FocusViT: Faithful Explanations for Vision Transformers via Gradient-Guided Layer-Skipping

Mohsin Ali[1]*, Haider Raza[1], John Q. Gan[1], Muhammad Haris[2]
[1] School of Computer Science and Electronics Engineering, University of Essex
[2]Mohamed Bin Zayed University of Artificial Intelligence
{ma22159, h.raza, jqgan}@essex.ac.uk, muhammad.haris@mbzuai.ac.ae

Abstract

Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs for various vision tasks, yet their token-based, attention-driven architecture makes interpreting their predictions challenging. Existing explainability methods, such as Grad-CAM and Attention Rollout, either fail to capture hierarchical semantic information or assume attention directly reflects importance, often leading to misleading explanations. We propose FocusViT, a novel explainability framework that integrates gradient-weighted attention attribution with dynamic, faithfulness-driven layer aggregation. By fusing attention maps with class-specific gradients and introducing per-head dynamic weighting, FocusViT highlights not only where the model attends but also how sensitive the prediction is to those attentions. Furthermore, our adaptive layer-skipping strategy ensures that only semantically meaningful layers contribute to the final explanation, enhancing both faithfulness and clarity. Extensive quantitative and qualitative evaluations on diverse benchmarks demonstrate that FocusViT improves over existing methods in faithfulness and sparsity, achieving competitive robustness and class sensitivity, and provides sharper, more reliable visual explanations for ViTs. The official implementation is publicly available at: https://github.com/game-sys/focusvit-aistats2026.git

1 Introduction

Vision Transformers (ViTs) Dosovitskiy et al. (2020); Vaswani et al. (2017) have recently emerged as a powerful alternative to convolutional neural networks (CNNs) for image classification and various computer vision tasks Carion et al. (2020); Zhang et al. (2022); Dosovitskiy et al. (2020). Unlike CNNs, which rely on convolutional filters to extract local features, ViTs divide input images into patches and process them using self-attention mechanisms, particularly multi-head self-attention (MHSA). This architecture enables ViTs to capture global contextual information, making them especially effective on large-scale datasets. However, the token-based and attention-driven nature of ViTs introduces new challenges for model interpretability, as their decision-making processes are less transparent compared to the spatially localised feature maps of CNNs. While numerous explainability techniques, such as Grad-CAM, have been developed for CNNs, explainability for ViTs remains an underexplored area.

Grad-CAM Selvaraju et al. (2016) generates explanations by computing the gradients of the predicted class logits with respect to the feature maps in the final convolutional layer and averaging these gradients spatially to highlight important regions. Although effective for CNNs, Grad-CAM's limitation lies in focusing solely on the last convolutional layer, thereby overlooking contributions from earlier convolutional and fully connected layers. Adaptations of Grad-CAM for ViTs Selvaraju et al. (2017); Draelos and Carin (2020) similarly emphasise only the last attention layer, missing the rich hierarchical information distributed across multiple transformer layers. Attention Rollout Abnar and Zuidema (2020) offers an alternative by leveraging the inherent attention mechanisms of ViTs. It aggregates attention weights from all transformer layers to estimate the influence of each input token on the final classification token. However, this method assumes that attention weights directly correspond to

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

feature importance, an assumption increasingly questioned by recent studies such as Layer-Wise Relevance Propagation (LRP) Montavon et al. (2019); Chefer et al. (2021), which provides layer-wise explanations by propagating relevance scores backwards from the output to the input. Despite its mathematical rigour, LRP depends on handcrafted propagation rules, which often struggle with nonlinear activations and complex architectural components.

Existing ViT explanation methods fall into four broad categories: propagation-based techniques such as LRP and AttnLRP Chefer et al. (2021); Achtibat et al. (2024), token-tracing approaches like TokenTM Wu et al. (2024) and AttCat Qiang et al. (2022), sensitivity-driven methods such as LeGrad Bousselham et al. (2024), and diagnostic studies revealing biases like ViT register tokens Darcet et al. (2023). While each provides valuable insight, they are limited by handcrafted propagation rules, over-reliance on the class token, heavy perturbation cost, or susceptibility to spurious tokens. To address these gaps, we propose **FocusViT**, a unified framework that fuses gradient sensitivity with attention attribution and dynamically aggregates layers based on faithfulness metrics.

The main contributions of this work are:

- **Gradient-weighted head attribution:** We weight attention heads by their gradient sensitivity to class logits, capturing not only where the model attends but also how those attentions influence predictions.
- **Faithfulness-driven skip-layer selection:** We introduce a dynamic criterion that selects the most semantically aligned layers for aggregation, improving on heuristic midpoint strategies such as SkipPLUS Mehri et al. (2024).
- **Additive aggregation across layers:** Unlike multiplicative Rollout, which suffers from vanishing attributions, our additive scheme preserves signal strength and produces clearer explanations.
- **Plug-and-play applicability:** FocusViT requires no retraining, avoids handcrafted propagation rules (AttnLRP), and mitigates register-token bias, making it lightweight and adaptable across datasets and tasks.

