https://anonymous.4open.science/r/C2MF-E760/

3.2 Models and Baselines

To evaluate the proposed C2MF framework, we adopt two representative backbone architectures: **NF-ResNet50** (Brock et al., 2021) and **ViT-B/32** (Dosovitskiy et al., 2021), and compare them against three baselines, including our own method (**Ours**) and two competitive baselines: **CRAFT** and **PCA**.

3.3 Evaluation Metrics

Concept faithfulness evaluation based on insertion & deletion scores (Petsiuk et al., 2018) To assess the importance and faithfulness of each concept, we first perform a ranking based on their relative significance. Specifically, we employ either Grad-CAM (Selvaraju et al., 2017) or Integrated Gradients (Sundararajan et al., 2017) to compute the contribution of each concept, and then evaluate their importance and faithfulness using the Insertion & Deletion scores. Concretely, Grad-CAM and IG estimate the impact of each concept on the model's output by leveraging gradient information, thereby producing spatial importance weights for concepts. Based on these weights, we rank the concepts in descending order of importance. In the insertion test, concepts are sequentially added to the input according to their importance (with the most important concepts inserted first), and changes in model accuracy are recorded. Conversely, in the deletion test, concepts are progressively removed. Ideally, inserting highly important concepts should significantly improve accuracy (with a larger area under the insertion curve being preferable), while their removal should substantially reduce accuracy (with a smaller area under the deletion curve being better).

Concept sparsity evaluation based on pearson correlation and cosine similarity Ensuring low correlation between concepts during the extraction and selection process is crucial. This helps reduce redundancy among concepts, enhances the independence of each concept, and improves both the interpretability and generalization ability of the model. Therefore, we require the extracted concepts to be as orthogonal as possible, meaning their similarity should be minimized so that each concept focuses on distinct visual factors. To assess the correlation between extracted concepts, we measure their pearson correlation and cosine similarity. Here lower pearson correlation and cosine similarity indicate lower correlation between concepts.

Concept reconstruction accuracy in each category For each category, we use the concept basis W^* obtained from the training set to project the activation matrix of the test set A_{test} onto this basis (i.e., by fixing W^* and solving for the new coefficient matrix U_{test} such that $A_{\text{test}} \approx U_{\text{test}} W^*$). The reconstructed features $U_{\text{test}} W^*$ are then fed into the classifier, and its classification accuracy is measured. This accuracy reflects the ability of the concept reconstruction to preserve class-specific feature information. We compute the reconstruction accuracy for each category on the test set and report the mean and standard deviation as statistical measures.

Decomposition consistency evaluation based on mean square error To evaluate the consistency of the decomposition between the training and test sets, we compute a distance metric between the original activation matrix A and the reconstructed matrix UW, using the Frobenius norm as the reconstruction loss. Since our decomposition objective is to minimize this reconstruction loss, a lower loss indicates higher decomposition quality and more expressive concept representations. Ideally, if the concept basis W generalizes well to the test set, the reconstruction error on the test set should be close to that on the training set. We compute the Mean Squared Error (MSE) between the original activation matrix A and the reconstructed matrix UW separately for the training and test sets, and calculate the difference between training and test MSEs to compare the train-test discrepancy of each method. A smaller difference indicates stronger generalization of the extracted concepts between the training and test sets, meaning the decomposition is more consistent.

Ablation evaluation of multi-class concept clustering based on concept reconstruction accuracy We conduct an ablation experiment on multi-class concept clustering to validate the rationale of using the *Sigmoid* function as a concept weight activation in the context of multi-class image inputs. Methods such as CRAFT typically extract concepts independently for each class (e.g., extracting k concepts per class), then cluster the extracted concepts into a number of clusters equal to the number of classes. The concepts closest to the cluster centroids are selected as the corresponding concepts for each class. In contrast, our method enables cross-class concept sharing through the *Sigmoid* activation function, while directly unifying the training and extraction of concepts that correspond one-to-one with each class.

3.4 Experimental results

Experimental results of insertion & deletion scores In this experiment, we comprehensively compared the average Insertion and Deletion scores of different baseline methods under two concept importance ranking criteria, Grad-CAM and IG, across two datasets and two models. Specifically, as shown in Figure 2, under the Grad-CAM ranking, the Deletion curves of all three baseline methods decrease as the number of deletion steps increases, indicating that the model's confidence drops when important features are progressively removed. The decline rate of the Ours curve is slower than that of CRAFT, but it clearly outperforms PCA and achieves better performance than CRAFT after the 10th step. Regarding the Insertion curves, our method increases rapidly and approaches nearly 1.0, demonstrating that progressively inserting features can quickly restore model confidence, whereas CRAFT and PCA perform poorly.

