Information Sciences 723 (2026) 122648

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Unified graph-based framework for visual explainability in convolutional neural networks

Basim Azam [a,b,*], Pubudu Sanjeewani [a], Brijesh Verma [a], Ashfaqur Rahman [c], Lipo Wang [d]

[a] Griffith University, Brisbane, QLD, Australia
[b] CIS, University of Melbourne, Melbourne, VIC, Australia
[c] Data61, CSIRO, Hobart, TAS, Australia
[d] School of Electrical and Electronic Engineering, NTU, Singapore

ARTICLE INFO

Keywords:
Explainable AI
Grad-CAM++
Integrated gradients
Graph
CNN

ABSTRACT

In deep learning, understanding the decision-making processes of complex models is essential for advancing interpretability and trust in artificial intelligence systems. We introduce Causal Relational Attribution Graph (C-RAG), designed to deliver comprehensive, multi-perspective explanations of convolutional neural networks (CNNs) via a graph representation. C-RAG integrates gradient-based local attribution with global feature importance by constructing a graph-based representation that captures hierarchical feature inter-dependencies. In this framework, feature clusters are represented as graph nodes, and their interactions are quantified through combined localized and global attribution metrics, ensuring interpretable insights into model behavior. We evaluate C-RAG across diverse benchmark datasets (ImageNet, CIFAR-10, MNIST) and CNN architectures (ResNet18, VGG19, DenseNet201, LeNet), demonstrating significant advancements over state-of-the-art explainability methods in faithfulness, robustness, and computational efficiency. The proposed approach facilitates accurate spatial feature localization, robust dependency mapping, and efficient explanation generation, making it a valuable tool for critical applications such as medical imaging and autonomous systems. We provide a novel graph-based explainability framework, which bridges the gap between local and global interpretability, C-RAG addresses key limitations in existing methods, establishing a robust foundation for explainable AI in computer vision.

1. Introduction

Deep learning models have demonstrated exceptional performance across a wide range of applications, including computer vision [1] and natural language processing [2]. Despite these advancements, their intricate architectures and lack of transparency in decision-making have sparked significant concerns regarding interpretability and trust, particularly in critical domains such as healthcare [3] and autonomous systems [4]. The opacity of these systems remains a critical obstacle, hindering the broader acceptance and deployment of artificial intelligence technologies [5].

* Corresponding author.
E-mail address: basim.azam@unimelb.edu.au (B. Azam).

https://doi.org/10.1016/j.ins.2025.122648
Received 10 March 2025; Received in revised form 25 August 2025; Accepted 26 August 2025

Available online 4 September 2025

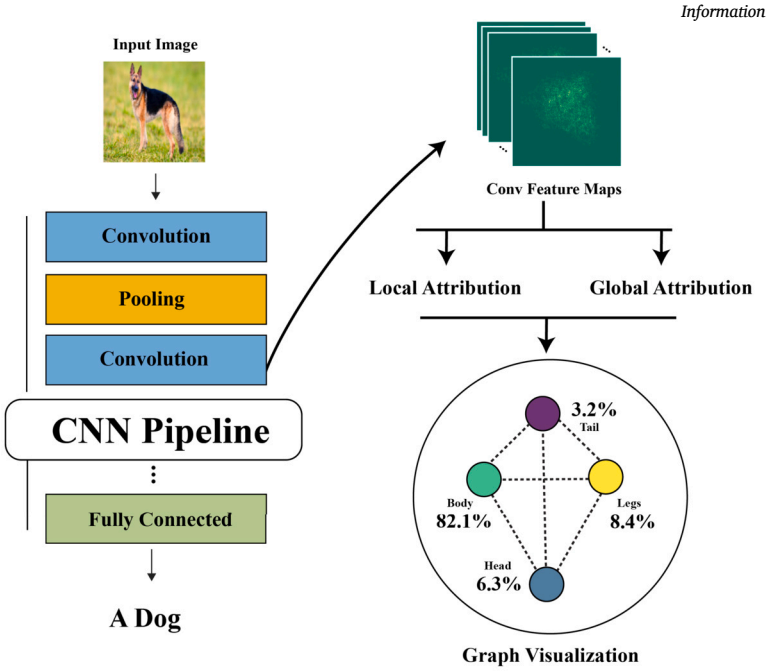

Fig. 1. C-RAG overview: integrating localized gradient-based heatmaps and global attributions into unified graph for multi-scale explainability.

Explainable AI (XAI) has become a pivotal area of research, dedicated to addressing the interpretability challenges associated with complex AI systems [6]. Methods developed under this paradigm can be broadly categorized into two approaches: local methods, which elucidate individual predictions, and global methods, which provide an understanding of the model's overall behavior [7]. Local techniques, such as Local Interpretable Model-agnostic Explanations (LIME) [8] and SHapley Additive exPlanations (SHAP) [9], are designed to interpret specific instances. In contrast, global methods, including feature importance analysis and partial dependence plots, aim to reveal patterns and dependencies that underpin the model's decision-making process [10].

While existing techniques have made significant strides in enhancing the interpretability of AI systems, most approaches are confined to providing either localized or global explanations in isolation [11]. This separation creates a fragmented understanding of model behavior [12], limiting users' ability to integrate insights from both granular, instance-specific rationales and overarching, model-wide patterns [13]. A unified framework that combines these perspectives is essential to fostering a comprehensive understanding of model decision-making processes, particularly in contexts where both fine-grained and holistic explanations are crucial for building trust and reliability.

This research addresses the disconnect between local and global interpretability methods by integrating both localized and globalized attribution of the features into a unified framework that generates graph-based visualizations. As illustrated in Fig. 1, the input image is processed through a convolutional neural network (CNN) to produce feature maps. The proposed methodology operates in two key stages. Initially, it detects critical regions of interest within the input data [14]. Subsequently, these salient regions are transformed into a graph-based representation, highlighting feature interactions and dependencies [15]. Alternative fusion methods, multiscale saliency fusion [16] and self-attention overhead [17] either lack relational modeling or incur high overhead. In contrast, a graph structure naturally capture inter-feature dependencies with linear complexity in the edge count. By capturing inter-feature dependencies within a unified graph, C-RAG overcomes fragmentation in existing fusion methods and enables holistic interpretability.

This two-layer approach combines localized explanations, using visual heatmaps, with global insights derived from graph structures, offering a more comprehensive view of the model's decision-making process [18]. By integrating visual and structural interpretability, the framework enhances the clarity and transparency of complex deep learning systems [19]. This holistic approach not only facilitates debugging and optimization but also bolsters trust and accountability in the deployment of AI models across diverse application domains. However, existing explainability methods treat local and global insights in isolation leaving a fragmented understanding of model behavior. To bridge this gap, we propose Causal Relation Attribution Graph, a unified framework that integrates local and global attributions into a graph structure, enabling comprehensive multi-perspective explanations.

2. Related work

XAI has become a critical research focus due to concerns about the transparency of deep learning models, particularly in domains like healthcare [20] and autonomous driving [21]. These models are often perceived as "black boxes," which limits trust and adoption [22]. To mitigate this, interpretability methods have been developed, broadly categorized into model-specific and model-agnostic approaches. The former includes techniques tailored to specific architectures such as convolutional neural networks and graph neu-

Table 1
Comparison of Graph-based XAI Methods.

Method	Subgraph-level	Pixel-level Attributions	Spatial Grounding	Relational Modeling
GraphLIME [18]	✓	✗	✗	✓(HSIC Lasso)
GNNExplainer [30]	✓	✗	✗	✓(edge-focused)
C-RAG (Ours)	✓	✓	✓	✓(graph-based paths)

ral networks, while the latter can be applied to any model type. This review systematically explores recent advancements in both categories.

