A report for Deep Vision

A Comparative Study of Explanable Ai methods

Sohambhai Joita

Supervised by
Prof. Dr. phil. Tatyana Ivanovska

Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

July 3, 2025

Abstract

This study presents a comparative evaluation of three attention-based interpretability methods for Vision Transformers (ViTs): Classical Attention Rollout, Gradient-weighted Multi-head Attention Rollout (GMAR) with L1 and L2 norms, and the Layerwise Gradient-based Explanation method (LeGrad). Each method is applied to a fixed pretrained transformer model to assess how effectively it highlights class-relevant regions that influence model predictions. I perform both qualitative visualization and quantitative evaluation using standard metrics: Pixel Accuracy, mean Intersection over Union (mIoU), and mean Average Precision (mAP), where heatmaps are thresholded at 0.5 for binarization. Experimental results show that LeGrad consistently outperforms GMAR and Classical Rollout across all metrics, producing more precise, class-specific, and spatially aligned explanations. These findings underscore the effectiveness of gradient-based token-level attribution in enhancing the interpretability of attention mechanisms in transformer-based vision models.

1 Introduction and Motivation

Vision Transformers (ViTs) have emerged as a compelling alternative to convolutional neural networks (CNNs) in a wide array of computer vision tasks, owing to their ability to model long-range dependencies using self-attention mechanisms. However, ViTs present significant challenges in terms of interpretability. Unlike CNNs, which leverage spatially localized operations that naturally lend themselves to visualization techniques, ViTs process images as sequences of tokenized patches, leading to more abstract and distributed feature representations. Many traditional explainability methods such as Grad-CAM, LRP, and SmoothGrad were originally developed for CNN architectures and depend heavily on the presence of convolutional feature maps [3]. Consequently, their direct application to ViTs often yields explanations that are neither spatially coherent nor semantically meaningful. To address this limitation, transformer-specific interpretability methods have been developed. One of the earliest among these is the Classical Attention Rollout method proposed by Abnar and Zuidema [1], which recursively aggregates attention across layers to visualize how information propagates through the model. Although intuitive and computationally efficient, this approach is class-agnostic and assumes equal contribution from all attention heads, which often results in diffuse and noisy heatmaps. To enhance class-awareness and head selectivity, the Gradient-weighted Multi-head Attention Rollout (GMAR) method introduces gradient-based weighting of attention heads [5]. This enables the model to prioritize heads that contribute more significantly to the predicted class, thereby producing more discriminative explanations. GMAR is evaluated in two variants: using L1-norm and L2-norm to compute head importance weights. Despite its improvements over the classical method, GMAR remains limited by factors such as the global nature of attention aggregation and sensitivity to gradient scaling. To overcome these limitations, LeGrad [2] proposes a layerwise gradient-based approach that directly computes the gradient of the predicted class score with respect to each attention map. These gradients are used to weigh and aggregate attention responses across layers, producing class-specific, spatially aligned visualizations that more faithfully reflect the decision-making process of the model.

In this work, three interpretability methods—Classical Attention Rollout, GMAR (L1 and L2), and LeGrad—are evaluated on a fixed ViT model ViT-B/16 pretrained on ImageNet-1K. The evaluation is carried out on a subset of 14 diverse images from the pascal-voc-2012 dataset, which includes manually annotated binary segmentation masks. Both qualitative and quantitative analyses are performed. For quantitative analysis, heatmaps are thresholded at 0.5 and compared against the ground truth masks using standard metrics: Pixel Accuracy, mean Intersection-over-Union (mIoU), and mean Average Precision (mAP). These evaluations aim to assess not only the visual coherence of each explanation method but also their ability to localize class-relevant regions in a statistically reliable manner.

2 Method

In this section, I first introduce ViT's mechanics and the feature aggregation mechanisms used for this architecture. I then explain the details of Classical Attention Rollout, GMAR, and LeGrad, starting with how they operate at a single layer and extending the explanation to their full multi-layer formulations.