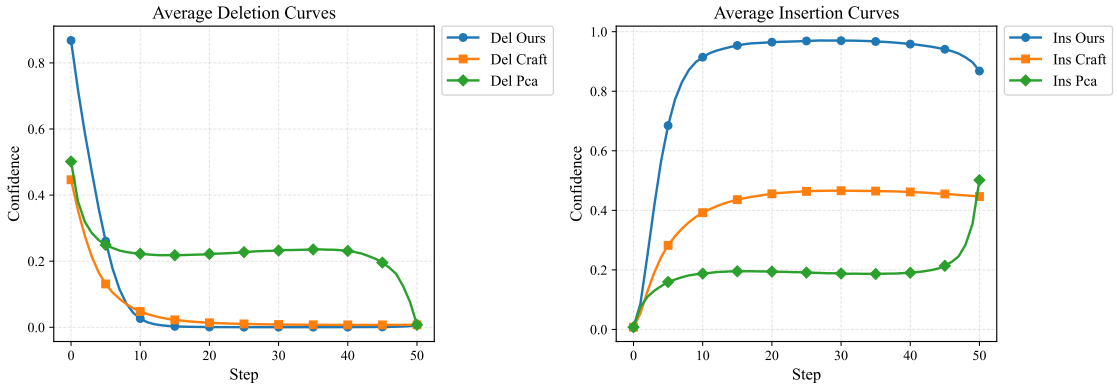


Figure 2: Average Insertion curve and average Deletion curve for Grad-CAM concept importance ranking.

On the other hand, as shown in Figure 3, under the IG ranking, the Deletion curves of all three baseline methods stabilize and approach the CRAFT and PCA curves after the 10th step. For the Insertion curve, our method also achieves the largest AUC. In summary, our method leads in the Insertion score and performs consistently in the Deletion score. The specific ROC value of Insertion & Deletion score can be found in our App D.

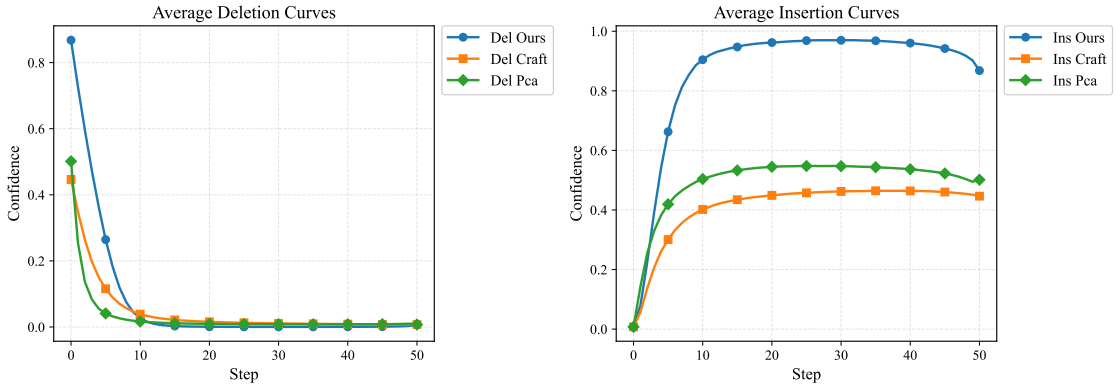


Figure 3: Average Insertion curve and average Deletion curve for IG concept importance ranking.

Experimental results of pearson correlation and cosine similarity In this experiment, we present the experimental results for concept sparsity in Table 1. For the CRAFT method, the pearson correlation and cosine similarity remain within the range of 0.1 to 0.3 across all datasets and models, indicating a high degree of redundancy in the learned concepts. Most of the concepts capture similar visual features. In contrast, for our method, the pearson correlation ranges from 0.068 to 0.089, and the cosine similarity ranges from 0.065 to 0.096. Compared to the sparsity metrics of CRAFT, our method reduces these values by 35-65%. This demonstrates that our method reduces redundant features, maintaining the independence

Table 1: Experimental results for concept sparsity (lower is better). We report mean ± std of pairwise similarity between learned concepts using Pearson correlation and Cosine similarity across datasets and backbones. The best results are highlighted in bold.

	CUB / NF-ResNet50		CUB / ViT-B/32		ImageNet / NF-ResNet50		ImageNet / ViT-B/32	
Variant	Pearson	Cosine	Pearson	Cosine	Pearson	Cosine	Pearson	Cosine
Ours	**0.089 ± 0.018**	**0.096 ± 0.020**	**0.079 ± 0.023**	**0.079 ± 0.023**	**0.068 ± 0.018**	**0.065 ± 0.015**	**0.069 ± 0.021**	**0.069 ± 0.021**
Craft	0.112 ± 0.015	0.256 ± 0.031	0.226 ± 0.054	0.319 ± 0.051	0.097 ± 0.026	0.191 ± 0.047	0.168 ± 0.060	0.243 ± 0.060
PCA	0.007 ± 0.008	0.000 ± 0.000	0.001 ± 0.000	0.000 ± 0.000	0.007 ± 0.007	0.000 ± 0.000	0.001 ± 0.000	0.000 ± 0.000

between features. Notably, the pearson correlation and cosine similarity for the PCA method are almost zero, which nearly eliminates the correlation between concepts. This phenomenon arises because PCA enforces orthogonality between the principal components, ensuring "sparsity." However, these components lack clear semantic interpretation, meaning that although the sparsity metrics are the lowest, they do not provide practical interpretability. We also use a case study in the App. E to demonstrate the semantic interpretation of concepts extracted by the C2MF method using the CLIP model.