2 Related Work

Classical XAI Methods. Early model-agnostic approaches include LIME (Local Interpretable Model-agnostic Explanations) Ribeiro et al. (2016), which fits a simple surrogate model around perturbed samples to approximate local decision boundaries. It generates superpixel-level importance maps but is unstable due to its dependence on segmentation and independence assumptions. SHAP (SHapley Additive Explanations) Lundberg and Lee (2017) instead attributes importance using Shapley values from cooperative game theory, providing solid theoretical guarantees but at high computational cost, which limits its scalability in vision tasks.

Grad-CAM (Gradient-weighted Class Activation Mapping) Selvaraju et al. (2017) adapts gradient backpropagation to CNNs by linking class scores to convolutional feature maps, producing class-discriminative heatmaps. While widely adopted in CNN-based vision, its reliance on convolutional structures prevents direct extension to ViTs.

Transformer-Specific Explanations. Attention Rollout Abnar and Zuidema (2020) proposed aggregating token contributions by recursively multiplying attention matrices across layers. This provides a global view of how input tokens influence the class token, but it assumes that the attention weights directly represent importance: an assumption that is increasingly challenged in subsequent work.

Layer-Wise Relevance Propagation (LRP) was adapted to Transformers by Chefer et al. Chefer et al. (2021), who introduced gradient-based propagation rules for self-attention. This enabled token-level explanations, but their reliance on handcrafted rules makes generalisation to diverse architectures and residual connections difficult. AttnLRP Achtibat et al. (2024) refined this approach by redistributing relevance proportionally to actual attention weights. By explicitly incorporating attention into the propagation process, it aligned explanations more closely with the model's internal dynamics, though it still required full access to all intermediate activations.

Wu et al. Wu et al. (2024) highlighted that token transformations such as projections, residual connections, and MLP layers play a central role in shaping model behaviour, and ignoring them distorts explanations. Their method, TokenTM, explicitly traces these transformations to preserve faithfulness, producing more reliable explanations at the cost of added computation.

AttCat Qiang et al. (2022) leveraged the class token itself to generate explanations. By tracing attention flows from the class token back to input patches, it produced class activation maps without the need for gradients. The method is efficient but depends strongly on the class token being the dominant driver of predictions, which may not hold in all ViT variants or tasks.

LeGrad Bousselham et al. (2024) proposed explanation via feature-formation sensitivity, measuring how perturbations to the input affect the construction of hidden features. This approach mitigates biases from spurious attention spikes and provides stable, less noisy maps, though it requires repeated perturbation analysis. Complementarily, Darcet et al. Darcet et al. (2023) revealed the presence of "register tokens" in ViTs, which act as global context storage and often dominate attribution scores. This diagnostic finding highlighted why many gradient-based methods overemphasise these tokens, producing misleading saliency.

Taken together, these methods illustrate both the richness and the challenges of explaining ViTs. Propagation-based approaches such as LRP and AttnLRP rely on carefully designed rules, while token-tracing methods, e.g. TokenTM and AttCat, focus on structural components such as transformations or class tokens. LeGrad and the ViT Register analysis further show that stability and token bias remain open challenges. In contrast, our proposed FocusViT addresses these gaps by weighting attention heads according to gradient sensitivity and adaptively skipping noisy early layers. This design avoids handcrafted propagation rules, mitigates register-token bias, and yields sharper explanations by combining gradient information with attention in a unified, faithfulness-driven framework.

Evaluation Metrics: The evaluation metrics used to assess the quality of model explainability are summarised in Table 1. These metrics include:

Metric	Direction	Description
Faithfulness Correlation Bhatt et al. (2020)	↑ Higher	Measures the alignment of the explanation with the model's true decision-making.
Max-Sensitivity Ye et al. (2019)	↓ Lower	Evaluates the robustness of the explanation by measuring its stability under small input perturbations.
SparsenessCl et al. (2020)	↓ Lower	Measures the concentration of the explanation across input features.
Parameter Randomisation Sixt et al. (2020)	↓ Lower	Assesses whether the explanation remains consistent when the model's parameters are shuffled.