2.1 Background: Architecture of ViT

The Vision Transformer (ViT) is a transformer-based architecture designed for image recognition tasks by adapting the principles of the transformer encoder, originally proposed for natural language processing, to visual data [4]. In ViT, an input image is divided into fixed-size non-overlapping patches. Each patch is flattened into a vector and then projected into a high-dimensional embedding space through a learnable linear transformation. These patch embeddings form a sequence of tokens that represent the image in a spatially abstracted form. To this sequence, a learnable classification token—denoted as `[CLS]`—is prepended. This special token is used by the model to aggregate information from all patches and serves as the final representation used for classification. Additionally, positional encodings are added to the tokens to preserve the spatial ordering of patches, as the transformer architecture itself is permutation-invariant and lacks inherent spatial structure. The resulting sequence of embedded tokens is passed through a standard transformer encoder, which consists of multiple identical layers. Each layer includes two primary components: multi-head self-attention (MSA) and a feed-forward neural network (MLP), both equipped with residual connections and layer normalization. The MSA mechanism allows each token to attend to every other token in the sequence, enabling the model to capture global relationships across the image.

Figure 1 shows a high-level overview of the Vision Transformer architecture, including patch embedding, token formation, positional encoding, and the transformer encoder block.

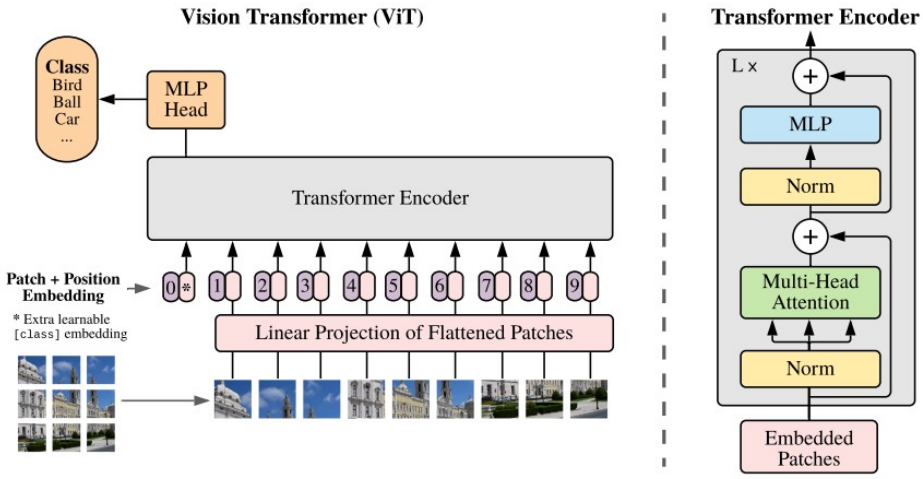

Figure 1: Architecture of the Vision Transformer (ViT) [4].

Within each transformer layer, the input token matrix is projected into query (Q), key (K), and value (V) matrices. For each head in the multi-head attention module, attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where d_k is the dimensionality of each head's key vector. The outputs of all heads are concatenated and passed through a final linear projection. This attention output is then

added to the input via a residual connection and normalized. The result is further processed by a position-wise feed-forward network, again with residual connections and layer normalization. The final output of the `[CLS]` token, after passing through all transformer layers, serves as the global image representation and is used for classification. The intermediate attention maps and token interactions across layers form the core interpretability targets for the methods I explore in this report, namely Classical Attention Rollout, GMAR, and LeGrad.

2.2 Classical Attention Rollout

To interpret the predictions of Vision Transformers, one of the earliest and most intuitive methods proposed is the Classical Attention Rollout [1]. This method is based on the idea of tracing how information flows from the input patches to the output classification token `[CLS]` across the layers of the transformer. Each transformer layer outputs a set of attention matrices that describe how much each token attends to every other token. Classical Rollout uses these attention matrices to compute an aggregated attention map that reflects how strongly each input token contributes to the model's final decision.