Table 2: Per-class reconstruction accuracy (%) across datasets and backbones. Mean ± std. Best results are highlighted in bold.

	CUB / NF-ResNet50			CUB / ViT-B/32			ImageNet / NF-ResNet50			ImageNet / ViT-B/32		
#Concepts	Ours	Craft	PCA	Ours	Craft	PCA	Ours	Craft	PCA	Ours	Craft	PCA
10	**98.6 ± 5.4**	71.8 ± 38.0	34.5 ± 40.3	**98.3 ± 6.5**	43.1 ± 40.1	65.6 ± 34.5	**98.7 ± 3.7**	94.4 ± 19.4	94.5 ± 19.0	**95.1 ± 11.1**	88.6 ± 20.5	90.7 ± 19.5
20	**97.0 ± 7.1**	74.9 ± 35.4	44.6 ± 38.8	**99.6 ± 1.4**	44.7 ± 38.5	64.7 ± 31.8	**97.9 ± 5.0**	95.1 ± 17.4	93.5 ± 20.0	**96.9 ± 6.0**	86.3 ± 21.8	89.8 ± 16.8
30	**95.7 ± 8.3**	75.8 ± 33.2	52.3 ± 36.4	**99.7 ± 1.1**	47.7 ± 36.7	65.2 ± 29.2	**97.5 ± 5.8**	95.2 ± 17.1	95.1 ± 15.8	**96.6 ± 5.8**	86.5 ± 20.0	90.5 ± 12.8
40	**94.5 ± 9.3**	76.2 ± 33.0	54.1 ± 36.3	**99.5 ± 1.6**	49.1 ± 36.6	65.1 ± 29.1	**97.5 ± 6.1**	95.0 ± 17.2	95.3 ± 12.7	**96.7 ± 6.2**	86.0 ± 19.6	88.1 ± 16.2
50	**93.9 ± 10.0**	76.6 ± 32.8	56.1 ± 34.6	**99.3 ± 1.9**	50.9 ± 36.0	66.5 ± 27.8	**97.1 ± 7.2**	95.1 ± 17.1	95.4 ± 13.8	**96.6 ± 6.2**	86.2 ± 18.9	88.1 ± 16.7

Experimental results of concept reconstruction accuracy in each category In this experiment, we evaluated the per-class reconstruction accuracy of three baseline methods across four different experimental setups. As shown in Table 2, for the CUB/NF-ResNet50 setup, as the number of extracted concepts increased, the reconstruction accuracy of our method decreased from 98.6% to 93.9%. The reconstruction accuracy of the CRAFT method increased from 71.8% to 76.6%, but still remained below the minimum accuracy of our method. The reconstruction accuracy of the PCA method rose from 34.5% to 56.%, showing the lowest overall accuracy, indicating that it nearly failed to reconstruct and discriminate features. Under optimal conditions, our method outperformed CRAFT by 26.8% and PCA by 64.1%. For the CUB/ViT-B/32 setup, as the number of extracted concepts increased, the reconstruction accuracy of our method increased from 98.3% to 99.3%. The reconstruction accuracy of the CRAFT method increased from 43.1% to 50.9%, while the PCA method's reconstruction accuracy slightly decreased from 65.6% to 65.5%. Under optimal conditions, our method outperformed CRAFT by more than 50% and led PCA by about 35%. For the ImageNet/NF-ResNet50 setup, as the number of extracted concepts increased, our method's reconstruction accuracy decreased from 98.7% to 97.1%. The CRAFT method's reconstruction accuracy increased from 94.9% to 95.1%, while the PCA method's reconstruction accuracy rose from 94.5% to 95.4%. Despite the smaller accuracy differences among the three methods, our approach still outperformed the baselines by 2%–3%, maintaining the best performance. For the ImageNet/ViT-B/32 setup, as the number of extracted concepts increased, our method's reconstruction accuracy increased slightly from 96.5% to 96.6%. The CRAFT method's reconstruction accuracy decreased from 86.3% to 86.2%, and the PCA method's reconstruction accuracy decreased from 90.8% to 88.8%. Under optimal conditions, our method outperformed CRAFT by about 10% and PCA by 7%–8%.

Table 3: Consistency check (MSE) across datasets and backbones. Mean ± std. Variation = Train - Test (lower is better). Best results are highlighted in bold.

	CUB / NF-ResNet50			CUB / ViT-B/32			ImageNet / NF-ResNet50			ImageNet / ViT-B/32		
Variant	Train	Test	Variation	Train	Test	Variation	Train	Test	Variation	Train	Test	Variation
Ours	83.47 ± 19.56	54.29 ± 18.91	**29.18**	1.21 ± 0.18	0.91 ± 0.21	**0.30**	0.84 ± 0.19	0.73 ± 0.18	**0.11**	1.36 ± 0.19	1.22 ± 0.21	**0.14**
Craft	158.21 ± 36.66	1.87 ± 0.46	156.34	2.55 ± 0.09	1.01 ± 0.06	1.54	1.14 ± 0.19	0.10 ± 0.02	1.04	2.49 ± 0.10	1.14 ± 0.07	1.35
PCA	176.33 ± 40.48	5.56 ± 1.10	170.77	1.80 ± 0.21	0.93 ± 0.30	0.87	1.20 ± 0.20	0.19 ± 0.03	1.01	1.85 ± 0.20	1.06 ± 0.29	0.79