Table 1: Evaluation Metrics for Model Explainability

3 FocusViT

Traditional interpretability methods, such as Grad-CAM and Attention Rollout, struggle with ViTs due to architectural differences compared to CNNs. ViTs operate on tokenised patches and distribute information globally via multi-head self-attention. As a result, early layers produce noisy, non-discriminative attention maps and naive aggregation of attention scores across all layers. Using multiplicative Rollout leads to vanishing attributions or over-smoothed outputs. To resolve this, we make three critical design choices:

- We extract both attention and gradient maps to capture not only what the model looks at, but also how sensitive the output is to these attentions.
- We avoid early layers in the computation of the attribution map, which empirical studies (e.g., SkipPLUS) Mehri et al. (2024) show to contain noisy or semantically diffuse information.
- We aggregate CAMs additively rather than multiplicatively, which preserves attribution signal strength and avoids exponential decay of influence.

3.1 Gradient-Weighted Attention Attribution

In ViTs, attention weights describe how each patch token attends to others in the input sequence. However, raw attention weights alone do not capture the model's sensitivity to class-specific decisions; they do not accurately reflect the spatial distribution of influence. This makes them insufficient for faithful explanation, especially in high-stakes domains such as medical imaging or scene classification. To address this, we combine the attention maps with their gradients with respect to the loss. This combination highlights not only where the model is looking, but also how much each region contributes to the output decision Xu et al. (2019).

Let $\mathbf{A}^{(l)} \in R^{H\times N\times N}$ be the attention matrix at layer l, where H is the number of attention heads and N is the number of tokens (including the class token). During the backwards pass, we compute the gradient of the loss $\mathcal{L}$ with respect to the attention weights:

$$\nabla\mathbf{A}^{(l)} = \frac{\partial\mathcal{L}}{\partial\mathbf{A}^{(l)}}$$

In our base implementation, we compute per-head CAMs using the gradient-weighted attention fusion:

$$\mathbf{CAM}_h^{(l)} = \mathrm{ReLU}\left(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\right)$$

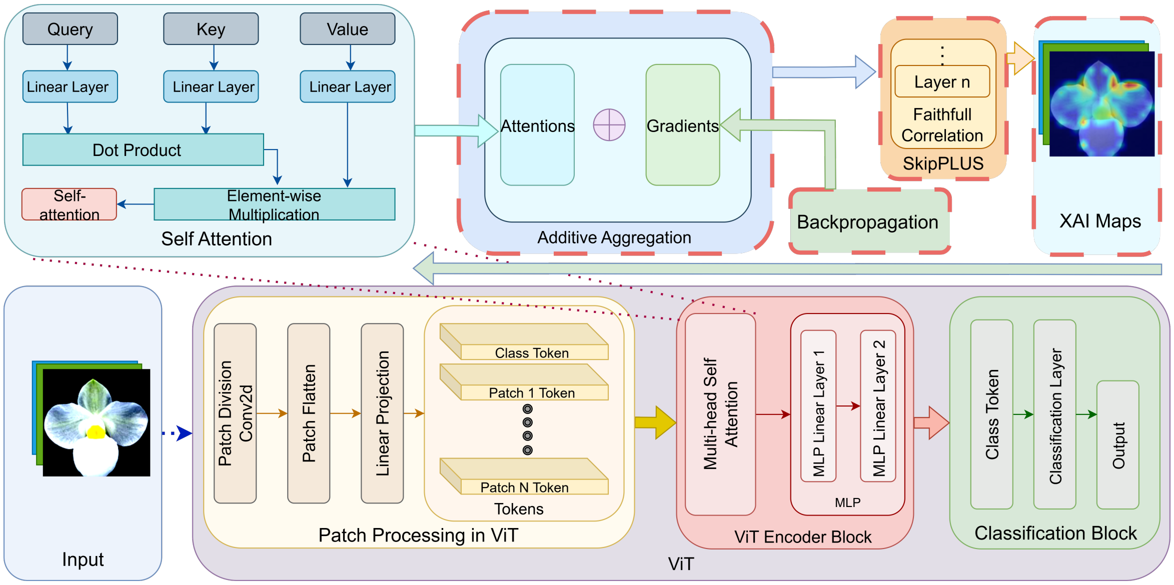

Figure 1: Overview of FocusViT: A hybrid approach combining gradient information and attention mechanisms, with a faithfulness-driven selection strategy and layer-skipping to enhance interpretability

Each attention head produces an individual CAM, and the final attribution map for the layer is obtained by averaging across heads:

$$\mathbf{CAM}^{(l)} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{CAM}_h^{(l)}$$

While this yields interpretable maps, it assumes that all heads contribute equally to the final decision. This assumption does not hold in practice, as certain heads focus on semantically important regions while others capture irrelevant structures. To address this limitation, we introduce a dynamic weighting scheme for each attention head based on the magnitude of its gradients. The weight for head h is defined as:

$$w_h = \frac{\sum_{i,j}\left|\nabla\mathbf{A}_h^{(l)}(i,j)\right|}{\sum_{h'=1}^{H}\sum_{i,j}\left|\nabla\mathbf{A}_{h'}^{(l)}(i,j)\right|}$$

The final layer-wise CAM is then computed as a weighted sum over heads:

$$\mathbf{CAM}^{(l)} = \sum_{h=1}^{H} w_h \cdot \mathrm{ReLU}\left(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\right)$$

This head-weighted attribution mechanism emphasises heads that are more sensitive to the loss function and thereby more relevant to the model's decision. Empirically, it produces sharper and more class-discriminative explanations than uniform averaging.