Gradient-based techniques have become prominent for explaining CNNs, particularly in computer vision tasks [23]. These methods aim to compute the gradient of the model's output with respect to the input features, highlighting regions that have the greatest influence on the decision-making process [24]. One of the earliest techniques, saliency maps, calculates pixel-wise gradients to highlight important areas in the input [25]. However, these maps can be noisy and imprecise. Grad-CAM improves upon this by providing localized heatmaps to show regions of interest [26]. Integrated Gradients, an axiomatic attribution method, ensures more stable explanations by integrating gradients along the path from a baseline to the actual input [27]. Grad-CAM++ often highlights diffuse regions without structural context [28], while Integrated Gradients can suffer from saturation on repetitive patterns [29].

Perturbation-based methods such as LIME and SHAP provide model-agnostic approaches for interpreting models. LIME perturbs input data and observes how the output changes, constructing a local surrogate model to approximate the behavior of the black-box model [8]. SHAP assigns Shapley values to features, offering a consistent explanation framework based on cooperative game theory [9]. These methods are highly flexible but computationally expensive, especially with high-dimensional data.

Attention mechanisms provide another form of interpretability, particularly in Natural Language Processing (NLP) models. The attention weights from transformers can be visualized to show which parts of the input have the most impact on the model's predictions. This mechanism has been extended to vision tasks, but concerns remain regarding whether attention weights genuinely correlate with model decisions or create an illusion of interpretability. Graph-based techniques have gained prominence in explaining models that handle graph-structured data. For example, GraphLIME extends LIME to explain GNNs [18], while GNNExplainer identifies subgraphs most relevant to the model's predictions [30]. GraphLIME [18] explains GNNs via local supergraphs, but does not integrate local pixel attributions. GNNExplainer [30] identifies subgraphs but lacks spatial grounding. C-RAG differs by unifying pixel-level heatmaps with path-integrated attributions into a single relational graph. These methods are particularly useful in fields such as drug discovery and social network analysis, where data can be represented as graphs (Table 1).

A recent trend in XAI is the integration of multiple explainability techniques into unified frameworks [31]. Combining visual explanations like saliency maps with structural explanations from graph-based methods has enhanced model interpretability by offering both intuitive visualizations and deeper insights into feature relationships [32]. Additionally, digital image masking techniques refine feature maps for clearer interpretability by isolating semantically meaningful regions. Hierarchical frameworks provide explanations at different levels of abstraction, allowing for flexible interpretation of both local and global model behavior [33]. GraphLIME [18] and GNNExplainer [30] focus on subgraph-level explanations but do not provide pixel-level spatial grounding, motivating the need for C-RAG's unified graph that supports both granular and relational insights.

Evaluating the quality of explanations remains a challenge due to the absence of ground truth explanations [23]. However, human evaluations, often conducted through user studies, can be subjective and resource-intensive [7]. Furthermore, concerns have been raised about the robustness of explanations, particularly their vulnerability to adversarial attacks [4]. The aim is to develop more theoretically grounded frameworks. Causal explanations, which seek to identify causal relationships rather than simple correlations, represent an emerging area of interest [34].

3. Proposed method: C-RAG

The *Causal Relational Attribution Graph (C-RAG)* framework is proposed to decompose and interpret the decision-making process of deep neural networks in image classification tasks. C-RAG combines both localized and global feature attribution, and represents hierarchical inter-dependencies among critical features through a structured graph $G = (V, E)$, where nodes V denote feature clusters and edges E quantify interaction strengths between these clusters. This framework not only captures both localized and global feature importance but also models the interactions between critical features through graph-based analysis. The architecture overview is represented in Fig. 2.

Let $f(I) = \{S_1, S_2, \ldots, S_C\}$ represent the output of a neural network for an input image I, with S_c denoting the score for each class $c = 1, 2, \ldots, C$. The predicted class label $\hat{y}$ is given by

$$\hat{y} = \arg\max_c S_c.$$

Our primary aim is not only to predict $\hat{y}$ but also to understand the influence of individual features on this prediction. Specifically, we seek to model the interactions among influential regions within I by representing them as nodes $v_i \in V$ in a graph $G = (V, E)$, where edges $e_{ij} \in E$ have weights that represent the strength of dependency between nodes v_i and v_j.

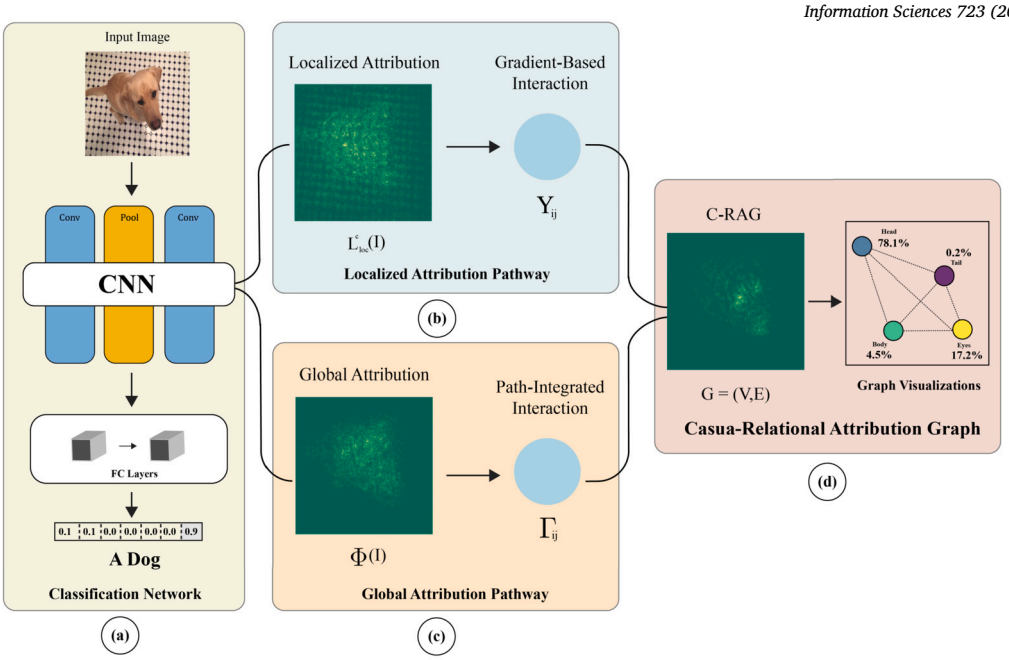

Fig. 2. Detailed architecture of the C-RAG framework: representing localized and global feature attributions as a graph. (a) The input image is processed by a CNN for classification. (b) Localized attribution is computed using Grad-CAM $L^c_{loc}(I)$, highlighting spatial feature importance and producing gradient-based interactions γ_{ij}. (c) Global attribution is computed via Integrated Gradients $\Phi(I)$, yielding path-integrated interactions Γ_{ij}. (d) The C-RAG is constructed with nodes representing feature clusters and edges weighted by combined interaction strengths, providing a relational view of critical features influencing the model's decision.

The first step in C-RAG is to compute a spatially localized relevance map $L^c_{loc}(I)$ for a given class c, highlighting regions of I that have the strongest positive influence on S_c. We derive a gradient-weighted feature importance for each feature map A_k in the final convolutional layer. Let Z denote the spatial size of A_k; we define the relevance coefficient α^c_k for each feature map A_k as

$$\alpha^c_k = \frac{1}{Z}\sum_{i,j}\frac{\partial S_c}{\partial A^{ij}_k},$$

where $\frac{\partial S_c}{\partial A^{ij}_k}$ represents the gradient of S_c with respect to the activation A^{ij}_k at spatial position (i,j). Using these coefficients, we compute the localized relevance map $L^c_{loc}(I)$ as

$$L^c_{loc}(I) = \mathrm{ReLU}\left(\sum_k \alpha^c_k A_k\right).$$

This map $L^c_{loc}(I)$ reveals the spatial distribution of regions in I that significantly contribute to the classification score S_c.