Experimental results of mean square error In this experiment, we compared the training-test Mean Squared Error (MSE) differences across various methods. Specifically, as shown in Table 3, for the CUB/NF-

ResNet50 setup, our method exhibited the smallest training-test MSE variation, with a value of 29.18. In contrast, both CRAFT and PCA showed severe overfitting, with variations of 156.34 and 170.77, respectively. In the ImageNet/NF-ResNet50 setup, the training-test MSE variations for all three baseline methods were relatively small, but our method still achieved the optimal result with a difference of 0.30. For the ImageNet/ViT-B/32 setup, our method's difference was 0.14, which is lower than CRAFT's 1.35 and PCA's 0.79. This demonstrates that even with more complex models, our method still exhibits stronger generalizability. In summary, given that our method consistently maintains the smallest training-test MSE variation across all datasets and backbones, it effectively balances generalization and consistency, helping to avoid overfitting.

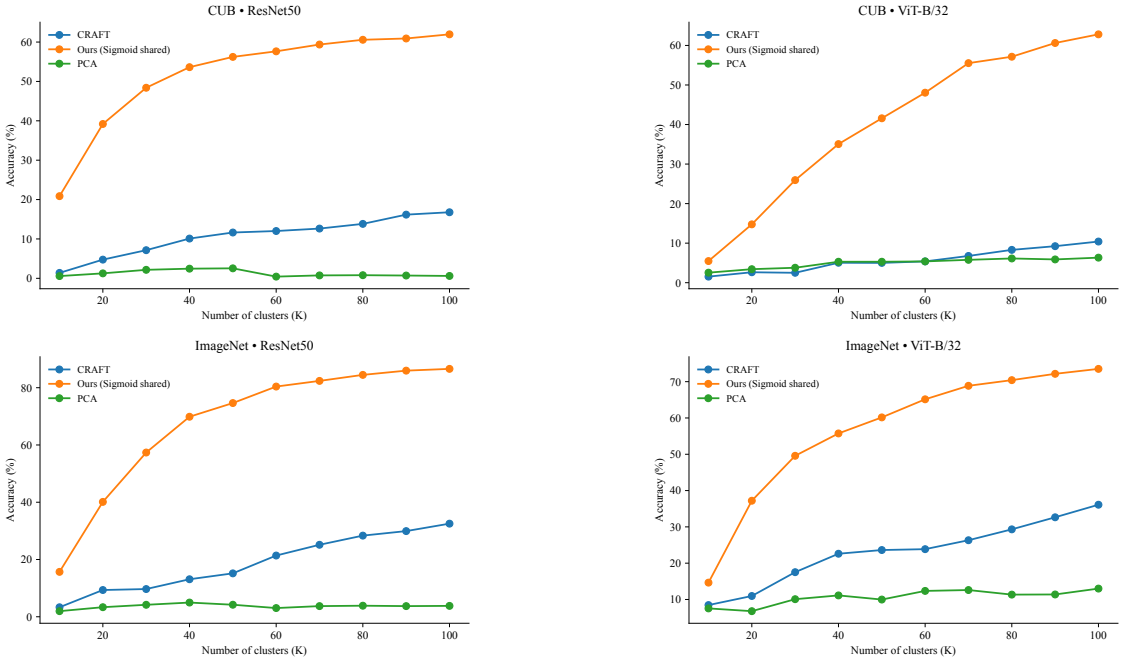


Figure 4: Multi-class concept reconstruction accuracy under different experimental settings.

Multi-class concept clustering ablation analysis In this experiment, we present in Figure 4 the variation of concept reconstruction accuracy with the number of clusters (or the number of extracted concepts). It is evident that the accuracy curve of our method shows the steepest increase and continues to rise as the number of concepts increases. Although CRAFT also improves as the number of clusters increases, its performance remains inferior. The accuracy curve for PCA remains the lowest across all four experimental setups, with minimal fluctuation. This indicates that our shared mechanism based on the *Sigmoid* activation function significantly enhances the performance of multi-class concept reconstruction. Furthermore, the minimal fluctuation in PCA's accuracy as the number of clusters increases corroborates the inability of PCA to form effective discriminative concepts.

4 Conclusion

In this work, we presented C2MF, a novel framework for interpretable concept learning that addresses key limitations of existing methods. By maintaining distributional consistency through full-image representations, relaxing restrictive non-negativity constraints, and introducing a shared global concept dictionary, C2MF achieves both higher flexibility in concept activation and stronger cross-category interpretability. Extensive experiments on large-scale benchmarks confirm its superior performance in terms of concept fidelity, classification accuracy, and generalization ability. These results highlight the potential of C2MF as a principled approach for enhancing transparency and trustworthiness in deep neural networks, particularly in domains where reliable interpretability is essential.

Ethics Statement

We have read and will adhere to the ICLR Code of Ethics. This work uses only public data, involves no human subjects or personally identifiable information, and therefore does not require IRB review. Results are reported for research purposes only; we release anonymized code/configurations to support verification, and will disclose any funding sources and potential conflicts of interest upon acceptance.