Observation (Gradient-Weighted Attention Attribution). For each attention head h in layer l, we define the gradient-weighted attribution map as

$$\mathbf{M}_h^{(l)} = \mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)},$$

where $\mathbf{A}_h^{(l)}$ are the attention weights and $\nabla\mathbf{A}_h^{(l)}$ their gradients with respect to the class loss. This formulation captures not only where the model attends but also how sensitive the prediction is to those attentions, offering a first-order view of local importance. Aggregating these maps across heads and layers therefore yields explanations that reflect both attention allocation and its causal impact on the output.

Observation (Faithfulness-Preserving Skip Aggregation). In practice, early ViT layers often introduce noise, while deeper layers encode more semantic features. Therefore, there exists a skip point m^* from which aggregating class attribution maps produces the most faithful explanations. We select m^* automatically by maximising a faithfulness score (e.g., Insertion or Deletion AUC) on a validation set. This ensures that only semantically relevant layers contribute to the final explanation.

4 Experiments and Results

Experimental setup: All experiments were conducted using PyTorch with the `timm` library to implement ViT models, and computations were performed on a single NVIDIA RTX 2080Ti GPU with 12GB memory. FocusViT introduces minor computational overhead, generating one explanation is approximately 12% slower than Attention Rollout due to the additional gradient computation, We evaluated FocusViT on five diverse image classification datasets: Oxford Flowers-102 Nilsback and Zisserman (2008), Oxford-IIIT Pets Parkhi et al. (2012), Stanford Dogs Khosla et al. (2011), Caltech-101 Fei-Fei et al. (2004), and MiT Indoor-67 Quattoni and Torralba (2009). These datasets cover a wide range of visual domains, object types, and scene complexities. For training, we employed the ViT-Base model (patch size 16×16) pre-trained on ImageNet, modifying the classification head to match the number of classes in each dataset. Images were preprocessed with resizing to 256, random cropping to 224×224, and augmented using random horizontal flips, ±15° rotations, and colour jitter. Normalisation was applied using ImageNet mean and standard deviation. The model was trained using the AdamW optimiser with a learning rate of 1×10^{-4}, a StepLR scheduler (decay factor 0.7 every 10 epochs), and a batch size of 32. Cross-entropy loss was used as the objective function, and early stopping was applied with a patience of 5 epochs to prevent overfitting. For explanation quality evaluation, we used QuantusLab (2023); Hedström et al. (2023) to compute four widely adopted XAI Guidotti et al. (2018) metrics: Faithfulness Correlation, Max-Sensitivity, Sparseness, and Model Parameter Randomisation. These evaluations were applied to FocusViT and Grad-CAM, LRP, SHAP, LIME, and Attention Rollout for comparative analysis.

4.1 Quantitative Evaluation Metrics

The table 2 show a detailed comparison of the performance of six XAI Chatzimparmpas et al. (2020) techniques that include our method, Grad-CAM Selvaraju et al. (2017), LRP Chefer et al. (2021), AttentionAbnar and Zuidema (2020), LIME Ribeiro et al. (2016), and SHAP Lundberg and Lee (2017) across five different datasets (Flower Nilsback and Zisserman (2008), DogKhosla et al. (2011), MiT, Caltech , and PetParkhi et al. (2012)) using four evaluation metrics: Faithfulness Correlation, Max-Sensitivity, Sparseness (Complexity), and Model Parameter Randomisation. The Faithfulness Correlation metric measures the alignment of an explanation with the model's decision-making process, revealing that our method consistently outperforms all other techniques across the datasets. In the Flower dataset, our method achieves a score of 0.0350, surpassing SHAP (0.0009) and LIME (0.0006), both of which show poor alignment. Grad-CAM (0.0293) and LRP (0.0323) provide moderate performance but still fall short of our method. A similar trend is observed in the Dog dataset, where our method outperforms other XAI methods by 0.0336. In the MiT dataset, our method scores 0.0500, with SHAP and LIME performing poorly at 0.0000, while Grad-CAM and Attention show moderate results. Finally, in the Caltech and Pet datasets, Our method continues to excel, with SHAP and LIME performing the worst, particularly at 0.0016 and 0.0005 in the Caltech and Pet datasets.