For each layer in the transformer, I first average the attention weights across all heads to obtain a single square matrix $A^{(l)} \in \mathbb{R}^{N \times N}$, where N is the number of tokens. Since transformers also utilize residual connections, I follow the rollout formulation by adding an identity matrix to each attention matrix and applying row-wise normalization:

$$\tilde{A}^{(l)} = \text{normalize}\left(A^{(l)} + I\right),$$

where I is the identity matrix and normalization ensures that each row sums to one. The full attention rollout is computed by recursively multiplying the normalized matrices from the first to the last layer:

$$A_{\text{rollout}} = \tilde{A}^{(1)} \cdot \tilde{A}^{(2)} \cdot \ldots \cdot \tilde{A}^{(L)},$$

where L is the total number of transformer layers. The first row of A_{rollout}, corresponding to the `[CLS]` token, quantifies how much it attends to each of the input patches after the complete forward pass. This row is used to generate a visual explanation by projecting the attention values back onto the spatial positions of the original image patches.

Although Classical Attention Rollout provides a simple and architecture-aware explanation method, it has notable limitations. It treats all attention heads equally and ignores the influence of the model's predicted class. As a result, the resulting explanations are class-agnostic and often spatially diffuse, which may reduce their utility for fine-grained interpretability. Nevertheless, this method serves as a foundational approach for understanding attention dynamics in Vision Transformers and provides a strong baseline for more advanced techniques such as GMAR and LeGrad.

2.3 Gradient-weighted Multi-head Attention Rollout (GMAR)

To overcome the limitations of Classical Attention Rollout, particularly its inability to differentiate between the importance of different attention heads, I incorporate Gradient-driven Multi-head Attention Rollout (GMAR) into this study. GMAR is a recently proposed interpretability method specifically designed for Vision Transformers [5]. It introduces a more refined attention rollout framework by leveraging class-specific gradient

information to weight individual attention heads based on their relative contribution to the model's prediction. This method combines the structural insights of attention rollout with the specificity of gradient-based explanations to produce sharper and more meaningful attention maps.

An overview of the GMAR framework is illustrated in Figure 2. As shown, the method begins by computing the gradient of the predicted class logit with respect to each attention head in the multi-head self-attention layers. These gradients are used to assign a scalar importance score to each head, quantifying its influence on the model's decision. The scores are computed using either the L_1 or L_2 norm:

$$\mathrm{GR}_i = \begin{cases} \sum |G_{hi}|, & \text{for } L_1 \text{ norm} \\ \sqrt{\sum G_{hi}^2}, & \text{for } L_2 \text{ norm} \end{cases}, \quad w_i = \frac{\mathrm{GR}_i}{\sum_j \mathrm{GR}_j}$$

where G_{hi} denotes the gradient of the attention map for head i.

Once the weights are computed and normalized, they are used to create a weighted attention matrix at each layer:

$$\hat{A}^{(\ell)} = \sum_{h=1}^{H} w^{(\ell,h)} A^{(\ell,h)}$$

where $A^{(\ell,h)}$ is the attention map for head h in layer ℓ, and $w^{(\ell,h)}$ is the corresponding importance weight. These matrices are then recursively aggregated across layers using the following rollout formula:

$$A_{\mathrm{GMAR}} = \prod_{\ell=1}^{L} \left(\hat{A}^{(\ell)} + \alpha I \right)$$

where α controls the residual contribution and I is the identity matrix. This formulation preserves hierarchical token interactions while incorporating head-specific contributions into the explanation.

