Universidade Federal de Pernambuco

Centro de Informática

Graduação em Sistemas de Informação

QUANTIFYING EXPLANATION SENSITIVITY IN CLIP: A Benchmark of Interpretability Robustness under Perturbations

Trabalho de Conclusão de Curso de Graduação

por

Diego Henrique Vilaça Calixto

Orientador: Prof. Flávio Arthur Oliveira Santos

Recife, Setembro de 2025

Diego Henrique Vilaça Calixto

QUANTIFYING EXPLANATION SENSITIVITY IN CLIP:

A Benchmark of Interpretability Robustness under Perturbations

Monografia apresentada ao Graduação em Sistemas de Informação, como requisito parcial para a obtenção do Título de Bacharel em Sistemas de Informação, Centro de Informática da Universidade Federal de Pernambuco.

Orientador: Prof. Flávio Arthur Oliveira Santos

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Calixto, Diego Henrique Vilaça.
Quantifying explanation sensitivity in CLIP: a benchmark of interpretability robustness under perturbations / Diego Henrique Vilaça Calixto. - Recife, 2025.
p. 49 : il.

Orientador(a): Flávio Arthur Oliveira Santos
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2025.
Inclui referências.

1. IA Explicável. 2. Modelos de Linguagem Visual. 3. CLIP. 4. Interpretabilidade. 5. Robustez. I. Santos, Flávio Arthur Oliveira. (Orientação). II. Título.

000 CDD (22.ed.)

Diego Henrique Vilaça Calixto

QUANTIFYING EXPLANATION SENSITIVITY IN CLIP:

A Benchmark of Interpretability Robustness under Perturbations

Trabalho de Conclusão de Curso apresentado ao Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE), como requisito parcial para a obtenção do título de bacharel em Sistemas de Informação.

Aprovado em: 05 de Agosto de 2025

Banca Examinadora:

Prof. Dr. Flávio Arthur Oliveira Santos (Orientador)
Universidade Federal de Pernambuco

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio (Examinador interno)
Universidade Federal de Pernambuco

Recife

2025

Agradecimentos

Agradeço profundamente à minha família, que serviu como base para toda a minha trajetória e me fez ser quem sou.

Aos meus amigos, que desde o começo compartilharam essa jornada comigo, tornando os desafios mais leves e as conquistas mais celebradas.

À minha metade da laranja, Júlia, que, literalmente, em todos os momentos esteve ao meu lado. Palavras não conseguem expressar minha gratidão por ter você comigo.

A todos os professores que cruzaram minha trajetória no CIn. Especialmente ao meu orientador, que confiou no meu potencial e dedicou seu tempo e conhecimento para me guiar, tornando este trabalho possível.

A todos vocês, minha sincera gratidão.

"I heard this story about a fish. He swims up to this older fish and says, "I'm trying to find this thing they call the ocean." "The ocean?" says the older fish. "That's what you're in right now." "This?" says the young fish. "This is water. What I want is the ocean."

Soul

RESUMO

Modelos de Linguagem Visual (VLMs), como o CLIP, são amplamente utilizados em aplicações de inteligência artificial multimodal por integrarem representações de imagens e textos por meio de codificadores de diferentes arquiteturas. Apesar do desempenho robusto, a complexidade desses modelos impõe desafios à compreensão e confiabilidade de suas decisões. Pensando em mitigar esse aspecto, métodos de interpretabilidade têm sido desenvolvidos para poder extrair mapas de atribuição visando entender quais regiões da imagem o modelo está utilizando para inferência. Entretanto, muitos desses métodos de interpretabilidade alteram seu resultado diante de uma pequena alteração na imagem de entrada. Este trabalho se propõe a investigar a robustez de métodos de interpretabilidade aplicados ao CLIP, com ênfase na sensibilidade dessas técnicas a pequenas perturbações nas entradas, aspecto que pode comprometer a confiabilidade das explicações geradas. Para isso, foi proposto um pipeline de avaliação baseado em perturbações controladas, além de um conjunto de métricas que inclui correlação de postos de Spearman, Índice de Similaridade Estrutural (SSIM), Interseção Top-K e Diferença de Similaridade. Foram avaliados nove métodos de interpretabilidade, observando-se variabilidade significativa em termos de estabilidade. Técnicas de interpretação como Grad-ECLIP e CLIP Surgery apresentaram maior robustez e coerência semântica frente às perturbações, enquanto abordagens como RISE e Self-Attention demonstraram instabilidade considerável. Os resultados indicam a importância de se considerar não apenas a capacidade informativa das explicações, mas também sua robustez em diferentes condições.

Palavras-chave: IA Explicável, Modelos de Linguagem Visual, CLIP, Interpretabilidade, Robustez

ABSTRACT

Visual Language Models (VLMs), such as CLIP, are widely used in multimodal artificial intelligence applications due to their ability to integrate image and text representations through encoders with different architectures. Despite their strong performance, the complexity of these models presents challenges to understanding and trusting their decisions. To mitigate this issue, interpretability methods have been developed to extract attribution maps in order to understand which regions of the image the model is using for inference. However, many of these interpretability methods produce different results when the input image is slightly altered. This work aims to investigate the robustness of interpretability methods applied to CLIP, with an emphasis on the sensitivity of these techniques to small input perturbations, an aspect that can undermine the reliability of the generated explanations. To this end, an evaluation pipeline based on controlled perturbations was proposed, along with a set of metrics including Spearman's rank correlation, Structural Similarity Index (SSIM), and Top-K Intersection. Nine interpretability methods were evaluated, revealing significant variability in terms of stability. Interpretation techniques such as Grad-ECLIP and CLIP Surgery showed greater robustness and semantic coherence in the face of perturbations, while approaches like RISE and Self-Attention demonstrated considerable instability. The results highlight the importance of considering not only the informativeness of the explanations, but also their robustness under different conditions.

Keywords: Explainable AI, Visual Language Models, CLIP, Interpretability, Robustness

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUCTION

Artificial Intelligence (AI) [3] has become one of the most impactful technological domains in recent decades, driven by advances in deep learning [4], which have enabled the overcoming of previous limitations by leveraging large volumes of data and computational resources at scale . Among the most significant contributions is the role of deep neural networks, especially convolutional neural networks (CNNs) [5], which have transformed the field of computer vision by enabling the automation of tasks such as image classification, object detection, and semantic segmentation, often achieving performance levels that rival or even surpass humans in various domains [6–8].

More recently, the integration of distinct modalities, such as natural language and computer vision, has led to the development of Visual Language Models (VLMs) [9], such as CLIP (Contrastive Language–Image Pre-training) [1], a model that consolidated the use of contrastive techniques by relating image and text information. These models combine the contextual and visual representation, enabling multimodal processing of images and text in tasks such as zero-shot learning [10], semantic retrieval, and visual command interpretation [1, 11]. This integration not only broadens the application scope of AI but also opens up new possibilities in interactive systems, content recommendation, and creative tools based on language and vision.

However, the increasing capability of these models brings a new challenge: understanding and trusting their decisions. In sensitive applications, such as medical diagnosis [12, 13] or judicial [14, 15] and financial systems [16], the need to justify the decisions made by automated systems is as critical as the model's own accuracy [17]. In this context, interpretability has emerged as an essential requirement to ensure transparency, auditability, and trust, allowing users and experts to understand which aspects of the inputs most influence the predictions [18].

Although several interpretability methods have been proposed for deep neural networks, recent studies have raised concerns about the fragility of these techniques, showing that the explanations generated can be highly sensitive to small variations in the inputs or the configuration of the interpretability algorithms [19]. This instability not only undermines trust in the explanations but also raises doubts on the validity of interpretive methods, since trivial differences can lead to conflicting interpretations, even when the model's prediction remains unchanged.

Despite growing attention to the robustness of interpretability in CNNs and transformers applied to unimodal domains (vision or language) [8, 20], the literature lacks systematic studies on the sensitivity of interpretations in multimodal models, such as CLIP. Given the widespread adoption of these models in high-impact scenarios, this gap represents both a practical and scientific risk that must be addressed.

1.1 Objectives

In this context, this work aims to investigate the sensitivity of interpretability methods when applied to the CLIP model, a representative vision-language transformer. The study focuses on evaluating how small perturbations in the input data impact the consistency and reliability of the resulting explanations. By conducting a systematic experimental analysis, the objective is to uncover the current limitations of interpretability techniques in VLMs and to offer empirical insights that inform the development of more stable and trustworthy approaches. Ultimately, the findings seek to contribute to the broader goal of ensuring safer and more transparent deployment of multimodal models in real-world scenarios.

2 STATE OF THE ART

This chapter aims to provide an overview of the relevant literature that forms the foundation for this work. It is organized into three main sections. **Section 2.1** introduces Visual Language Models (VLMs), with a particular emphasis on CLIP, outlining their architectures, training paradigms, and recent developments in the field. **Section 2.2** reviews the landscape of interpretability methods in artificial intelligence, focusing on techniques commonly applied to deep neural networks and multimodal models. Finally, **Section 2.3** presents recent findings on the sensitivity of interpretation methods, highlighting the challenges related to their stability and robustness when applied to models like CLIP. Together, these sections establish the theoretical and empirical background necessary for understanding the motivations and contributions of this work.

2.1 Visual Language Models

The intersection between computer vision and natural language processing has given rise to a new class of machine learning systems known as Visual Language Models (VLMs) [9]. Unlike traditional systems that operate on either images or text in isolation, these models are designed to jointly process and integrate information from both visual and textual modalities. VLMs enable cross-modal reasoning, making it possible for machines to understand and generate connections between what they see and what they read. [1, 9] This opens the door to a variety of tasks such as image captioning [21], visual question answering [22], semantic retrieval [23], and zero-shot classification [10].

The growing interest in VLMs reflects their potential to serve as general-purpose models in a wide range of domains [9]. Their capacity to interpret images in the context of natural language makes them especially useful in applications where labeled data is limited or impractical to obtain [24]. More importantly, the ability to perform zero-shot learning, where the model can classify or describe unseen data without task-specific fine-tuning [10], makes VLMs a powerful solution for scalable and adaptable AI systems.