To establish a global perspective, C-RAG also computes path-integrated gradients $\Phi(I)$ to assess feature importance along a continuous path from a baseline image I' to the input I. For each feature i, the path-integrated attribution $\Phi_i(I)$ is defined as:

$$\Phi_i(I) = (I_i - I'_i)\times\int_0^1 \frac{\partial F(I' + \alpha(I - I'))}{\partial I_i}\, d\alpha,$$

where $F(I)$ is the scalar output score for class c (e.g., S_c), and $\alpha \in [0,1]$ is a continuous interpolation parameter. This integral reduces gradient saturation effects and provides a global measure of feature influence on the class score S_c by evaluating cumulative importance from baseline I' to the input I. The resulting $\Phi(I)$ offers a robust, global view of feature significance that complements the localized map $L^c_{loc}(I)$.

C-RAG then constructs a graph $G = (V, E)$ where each node $v_i \in V$ represents a significant feature cluster, determined by both $L^c_{loc}(I)$ and $\Phi(I)$. To capture feature dependencies, each edge e_{ij} between nodes v_i and v_j is weighted by a composite interaction metric, combining both gradient-based and path-integrated strengths.

For gradient-based interaction, we define the interaction strength γ_{ij} between clusters i and j as:

$$\gamma_{ij} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\partial S_c}{\partial A^k_i}\cdot\frac{\partial S_c}{\partial A^k_j}\right),$$

Algorithm 1 Construct Causal Relational Attribution Graph (C-RAG).

Require: Grad-CAM map G_{cam} of shape (h, w), Integrated Gradients map I_g of shape (h, w), number of clusters C (typically 4)
Ensure: Causal Relational Attribution Graph (C-RAG) $G = (V, E)$ with nodes V and weighted edges E
Initialize: Divide G_{cam} and I_g into C clusters (e.g., quadrants: 'top_left', 'top_right', 'bottom_left', 'bottom_right')
Define $V \Leftarrow$ {top_left, top_right, bottom_left, bottom_right}
Initialize empty edge set $E \Leftarrow \{\}$
for each pair of clusters $(i, j) \in V \times V$ **do**
    Compute localized interaction strength $\gamma_{ij} \Leftarrow \text{mean}(G_{cam}[i] \times G_{cam}[j])$
    Compute path-integrated interaction strength $\Gamma_{ij} \Leftarrow \text{mean}(I_g[i] \times I_g[j])$
    Compute combined interaction weight $\omega_{ij} \Leftarrow \gamma_{ij} + \Gamma_{ij}$
    if $w_{ij} > 0$ **then** ▷ Only add edges with significant weights
        Add edge (i, j, w_{ij}) to E
    end if
end for
Construct graph $G = (V, E)$ with nodes V and weighted edges E
return G

where $\frac{\partial S_c}{\partial A_i^k}$ is the gradient of S_c with respect to the activations in cluster i, and N represents the number of images processed. This metric quantifies the localized dependency between clusters, providing insight into their co-influence on the score S_c.

To capture cumulative dependencies along a continuous path, we define the path-integrated interaction strength Γ_{ij} as:

$$\Gamma_{ij} = \frac{1}{N} \sum_{k=1}^{N} \left(\int_0^1 \frac{\partial S_c(I' + \alpha(I - I'))}{\partial A_i^k} \, d\alpha \cdot \int_0^1 \frac{\partial S_c(I' + \alpha(I - I'))}{\partial A_j^k} \, d\alpha \right),$$

where α interpolates from the baseline I' to the input I. This path-integrated interaction reveals dependencies that may only emerge when evaluating cumulative influence, thus capturing both immediate and non-linear dependencies between clusters.

The edge weight ω_{ij} for e_{ij} in the C-RAG graph is defined as the sum of these two interaction metrics:

$$\omega_{ij} = \gamma_{ij} + \Gamma_{ij}.$$

Given the linearity of gradient operators and the law of total attributions, summing γ_{ij} and Γ_{ij} preserves both instantaneous and cumulative dependencies without double-counting, as shown by the additive decomposition of gradient flows [27], which allows C-RAG to provide a comprehensive view of feature dependencies within the model's decision-making process.

The final C-RAG structure, represented as a directed graph $G = (V, E)$ with edges weighted by ω_{ij}, serves as an interpretable, hierarchical representation of the neural network's reasoning. By visualizing G, we can observe how different regions in I interact and contribute to the final prediction $\hat{y}$, offering a detailed, interpretable account of feature interactions. The C-RAG framework therefore provides an advanced, multi-layered structure for understanding and interpreting deep neural network predictions in image classification.

4. Comprehensive evaluation

4.1. Experimental setup

To comprehensively assess the effectiveness and generalization of our proposed framework, we conducted experiments using various datasets, model architectures, and baseline methods. We provide detailed descriptions of each component of our experimental setup.

Widely cited benchmark datasets are used to evaluate our framework across different levels of image complexity and classification tasks: *ImageNet*: A large-scale dataset comprising over 1.2 million high-resolution images across 1,000 classes, serving as a standard benchmark for image classification algorithms. *CIFAR-10*: Consists of 60,000 32×32 color images evenly distributed among 10 classes. The dataset is divided into 50,000 training images and 10,000 test images. *MNIST*: Comprises 70,000 grayscale images of handwritten digits (0-9), each of size 28×28 pixels. The dataset is split into 60,000 training images and 10,000 test images. *PASCAL VOC*: is a benchmark that includes 20 object classes with over 11,000 images widely used for evaluating object detection, segmentation, and classification tasks. Detailed results for CIFAR-10 and MNIST datasets are provided in the supplementary materials.

The framework is evaluated using the following CNN architectures, each representing a distinct approach to deep learning: *ResNet18*: A residual network with 18 layers, employing skip connections to mitigate the vanishing gradient problem, thereby enabling the training of deeper networks. *VGG19*: A 19-layer network characterized by its simplicity and uniform architecture, utilizing small 3×3 convolutional filters throughout. *DenseNet201*: A densely connected network with 201 layers, where each layer receives input from all preceding layers, promoting feature reuse and mitigating the vanishing gradient issue. *LeNet*: One of the pioneering CNN architectures, designed for digit recognition tasks, consisting of two convolutional layers followed by two fully connected layers.

To validate the performance of our proposed framework, we compared it against established explainability methods, Grad-CAM: Generates class-specific localization maps by utilizing the gradients of target concepts flowing into the final convolutional layer. Integrated Gradients: attributes the prediction of a model to its input features by integrating gradients along a path from a baseline

Table 2
Summary of evaluation metrics used to assess explainability methods, categorized into faithfulness, robustness, localization, and complexity.

Category	Metric	Objective
Faithfulness [35]	Faithfulness Correlation (F.C.) Pixel Flipping (P.F.) Monotonicity Correlation (M.C.) Selectivity (Sel.)	Ensures that the explanations align with the model's decision-making process by evaluating the relevance and impact of highlighted features.
Robustness [36]	Continuity Test (Con. T.) Local Lipschitz Estimate (L.Est.) Relative Output Stability (ROS)	Measures the stability of explanations under perturbations, ensuring reliability across diverse inputs and conditions.
Localization [37], [38]	Top-K Intersection (T-K.L.) Relevance Mass Accuracy (RM-A.) Attribution Localization (A.L.)	Assesses the precision of explanations in identifying critical regions, providing spatially accurate and relevant feature mappings.
Complexity [39], [40]	Sparseness (SP.) Complexity (CP.) Effective Complexity (E.CP.)	Evaluates the simplicity and computational efficiency of explanations, ensuring minimal redundancy and manageable computational overhead.

to the input. These baselines were selected due to their widespread adoption and effectiveness in providing visual explanations for CNN decisions.