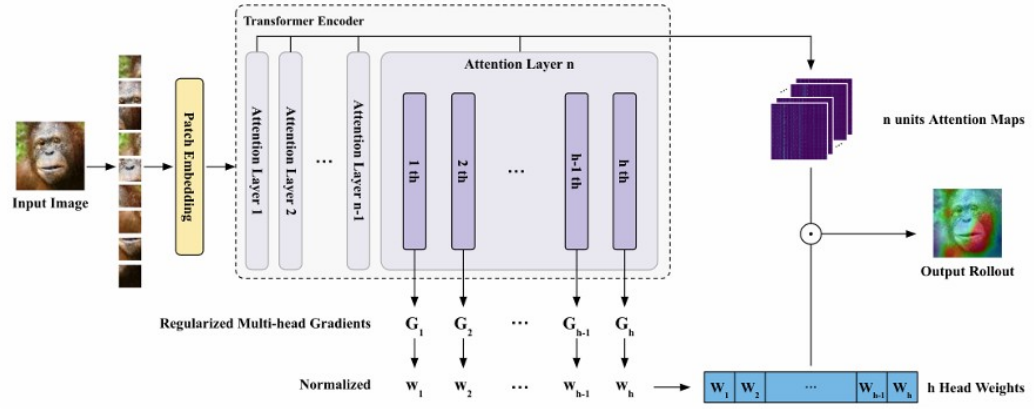

Figure 2: Workflow of the GMAR (Gradient-Driven Multi-Head Attention Rollout) algorithm [5].

By integrating head-level gradient signals, GMAR generates class-specific attention maps that are both spatially focused and semantically aligned with the model's decision.

Compared to Classical Attention Rollout, it produces more interpretable and accurate visualizations while maintaining compatibility with the original transformer architecture. Although slightly more computationally demanding due to the gradient calculations, GMAR's improvements in interpretability make it a valuable step toward more transparent Vision Transformer models.

2.4 Layerwise Gradient-based Explanation (LeGrad)

To build on the strengths of attention-based interpretability while addressing their limitations, I incorporate LeGrad [2], a state-of-the-art explainability method specifically designed for Vision Transformers. Unlike classical rollout or GMAR, which rely solely on attention weights or gradient-weighted attention, LeGrad uses the gradient of the model's prediction with respect to the attention maps as the explanation signal itself. This enables it to measure how sensitive each attention score is to changes in the final prediction, providing a more direct and model-aligned explanation.

The overall framework of LeGrad is illustrated in Figure 3. For a given image and prediction class, LeGrad computes the gradient of the class score with respect to each layer's attention map. The result is a set of class-specific relevance maps for every layer, which are then aggregated to form the final heatmap. The method operates independently at each transformer block, extracting gradients of the class activation with respect to the self-attention weights.

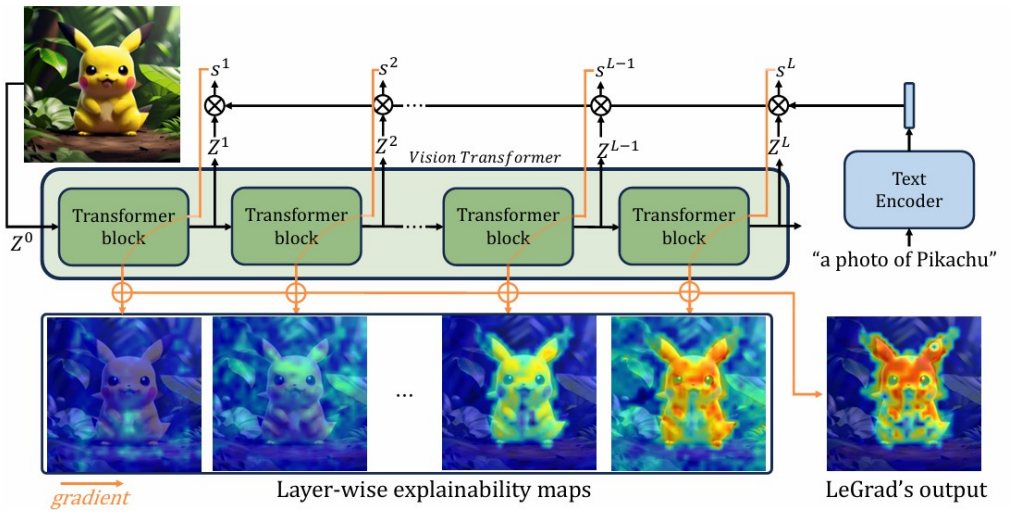

Figure 3: Overview of LeGrad: [2]

To better understand how LeGrad operates at the layer level, consider the computation for a single transformer block, as shown in Figure 4. Let $A^{(l)} \in \mathbb{R}^{H \times N \times N}$ denote the raw attention map of layer l, where H is the number of attention heads and N is the number of tokens. LeGrad computes the gradient of the class score $s^{(l)}$ with respect to the attention map and clips negative gradients using ReLU:

$$\frac{\partial s^{(l)}}{\partial A^{(l)}} \longrightarrow \left(\frac{\partial s^{(l)}}{\partial A^{(l)}} \right)_+$$