Enabled by the large data available on the internet, authors began to train a visual language multi-modal system to predict captions on images, first

One of the most influential VLMs to date is CLIP (Contrastive Language–Image Pretraining), introduced by Radford et al. [1]. CLIP departs from traditional supervised

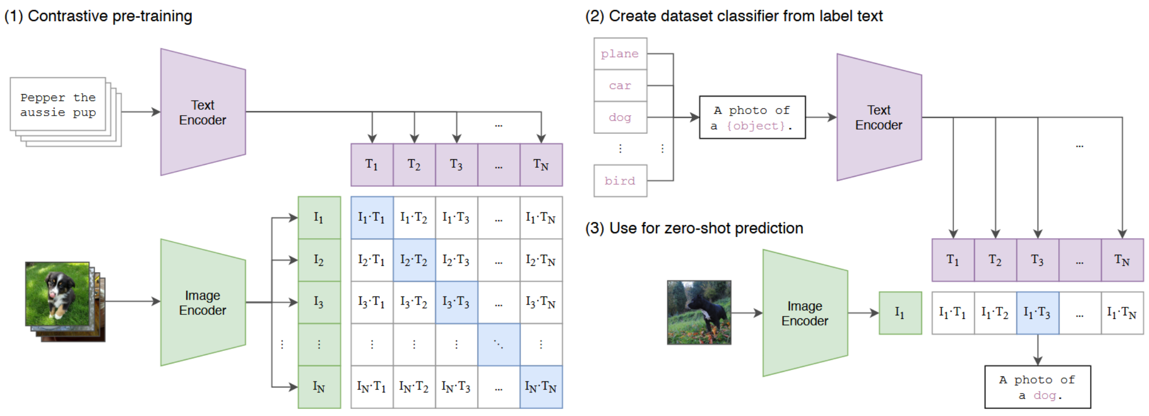
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1: **Summary of CLIP approach to a multi model caption predictor.** For instance, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training example. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. Figure reproduced from [1].

learning paradigms by leveraging a dataset of 400 million image-text pairs collected from the web. Rather than being trained to recognize a fixed set of labels, CLIP learns to align visual and textual information in a shared embedding space. This strategy allows the model to generalize effectively to a wide variety of downstream tasks using only natural language prompts, without requiring task-specific retraining [1].

At the core of CLIP's design are two separate but jointly trained components: an image encoder and a text encoder. The image encoder processes input images to generate visual embeddings, while the text encoder transforms textual descriptions into corresponding language embeddings. These embeddings are projected into a common multimodal space, where similarity is computed, typically using cosine similarity. During inference, CLIP determines the relationship between an image and a set of candidate texts by measuring their proximity in this shared space.

The image encoder used in CLIP is most commonly based on the Vision Transformer (ViT) architecture. CNN approaches were used initially but encountered difficulties efficiently scaling this method. Unlike CNNs, which rely on spatial hierarchies, ViTs divide the input image into fixed-size patches and treat each patch as a token in a sequence. These tokens are embedded and passed through a series of self-attention layers, allowing the model to capture long-range dependencies and contextual relationships across the entire image [25]. This design not only offers flexibility in representing complex visual features but also makes it possible for CLIP to improve training efficiency, scaling to bigger datasets.

On the textual side, CLIP employs a transformer model similar to those used by Vaswani et al. [26] in LLMs. The input text is tokenized and passed through multiple layers of self-attention and feed-forward networks. The model outputs a dense vector representation of the sentence that summarizes the semantic content of the sequence. This representation is then projected into the shared embedding space, making it directly comparable to the visual embedding of an image [1].

A key innovation of CLIP is its training objective, which relies on contrastive learning. Given a batch of image-text pairs, the model is trained to maximize the similarity between the correct pairs and minimize it for all others. This is formalized using a contrastive loss function based on the InfoNCE objective [27]. By training in this way, CLIP effectively learns a bidirectional alignment between vision and language, enabling robust zero-shot generalization across a wide range of datasets and tasks.

Since the introduction of CLIP, Visual Language Models (VLMs) have made significant progress in various aspects. These advancements can be divided into three main areas: (1) Pre-training objectives have evolved from using a single contrastive objective to integrating multiple hybrid objectives. Early VLMs like CLIP relied primarily on contrastive learning to align images and text, but recent models such as FLAVA [28] and FIP [29] combine contrastive, alignment, and generative objectives to improve performance and robustness in downstream tasks. (2) Pre-training frameworks have also advanced, with early models using two-tower architectures for image and text processing. More recent approaches, such as Single-Tower Transformers [30], now use unified networks, reducing GPU memory usage and improving the efficiency of communication between modalities. (3) Downstream tasks have shifted from simple image-level recognition to more complex tasks such as object detection and semantic segmentation. Models like DETECTCLIP [31] and SEGCLIP [32] demonstrate CLIP's ability to handle dense prediction tasks that require understanding spatial relationships and fine-grained details in images.

These developments highlight how recent VLMs have expanded upon CLIP's framework, improving their flexibility and performance, especially in more complex and diverse real-world applications. In addition, the introduction of SigLIP 2 [33] builds upon the success of CLIP by incorporating advancements such as multilingual training, enhanced localization, and dense feature extraction, providing improvements in both vision-language understanding and dense prediction tasks.

Despite the emergence of several newer VLMs, CLIP remains one of the most influential and widely adopted models in the field, largely due to its strong performance in zero-shot tasks across a variety of downstream applications. As highlighted in recent surveys [9], CLIP has set the benchmark for vision-language understanding and continues to be a foundational model for both academic research and industrial applications. This enduring relevance and its ability to generalize across diverse tasks make CLIP the ideal model for our experiment, allowing us to assess the sensitivity and robustness of interpretability methods within a well-established framework.

2.2 Interpretability in artificial intelligence

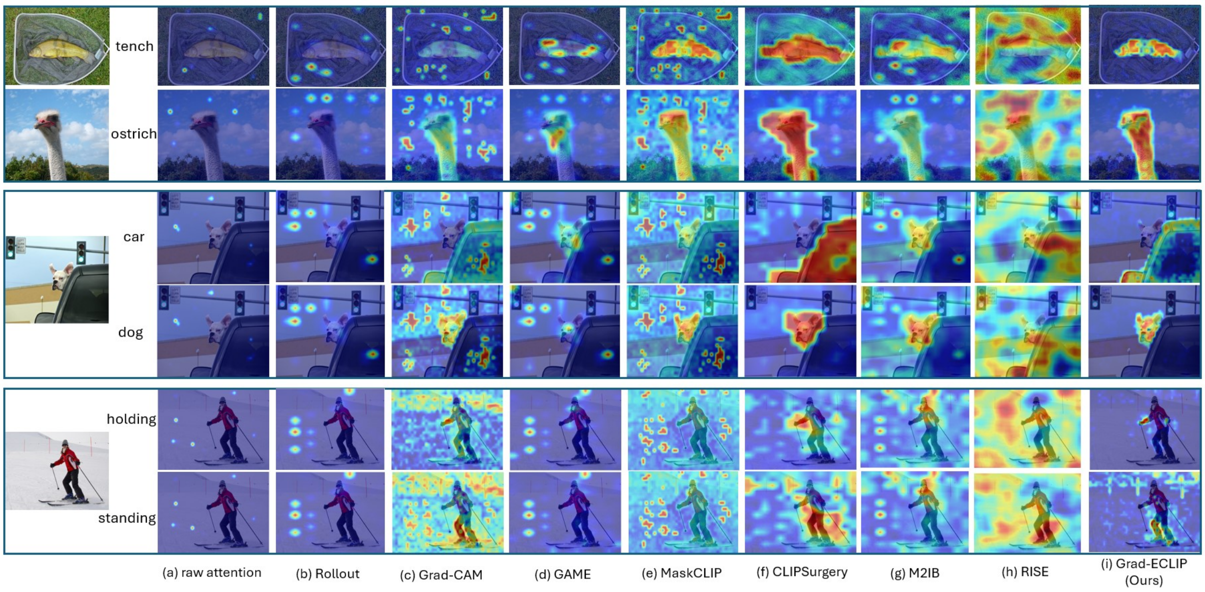

Figure 2: **Comparison of heat maps from different visual explanations.** They are provided for the matching score between the image and the specific text prompts, which can be nouns (e.g., car, dog) or verbs (e.g., holding, standing). Figure reproduced from [2]

Interpretability in artificial intelligence (AI) refers to the ability to understand and explain the decision-making processes of a model. It is a critical aspect, particularly in the context of deep learning models, which often operate as ”black boxes” [34]. These models, including convolutional neural networks (CNNs), transformers, and multimodal models, are typically too complex for humans to intuitively grasp how decisions are made.

Interpretability ensures that AI systems can be trusted, and it enables practitioners to identify errors, understand limitations, and increase user confidence [17].

For complex AI systems, such as CLIP, interpretability plays a pivotal role in ensuring the reliability and safety of their deployment, especially in high-stakes applications like healthcare, finance, and law, where incorrect decisions can have severe consequences [17]. By making model decisions understandable and transparent, interpretability methods help build trust among users and stakeholders, facilitate compliance with regulations, and improve the reliability and fairness of AI-driven decisions [34, 35].

Interpretability methods can be categorized based on the scope of their explanations: global explanations aim to clarify the model's overall behavior, while local explanations focus on individual predictions [36]. Among local methods, CAM-derived techniques, including CAM [37], Grad-CAM [38], and Grad-CAM++ [39] generate saliency maps by linearly combining activation maps from intermediate layers, using feature importance scores as weights. In particular, Grad-CAM computes these weights via global average pooling of the gradients flowing from the prediction layer. In contrast, perturbation-based approaches, like RISE [40] and LIME [35], estimate feature relevance by systematically modifying input regions and observing the resulting changes in model output, offering an architecture-agnostic yet often more computationally intensive alternative.

While intuitive and architecture-agnostic, these methods exhibit high computational overhead and sensitivity to perturbation design. Shapley-value approaches [41] are grounded in cooperative game theory and provide theoretically rigorous attribution but face scalability constraints in large-scale vision-language models. Similarly, attribution propagation methods, such as Layer-wise Relevance Propagation (LRP) [42], decompose predictions recursively based on Deep Taylor Decomposition principles, propagating relevance from outputs to inputs, though they too encounter scalability challenges in complex models.

The emergence of transformer architectures has led to specialized interpretability techniques that leverage self-attention mechanisms, including attention rollout [43] and gradient-based adaptations such as Transformer interpretability [44] and GAME [45]. However, when applied to multimodal transformers like CLIP, these methods often fall short due to the sparse softmax attention patterns, resulting in fragmented or misleading explanations [1]. Moreover, CLIP-specific interpretability methods face significant limitations: cosine similarity maps between localized image features and text embeddings often reflect bottom-up feature alignment without verifying their actual contributions to the final predictions, while information bottleneck approaches, such as M2IB [46], rely heavily on hyperparameter tuning, hindering their practical deployment.

To address these challenges, gradient-driven methods like Grad-ECLIP [2] offer a principled solution by using top-down gradient attribution to identify the features that have the most influence on predictions, thereby bypassing the need for attention maps or cosine similarity. Similarly, the CLIP Surgery [47] approach aligns with this methodology by providing an architecture modification that addresses the inconsistency in self-attention and eliminates redundant features, further improving CLIP's explainability without fine-tuning the model. While these methods offer promising directions, their effectiveness depends not only on the accuracy of their explanations but also on their robustness to variations in input and model configurations—an aspect increasingly recognized as critical in recent studies and further discussed in the next section.