Reproducibility Statement

To support reproducibility, we release an anonymized repository with all experiment details including training/evaluation scripts, default hyperparameters, configuration files, and software/hardware environment.

References

Ighoyota Ben Ajenaghughrure, Sonia Claudia da Costa Sousa, and David Lamas. Risk and trust in artificial intelligence technologies: A case study of autonomous vehicles. In *2020 13th international conference on human system interaction (HSI)*, pp. 118–123. IEEE, 2020.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pp. 1059–1071. PMLR, 2021.

Maelenn Corfmat, Joé T Martineau, and Catherine Régis. High-reward, high-risk technologies? an ethical and legal account of ai development in healthcare. *BMC medical ethics*, 26(1):4, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53(8):5455–5516, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Neehar Kondapaneni, Markus Marks, Oisin Mac Aodha, and Pietro Perona. Less is more: Discovering concise network explanations. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. *arXiv preprint arXiv:2503.15699*, 2025.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

Jae Hee Lee, Georgii Mikriukov, Gesina Schwalbe, Stefan Wermter, and Diedrich Wolter. Concept-based explanations in computer vision: Where are we and where could we go? In *European Conference on Computer Vision*, pp. 266–287. Springer, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms. In *International Conference on Machine Learning*, pp. 30646–30688. PMLR, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Hengyi Wang, Shiwei Tan, and Hao Wang. Probabilistic conceptual explainers: Towards trustworthy conceptual explanations for vision foundation models. In *Proceedings of the Forty-First International Conference on Machine Learning (ICML)*, 2024.

Yue Wang and Sai Ho Chung. Artificial intelligence in safety-critical systems: a systematic review. *Industrial Management & Data Systems*, 122(2):442–470, 2022.

Jiayu Zhang, Xinyi Wang, Zhibo Jin, Zhiyu Zhu, Jianlong Zhou, Fang Chen, and Huaming Chen. Splitting & integrating: Out-of-distribution detection via adversarial gradient attribution. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=b1Bae7TKmw`.

Zhiyu Zhu, Huaming Chen, Xinyi Wang, Jiayu Zhang, Zhibo Jin, Jason Xue, and Jun Shen. Iterative search attribution for deep neural networks. In *Forty-first International Conference on Machine Learning*, 2024a.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Jason Xue, and Flora D Salim. Attexplore: Attribution for explanation with model parameters exploration. In *The Twelfth International Conference on Learning Representations*, 2024b.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17228–17236, 2024c.

LLM Usage Disclosure

We used large language models (OpenAI GPT-4o and GPT-5) as auxiliary tools for grammar checking and language polishing of the manuscript. These models were not involved in research ideation, experimental design, implementation, or analysis. The authors take full responsibility for all content.

A Related Work

A.1 Traditional attribution methods

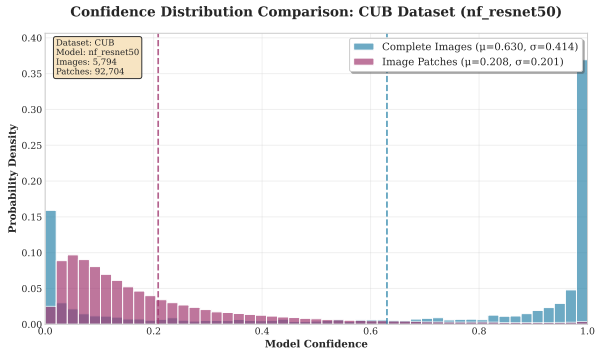
Traditional neural network interpretability methods primarily rely on feature attribution to generate pixel-level or region-level heatmaps. For instance, saliency maps, Grad-CAM (Selvaraju et al., 2017), and surrogate model-based methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) explain model decisions by assigning importance scores to input pixels or regions. These attribution methods provide intuitive insights into where the input contributes to the model's output, but they typically fail to clarify which high-level semantic concepts the model has focused on. In other words, these methods reveal the input regions that influence decision-making but do not explain the concepts identified by the model within these regions. As a result, they cannot uncover the abstract concepts that the model attends to. This limitation is also present in state-of-the-art attribution methods such as MFABA (Zhu et al., 2024c), ISA (Zhu et al., 2024a) and AttEXplore (Zhu et al., 2024b).