Then, the explainability vector $\hat{E}^{(l)} \in \mathbb{R}^N$ is computed by summing over heads and query

positions:

$$\hat{E}^{(l)} = \frac{1}{H \cdot N} \sum_{h=1}^{H} \sum_{i=1}^{N} \left(\frac{\partial s^{(l)}}{\partial A^{(l)}_{h,i,:}} \right)_+$$

After removing the entry corresponding to the [CLS] token, this vector is reshaped into a $W \times H$ grid and min-max normalized:

$$E^{(l)} = \text{norm}\left(\text{reshape}\left(\hat{E}^{(l)}_{1:}\right)\right) \in \mathbb{R}^{W \times H}$$

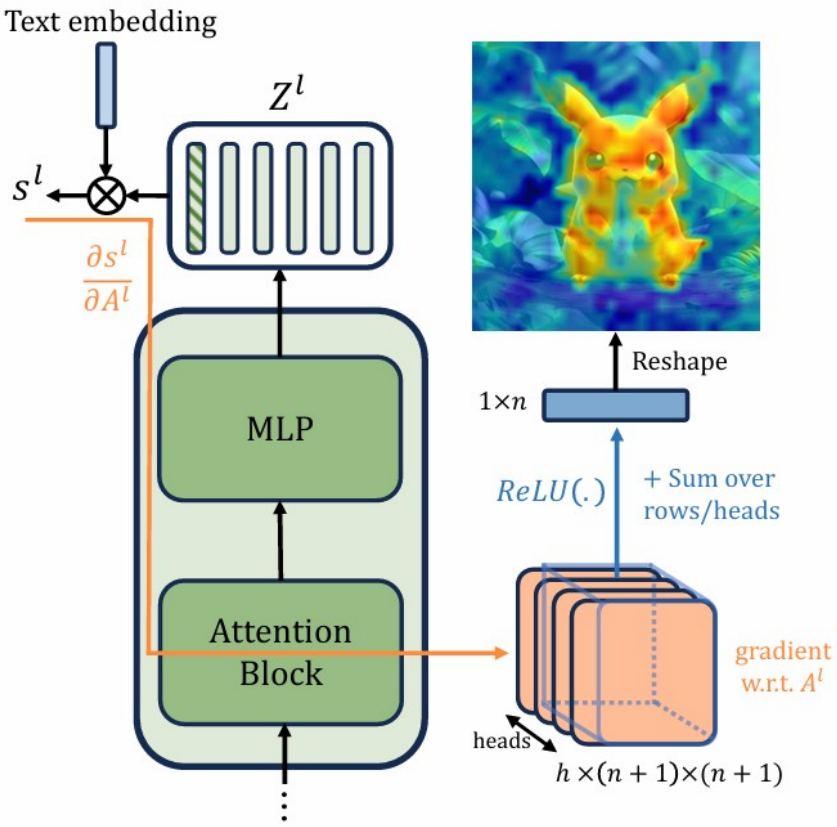

Figure 4: LeGrad applied to a single layer [2]

To compute the final heatmap E, the layerwise maps are averaged:

$$E = \text{norm}\left(\text{reshape}\left(\frac{1}{L}\sum_{l=1}^{L} \hat{E}^{(l)}_{1:}\right)\right) \in \mathbb{R}^{W \times H}$$

This aggregation captures contributions from each layer, ensuring that both low-level and high-level attention interactions are represented in the final visual explanation. Compared to other methods, LeGrad is highly scalable, architecture-agnostic, and applicable across various ViT tasks such as zero-shot localization, segmentation, and audio-visual grounding. In my experiments, I find that LeGrad consistently produces sharper, more class-aligned explanations than GMAR or Classical Rollout.