2.3 Sensitivity of interpretability methods

As interpretability methods become more integrated into high-stakes AI applications, a growing body of research has exposed a concerning limitation: their sensitivity to small perturbations in input data or model settings [19, 48]. This phenomenon, known as explanation sensitivity, refers to the degree to which an interpretability method's output changes in response to minor, non-adversarial perturbations, assuming the predicted label remains constant.

Such instability can lead to significant variations in the generated explanations, even when the model's prediction remains unchanged, raising doubts about the reliability and stability of these techniques. Consequently, interpretability must be evaluated not only in terms of informativeness but also in terms of robustness, especially when explanations are used to support decisions in domains such as medicine, finance, and autonomous systems [17].

In the field of deep learning, several studies have highlighted the fragility of interpretation methods. Ghorbani et al. [19] demonstrated that for conventional deep neural networks (DNNs), small adversarial perturbations to the input could drastically alter the attribution maps, even when the model's output label remained unchanged. For instance, in the case of methods like Grad-CAM and Integrated Gradients, adversarial noise can cause a significant shift in the saliency maps, undermining the reliability of the explanation. Similarly, Bansal et al. [49] pointed out that attribution methods are highly sensitive to hyperparameters such as the random seed, the number of perturbations, or the sample size. A change in these settings can lead to drastically different explanations, which not only complicates the reproducibility of experiments but also casts doubt on the correctness of the explanations.

Interestingly, most existing sensitivity studies focus on the models themselves and their robustness to perturbations, rather than on the interpretation methods. While many works examine how adversarial attacks affect model performance, fewer address how these attacks impact the interpretability of the models' decisions [50]. For example, [51] studied the robustness of CLIP models to visual shifts, such as changes in pose, texture, or lighting. They found that CLIP models are not uniformly robust across different visual factors, which directly affects the interpretability of the model. When applied to CLIP, traditional interpretability methods like saliency maps can produce inconsistent or misleading results, especially under perturbations. This suggests that while models like CLIP show strong zero-shot performance across various shifts, their interpretability is far more fragile and sensitive to input variations.

Several interpretability methods, such as LIME, SmoothGrad, and Meaningful Perturbation have been shown to be highly sensitive to hyperparameter choices like random seeds, iteration counts, and patch sizes—leading to inconsistent explanations across runs [49]. In multimodal models like CLIP, the problem is exacerbated by sparse and fragmented attention patterns, making explanations even more unstable [2]. Although methods like Grad-ECLIP and CLIP Surgery aim to improve attribution quality, their explanations still fluctuate under minor input changes or model variations. While adversarial training improves the robustness of predictions and partially stabilizes explanations, ensuring robust interpretability remains an open challenge.

Moving forward, more attention must be given to understanding and mitigating the fragility of interpretation techniques, especially in multimodal models like CLIP. Developing sensitivity benchmarks and introducing methods that are less sensitive to input changes and hyperparameter variations will be key in advancing the trustworthiness of interpretability in deep learning systems.

In summary, while there has been significant progress in developing interpretability methods for deep learning models, the sensitivity of these techniques remains a critical challenge. The sensitivity of CLIP and similar multimodal models presents unique challenges for producing robust, reliable, and consistent explanations. Addressing these issues is crucial for the deployment of AI systems in real-world applications where trust and safety are foremost.

3 SENSITIVITY EVALUATION PIPELINE FOR CLIP INTERPRETABILITY

This section describes the experimental pipeline designed to assess the sensitivity of interpretation methods applied to CLIP-based models. The pipeline, shown in the Figure 3, evaluates how consistent each explanation method remains under small perturbations, using a multimodal dataset, a diverse set of interpretability techniques, and multiple similarity metrics.

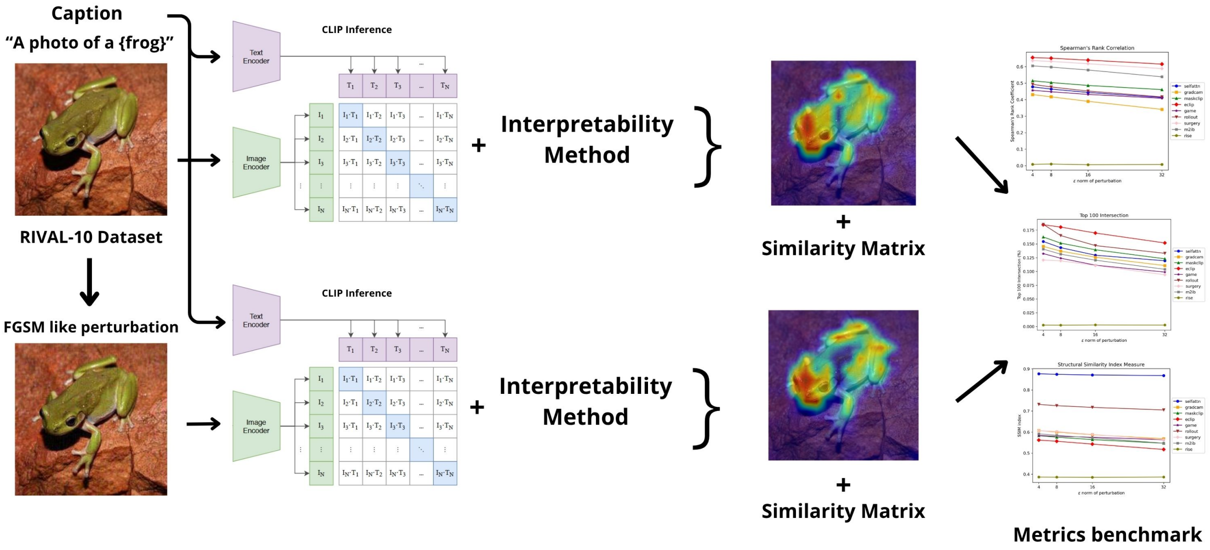

Figure 3: **Visualization of the method pipeline.** The workflow begins with a pair of original and perturbed images from the RIVAL-10 dataset being input into CLIP along with the explanation method. This generates a heatmap and a similarity matrix, which are then used to compute metrics for comparing different methods.

A subset of the RIVAL10 dataset [52] was selected for the experiments, as it provides a rich multimodal benchmark specifically designed to support model interpretability and robustness analysis. RIVAL10 consists of over 26,000 high-resolution images derived from 20 ImageNet-1k [53] classes, organized into 10 semantically distinct categories aligned with CIFAR-10 [54] (e.g., bird, dog, truck, ship). A representative subset was chosen, consisting of 5,285 images of this dataset, to maintain computational tractability while preserving visual diversity across classes and attribute types. This subset enables probing the sensitivity of CLIP explanations with respect to meaningful visual concepts across multiple attribution methods. Furthermore, since the dataset includes paired image and attribute-level localization, it allows for more precise evaluation of explanation consistency under perturbed inputs, serving as an ideal benchmark for vision-language model interpretability.

To evaluate the sensitivity of interpretability methods for CLIP, nine techniques that reflect a variety of attribution strategies and theoretical motivations were selected.

Grad-CAM [38] generates visual explanations by combining feature maps from intermediate layers with the gradients of the output, highlighting regions deemed important for the prediction. While originally designed for CNNs, its application to CLIP's ViT-based architecture often results in spatially imprecise and noisy heatmaps.

GAME [45] extends relevance propagation to Transformer-based models by integrating gradients across all attention heads and layers. It employs Layer-wise Relevance Propagation (LRP) to trace the influence of input tokens through the model, offering a general framework for interpreting attention-based architectures.

Grad-ECLIP [2], developed specifically for CLIP, decomposes the image-text similarity score and backpropagates it through spatial and channel-wise features. It mitigates the sparsity and instability of standard attention maps by introducing a loosened attention mechanism, yielding high-resolution and semantically coherent saliency maps.

Self-Attention explanations visualize raw attention weights from the Transformer encoder, assuming that these reflect the model's focus. However, attention patterns in deeper layers tend to become diffuse or uniform [43], limiting their interpretability.

Attention Rollout [43] attempts to overcome this by aggregating attention weights across layers, estimating the influence of input tokens on the output. While it improves alignment with early layer signals, it cannot distinguish between positive and negative contributions, potentially producing ambiguous maps.

CLIP Surgery [47] modifies the inference process of CLIP to enhance attribution stability. It introduces consistent self-attention mechanisms and a dual-path structure that reduces noisy activations and improves interpretability without retraining the model.

RISE [40] generates saliency maps by applying random binary masks to the input and aggregating the model's outputs. As a black-box method, it is architecture-agnostic but computationally expensive and prone to producing noisy or unstable heatmaps under small perturbations.

M2IB [46] applies the information bottleneck principle to multimodal attribution. It learns latent representations that preserve only the features most relevant to the image-text alignment, enabling semantically meaningful explanations. However, it is highly sensitive to hyperparameter settings and optimization heuristics.

MaskCLIP [23] extracts dense patch-level features from CLIP's image encoder and applies the text embeddings as fixed classifiers for pixel-wise predictions. By directly leveraging internal feature representations, it enables interpretable outputs without altering model weights or requiring fine-tuning.

These methods were selected for their diversity in approach and their prominence in recent literature, in addition to the available implementation on the GRAD-ECLIP [2] repository, enabling a systematic comparison of their behavior under controlled perturbations.

To evaluate the sensitivity and stability of the interpretability methods under controlled perturbations, we employed a combination of similarity metrics commonly used in the literature. Following the methodology introduced by Ghorbani et al. [19], we used Spearman's rank correlation and Top-K Intersection to assess changes in the relative and absolute importance of input features before and after perturbation.

Spearman's Rank Correlation (ρ) measures the monotonic relationship between two ranked variables, reflecting the consistency in the ordering of attribution scores. Given two attribution vectors $\mathbf{x}$ and $\mathbf{y}$ with n elements, their ranks are denoted by $R(\mathbf{x}_i)$ and $R(\mathbf{y}_i)$. The correlation is computed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \left(R(\mathbf{x}_i) - R(\mathbf{y}_i)\right)^2}{n(n^2 - 1)} \tag{3.1}$$

This metric ranges from -1 to 1. A value of $\rho = 1$ indicates perfect agreement in ranking (high stability), $\rho = 0$ denotes no correlation, and $\rho = -1$ indicates perfect inverse ranking (complete instability).

Top-K Intersection quantifies the proportion of overlap between the K most salient features before and after perturbation. Given the sets of top-K indices $\mathcal{T}_x$ and $\mathcal{T}_y$ from the original and perturbed attribution maps respectively, the Top-K Intersection is defined as:

$$\text{TopK} = \frac{|\mathcal{T}_x \cap \mathcal{T}_y|}{K} \tag{3.2}$$

This metric ranges from 0 (no overlap) to 1 (identical top-K features), directly reflecting the stability of the most relevant input regions.