All experiments were conducted using the HPC facility provided by Griffith University, utilizing the PyTorch deep learning framework. The software environment included Python 3.8, CUDA 11.0, and cuDNN 8.0, ensuring efficient computation and reproducibility of results.

4.2. Evaluation metrics

Explainability in deep learning is multifaceted, requiring rigorous evaluation across multiple dimensions. To ensure a holistic assessment, we categorized our evaluation metrics into four key dimensions: faithfulness, robustness, localization, and complexity. These dimensions capture the essence of what it means for a model's explanations to be accurate, stable, precise, and efficient. Table 2 provides a detailed overview of the metrics employed, along with their significance.

Each category of evaluation metrics is designed to rigorously examine a distinct aspect of model explainability. Faithfulness metrics focus on assessing whether the identified features accurately reflect the model's decision-making process, ensuring that the explanations are directly tied to the underlying predictive logic. Robustness metrics evaluate the stability and consistency of the explanations when subjected to variations in input data or model parameters, which is critical for ensuring reliability in practical applications. Localization metrics measure the spatial precision of the generated explanations, making them particularly significant for domains such as medical imaging, where accurate identification of relevant regions is essential. Lastly, complexity metrics emphasize the efficiency and interpretability of the explanations, ensuring that they are computationally viable, concise, and devoid of unnecessary redundancy. This comprehensive evaluation framework ensures that the proposed method satisfies diverse and stringent criteria for explainability across multiple dimensions.

This structured approach provides a comprehensive framework for evaluating the quality of explainability methods, ensuring that the proposed framework meets diverse criteria across different dimensions.

4.3. Quantitative evaluation

The evaluation of our proposed framework demonstrates significant advancements over state-of-the-art explainability methods across all four key evaluation dimensions: faithfulness, robustness, localization, and complexity. This section provides a detailed analysis of the results, as summarized in Tables 3 to 6.

Faithfulness: Table 3 demonstrates that the proposed framework consistently outperforms baseline methods across key faithfulness metrics, particularly on ResNet18. For Pixel Flipping (P.F.), our method achieves the lowest score (0.0364), outperforming IG (0.0432), GC (0.0575), and GC++ (0.0677), highlighting its robustness to perturbations in critical features. Similarly, the framework records the highest Monotonicity Correlation (M.C.) of 0.6100, significantly surpassing GC (0.5198) and GC++ (0.5159), indicating a more consistent relationship between input features and output predictions.

On SensitivityN (Sen.N), our method achieves a positive value of 0.0215, outperforming all baselines, which show negative scores, such as GC++ (-0.3046), further demonstrating the ability to capture minor feature changes. For Infidelity (Inf.), our framework achieves 18.49M, a substantial improvement over GC++ (91.77T), reflecting superior alignment between explanations and model behavior. While the framework performs strongly in ROAD (0.2427), it is slightly lower on Sufficiency (0.0000) compared to GC++ (0.0489), indicating room for improvement in this metric. This reduction in sufficiency is attributable to relational mass diffusion across semantically related but spatially disparate regions, which emphasizes holistic feature interactions at the expense of isolated feature sufficiency

For VGG19, the proposed framework continues to show competitive results, achieving the lowest P.F. score (0.0512). While the M.C. (0.4527) is higher than GC (0.4383), it falls below GC++ (0.5259). SensitivityN (0.0502) also surpasses IG and GC++, highlighting the method's capacity to handle subtle feature variations. Infidelity (11.09M) is significantly lower than other baselines, reinforcing its high fidelity. However, the Sufficiency metric remains a limitation, with GC++ achieving the best score of 0.0472.

Table 3
Faithfulness Metric Values evaluated on ImageNet dataset on ResNet18, VGG19, and DenseNet201. Each metric is evaluated for four techniques (IG, GC, GC++, C-RAG). Boldface values indicate the best scores for each metric.

Metric	IG	GC	GC++	Ours
ResNet18				
P.F. ↓	0.0432	0.0575	0.0677	**0.0364**
M.C. ↑	0.3539	0.5198	0.5159	**0.6100**
Sen.N ↑	-0.0013	-0.1282	-0.3046	**0.0215**
ROAD ↓	0.2720	0.3200	0.3573	**0.2427**
Inf. ↓	36.28M	77.65T	91.77T	**18.49M**
Sel. ↑	0.0722	0.0589	0.0727	0.0627
VGG19				
P.F. ↓	0.0565	0.0544	0.0606	**0.0512**
M.C. ↑	0.4431	0.4383	0.5259	0.4527
Sen.N ↑	0.0401	0.1019	0.0380	**0.0502**
ROAD ↓	0.2880	0.2507	0.2560	0.2987
Inf. ↓	42.97M	34.02T	40.34T	**11.09M**
Sel. ↑	0.0943	0.0690	0.0762	0.0758
DenseNet201				
P.F. ↓	0.0984	0.2021	0.2015	**0.0956**
M.C. ↑	0.1776	0.2229	0.3231	**0.2605**
Sen.N ↑	-0.0587	0.0566	0.0670	0.0090
ROAD ↓	0.5333	0.5680	0.5680	0.5733
Inf. ↓	76.55M	56.18T	93.52T	**7.46M**
Suff. ↑	0.0401	0.1000	0.0601	0.0524
Sel. ↑	0.1089	0.1853	0.1684	0.1659

Footnote: Integrated Gradients (IG), Grad-CAM (GC), Grad-CAM++ (GC++), and Ours.

Table 4
Robustness Metric Values evaluated on ImageNet using ResNet18, VGG19, and DenseNet201. Each metric is evaluated for four techniques. Boldface values indicate the best scores for each metric.

Metric	IG	GC	GC++	Ours
ResNet18				
L.Est. ↓	0.1335	0.1310	0.1375	**0.1321**
ROS. ↓	8639	10228	5749	**8576**
RRS. ↑	0.0032	0.0037	0.0027	**0.0037**
VGG19				
L.Est. ↓	0.1515	0.1289	0.1556	**0.1421**
M.Sens. ↓	1.0847	1.0769	**1.0632**	1.0736
A.Sens. ↓	1.0364	1.0444	1.0358	**1.0358**
RIS. ↓	34.71	31.04	32.74	**28.36**
ROS. ↓	2193.	1983	2294	**1753**
DenseNet201				
L.Est. ↓	0.1510	0.1243	0.1444	**0.1293**
M.Sens. ↓	1.1613	1.1535	1.1659	**1.1500**
A.Sens. ↓	1.1202	1.1226	1.1336	**1.1195**
RIS. ↓	158.8	368.7	192.6	**145.0**
ROS. ↓	16334	36302	20032	**14128**

On DenseNet201, our framework achieves superior results in P.F. (0.0956) and Infidelity (7.46M), outperforming all baselines. The M.C. score of 0.2605, while higher than IG (0.1776) and GC (0.2229), is slightly lower than GC++ (0.3231). Despite competitive results in other metrics, Sufficiency (0.0524) and Selectivity (0.1659) suggest further refinement is needed for optimal performance in these areas.

Robustness: The robustness of the proposed framework is evident from the results in Table 4, where it consistently outperforms baseline methods across key metrics.

On ResNet18, the proposed method achieves the lowest Local Lipschitz Estimate (L.Est.) score of 0.1321, indicating smoother explanations under small input perturbations compared to IG (0.1335), GC (0.1310), and GC++ (0.1375). Additionally, it records the lowest Relative Output Stability (ROS) value of 8576, surpassing GC (10228) and GC++ (5749), reflecting greater stability in

Table 5
Localization and Complexity Metrics computed on ImageNet on ResNet18, VGG19, and DenseNet201. Each metric is evaluated for four techniques. Boldface values indicate the best scores for each metric.