A.2 Typical attribution decomposition methods: TCAV and CRAFT

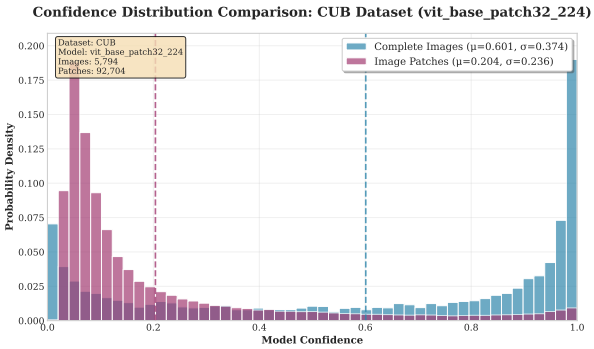
In recent years, a class of methods has attempted to decompose model explanations into interpretable concepts. Kim et al. (2018) proposed the Testing with Concept Activation Vectors (TCAV) method, which trains concept vectors (CAVs) based on a set of concept examples provided by the user, and uses the directional derivative to measure the model's sensitivity to a specific concept. For example, TCAV can quantify the model's response to the concept of "stripes" when predicting "zebra." This method tests the model with a predefined set of concepts but requires manually collecting concept examples and training the concept vectors. The Concept Recursive Activation Factorization (CRAFT) method, proposed by Fel et al. (2023), further automates the extraction of concepts and generates concept-based explanations. CRAFT recursively factorizes activations across network layers to detect concepts and uses the Sobol index to more accurately estimate the importance of each concept. In addition, CRAFT simultaneously generates attribution maps in the concept space, answering both *what* the model attends to and *where*, thereby partially addressing the limitation of traditional heatmap methods that can only answer *where* the model focuses.

A.3 Explored Concept-Based Interpretability Methods

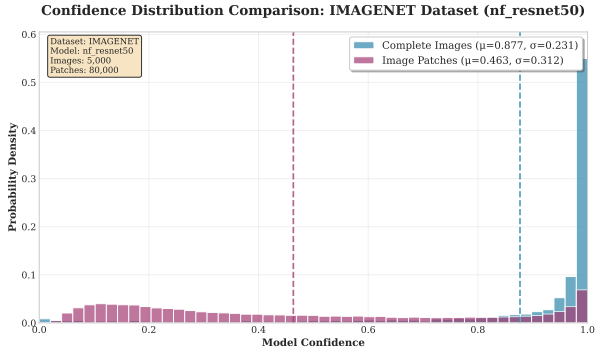
In recent years, researchers have proposed various concept-based interpretability methods to understand and compare the internal representations of deep models. An existing review work (Lee et al., 2024) have systematically examined this field and suggested that future developments should focus on three dimensions: concept types, representation methods, and concept control, with a particular emphasis on integrating knowledge representation and logical reasoning to further enhance model interpretability and controllability. More specifically, the RSVC method (Kondapaneni et al., 2025) decomposes model activations into concept bases and corresponding coefficients, and measures the similarity between model representations at the concept level by comparing the consistency of concept responses across different models on a shared image set. To further simplify concept explanations, Kondapaneni et al. (2024) proposed the DCNE framework, which leverages a large number of neuron attribution maps generated by CRP, and refines them into a small set of representative concepts through NMF compression and DBSCAN clustering, thereby achieving a more concise explanation. In the area of concept modeling, Wang et al. (2024) introduced the PACE framework, which represents the features of Vision Transformers as multi-level Gaussian mixture concepts using hierarchical Bayesian modeling, providing stable and sparse explanations. Meanwhile, Rao et al. (2024) proposed the DN-CBM method, which automatically discovers concepts from CLIP features using sparse autoencoders, aligns them with the text space, and generates natural language labels, achieving task-agnostic concept discovery without manual pre-definition. On the other hand, to compare the differences between models, Kondapaneni et al. (2025) introduced the RDX method, which locates sample clusters that are similar in one model but dissimilar in another by analyzing the differences in the similarity matrix using spectral clustering, thus revealing differences at the representation level. Shah et al. (2023) proposed the



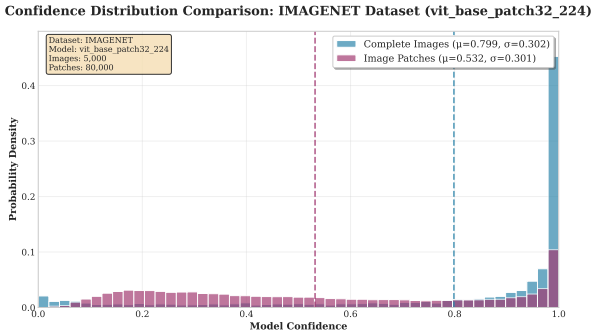
(a) CUB — NF-ResNet50



(b) CUB — ViT-B/32



(c) ImageNet — NF-ResNet50



(d) ImageNet — ViT-B/32

Figure 5: Confidence distributions of *complete images* vs. *image patches* on CUB and ImageNet with NF-ResNet50 and ViT-B/32. Complete-image mass concentrates in a high-confidence region, whereas patch scores shift toward lower confidence, yielding little overlap.

ModelDiff framework, which combines data models and input transformation designs to identify and validate the key features relied upon by different learning algorithms.These frameworks demonstrate the significant potential of concept decomposition in practical applications.

B OOD Detection Evaluation

Setup and metrics. We report standard score-only OOD detection metrics on *CUB* and *ImageNet* with two backbones (NF-ResNet50, ViT-B/32). Higher is better for **AUROC** and **AUPR** (In/Out); lower is better for **FPR@95TPR**. "AUPR-In" uses in-distribution as the positive class, while "AUPR-Out" treats OOD as positive.

Dataset	Backbone	AUROC↑	FPR@95TPR↓	AUPR-In↑	AUPR-Out↑
CUB	NF-ResNet50	0.8701	0.2442	0.7354	0.9790
CUB	ViT-B/32	0.9282	0.4249	0.4944	0.9951
ImageNet	NF-ResNet50	0.8804	0.7284	0.2348	0.9910
ImageNet	ViT-B/32	0.8400	0.5872	0.3053	0.9854

Table 4: OOD detection results (higher is better except FPR@95TPR).