Additionally, motivated by the analysis in Bansal et al. [49], we included the Structural Similarity Index (SSIM) to capture perceptual and spatial inconsistencies in the attribution maps, particularly under hyperparameter variations or input noise. SSIM compares local patterns of pixel intensities between two images $\mathbf{x}$ and $\mathbf{y}$ and is computed as:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.3}$$

where μ_x, μ_y are the means, σ_x^2, σ_y^2 the variances, and σ_{xy} the covariance between $\mathbf{x}$ and $\mathbf{y}$; C_1 and C_2 are constants to stabilize the division. SSIM values range from -1 to 1, with 1 indicating perfect structural similarity.

Finally, we computed the Cosine Similarity Difference, which captures the variation in semantic alignment between the image and text caused by perturbations. Given a similarity score s_{orig} for the original input and s_{pert} for the perturbed input, the metric is defined as:

$$\Delta_s = \frac{s_{\text{pert}} - s_{\text{orig}}}{s_{\text{orig}}} \tag{3.4}$$

This relative change quantifies how much the matching strength between the image and its associated text shifts due to perturbation. Higher absolute values of Δ_s indicate greater semantic instability, while values close to zero suggest high robustness of the model's multimodal representation.

Together, these four metrics provide a robust and comprehensive framework for quantifying the sensitivity of interpretability methods in vision-language models, capturing ranking stability, feature overlap, spatial coherence, and semantic alignment consistency.

4 EXPERIMENTS AND RESULTS

This section will introduce the experiments setup that was conducted to investigate the state-of-the-art interpretability methods to understand CLIP decisions, followed by the results of each metric analysis. The aim is to evaluate the visual explanation quantitatively by providing a pair of an original picture and one attacked by FGSM-like perturbation.

4.1 Experiment setup

The repository Grad-ECLIP, developed by Zhao et al. [2], was the starting point to generate explanations for the CLIP model. This choice was made because the repository provided the necessary tools for computing the gradient maps. It is important to mention that the CLIP model used was the ”ViT-B/16”.

The image input perturbation was inspired by the Fast Gradient Sign Method (FGSM) [55] implementation by the torchattacks library[1]. While FGSM focuses on generating adversarial examples to maximize the model's classification loss, the perturb function on this work pipeline is tailored for multimodal models like CLIP. Our perturbation function computes the gradient of the image-text similarity score and modifies the image in the direction that increases this similarity. This design choice allows us to introduce small, targeted changes that do not flip the model's prediction, but can still significantly affect the generated explanations. The goal is to assess the sensitivity and robustness of interpretability methods under subtle, semantically consistent perturbations.

The image input perturbation was inspired by the Fast Gradient Sign Method (FGSM) [55] implementation from the torchattacks library. While FGSM traditionally generates adversarial examples by maximizing the classification loss, in this work we adapt its principle to multimodal models like CLIP. Our perturbation function computes the gradient of the image-text similarity score and modifies the image in the direction that increases this similarity. This design choice allows us to introduce small, targeted changes that do not flip the model's prediction, but can still significantly affect the generated explanations. The goal is to assess the sensitivity and robustness of interpretability methods

[1] https://adversarial-attacks-pytorch.readthedocs.io/en/latest/_modules/torchattacks/attacks/fgsm.html#FGSM

under subtle, semantically consistent perturbations. Future work may also explore perturbations that decrease the similarity to evaluate contrastive behavior.

The perturbation coefficient (ϵ) defines the degree to which the image is modified. Common values used in the literature are applied in this work, such as the default value of $\epsilon = 8/255$ in the torchattacks implementation. Four different values of ϵ (4/255, 8/255, 16/255, and 32/255) were used in the experiments to observe how heatmaps changed across different interpretation methods. By exploring the impact of various levels of adversarial perturbations on CLIP inputs, it is possible to identify which methods are more robust and what characteristics contribute to their resilience.

The results are organized by evaluation metric. **Section 4.2** presents the Spearman's rank correlation results, which measure the consistency of feature rankings under perturbation. **Section 4.3** describes the SSIM scores, evaluating the structural similarity of saliency maps. **Section 4.4** reports the Top-100 Intersection scores, focusing on the stability of the most important regions. **Section 4.5** analyzes the relationship between Spearman correlation and attribution similarity. **Section 4.6** discusses the relationship between SSIM and similarity difference. Finally, **Section 4.7** presents the relationship between Top-K intersection and similarity difference, providing further insight into the trade-offs between visual stability and semantic sensitivity.

4.2 Spearman's rank correlation per norm of perturbation

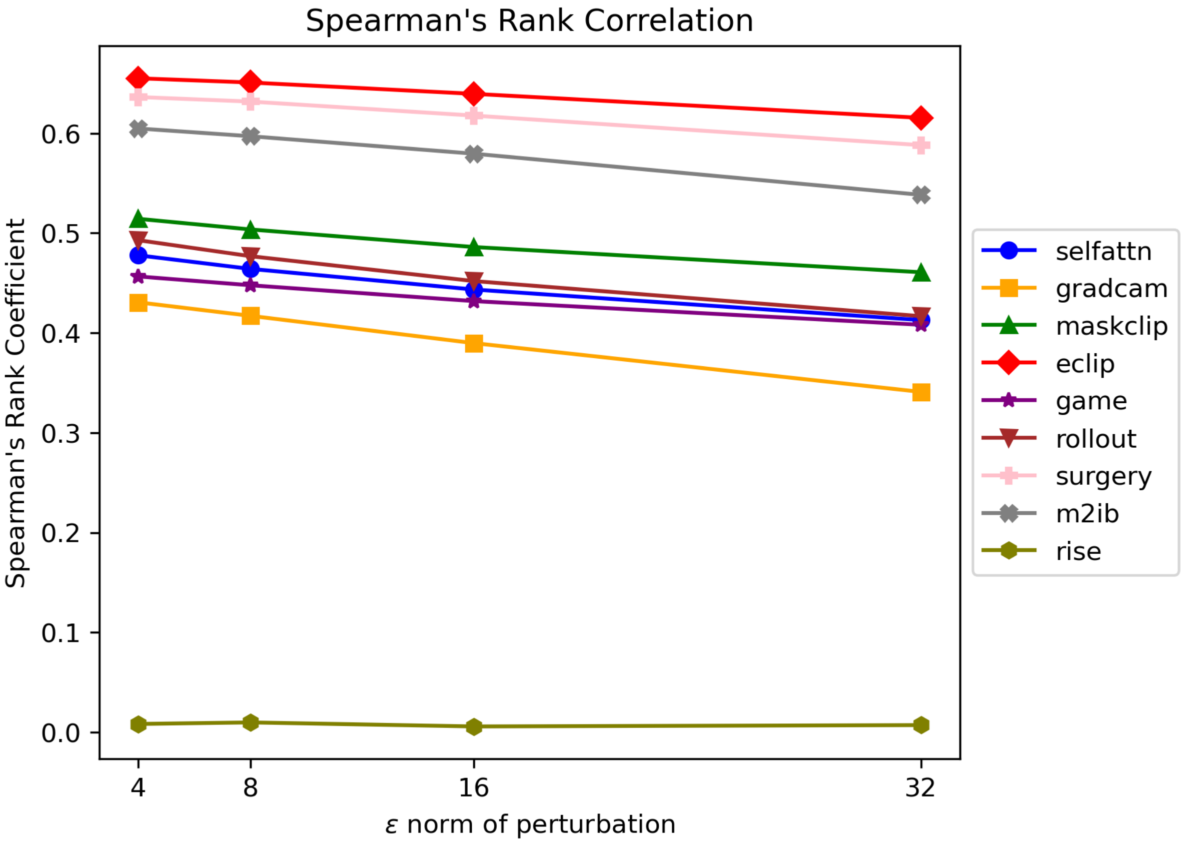

Figure 4: Interpretation methods spearman's rank correlation per norm of pertubation.

Spearman's rank correlation coefficient (ρ) quantifies the monotonic relationship between two ranked variables, providing a non-parametric measure of explanation stability under perturbations. Following Ghorbani et al. [19], we employed this correlation metric to evaluate how consistently interpretability methods rank salient features in CLIP's saliency maps when subjected to adversarial noise of varying magnitudes ($\epsilon = 4/255$, $8/255$, $16/255$, $32/255$). This metric is particularly suitable for three interrelated reasons.

First, its focus on ordinal consistency rather than linear relationships makes it ideal for comparing saliency maps, where the exact magnitude of activation values may vary significantly between methods, but the relative ranking of important regions carries meaningful interpretation. This property allows fair comparison across different explanation techniques that might use disparate activation scales. Second, the rank-based approach demonstrates inherent robustness to outliers in saliency map activations, a critical feature given that explanation methods frequently produce extreme values in isolated pixels due to artifacts in gradient computation or attention mechanisms. Third, and most fundamentally, Spearman correlation directly assesses whether perturbations preserve the hierarchical importance of image regions, which aligns with the cognitive principle that humans prioritize relative rather than absolute feature importance when interpreting visual explanations. This last characteristic proves especially valuable for CLIP's multimodal context, where text-guided attention often creates non-uniform importance distributions across image regions.

As shown in Figure 4, the *ECLIP*, *Surgery*, and *M2IB* methods displayed the highest robustness among all evaluated techniques. For example, *ECLIP* started with $\rho = 0.655$ at the lowest perturbation level and declined gradually to $\rho = 0.615$ at the highest. *CLIP Surgery* and *M2IB* followed a similar trend, with *Surgery* decreasing from 0.636 to 0.588 and *M2IB* from 0.605 to 0.538. These results suggest a moderate but consistent degradation in feature ranking, indicating that these methods are sensitive to perturbations while still retaining a relatively stable attribution hierarchy.

A group of methods including *Grad-CAM*, *GAME*, *Rollout*, *MaskCLIP*, and *Self-Attention* exhibited intermediate robustness. Their Spearman correlation scores declined gradually across perturbation levels, with variations in the range of approximately 0.05 to 0.09. This indicates moderate sensitivity to perturbations: while the ranking of salient regions is not fully preserved, it does not collapse entirely. These methods are thus partially robust, though not as stable as *ECLIP*, *Surgery*, or *M2IB*.

The *RISE* method demonstrated the lowest correlation across all levels, with ρ values consistently below 0.01. This implies that even small perturbations drastically alter the order of salient regions, rendering RISE highly unstable and unreliable in terms of ranking consistency.

Overall, the results suggest that these methods appear to retain some structural consistency in their explanations, yet still respond to input perturbations in ways that may affect interpretability trust. Notably, the extreme instability of *RISE*, with near-zero rank consistency, highlights the limitations of certain perturbation-based approaches in multimodal contexts.