3 Experiment

The primary task in this study is image classification. Given an input image, the goal is to predict the correct category label from a fixed set of classes. To perform this task, I

use the Vision Transformer Base model with a patch size of 16×16, commonly referred to as ViT-B/16. This model is pretrained on the ImageNet-1K dataset. ViT-B/16 consists of 12 transformer encoder layers, each with 12 self-attention heads, a hidden embedding dimension of 768, and a feed-forward network of intermediate size 3072. I use ViT-B/16 solely for inference. No fine-tuning or retraining is performed.

To evaluate the interpretability methods, I use a curated subset of 14 images from the PASCAL VOC 2012 dataset. This dataset provides both natural images and corresponding pixel-wise segmentation masks, which serve as ground truth for object localization. The selected images cover diverse object categories and serve as a consistent benchmark for both qualitative and quantitative evaluations. Each image is passed through the pretrained ViT-B/16 model, and interpretability heatmaps are generated using each method.

For the quantitative analysis, I employ three metrics: Pixel Accuracy, Mean Intersection over Union (mIoU), and Mean Average Precision (mAP). To compute these metrics, I generate a heatmap for each class object present in the image, binarize the heatmap using a threshold of 0.5, and compare it against the ground truth mask. This thresholding step ensures that only the most salient regions identified by the model are considered as object-localization regions. This enables a rigorous evaluation of how well each interpretability method identifies the spatial extent of relevant objects. For qualitative analysis, I visualize the attention heatmaps overlaid on the original image to assess the interpretability and focus of each method in human-perceptible terms. The interpretability techniques evaluated include Classical Attention Rollout, Gradient-weighted Multi-head Attention Rollout (GMAR) with both L1 and L2 norm variants, and Layerwise Gradient-based Explanation (LeGrad). Each method provides a visual mechanism to highlight the regions of the image that contribute most to the classification decision. These interpretability methods are all applied to the same ViT-B/16 model under identical classification conditions. The objective of this work is not to improve the predictive performance of the model, but rather to examine how effectively and faithfully each method can uncover the internal decision-making behavior of a pretrained Vision Transformer.

4 Results and Discussion

To assess the quality and interpretability of attention-based explainability methods for Vision Transformers (ViT), I evaluate and compare four distinct techniques: Classical Attention Rollout, GMAR (with both L1 and L2 normalization variants), and LeGrad. The evaluation is conducted using a combination of quantitative metrics—namely, Pixel Accuracy, Mean Intersection over Union (mIoU), and Mean Average Precision (mAP)—alongside qualitative visualizations that help capture the spatial relevance and fidelity of each generated heatmap. My objective is to analyze how well each method identifies and localizes the most salient regions contributing to the model's classification decision.

4.1 Quantative Comparison

To assess the interpretability of the generated attention maps, I evaluate four explainability techniques applied to Vision Transformers (ViT): Classical Attention Rollout, GMAR with L1 normalization, GMAR with L2 normalization, and LeGrad. Each method produces a heatmap that is compared with ground truth binary segmentation masks using standard evaluation metrics.

Method	mPA (↑)	mIoU (↑)	mAP (↑)
Classical Rollout	79.12	2.24	37.71
GMAR-L1	79.63	5.00	43.45
GMAR-L2	79.79	4.93	43.48
LeGrad	88.88	42.96	85.23

Table 1: Quantitative Comparison of ViT Explainability Methods

Among all methods, LeGrad achieves the highest values across all evaluation metrics, demonstrating strong performance in localizing class-relevant regions. It reaches 88.88 in mean pixel accuracy, 42.96 in mean Intersection over Union, and 85.23 in mean Average Precision, making it the most reliable technique for generating faithful visual explanations. GMAR-L1 and GMAR-L2 both show moderate improvements over Classical Rollout, with GMAR-L2 slightly outperforming L1 in precision. However, their mIoU scores remain substantially lower than that of LeGrad, indicating weaker spatial localization. Classical Rollout, despite its computational simplicity, performs the worst in all metrics, highlighting its limitations in generating precise and semantically aligned heatmaps.