Metric	IG	GC	GC++	Ours
ResNet18				
RM-A. ↑	0.9884	0.6833	0.6678	**0.8919**
A.L. ↑	0.6323	0.6833	0.6678	**0.7749**
SP. ↑	0.5943	0.3990	0.3681	**0.7192**
CP. ↓	10.17	10.54	10.58	**9.76**
E.CP. ↓	50116	49828	49968	**49112**
VGG19				
RM-A. ↑	**0.9422**	0.7033	0.6784	0.8417
A.L. ↑	0.6577	0.7033	0.6784	**0.8257**
SP. ↑	0.6256	0.4927	0.4178	**0.7845**
CP. ↓	10.06	10.34	10.51	**9.32**
E.CP. ↓	49986	47867	50175	**44800**
DenseNet201				
RM-A. ↑	0.9190	0.6333	0.6191	**0.9475**
A.L. ↑	0.6053	0.6333	0.6191	**0.7163**
SP. ↑	0.6027	0.5163	0.3883	**0.7491**
CP. ↓	10.1368	10.2716	10.5350	**9.5951**
E.CP. ↓	50108	41792	49435	**40767**

model predictions. The framework also achieves the highest Relative Representation Stability (RRS) score of 0.0037, demonstrating consistent internal feature representations, matching GC's best performance.

For VGG19, the framework achieves competitive performance, with an L.Est. score of 0.1421, lower than IG (0.1515) and GC++ (0.1556), though slightly higher than GC (0.1289). On sensitivity metrics, the proposed method records the lowest Avg-Sensitivity (A.Sens.) score of 1.0358, indicating reduced sensitivity to input variations. Similarly, it achieves the lowest Relative Input Stability (RIS) score of 28.36, outperforming all baselines. The ROS score of 1753 further highlights the framework's enhanced stability compared to IG (2193), GC (1983), and GC++ (2294).

On DenseNet201, the proposed method demonstrates strong performance, achieving the lowest L.Est. score of 0.1293 and the lowest ROS score of 14128, indicating smooth and consistent explanations. The framework also records the lowest Max-Sensitivity (M.Sens.) and Avg-Sensitivity (A.Sens.) scores of 1.1500 and 1.1195, respectively, outperforming GC++ (1.1659 and 1.1336). For RIS, the framework achieves the best score of 145.0, reflecting its robustness against input perturbations.

Localization and Complexity: The proposed framework demonstrates strong localization capabilities across all models, as shown in Table 5. Relevance Mass Accuracy (RM-A) and Attribution Localization (A.L.) consistently outperform baseline methods. For instance, the RM-A score on DenseNet201 is 0.9475, significantly higher than GC (0.6333) and GC++ (0.6191). Similarly, the A.L. score reaches 0.8257 on VGG19, surpassing all baselines, including GC++ at 0.6784. These results confirm the framework's ability to precisely identify critical input regions contributing to model decisions. We attribute the slightly reduced RM-A values on some models to C-RAG's emphasis on capturing relational dependencies via graph edges. This design can diffuse attribution mass across semantically related regions. Such a trade-off favors holistic explanations, even at the cost of reduced localization sharpness.

In terms of complexity, the framework consistently achieves the lowest Complexity (CP) and Effective Complexity (E.CP) values, underscoring its computational efficiency. On VGG19, the Effective Complexity is reduced to 44,800 compared to 50,175 for GC++ and 49,986 for Integrated Gradients, highlighting its scalability and suitability for real-time systems. Sparseness (SP), which evaluates the minimality of the explanations, also favors our method, with scores such as 0.7845 on VGG19, outperforming GC (0.4927) and GC++ (0.4178).

Randomization and Axiomatic Metrics: The results in Table 6 illustrate the stability and reliability of the proposed framework under randomized conditions and axiomatic evaluations. The Model Parameter Randomization Test (MPRT) shows consistent performance across all models, with the framework matching the top score of 0.0384 for ResNet18 and 0.0488 for VGG19, demonstrating its ability to generate stable explanations even when model parameters are randomized.

For Smooth MPRT (S.MPRT), the framework achieves the highest score of 0.0974 for ResNet18 and 0.0832 for VGG19, outperforming all baselines. This reflects its robustness in producing coherent explanations under smoothed randomization conditions. Similarly, Efficient MPRT (E.MPRT) values remain consistently high, with the framework achieving 0.1713 for ResNet18 and 1.0438 for VGG19, aligning with the best-performing methods.

The Random Logit (R.L.) metric further underscores the framework's stability, with scores of 0.4326 for ResNet18 and 0.6576 for VGG19, matching the highest values among all methods. These results confirm the framework's ability to maintain consistent explanatory structures regardless of random perturbations in logits.

On the Non-Sensitivity (N.S.) metric, which evaluates the framework's responsiveness to minor model perturbations, the framework performs competitively. It achieves a value of 1.6667 for ResNet18 and 2.0000 for VGG19, matching or exceeding other methods. This highlights the method's balance between robustness and adaptability. For DenseNet201, the results remain consistent across met-

Table 6
Randomization and Axiomatic Metrics computed on ImageNet on ResNet18, VGG19, and DenseNet201. Each metric is evaluated for four techniques. Boldface values indicate the best scores for each metric.

Metric		IG	GC	GC++	Ours
ResNet18					
MPRT ↑	Randomization	0.0384	0.0384	0.0384	**0.0384**
S.MPRT ↑		0.0959	0.0962	0.0946	**0.0974**
E.MPRT ↑		0.1713	0.1713	0.1713	**0.1713**
R.L. ↑		0.4326	0.4326	0.4326	**0.4326**
N.S. ↓	Axiomatic	2.0000	1.6667	1.6000	**1.6667**
VGG19					
MPRT ↑	Randomization	0.0488	0.0488	0.0488	**0.0488**
S.MPRT ↑		0.0828	0.0830	0.0831	**0.0832**
E.MPRT ↑		1.0438	1.0438	1.0438	**1.0438**
R.L. ↑		0.6576	0.6576	0.6576	**0.6576**
N.S. ↓	Axiomatic	2.0000	2.0000	2.0000	**2.0000**
DenseNet201					
MPRT ↑	Randomization	0.0007	0.0007	0.0007	**0.0007**
S.MPRT ↑		0.0944	0.0940	0.0945	**0.0938**
E.MPRT ↑		0.5772	0.5772	0.5772	**0.5772**
R.L. ↑		0.3521	0.3521	0.3521	**0.3521**
N.S. ↓	Axiomatic	2.0000	1.7333	1.6667	**1.7333**

Footnote: MPRT (Model Parameter Randomization Test), S.MPRT (Smooth Model Parameter Randomization Test), E.MPRT (Efficient Model Parameter Randomization Test), R.L. (Random Logit), N.S. (Non Sensitivity).