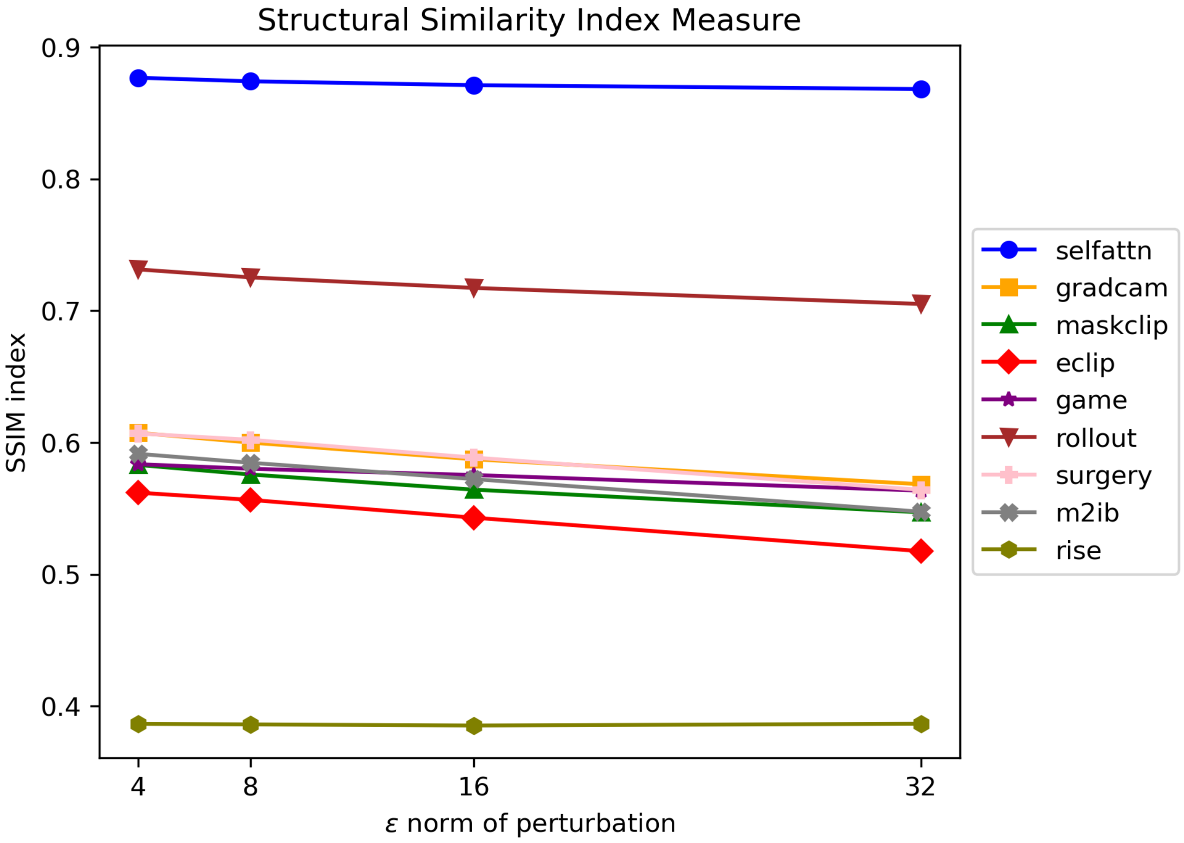

Figure 5: SSIM scores of interpretation methods under increasing perturbation norms.

4.3 SSIM score per norm of perturbation

The Structural Similarity Index Measure (SSIM) provides a perceptually grounded comparison between images, focusing on luminance, contrast, and structure. Unlike pixel-wise metrics, SSIM provides a more holistic and human-aligned assessment of visual changes. When applied to saliency maps, SSIM quantifies the degree to which the visual structure of an explanation is preserved after perturbation. This is especially important for interpretability in vision-language models, where the spatial coherence of highlighted regions influences human trust in the explanation.

Figure 5 illustrates how different interpretability methods behave under increasing perturbation levels ($\epsilon = 4/255$, $8/255$, $16/255$, and $32/255$). The *Self-Attention* method consistently achieved the highest SSIM scores among all techniques, starting at 0.877 and dropping slightly to 0.868. This indicates that its explanations are highly stable in terms of spatial layout. *Rollout* also demonstrated strong performance, with scores ranging from 0.731 to 0.705, suggesting that its visual outputs remain relatively consistent even under stronger perturbations. However, it is worth noting that high SSIM scores alone do not guarantee meaningful explanations; in some cases, these methods may simply produce static or low-resolution attribution maps that change little because they fail to capture input-specific information, rather than due to true robustness.

In contrast, methods like *Grad-CAM*, *GAME*, *Surgery*, *M2IB*, and *MaskCLIP* presented moderate SSIM values, generally fluctuating between 0.54 and 0.61. These results suggest partial robustness, where some structural degradation occurs, but not enough to severely distort the overall saliency map. *ECLIP*, on the other hand, showed slightly lower performance, with scores decreasing from 0.56 to 0.52, indicating comparatively higher sensitivity to perturbations in spatial structure.

Finally, *RISE* again exhibited the lowest SSIM values, consistently around 0.386 across all perturbation levels. This low score indicates a high degree of structural instability, confirming that its explanations are not spatially reliable under perturbations.

Overall, the SSIM analysis reveals that while some methods are able to preserve perceptual structure well, others degrade significantly, which can hinder their usefulness in contexts requiring stable and interpretable visual outputs. SSIM thus complements rank-based metrics by offering a perceptual perspective on the robustness of visual explanations.

4.4 Top-100 Intersection per norm of perturbation

The Top-K Intersection metric evaluates the overlap between the K most salient pixels or regions in the attribution maps before and after perturbation. In this experiment, it was fixed $K = 100$ to compare the top-100 most important pixels as ranked by each interpretability method. This metric directly reflects the stability of the most critical regions identified by the model and is particularly useful for assessing robustness in practical settings where explanations are typically visualized using thresholded heatmaps.

As shown in Figure 6, *ECLIP* achieved the most stable performance across perturbation levels, starting with a Top-K Intersection score of 0.185 at $\epsilon = 4/255$ and decreasing gradually to 0.152 at $\epsilon = 32/255$. *Rollout* followed closely, with scores ranging from 0.186 to 0.133. Notably, Rollout exhibited a drop of 0.053 in intersection score across the perturbation range, reflecting a moderate but steady decline in the consistency of high-importance regions.

MaskCLIP and *Self-Attention* also maintained relatively robust outputs, beginning at 0.163 and 0.154 respectively, and ending at 0.123 and 0.119. These results indicate that

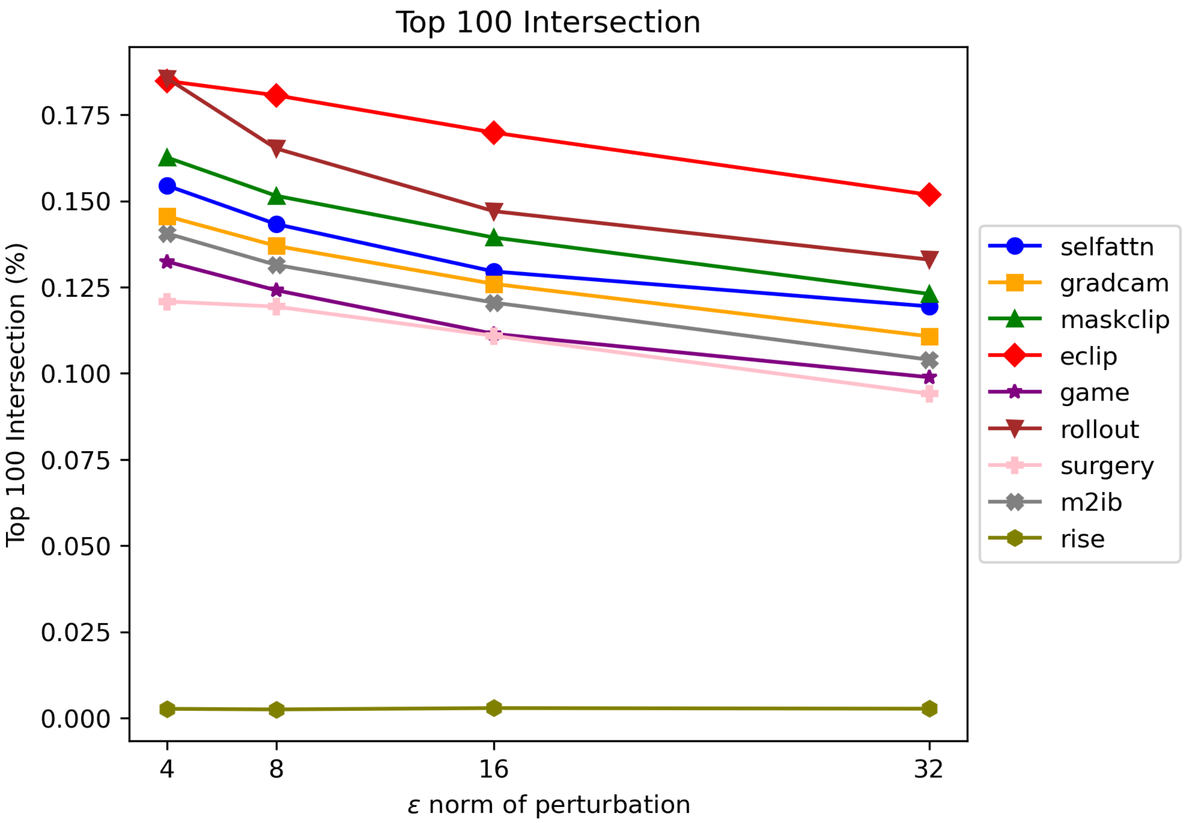

Figure 6: Top-100 intersection scores of interpretation methods under increasing perturbation norms.

both methods manage to retain a significant portion of their top-ranked features under noise, although slightly less stable than ECLIP or Rollout.

Grad-CAM, *GAME*, *M2IB*, and *Surgery* showed intermediate robustness. Their intersection scores started in the range of 0.13–0.14 and declined to values near 0.10 or below. These drops suggest moderate sensitivity, with some methods, like *Surgery* and *GAME*, exhibiting more pronounced degradation in attribution consistency.

RISE, once again, demonstrated the lowest performance. Its scores remained around 0.002 across all perturbation levels, indicating near-complete instability in preserving salient features. This reinforces previous findings that perturbation-based methods like RISE are particularly vulnerable in multimodal contexts.

In summary, Top-K Intersection analysis corroborates the results observed with Spearman and SSIM: gradient and attention based methods like *ECLIP* and *Rollout* tend to exhibit higher stability under perturbations, while methods such as *RISE* fail to retain consistency in highlighting important input regions.