The Max-Sensitivity metric quantifies the stability of explanation methods by measuring their response to minor input perturbations, with lower values indicating higher robustness. Across multiple datasets, our method consistently demonstrates low sensitivity, comparable to or outperforming other techniques. On the Flower dataset, Attention exhibits the lowest sensitivity (1.0004), followed by Grad-CAM (1.0020), LRP (1.0035), and our method (1.0037), while LIME (3.7378) and SHAP (13.3765) show substantial variability. For the Dog dataset, our method achieves the lowest score (1.2663), outperforming Grad-CAM (1.4135), LRP (1.3110), and Attention (1.3334), with LIME and SHAP reaching 34.2574 and 46.2796, respectively. In the MiT dataset, our method again leads with a sensitivity of 1.1240, followed by Grad-CAM (1.1362), Attention (1.2214), and LRP (1.2668), while LIME and SHAP show extreme sensitivity at 26.1842 and 58.5712. The Caltech dataset reveals minimal variation among our method, Grad-CAM, LRP, and Attention (all 1.00), contrasting with LIME (4.9136) and SHAP (10.0210). Finally, in the Pet dataset, Grad-CAM (1.0043) and LRP (1.0354) show low sensitivity, with our method (1.0510) and Attention (1.5483) slightly higher, whereas LIME (5.5832) and SHAP (37.8872) again show high sensitivity. These results highlight the robustness of our method across diverse datasets, particularly when compared to perturbation-based approaches like LIME and SHAP.

Sparseness, in the context of XAI, measures the concentration of the explanation across the input features, indicating how distributed or concentrated the importance values are across the input. In the Flower dataset, our method demonstrates the lowest sparsity (0.1248), outperforming SHAP (0.3449) and other methods, indicating its better sparsity. Similarly, in the Dog dataset, our method (0.4844) maintains a relatively low sparsity, in comparison to LRP and Grad-CAM (both 0.5265) show comparable sparsity, while LIME and SHAP exhibit significantly higher sparsity values (0.9975 and 0.6452, respectively). The MiT dataset also reveals our method (0.3121) as the most efficient, followed closely by LRP (0.3771), with LIME and SHAP showing higher sparsity at 0.9978 and 0.5027. In the Caltech dataset, our method (0.4298) and Attention (0.4412) lead in sparsity, while SHAP (0.7261) and LIME (0.9967) show the highest sparsity.

Table 2: XAI Evaluation Metrics across Datasets (Best highlighted in green, worst highlighted in pink)

Dataset	Metric	LIME	SHAP	LRP	Grad-CAM	Attention	Our
Flower	Faithfulness Correlation	0.0006	0.0009	0.0323	0.0293	0.0216	0.0350
	Max-Sensitivity	3.7378	13.3765	1.0035	1.0020	1.0004	1.0037
	Sparseness (Complexity)	0.9987	0.3449	0.4348	0.5466	0.5056	0.1248
	Model Randomisation	0.1854	0.0000	0.1265	0.1583	0.0016	0.0015
Dog	Faithfulness Correlation	0.0036	0.0012	0.0233	0	0	0.0336
	Max-Sensitivity	34.2574	46.2796	1.3110	1.4135	1.3334	1.2663
	Sparseness (Complexity)	0.9975	0.6452	0.5265	0.5265	0.5527	0.4844
	Model Randomisation	0.0003	0.0000	0.2452	0.3713	0.398	0.1931
MiT	Faithfulness Correlation	0.0000	0.0000	0.0186	0.0211	0.0463	0.0500
	Max-Sensitivity	26.1842	58.5712	1.2668	1.1362	1.2214	1.1240
	Sparseness (Complexity)	0.9978	0.5027	0.3771	0.6069	0.4987	0.3121
	Model Randomisation	0.0003	0.0000	0.1101	0.2680	0.1746	0.05800
Caltech	Faithfulness Correlation	0.0024	0.0016	0.0122	0.0087	0.0152	0.1558
	Max-Sensitivity	4.9136	10.0210	1.0060	1.0096	1.0055	1.0023
	Sparseness (Complexity)	0.9967	0.7261	0.5267	0.6090	0.4412	0.4298
	Model Randomisation	0.0006	0.0000	0.2052	0.2395	0.3437	0.1627
Pet	Faithfulness Correlation	0.0005	0.0018	0.0063	0.0197	0.0239	0.0635
	Max-Sensitivity	5.5832	37.8872	1.0354	1.0043	1.5483	1.0510
	Sparseness (Complexity)	0.9978	0.7212	0.5577	0.3749	0.5895	0.3675
	Model Randomisation	0.0010	0.0001	0.4169	0.2596	0.0010	0.05135