C Confidence Distributions: Complete Images vs. Image Patches

As illustrated in Fig. 5, the confidence distributions of complete images and image patches exhibit minimal overlap across datasets and backbones; see Figures 5a to 5d for per-setting views.

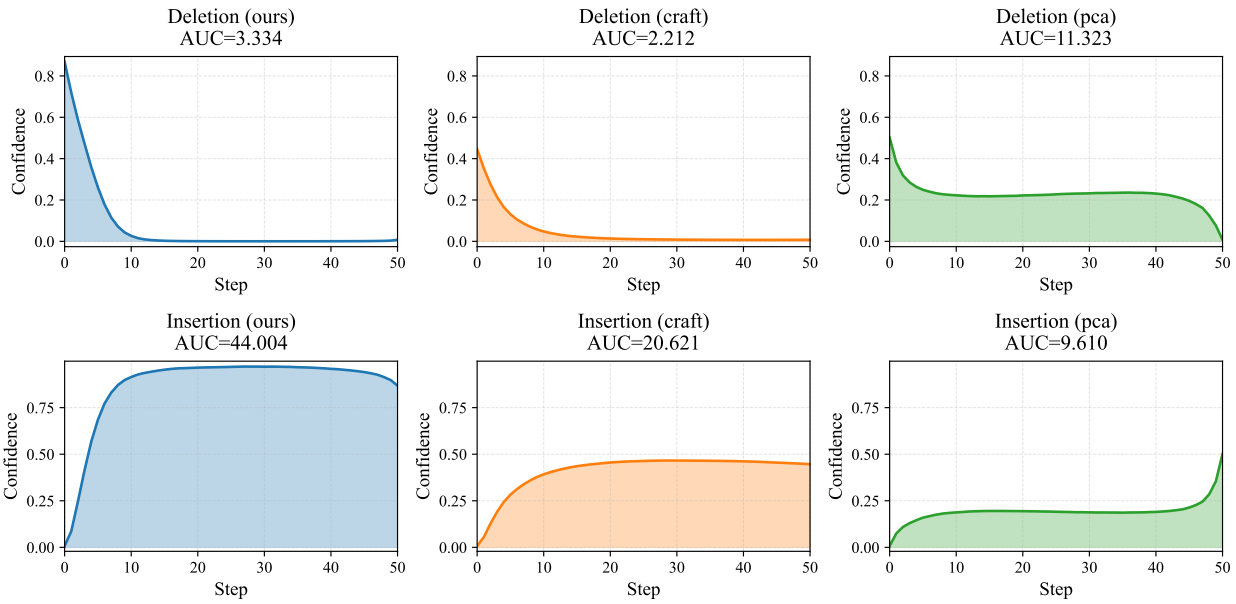


Figure 6: Average AUC for Grad-CAM concept importance ranking.

We compare the model confidence distributions of *complete images* and their *image patches* across two datasets (CUB, ImageNet) and two backbones (NF-ResNet50, ViT-B/32). Across all settings, the two distributions exhibit clearly separated modes with large mean gaps, indicating that patches occupy a markedly lower-confidence regime while full images concentrate in a high-confidence regime.

Observed separation. On **CUB**, the mean confidence of complete images is $\mu \approx 0.63$ (NF-ResNet50) and $\mu \approx 0.60$ (ViT-B/32), whereas patches concentrate near $\mu \approx 0.21$ in both cases; this yields mean gaps of $\Delta\mu \approx 0.42$ and $\Delta\mu \approx 0.40$, respectively. On **ImageNet**, complete images remain substantially higher ($\mu \approx 0.88$ for NF-ResNet50; $\mu \approx 0.80$ for ViT-B/32) than patches ($\mu \approx 0.46$ and $\mu \approx 0.53$), resulting in $\Delta\mu \approx 0.41$ and $\Delta\mu \approx 0.27$. In all four cases, the bulk of probability mass for patches lies in the low-to-mid confidence range, while complete-image mass is skewed toward high confidence, yielding minimal overlap of the dominant density regions.

Implication. The near-separable confidence profiles confirm that cropping to patches induces a systematic distribution shift toward lower confidence, making patches a practical proxy for "hard" or out-of-distribution-like inputs. This observation complements our OOD results (App. B) by explaining *why* patch-level inputs are more frequently rejected at high-recall operating points: their scores rarely intrude into the high-confidence region where complete images dominate.