4.2 Qualitative Comparison

The qualitative results offer a comprehensive visual assessment of the effectiveness of different attention-based explainability methods—namely Classical Attention Rollout, GMAR-L1, GMAR-L2, and LeGrad—across a diverse set of examples including Beagle, Boathouse, Sorrel, Couch, and Bittern. In the Beagle image, LeGrad distinctly focuses on the dog inside the kennel, accurately localizing the most relevant semantic regions while suppressing the background. In contrast, Rollout produces a relatively diffused activation map, spreading attention across the kennel structure and surrounding area, while GMAR-L1 and GMAR-L2 concentrate somewhat better but still lack the spatial sharpness of LeGrad. The Boathouse image illustrates a similar trend, where LeGrad highlights the boat structure and its surroundings with high confidence and clear separation from the water, whereas GMAR variants activate scattered patches across the water body, and Rollout's attention spreads uniformly, missing the core object of interest.

In the Sorrel (horse) image, LeGrad again succeeds in identifying the most salient features, such as the horse's face and upper torso, reflecting strong alignment with human perception. Rollout fails to offer clear boundaries, and GMAR-L2 focuses on multiple redundant regions, leading to interpretability ambiguity. The Couch example further emphasizes these disparities: while Rollout, GMAR-L1, and GMAR-L2 tend to dilute the focus and sometimes misplace the regions of importance toward background elements or edges, LeGrad sharply centers the attention on the sofa itself, revealing its capability to differentiate objects from the scene effectively. Finally, in the Bittern example, which includes a relatively small bird flying over a uniform background, LeGrad excels in capturing the bird's exact shape and position with minimal background interference. In comparison, both GMAR variants exhibit noisy activations spread throughout the image, and Rollout again distributes attention in a broad, unfocused manner.

This consistent pattern across all test cases suggests that LeGrad is significantly more effective at isolating and visualizing class-discriminative features than its counterparts. While Rollout ensures full attention propagation through layers, it lacks the ability to