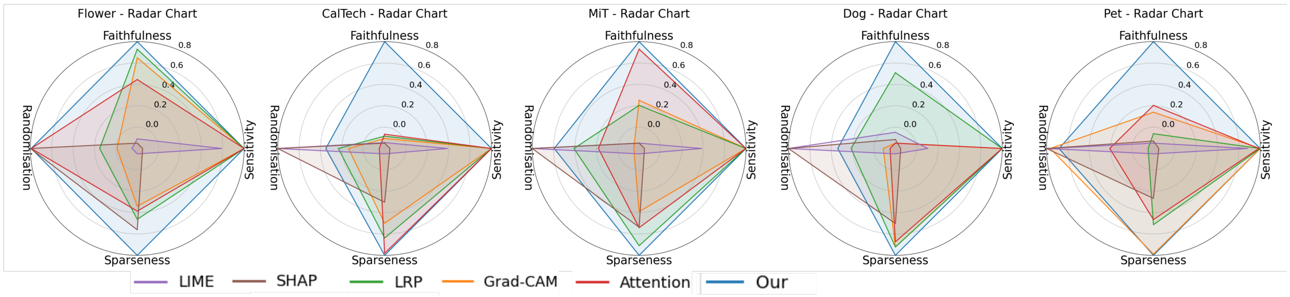

Figure 2: Radar Charts for Evaluating XAI Techniques Across Multiple Datasets

Similarly, in the Pet dataset, Grad-CAM (0.3749) and our method (0.3675) demonstrate the lowest sparsity, while SHAP (0.7212) and LIME (0.9978) exhibit the highest values.

The Model Parameter Randomisation metric assesses the robustness of explanation methods to perturbations in model parameters, with lower scores indicating greater stability. Across the five datasets, SHAP consistently demonstrates the highest robustness, achieving the lowest score in each case (ranging from 0.0000 to 0.0001). LIME also performs well, securing second place in most datasets, particularly in the Dog (0.0003), MiT (0.0003), Caltech (0.0006), and Pet (0.0010) datasets. Our method shows moderate resilience, ranking third overall and first in the non-perturbation base XAI technique with relatively low values (e.g., 0.0015 in Flower and 0.0514 in Pet), outperforming most gradient-based techniques. Grad-CAM, LRP and Attention display noticeably higher sensitivity, with values frequently exceeding 0.4, as seen in the Pet (0.4169 and 0.2596, respectively) and Caltech (0.3437 and 0.2395) datasets. These results suggest that, while perturbation-based methods like SHAP and LIME are highly stable under model randomisation, gradient- and attention-based methods exhibit greater sensitivity, with our method offering a balanced trade-off between stability and interpretability.

Area Under Radar (AUR): To quantitatively compare explanation methods across multiple evaluation metrics, we compute the area under the radar (AUR)

Table 3: Area Under Radar (AUR) for Each Method Across Datasets. The highest values per dataset are highlighted.

Dataset	LIME	SHAP	LRP	Grad-CAM	Attention	Our
Flower	0.0163	1.1656	1.0342	0.7785	1.1859	1.9916
Dog	0.2687	1.0714	1.1870	0.1371	1.3432	1.6311
MiT	0.2809	1.1006	1.3474	0.0000	1.3036	1.9332
Caltech	0.2818	0.9054	1.0339	0.5159	1.0578	1.6378
Pet	0.2886	0.8235	0.8447	1.6803	1.1440	1.9111

using the polar sector formula:

$$\text{Area} = \frac{1}{2}\sum_{i=0}^{n-1} r_i \cdot r_{i+1} \cdot \sin(\theta_{i+1} - \theta_i)$$

Where r_i represents the normalised score of the i^{th} metric and θ_i is the angle corresponding to that metric on the radar chart. This formula yields a scalar measure reflecting both the magnitude and consistency of the method's performance across dimensions.

The results presented in the radar charts and the AUR table illustrate the performance of FocusViT in comparison to several other XAI techniques, Grad-CAM, LRP, Attention, LIME and SHAP, in five datasets: Flower, Dog, MiT, Caltech, and Pet. Our method consistently outperforms all other techniques across all datasets. For example, in the Flower dataset, our method achieved an AUR value of 1.9916, which is higher than other methods. Similarly, in the Dog dataset, our method scored 1.6311, surpassing Grad-CAM's 0.1371 and LIME's 0.2687, further highlighting the strength of our approach. The radar charts clearly demonstrate that our method covers the largest area in all four metrics, indicating superior performance in providing well-rounded and reliable explanations. While LRP and Attention exhibit some improvement in certain datasets (e.g., MiT), with LRP achieving an AUR value of 1.3474, they still fail to match the overall performance of our method, which scored 1.9332 in the MiT dataset. Furthermore, LIME and SHAP consistently perform poorly, especially in Sparseness and Faithfulness, as evidenced by their low AUR scores across all datasets—SHAP, for instance, only scored 0.8235 in the Pet dataset. The AUR values confirm these observations, with our method achieving the highest values in all datasets, underscoring its ability to generate faithful, low-sensitive, and less sparse explanations across diverse tasks. The lower AUR scores for Grad-CAM, LIME, and SHAP further highlight the relative limitations of these techniques, particularly in comparison to our method. These results suggest that our method provides a more robust and balanced approach to explainable AI, offering significant advantages over other techniques.