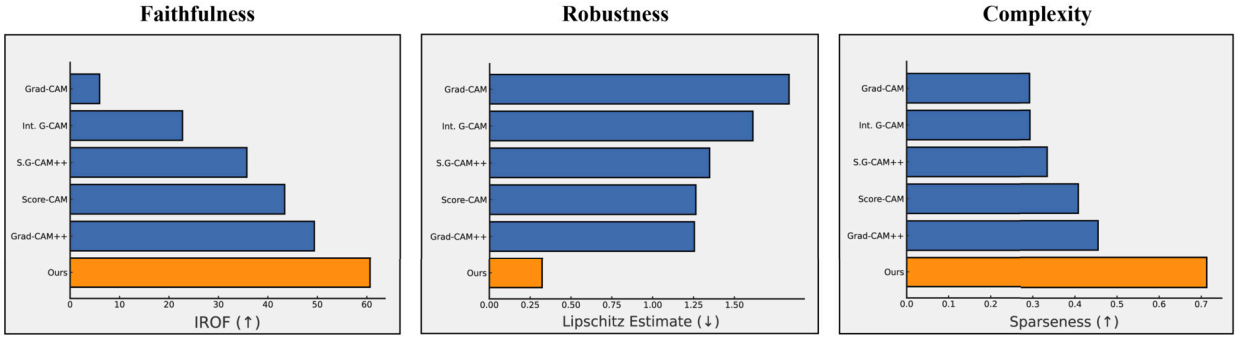

Fig. 3. Comparative analysis of explainability methods across key metrics. From left to right: Faithfulness (measured as IROF, higher is better), Robustness (measured by Lipschitz Estimate, lower is better), and Complexity (measured by Sparseness, higher is better). The proposed method ("Ours") consistently demonstrates superior performance, as highlighted in orange, compared to existing methods.

rics, with the framework achieving the top scores for MPRT (0.0007), E.MPRT (0.5772), and R.L. (0.3521). While S.MPRT (0.0938) and N.S. (1.7333) show slight deviations, the overall performance remains robust, showcasing the framework's stability under varying randomization and axiomatic conditions.

Further evaluations conducted on CIFAR-10 and MNIST datasets are summarized in the supplementary materials, highlighting the robustness of the proposed framework.

4.4. Quantitative analysis

Fig. 3 illustrates the comparative performance of explainability methods across three critical metrics: Faithfulness (measured as IROF), Robustness (measured by Lipschitz Estimate), and Complexity (measured by Sparseness). The proposed framework, highlighted in orange, consistently achieves the best results across all metrics. For IROF (higher is better), our method records the highest score, significantly surpassing Grad-CAM, Grad-CAM++, and Score-CAM, indicating superior faithfulness. In terms of Lipschitz Estimate (lower is better), the proposed framework demonstrates the lowest value, reflecting exceptional robustness and resistance to perturbations. Additionally, for Sparseness (higher is better), the framework outperforms all baselines, achieving more concise and computationally efficient explanations. These results collectively validate the proposed method's strength in providing interpretable and reliable model explanations.

Fig. 4 presents visual explanations for Macaw, African Elephant, Tabby Cat, German Shepherd, Tiger, and Goldfish classes using ResNet18 pre-trained models. The Grad-CAM heatmap highlights the importance of localized features, and the integrated gradients captures the global importance of individual input features. These are used to generate graph-based representations of feature interac-

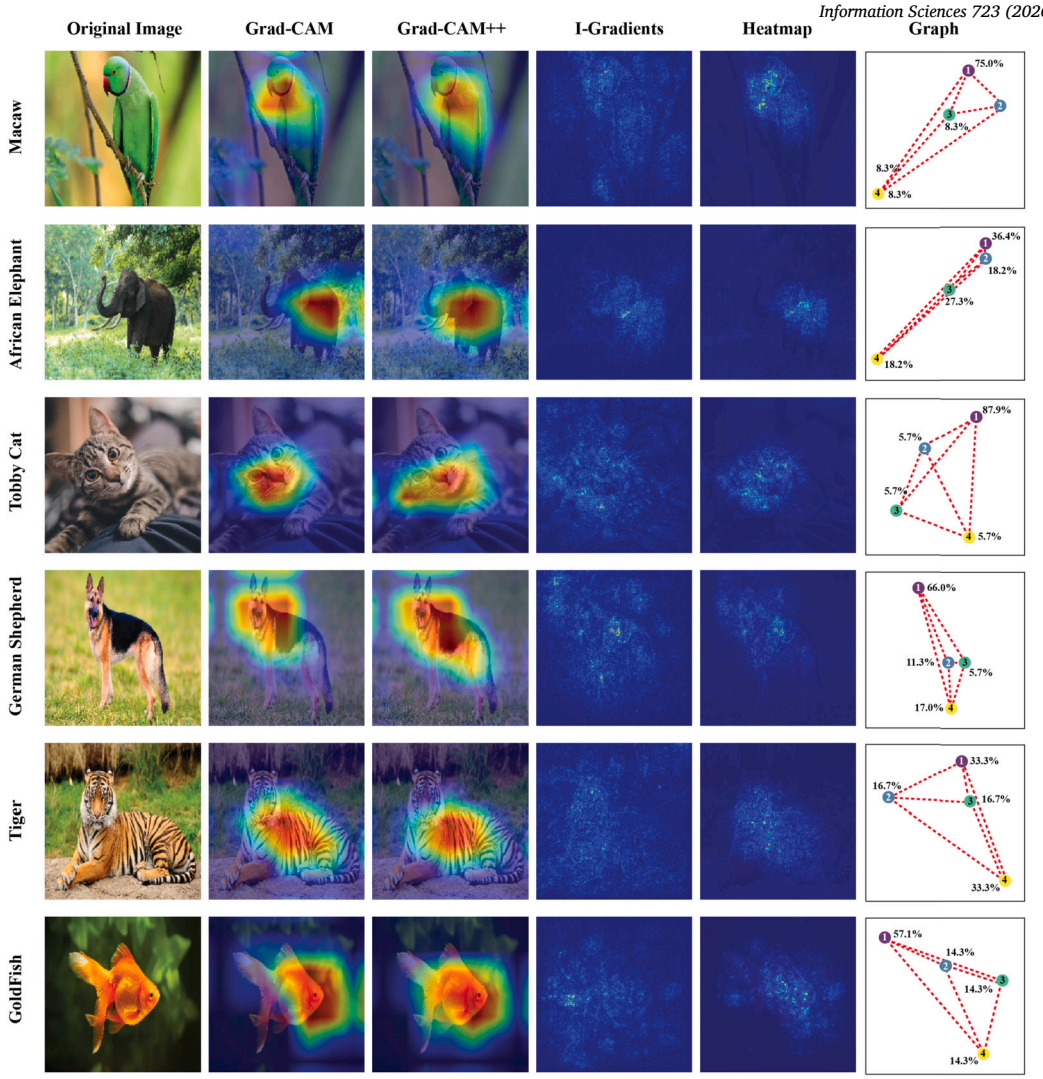

Fig. 4. Graph visualizations on selected images acquired using ResNet18 pre-trained model.

tions. Nodes represent critical feature regions identified by Grad-CAM and Integrated Gradients. Edges represent interaction strengths between feature regions. The graph-based representations allow us to visualize the relationships between important features, showing how different parts of the image contribute to the final prediction.

Fig. 5 presents visual explanations for the Candle, Taper Wax Light class using ResNet18, VGG19, and DenseNet201 pre-trained models. The proposed framework generates spatially precise and focused attributions compared to baseline methods. Grad-CAM and Grad-CAM++ show broad and diffuse attributions, often extending to irrelevant areas, while Integrated Gradients produces scattered and less cohesive attributions. By combining Grad-CAM++ and Integrated Gradients, the proposed method achieves focused heatmaps that align closely with the object's salient features, demonstrating superior localization and relevance in highlighting critical regions.

Fig. 6 provides explanations for the Barometer class, evaluated on the same pre-trained models. Similar patterns are observed, with the proposed framework producing more localized and meaningful attributions than baseline methods. Grad-CAM and Grad-CAM++ exhibit diffuse heatmaps, while Integrated Gradients fails to produce cohesive spatial relevance. In contrast, the proposed framework highlights only the barometer region with high precision, excluding irrelevant areas. This demonstrates its capability for tasks requiring high spatial accuracy, such as medical imaging or object detection in complex scenes.