4.5 Relationship between Spearman's rank correlation and similarity difference

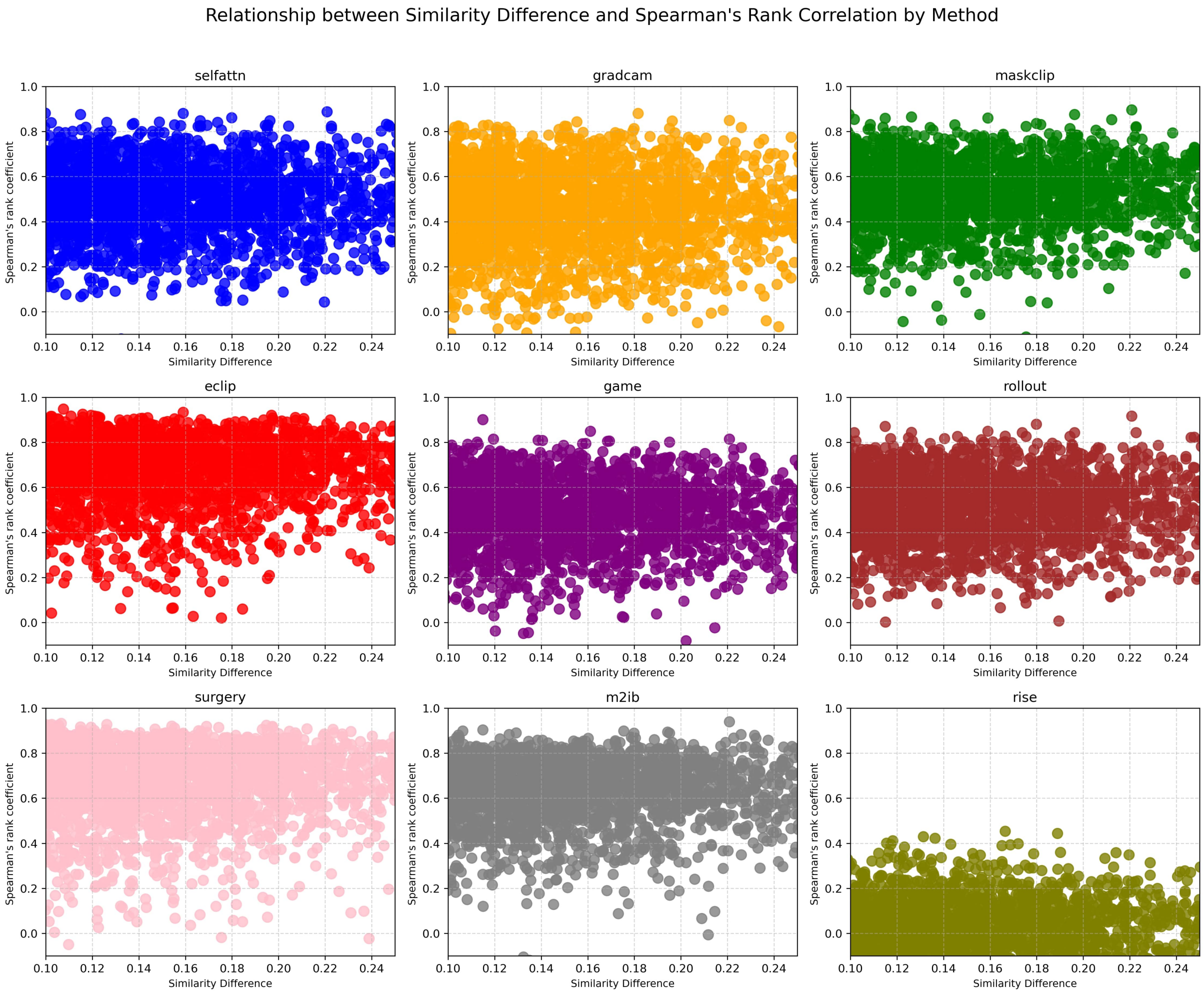

Figure 7: **Relationship between similarity difference and Spearman's rank correlation coefficient for each interpretability method.** Each point represents a perturbed sample. The analysis assesses the robustness of explanations by showing how changes in similarity affect the consistency of feature importance rankings. Higher Spearman values indicate more stable and reliable explanations despite perturbations

Sections 4.2, 4.3, and 4.4 evaluate different metrics independently. However, it is also important to examine how the same model behaves across multiple metrics. In this section, we explore the relationship between Spearman's rank correlation coefficient and the similarity difference. Figure 7 presents scatter plots that illustrate this relationship for each interpretability method. This analysis reveals how variations in the explanation maps, measured by vector-based similarity difference, affect the consistency of ranked feature importance across perturbations.

Analyzing the plots qualitatively the *ECLIP* and *Surgery* demonstrated the most stable behavior among all methods. Despite increasing similarity differences, the majority of its samples maintained a high Spearman coefficient, frequently above 0.6. This indicates that the relative ranking of salient regions remains preserved even when visual explanations differ significantly, supporting the method's robustness.

Grad-CAM, *MaskCLIP*, *M2IB*, *Rollout* and *Self-Attention* showed a more intermediate behavior. Their plots revealed broader dispersions and a tendency toward slightly lower Spearman values as similarity difference increased. Besides that, these methods present dense clusters between 0.5 and 0.7 Spearman's coefficient range across most levels of similarity difference. Notably, *Maskclip* showed less degradation in rank correlation than other methods, reflecting moderate resilience.

RISE was the least robust method. Its Spearman values remained predominantly below 0.3, independent of the similarity difference. Such poor correlation indicates that RISE fails to maintain stable rankings of important regions under minor input changes, undermining its reliability.

In summary, this analysis reinforces previous findings: methods such as *ECLIP* and *Surgery* are better suited for applications demanding stable interpretability, while techniques like *RISE* exhibit extreme sensitivity and should be applied with caution in high-stakes or reliability-critical contexts.

4.6 Relationship between SSIM score and similarity difference

Figure 8 shows scatter plots relating the SSIM score to the similarity difference for each interpretability method. This analysis helps clarify whether structural consistency in the saliency maps is preserved when the underlying attribution vectors diverge.

Self-Attention and *Rollout* methods stand out with consistently high SSIM scores across all similarity difference levels, rarely falling below 0.7. This stability indicates that, regardless of changes in attribution vectors, these methods maintain coherent visual explanations. A key reason for this behavior is that both are based directly on the attention maps generated by the Vision Transformer (ViT), which tend to be stable under small perturbations. These attention patterns reflect high-level contextual dependencies learned across layers and are generally less sensitive to localized changes. However, this same stability may also point to a limitation: the maps may reflect static or global attention patterns that do not adapt well to input-specific features, potentially overlooking subtle semantic shifts introduced by perturbations.

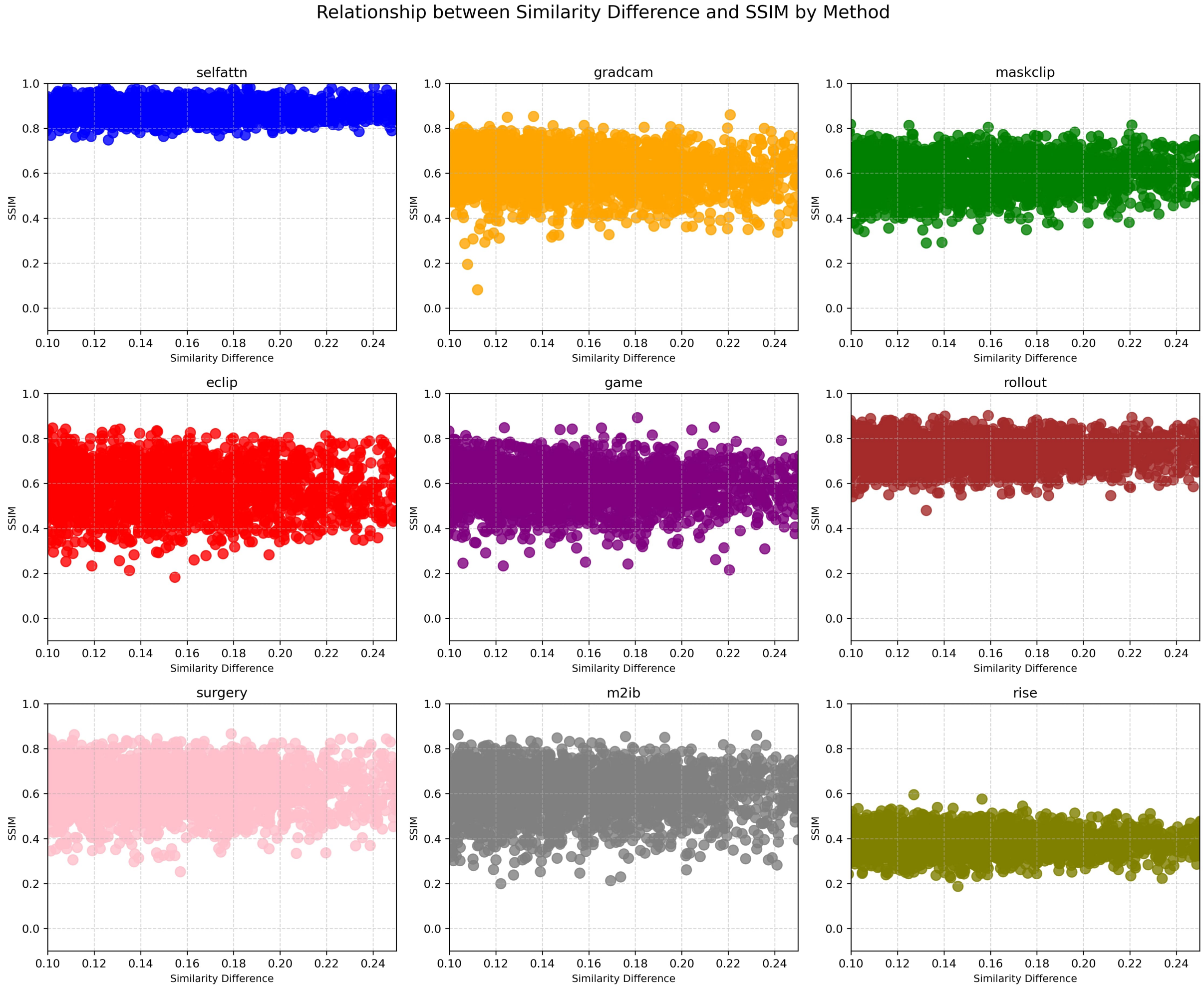

Figure 8: **Relationship between similarity difference and SSIM score for each interpretability method.** Each point represents a perturbed sample. This analysis evaluates how changes in attribution vectors affect the structural consistency of saliency maps. Higher SSIM scores indicate more stable visual explanations despite similarity variations.