4.2 Qualitative Analysis

In Figure 3, we present qualitative visualisations showing the output of each XAI method on various datasets under different conditions, highlighting both correct and incorrect predictions by the model. These feature maps, generated using different XAI techniques, aim to identify the parts of the image that the model considers most relevant to its prediction. Upon examining the figure, it is evident that, in most correctly predicted cases, our model successfully highlights the relevant image regions. For instance, in the Pet dataset, when the model correctly predicts, it focuses on the cat; similarly, in the Caltech dataset, it concentrates on the insect, and in the Flower dataset, it highlights the relevant parts of the flowers. However, in the Dog dataset, the model does not focus on the dog, and this issue is also observed across other XAI methods. Furthermore, when compared with other methods, permutation-based techniques mostly fail to explain the predictions effectively, as they tend to focus on irrelevant parts of the image. This observation is also corroborated by the quantitative analysis in Section 4.1. On the other hand, gradient-based and non-permutation methods show moderate performance, but they still struggle in more complex scenarios. For example, in the Pet dataset, these methods often focus on irrelevant areas, even when the model's prediction is correct. A similar pattern is observed in the Caltech dataset.

5 Conclusion

We proposed FocusViT, an explainability framework for ViTs that integrates gradient-weighted attention with dynamic layer-skipping. Quantitative results show that FocusViT surpasses methods such as Grad-CAM, LIME, and SHAP across core metrics of faithfulness, robustness, and sparsity. It achieves higher faithfulness correlation and lower sensitivity to perturbations, producing accurate and stable explanations. The method also yields sparser maps, highlighting the most relevant features for improved interpretability. Qualitatively, FocusViT consistently