D ROC value of Insertion & Deletion score

E Case study

In this section, since concepts in interpretable deep learning typically refer to abstract, reusable feature patterns, we conducted a case study to provide human-understandable explanations for the concepts extracted by our method. Specifically, we cropped local patches from the original images and fed them into the CLIP model (Radford et al., 2021). These patches, which visually instantiate the concepts, can then be matched with WordNet, allowing the output to provide the semantically closest textual representation of the concept. As shown in Figure 8, We can observe that for concept 0, the most representative semantic terms are "bird," "tern," and "penguin". For concept 1, the most representative semantic terms are "cock," "bird," and "blackcock". This result demonstrates that the extracted concepts are not only visually distinct but also semantically interpretable, aligning with human understanding. By mapping visual features to semantically meaningful words, our approach offers clear, traceable explanations for the learned concepts. This enhances

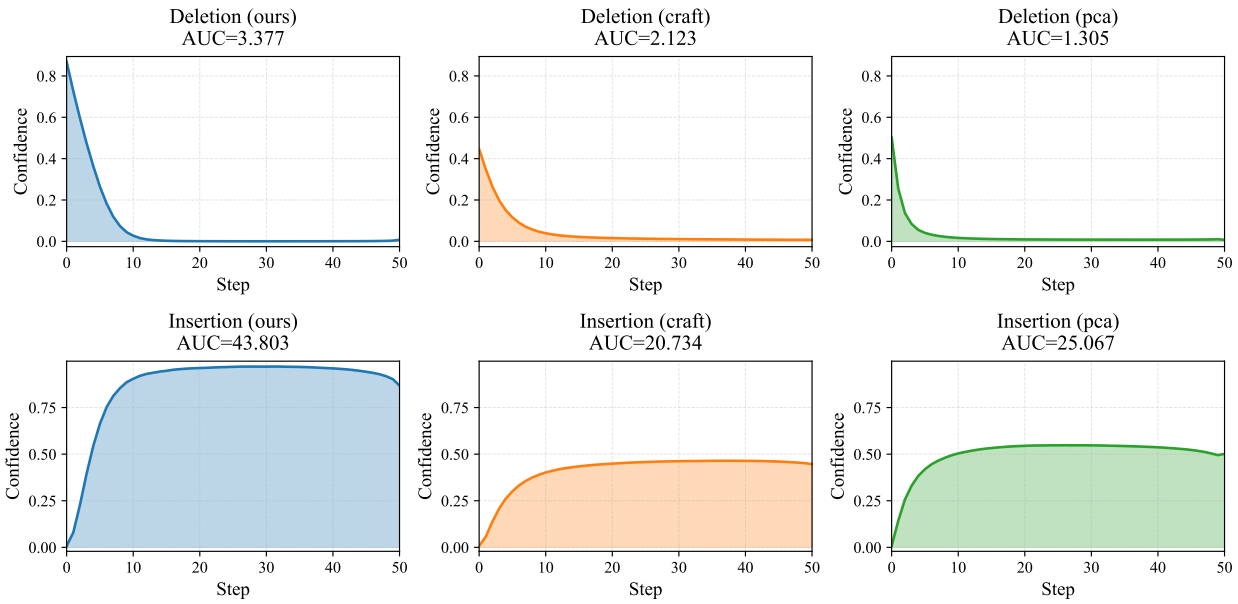


Figure 7: Average AUC for IG concept importance ranking.

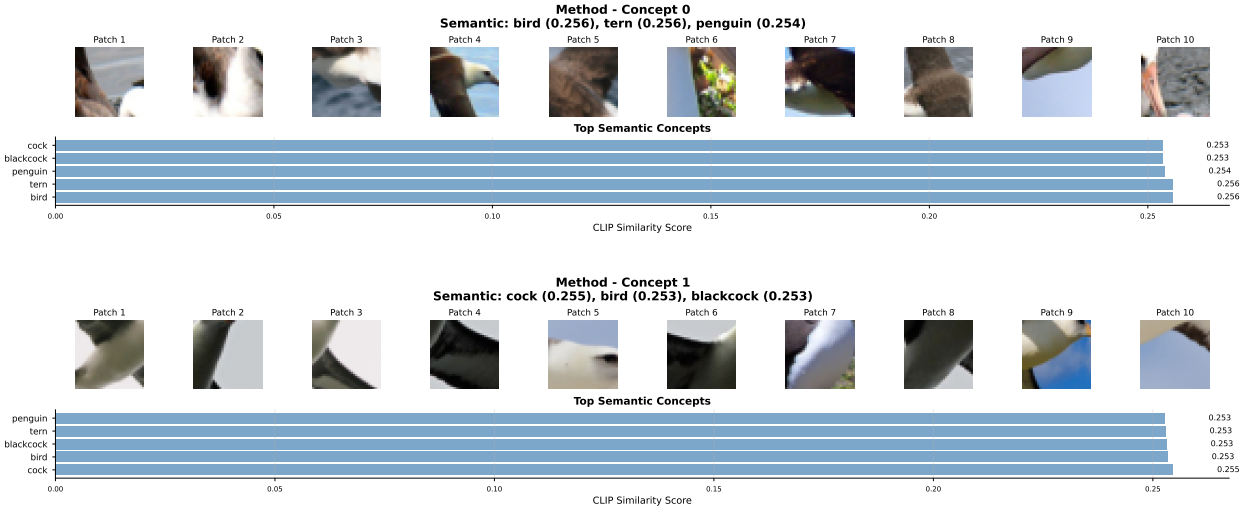


Figure 8: Semantic interpretation of extracted concepts using CLIP model

the transparency of the model and provides insight into how it interprets various features, thus increasing both its interpretability and trustworthiness in high-stakes applications.