Grad-CAM, *MaskCLIP*, *ECLIP*, *M2IB*, *GAME*, and *Surgery* exhibit more scattered patterns. Their SSIM values typically range between 0.4 and 0.8, showing a gradual decline as similarity difference increases. This indicates that these methods are moderately sensitive to vector-level changes, and while they often maintain structural features, their visual explanations can degrade under perturbation. Among these, *MaskCLIP* and *Surgery* appear to be slightly more resilient.

Finally, *RISE* displays the most unstable pattern. Its SSIM scores remain low and dispersed across the entire range of similarity differences, underscoring its high structural variability and weak robustness.

To sum up, this analysis shows that spatial consistency (SSIM) and attribution similarity do not always align. Methods like *Self-Attention* and *Rollout* preserve visual structure even with divergent attribution vectors, while others degrade both structurally and semantically. These findings confirm that some methods produce stable yet potentially uninformative maps, reinforcing the need for multi-perspective evaluation when assessing interpretability robustness.

4.7 Relationship between Top-K intersection and similarity difference

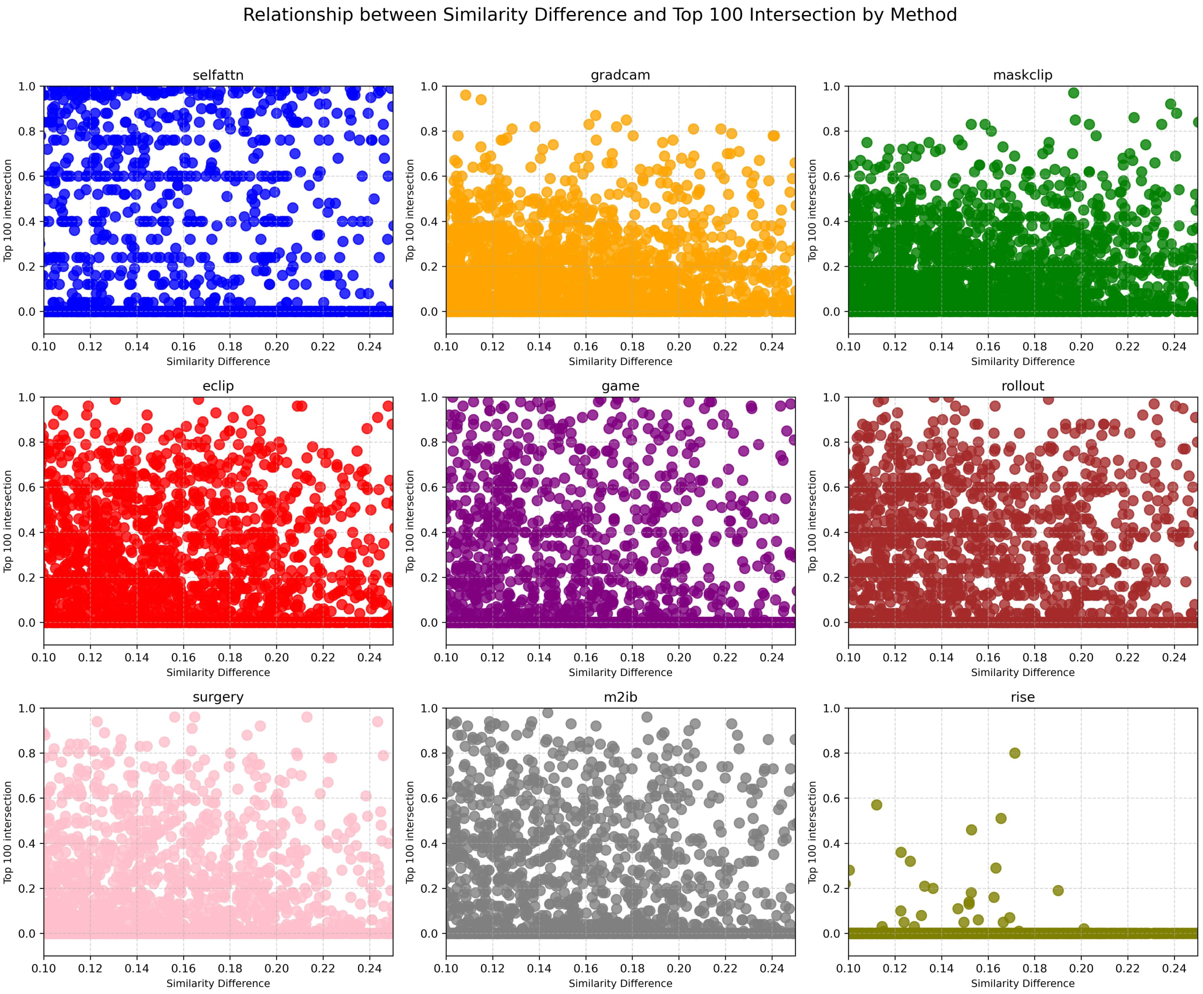

Figure 9: **Relationship between similarity difference and Top-100 intersection score for each interpretability method.** Each point represents a perturbed sample. This metric captures how well each method preserves the most salient regions (Top-100 pixels) under input perturbations. Higher intersection values indicate greater consistency in attribution.

Figure 9 presents scatter plots showing the relationship between similarity difference and Top-100 intersection score across interpretability methods. This analysis offers insight into how well each method maintains its most critical attribution regions as input similarity shifts.

Self-Attention stands out with a unique pattern. While its average intersection scores are moderate, its results show the widest dispersion, covering nearly the entire range of possible values. Notably, it frequently achieves perfect intersection scores (1.0), indicating that in certain cases it preserves exactly the same top-100 pixels despite input changes. However, the broad variation also suggests high inconsistency, where other samples experience dramatic shifts in attributions. This duality reflects a structurally unstable behavior: either fully aligned or highly divergent.

In contrast, methods like *ECLIP*, *GAME*, *Surgery*, *Rollout*, and *M2IB* exhibit more homogeneous behavior. Their scatter plots share similar shapes, characterized by moderate intersection values and significant dispersion. These methods tend to retain a subset of key regions across perturbations, but their consistency weakens as similarity difference grows. Although not as erratic as *Self-Attention*, these methods demonstrate only partial robustness in preserving salient regions.

Meanwhile, *Grad-CAM* and *MaskCLIP* exhibit lower and more concentrated intersection scores. Their values predominantly fall below 0.6, with less dispersion than the previously discussed methods. This indicates a more stable but less adaptive behavior, where saliency maps change modestly regardless of the degree of input variation. Such conservativeness might reflect a tendency to highlight similar regions repeatedly, possibly limiting sensitivity to nuanced input shifts.

Lastly, *RISE* performs the worst by a wide margin. Its intersection scores remain consistently near zero across all levels of similarity difference, confirming extreme instability and suggesting that it fails to preserve any consistent attribution patterns under perturbation.

These groupings help clarify the trade-offs between stability and sensitivity in interpretability methods. While some methods offer high visual overlap or structural consistency, they may do so at the cost of responsiveness to input variation, highlighting the need for careful selection based on application context.

5 DISCUSSION

This section analyzes the experimental findings regarding the sensitivity of interpretability methods applied to the CLIP model under input perturbations. The discussion is organized around four main aspects: the quantitative stability of methods, their comparative behavior, the nature of critical cases, and the practical implications of the observed sensitivity patterns.

5.1 Sensitivity Analysis Across Metrics

The combination of Spearman's rank correlation, SSIM, and Top-K Intersection metrics provided a multifaceted view of explanation stability. Overall, methods such as *ECLIP* and *Surgery* consistently demonstrated the highest resilience across all metrics. They maintained stable attribution rankings and preserved key salient regions, even under increasing levels of perturbation. This suggests that these techniques are less sensitive to small input variations, making them more reliable for use in practical applications.

In contrast, *RISE* showed extreme sensitivity. Its saliency maps were highly unstable across perturbation levels, with near-zero Spearman correlations and low spatial consistency. As a sampling-based black-box method, RISE is affected by hyperparameter choices and randomness [40], resulting in noisy and unreliable explanations for CLIP. This instability significantly undermines its usefulness in settings that demand consistent and interpretable outputs.

5.2 Behavioral Comparison Between Methods

The results revealed substantial differences in how each method responds to input perturbations. *ECLIP* preserved both the semantic and spatial structure of attribution maps, likely due to its use of gradient-guided token masking. Its architecture leverages the gradients of the image-text matching score to generate top-down explanations that are better aligned with the model's decision process. Furthermore, by replacing the sparse softmax attention with a normalized similarity map, ECLIP produces more continuous and interpretable heatmaps that maintain high fidelity under perturbations. It also combines spatial and channel weighting to enhance attribution clarity and aggregates layers strategically to maximize interpretability.

Surgery also demonstrated a strong ability to preserve focus on key regions despite noise, reinforcing its robustness. This method was explicitly designed to overcome common failure modes of CLIP, such as noisy activations and attention to background regions. Through a dual-path architecture and selective pruning of feature components (notably those associated with feed-forward networks), CLIP Surgery constructs more semantically coherent attention maps. Additionally, its inference-only modifications enable it to correct undesirable behaviors of CLIP without the need for fine-tuning, making it highly practical and effective.

M2IB and *MaskCLIP* showed intermediate robustness, supported by their architectural designs. *M2IB*, grounded in the multi-modal information bottleneck principle, learns to retain relevant information while suppressing irrelevant features. This results in focused and semantically aligned attributions across modalities, even in the presence of noise. Despite sensitivity to hyperparameter choices and occasional omission of complete object regions, its explanations remained meaningful in most cases. *MaskCLIP*, by leveraging dense features from the pretrained CLIP encoder and directly projecting them onto text embeddings, maintained strong localization of target concepts without requiring fine-tuning. Its ability to segment open-vocabulary phrases and resist noise makes it suitable for robust visual grounding, although its performance slightly declines under perturbation. Together, these methods demonstrated the capacity to produce coherent attributions with only moderate sensitivity to input perturbations.

Grad-CAM produced inconsistent results, frequently emphasizing background elements and demonstrating high sensitivity to input perturbations. This behavior can be attributed to several known limitations. When applied to CLIP, Grad-CAM often suffers from "opposite visualizations," where background regions are incorrectly prioritized over foreground objects. The method tends to generate noisy activations that reduce its class discriminativeness, especially due to its reliance on gradient information affected by architectural aspects such as ReLU activations. Although it occasionally highlights relevant regions, these are typically accompanied by substantial background noise. These weaknesses limit its applicability for robust interpretability in vision-language settings.