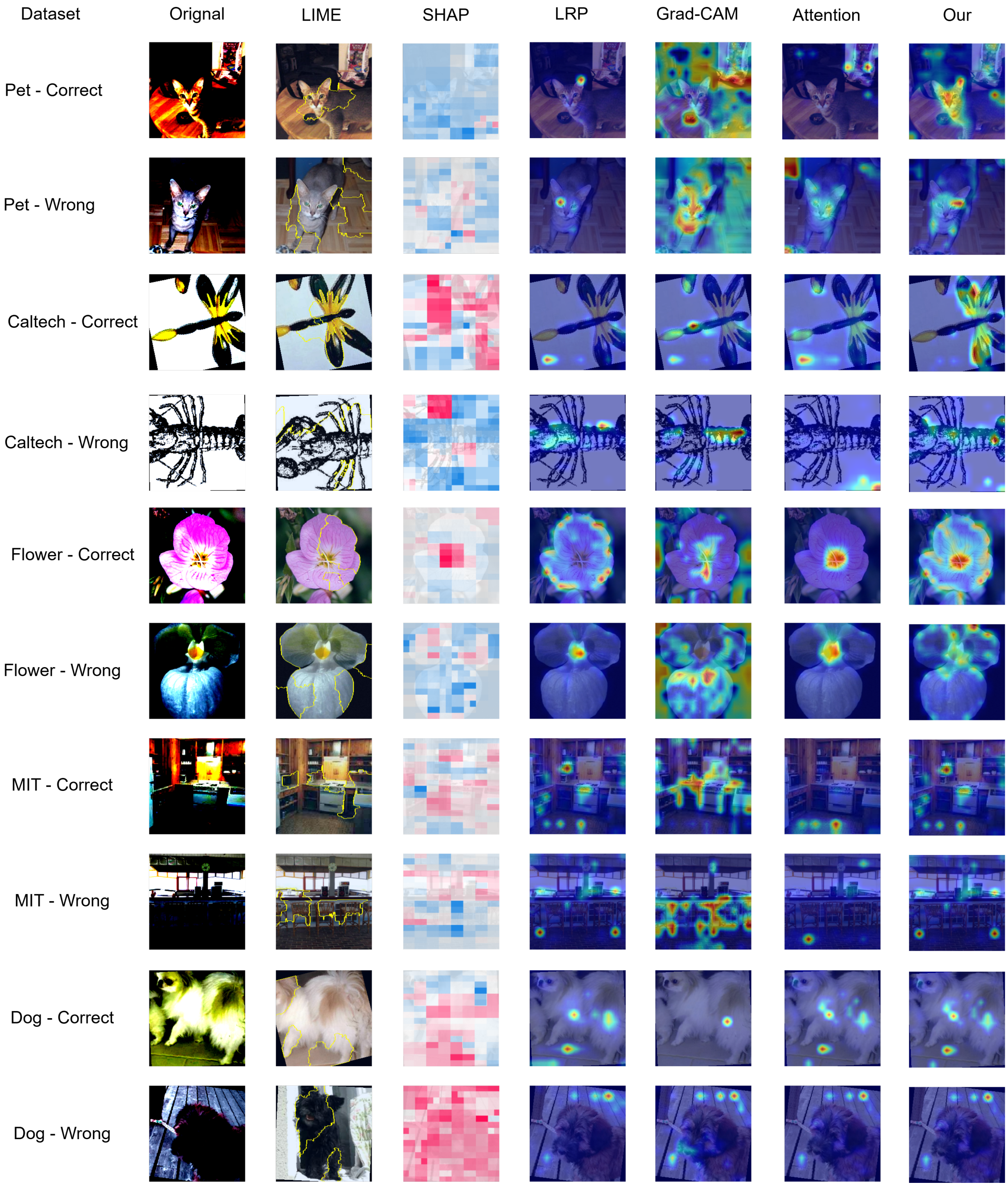

Figure 3: Qualitative comparison of different XAI methods (LIME Ribeiro et al. (2016), SHAP Lundberg and Lee (2017), Grad-CAM Selvaraju et al. (2017), Attention Rollout Abnar and Zuidema (2020), LRP Chefer et al. (2021), and our FocusViT) across multiple datasets: Flowers-102 Nilsback and Zisserman (2008), Stanford Dogs Khosla et al. (2011), Oxford-IIIT Pets Parkhi et al. (2012), Caltech-101 Fei-Fei et al. (2004), and MiT Indoor-67 Quattoni and Torralba (2009). FocusViT consistently highlights semantically relevant regions, while other methods often spread attention to background or irrelevant areas.

localises task-relevant regions better than competing techniques, making it a reliable tool for enhancing ViT transparency, particularly in applications where interpretability is critical.

References

S. Abnar and W. Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

R. Achtibat, S. M. V. Hatefi, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, and W. Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024.

U. Bhatt, A. Weller, and J. M. Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

W. Bousselham, A. Boggust, S. Chaybouti, H. Strobelt, and H. Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. *arXiv preprint arXiv:2404.03214*, 2024.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020.

A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020.

H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

R. L. Draelos and L. Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. M. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL `http://jmlr.org/papers/v24/22-0142.html`.

A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

U. M. I. Lab. Quantus: Xai metrics for machine learning, 2023. URL `https://github.com/understandable-machine-intelligence-l` Accessed: 2025-07-15.

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

F. Mehri, M. Fayyaz, M. S. Baghshah, and M. T. Pilehvar. Skipplus: Skip the first few layers to better explain vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–215, 2024.

G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, and D. Zhu. Attcat: Explaining transformers via attentive class activation tokens. *Advances in neural information processing systems*, 35:5052–5064, 2022.

A. Quattoni and A. Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009.

M. T. Ribeiro, S. Singh, and 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*, pages 1135–1144, 2016.

R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

L. Sixt, M. Granz, and T. Landgraf. When explanations lie: Why many modified bp attributions fail. In *International conference on machine learning*, pages 9046–9057. PMLR, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J. Wu, B. Duan, W. Kang, H. Tang, and Y. Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10926–10935, 2024.

J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, 35:4971–4982, 2022.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**] (See Section 3, and Observations 1 and 2.)

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] (Section 4)

   (c) (Optional) Anonymised source code, with specification of all dependencies, including external libraries. [**Yes**] (Will be provided on acceptance)

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [**Not Applicable**] (No formal theorems are presented; Observations 1 and 2 provide intuition only.)

   (b) Complete proofs of all theoretical results. [**Not Applicable**] (No formal proofs are required since no theoretical results are claimed.)

   (c) Clear explanations of any assumptions. [**Not Applicable**] (The paper does not make formal theoretical assumptions)

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**] (All datasets are publicly available; see Section 4. We will provide anonymized code and reproduction instructions upon acceptance.)

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**] (Section 4, Experimental Setup describes dataset splits, preprocessing, augmentations, optimizer, scheduler, and training settings.)

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**] (Section 3, Section 2 defines all evaluation metrics)

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**] (Section 4, Experimental Setup: single NVIDIA RTX 2080Ti GPU with 12GB memory.)

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [**Yes**] (All datasets and pretrained models are cited in Section 4.)

(b) The license information of the assets, if applicable. [**Yes**] (All public datasets are released for non-commercial academic research use.)

(c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]

(d) Information about consent from data providers/curators. [**Not Applicable**] (All datasets are publicly available and widely used in academic research.)

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]