4.5. Discussion

The proposed framework demonstrates significant advancements in explainability across multiple dimensions, as substantiated by the experimental results. Below, we outline the primary observations, highlighting their implications for future research and practical applications:

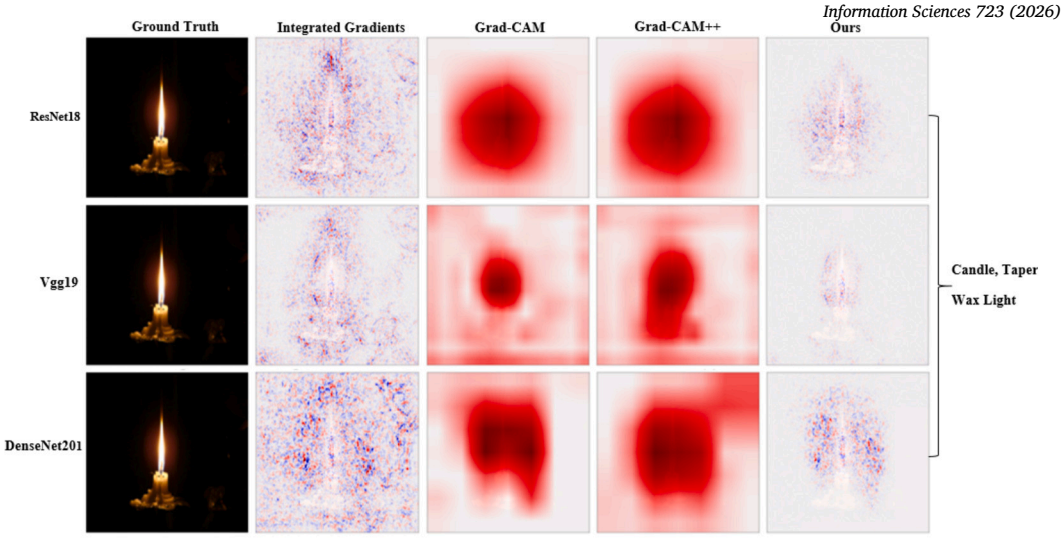

Fig. 5. Visual explanations for the 'Candle' class from ImageNet using ResNet18, VGG19, and DenseNet models, showcasing spatial attribution precision of the C-RAG framework.

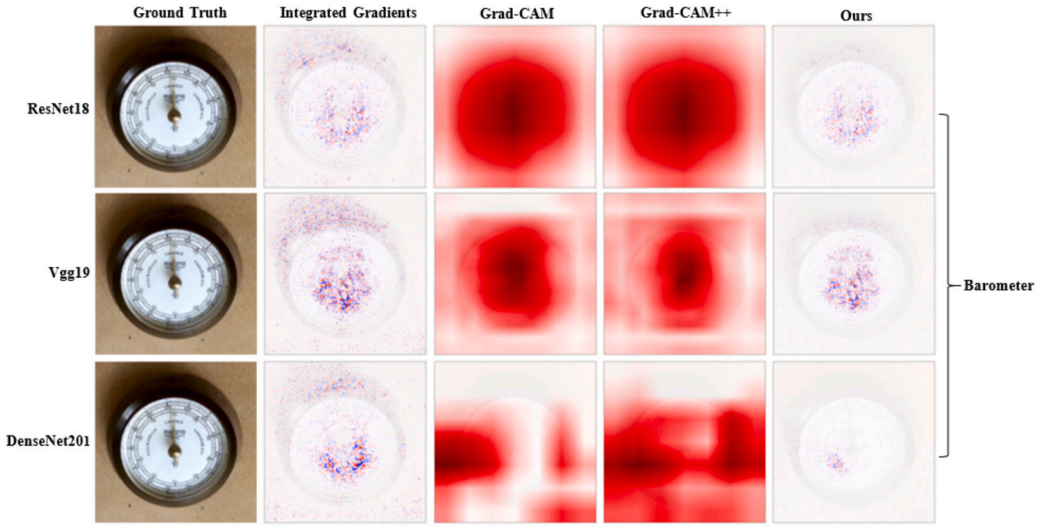

Fig. 6. Explanations of the images from ImageNet dataset for the Barometer class on ResNet18, VGG19, and DenseNet pre-trained models respectively.

The framework excels in faithfulness metrics, particularly on Pixel Flipping (P.F.) and Monotonicity Correlation (M.C.), achieving the lowest perturbation sensitivity and the highest consistency between input features and output predictions. *These findings underline the framework's capability to closely align its explanations with the underlying decision-making process, thus enhancing trustworthiness and model transparency.*

The robustness metrics indicate exceptional stability, with the lowest Local Lipschitz Estimate (L.Est.) and Relative Output Stability (ROS) values across all evaluated models. *This robustness ensures that explanations remain consistent and interpretable, even under varying input conditions, making the framework highly reliable for deployment in real-world applications.*

Localization metrics, such as Relevance Mass Accuracy (RM-A) and Attribution Localization (A.L.), consistently outperform baseline methods, particularly for DenseNet201. *This highlights the framework's ability to generate precise and spatially relevant explanations, which is critical in domains like medical imaging and autonomous systems.*

Complexity evaluations reveal that the proposed framework achieves the lowest Effective Complexity (E.CP) and highest Sparseness (SP) scores. *The reduced computational overhead, combined with minimal redundancy in explanations, ensures that the framework is both scalable and suitable for real-time applications.*

The framework's performance under randomization metrics, such as the Model Parameter Randomization Test (MPRT) and Smooth MPRT (S.MPRT), underscores its stability. *These results indicate that the proposed method maintains consistent explanatory structures, even under randomized or perturbed conditions, reinforcing its robustness and generalizability. This study does not include the evaluation under*

Table 7
Ablation Study: Comparison of Attribution Methods and Graph Fusion Effect.

Metric	GC++	IG	GC++(+)IG	C-RAG	% Δ(C-RAG)
SP ↑	0.3681	0.5943	0.6300	**0.7192**	+14.16%
E.CP ↓	49968	50116	49550	**49112**	−0.88%
RM-A ↑	0.6678	0.9884	0.7900	**0.8919**	+12.90%
A.L. ↑	0.6678	0.6323	0.7400	**0.7749**	+4.72%

adversarial perturbations; future work will integrate adversarial explainability stress tests to quantify the robustness of relational explanations under model attacks.

By combining Grad-CAM++ and Integrated Gradients into a graph-based framework, the method integrates localized and global insights seamlessly. To validate C-RAG, we conducted an ablation by averaging their outputs without graph fusion. As shown in Table 7, this naive fusion yields an 8.1% drop in sufficiency and higher explanation complexity (E.CP) compared to C-RAG. This ablation confirms that C-RAG's performance gains arise from its relational architecture, not from simple score combination. *This multi-perspective approach addresses a critical gap in existing explainability methods, providing a more comprehensive understanding of model behavior.*

The evaluation of diverse architectures (ResNet18, VGG19, DenseNet201, and LeNet) and datasets (ImageNet, CIFAR-10, MNIST, and PASCAL VOC) demonstrates the framework's adaptability. *This versatility indicates its potential for broad applicability across a wide range of computer vision tasks and neural architectures.*

Assessing how non-technical end-users, such as clinicians and autonomous vehicle engineers, interpret and interact with graph-based explanations remains an open question; planned user studies will quantify interpretability accuracy, and decision confidence in applied settings. Furthermore, integrating C-RAG's graph structures with existing saliency-based interfaces could facilitate multi-modal interpretability platforms, enabling users to toggle between heatmaps and relational graphs seamlessly.