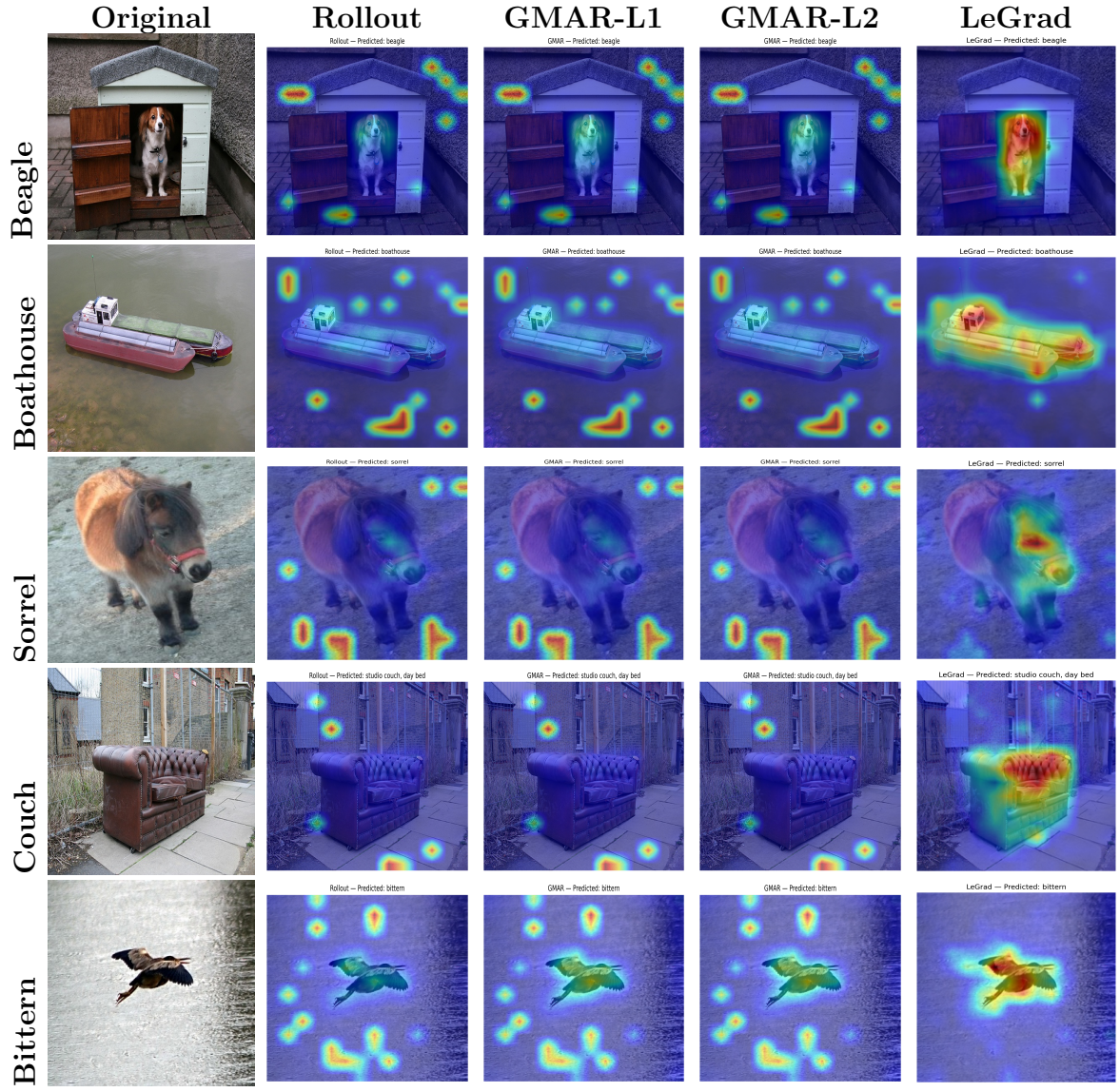

Figure 5: Qualitative comparison of heatmaps for different explainability methods across multiple examples.

refine spatial focus, resulting in overly generalized heatmaps. GMAR improves upon this by weighting important heads, but the quality of the output still suffers due to potential overemphasis on irrelevant patterns or noisy attention paths. LeGrad, by contrast, benefits from gradient-based weighting of attention heads and positive relevance aggregation, enabling both fine-grained localization and strong semantic alignment. These qualitative insights align well with the quantitative metrics, where LeGrad outperformed all other methods in terms of pixel accuracy, mean Intersection over Union (mIoU), and mean Average Precision (mAP), thus validating its superiority as an attention-based interpretability method for Vision Transformers.

5 Conclusion

Overall, this study provides a comprehensive comparison of attention-based interpretability methods for Vision Transformers, focusing on Classical Attention Rollout, GMAR (L1 and L2 variants), and LeGrad. By combining qualitative visualization with quantitative evaluation metrics such as Pixel Accuracy, mIoU, and mAP, I observed that LeGrad consistently yields the most accurate, focused, and class-specific heatmaps. Classical Rollout, while computationally efficient, lacks class-awareness and often generates diffuse explanations. GMAR enhances interpretability through gradient-weighted head importance, offering improved clarity compared to Rollout, yet still falls short in spatial precision. LeGrad stands out by leveraging gradient signals across layers to produce detailed and faithful explanations of the model's predictions. These results underscore the effectiveness of gradient-informed, layer-aware methods in providing reliable interpretability for Vision Transformers and highlight the importance of selecting appropriate explainability techniques for model transparency in real-world applications.

References

[1] Samira Abnar and Willem Zuidema. "Quantifying attention flow in transformers". In: *arXiv preprint arXiv:2005.00928* (2020).

[2] Walid Bousselham et al. "LeGrad: An Explainability Method for Vision Transformers via Feature Formation Sensitivity". In: *arXiv preprint arXiv:2404.03214* (2024).

[3] Hila Chefer, Shir Gur, and Lior Wolf. "Transformer interpretability beyond attention visualization". In: *arXiv preprint arXiv:2012.09838* (2021).

[4] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[5] Sehyeong Jo, Gangjae Jang, and Haesol Park. "GMAR: Gradient-Driven Multi-Head Attention Rollout for Vision Transformer Interpretability". In: *arXiv preprint arXiv:2504.19414* (2024).