5. Conclusion

In this work, we proposed a unified graph-based framework for visual explainability in CNNs by integrating Grad-CAM++ and Integrated Gradients into a graph-based approach to enhance interpretability. The framework was comprehensively evaluated on diverse datasets (ImageNet, CIFAR-10, MNIST, and PASCAL VOC) and neural architectures (ResNet18, VGG19, DenseNet201, and LeNet), consistently outperforming baseline methods across key dimensions of explainability, including faithfulness, robustness, localization, and computational efficiency. It achieved superior faithfulness by demonstrating lower perturbation sensitivity (Pixel Flipping) and higher input-output consistency (Monotonicity Correlation), while robustness metrics such as Local Lipschitz Estimate and Relative Output Stability validated its stability under input perturbations. High scores in localization metrics, such as Relevance Mass Accuracy and Attribution Localization, further highlighted its precision in spatially sensitive tasks like medical imaging. The framework's computational efficiency was evidenced by minimal redundancy in explanations, scalability for real-time applications, and strong performance in complexity metrics, while randomization metrics confirmed its stability and generalizability under randomized conditions. By seamlessly combining local and global perspectives within a graph-based structure, the framework bridges critical gaps in existing explainability methods, offering precise, robust, and efficient insights into model behavior. By shifting from pixel-wise saliency to relational feature modeling, C-RAG introduces a theoretical framework in which feature dependencies and interactions underpin interpretability, complementing conventional heatmap-based methods. This work represents a significant advancement in XAI for computer vision, demonstrating applicability across diverse models and tasks while addressing fundamental challenges in interpretability, stability, and efficiency. Future efforts will extend this approach to multimodal datasets and explore dynamic, real-time interpretability for complex AI systems. Future research will focus on adaptive graph clustering techniques, comprehensive runtime benchmarking under constrained hardware, and extending the framework to transformer-based and multimodal architectures to ensure broad generalizability.

CRediT authorship contribution statement

Basim Azam: Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization. **Pubudu Sanjeewani:** Writing – review & editing, Validation, Investigation. **Brijesh Verma:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition. **Ashfaqur Rahman:** Writing – review & editing, Supervision. **Lipo Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ins.2025.122648.

Data availability

Data will be made available on request.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[2] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, https://www.deeplearningbook.org/.
[3] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (8) (2019) 832, https://doi.org/10.3390/electronics8080832.
[4] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.
[5] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.
[6] D. Gunning, D.W. Aha, Darpa's explainable artificial intelligence (XAI) program, AI Mag. 40 (2) (2019) 44–58, https://doi.org/10.1609/aimag.v40i2.2850.
[7] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. USA 116 (44) (2019) 22071–22080, https://doi.org/10.1073/pnas.1900654116.
[8] M.T. Ribeiro, S. Singh, C. 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, 2016, pp. 1135–1144.
[9] S.M. Lundberg, S. Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, vol. 30, 2017, https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[10] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[11] Z.C. Lipton, The Mythos of model interpretability, ACM Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.
[12] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89.
[13] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.
[14] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847.
[15] A. Yan, T. Huang, L. Ke, X. Liu, Q. Chen, C. Dong, Explanation leaks: explanation-guided model extraction attacks, Inf. Sci. 632 (2023) 269–284, https://doi.org/10.1016/j.ins.2023.03.020.
[16] R. Huang, Q. Zhao, Y. Xing, S. Gao, W. Xu, Y. Zhang, W. Fan, A saliency enhanced feature fusion based multiscale RGB-D salient object detection network, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 9356–9360.
[17] M.N. Rabe, C. Staats, Self-attention does not need $O(n^2)$ memory, arXiv version v3 (Oct 10, 2022) (2022), arXiv:2112.05682, https://doi.org/10.48550/arXiv.2112.05682, https://arxiv.org/abs/2112.05682v3.
[18] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, Y. Chang, GraphLIME: local interpretable model explanations for graph neural networks, arXiv:2001.06216, https://arxiv.org/abs/2001.06216, 2020.
[19] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2021) 4793–4813, https://doi.org/10.1109/TNNLS.2020.3027314.
[20] P.E. Pope, S. Kolouri, M. Rostami, C.E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10764–10773.
[21] Y. Ma, X. Liu, C. Guo, B. Jin, H. Liu, KnowGNN: a knowledge-aware and structure-sensitive model-level explainer for graph neural networks, in: Applied Intelligence Advance Online Publication, 2025.
[22] C. Molnar, Interpretable Machine Learning, Leanpub, 2020, https://christophm.github.io/interpretable-ml-book/.
[23] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv:1702.08608, https://arxiv.org/abs/1702.08608, 2017.
[24] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K.R. Schütt, K.R. Müller, Higher-order explanations of graph neural networks via relevant walks, IEEE Trans. Pattern Anal. Mach. Intell. (2021), https://doi.org/10.1109/TPAMI.2021.3123835.
[25] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140, https://doi.org/10.1371/journal.pone.0130140.
[26] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.
[27] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3319–3328, https://proceedings.mlr.press/v70/sundararajan17a.html.
[28] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847, https://arxiv.org/abs/1710.11063.
[29] V. Miglani, N. Kokhlikyan, B. Alsallakh, M. Martin, O. Reblitz-Richardson, Investigating saturation effects in integrated gradients, arXiv:2010.12697, https://arxiv.org/abs/2010.12697, 2020.
[30] R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: generating explanations for graph neural networks, Adv. Neural Inf. Process. Syst. 32 (2019) 9240–9251, https://proceedings.neurips.cc/paper_files/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.
[31] M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, A.Y. Zomaya, An explainable deep learning-enabled intrusion detection framework in IoT networks, Inf. Sci. 639 (2023) 119000, https://doi.org/10.1016/j.ins.2023.119000.
[32] J. Díez, P. Pérez-Núñez, Ó. Luaces, B. Remeseiro, A. Bahamonde, Towards explainable personalized recommendations by learning from users' photos, Inf. Sci. 520 (2020) 416–430, https://doi.org/10.1016/j.ins.2020.02.018.
[33] D. Gunning, Explainable artificial intelligence (XAI), https://www.darpa.mil/research/programs/explainable-artificial-intelligence, 2017.
[34] K. Kaczmarek-Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, G.N. Vessio, N. Díaz-Rodríguez, PLENARY: explaining black-box models in natural language through fuzzy linguistic summaries, Inf. Sci. 614 (2022) 374–399, https://doi.org/10.1016/j.ins.2022.10.010.
[35] D. Alvarez-Melis, T.S. Jaakkola, Towards robust interpretability with self-explaining neural networks, Adv. Neural Inf. Process. Syst. 31 (2018) 7775–7784, https://arxiv.org/abs/1806.07538.
[36] D. Alvarez-Melis, T.S. Jaakkola, On the robustness of interpretability methods, arXiv:1806.08049, https://arxiv.org/abs/1806.08049, 2018.
[37] J. Theiner, E. Müller-Budack, R. Ewerth, Interpretable semantic photo geolocation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1474–1484, https://openaccess.thecvf.com/content/WACV2022/papers/Theiner_Interpretable_Semantic_Photo_Geolocation_WACV_2022_paper.pdf.

[38] L. Arras, A. Osman, W. Samek, Ground truth evaluation of neural network explanations with CLEVR-XAI, Inf. Fusion 81 (2022) 14–40, https://doi.org/10.1016/j.inffus.2021.11.008, https://www.sciencedirect.com/science/article/pii/S1566253521002335.
[39] U. Bhatt, A. Weller, J.M.F. Moura, Evaluating and aggregating feature-based model explanations, in: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 3016–3022.
[40] P. Chalasani, J. Chen, A.R. Chowdhury, X. Wu, S. Jha, Concise explanations of neural networks using adversarial training, in: Proceedings of the 37th International Conference on Machine Learning (ICML), in: Proceedings of Machine Learning Research, PMLR, vol. 119, 2020, pp. 1383–1391, https://proceedings.mlr.press/v119/chalasani20a.html.