In contrast, *Self-Attention*, *GAME*, *Rollout*, and *RISE* were the least effective due to a combination of intrinsic methodological limitations and challenges specific to explaining CLIP. *Self-Attention* maps derived from raw attention weights often fail to correlate with token importance and tend to become uniformly distributed across deeper layers. When applied to CLIP, they frequently focus on irrelevant or background regions due to sparse and semantically inconsistent attention patterns. *GAME*, designed primarily for self-attention modules, struggles in multimodal contexts like CLIP that rely heavily on cross-modal interactions. Its heatmaps are typically noisy and lack discriminative precision. *Rollout* accumulates attention across layers without distinguishing between positive and negative contributions, resulting in diffuse and sometimes contradictory relevance scores. Finally, *RISE*, a perturbation-based black-box method, generates highly sparse and computationally expensive explanations that are frequently misaligned with the target object. Its sampling-based approximations are particularly fragile under input variation and produce poor localization performance, especially in CLIP. Together, these methods suffer from an inability to consistently highlight semantically meaningful regions, reinforcing their unsuitability for robust interpretability in vision-language settings.

5.3 Critical Examples and Visual Evidence

To complement the quantitative evaluation, Figure 10 presents a qualitative comparison of saliency maps generated by each interpretability method, before and after perturbation. The examples include diverse visual contexts associated with the labels “a photo of a dog,” “a photo of a frog,” and “a photo of a plane.” These samples offer concrete insights into the ability of each method to localize semantically relevant regions.

Among all methods, *ECLIP* and *Surgery* provided the most coherent and consistent explanations. Their saliency maps remained stable under perturbations and accurately highlighted the core regions corresponding to the described objects, such as the dog’s face, the frog’s body, and the fuselage of the plane. Minor shifts in attention were observed, but the primary focus of the attribution was preserved.

M2IB and *MaskCLIP* followed closely in performance. While some variation in activation patterns occurred after perturbation, these methods still concentrated their attention around the main object mentioned in the label. Their explanations tended to be more spatially aligned with the target concepts compared to most other techniques.

Grad-CAM displayed more erratic behavior. Although it occasionally emphasized relevant regions, for instance, parts of the animal bodies, it also highlighted unrelated

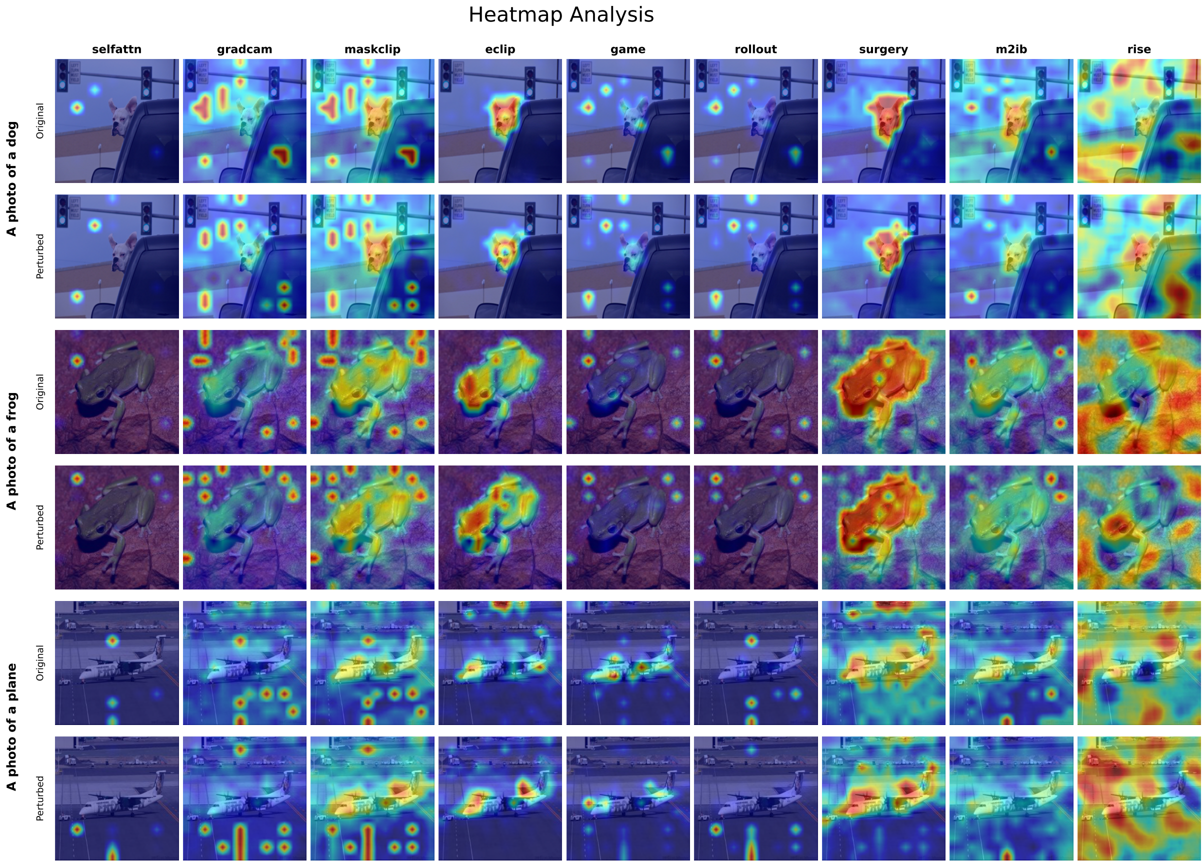

Figure 10: **Visual comparison of saliency maps before and after perturbation for each interpretability method.** Rows represent a original and perturbed input for three different labels and columns correspond to interpretability methods. The perturbation coefficient used was $\epsilon = 8/255$

elements in the background. This inconsistency compromises the reliability of its attributions in high-sensitivity scenarios.

On the other hand, *Self-Attention*, *GAME*, *Rollout*, and *RISE* failed to consistently identify the key object in the image. In multiple examples, their attributions were scattered across irrelevant areas or entirely ignored the regions associated with the label. Notably, *RISE* produced highly sparse heatmaps, with low activation over any meaningful region. The other three methods displayed isolated or misaligned points of attention that did not correspond to the semantic content of the prompt.

These qualitative findings support the previous metric-based analysis, confirming that only a subset of methods maintain robust, semantically meaningful explanations when exposed to input perturbations. In particular, *ECLIP* and *Surgery* emerge as strong candidates for reliable interpretability in vision-language models.

6 CONCLUSION

This work investigated the sensitivity of different interpretability methods when applied to the ”ViT-B/16” CLIP model, a representative vision-language transformer. Through a combination of quantitative metrics (Spearman’s rank correlation, SSIM, and Top-K Intersection) and qualitative heatmap analysis, the robustness of each method was systematically assessed under increasing levels of input perturbation.

The results indicate a significant variation in robustness across methods. *ECLIP* and *Surgery* consistently outperformed others, showing high stability and semantic alignment in their explanations. These methods effectively addressed limitations typical of CLIP, such as attention to background and noisy activations, through gradient-based strategies or architectural modifications. *M2IB* and *MaskCLIP* also demonstrated promising performance, producing explanations that remained meaningful even when perturbed, although with moderate degradation.

On the other hand, methods such as *Grad-CAM*, *Self-Attention*, *GAME*, *Rollout*, and *RISE* struggled to maintain relevance under noise. Their heatmaps frequently shifted toward irrelevant or dispersed regions, undermining their reliability in sensitive or high-stakes applications.

These findings underscore the need to critically assess not only the interpretability quality of a method in static conditions, but also its stability under real-world scenarios, where input variability is common. Interpretability techniques that are unstable can lead to misleading or inconsistent explanations, limiting their applicability in domains that demand transparency and trust.

Future research should aim to develop interpretability methods that are both semantically faithful and robust to perturbations, ideally grounded in the internal reasoning mechanisms of multimodal models. Studies could also expand the analysis to other vision-language models, such as SigLIP, which would provide valuable insights into whether the sensitivity patterns observed in CLIP generalize across architectures. Additionally, evaluating a broader set of interpretability approaches could help uncover strategies that are naturally more robust. This work presented experimental evidence, however, future investigations could also theoretically analyze each interpretation method to identify aspects that justify the results found. The establishment of standard benchmarks for sensitivity analysis, alongside human-in-the-loop assessments of explanation quality, would be instrumental in advancing the interpretability field toward more practical and reliable solutions.

Ultimately, this work contributes to the growing field of robust interpretability for vision-language models and provides a framework for evaluating explanation stability, a key requirement for trustworthy AI systems.

REFERÊNCIAS

[1] RADFORD, A. et al. *Learning Transferable Visual Models From Natural Language Supervision.* arXiv, 2021. Version Number: 1. Available at: <https://arxiv.org/abs/2103.00020>.

[2] ZHAO, C. et al. Grad-eclip: Gradient-based visual and textual explanations for clip. *arXiv preprint arXiv:2502.18816*, 2025.

[3] RUSSELL, S. J.; NORVIG, P.; DAVIS, E. *Artificial intelligence: a modern approach.* 3rd ed. ed. Upper Saddle River: Prentice Hall, 2010. (Prentice Hall series in artificial intelligence). ISBN 978-0-13-604259-4.

[4] GOODFELLOW, Y. B. I.; COURVILLE, A. *Deep Learning.* [S.l.]: MIT Press, 2016. `http://www.deeplearningbook.org`.

[5] LECUN, Y.; BENGIO, Y. et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, Cambridge, MA USA, v. 3361, n. 10, p. 1995, 1995.

[6] SHINDE, P. P.; SHAH, S. A Review of Machine Learning and Deep Learning Applications. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).* Pune, India: IEEE, 2018. p. 1–6. ISBN 978-1-5386-5257-2. Available at: <https://ieeexplore.ieee.org/document/8697857/>.

[7] ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, v. 8, n. 4, p. e1253, jul. 2018. ISSN 1942-4787, 1942-4795. Available at: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1253>.

[8] CHAI, J. et al. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, v. 6, p. 100134, dez. 2021. ISSN 26668270. Available at: <https://linkinghub.elsevier.com/retrieve/pii/S2666827021000670>.

[9] ZHANG, J. et al. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 8, p. 5625–5644, ago. 2024. ISSN 0162-8828, 2160-9292, 1939-3539. Available at: <https://ieeexplore.ieee.org/document/10445007/>.

[10] YANG, H. et al. Application of CLIP for efficient zero-shot learning. *Multimedia Systems*, v. 30, n. 4, p. 219, ago. 2024. ISSN 0942-4962, 1432-1882. Available at: <https://link.springer.com/10.1007/s00530-024-01414-9>.

[11] LIN, J. et al. VILA: On Pre-training for Visual Language Models. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2024. p. 26679–26689. ISBN 979-8-3503-5300-6. Available at: <https://ieeexplore.ieee.org/document/10657989/>.

[12] GIPIŠKIS, R.; TSAI, C.-W.; KURASOVA, O. Explainable ai (xai) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express*, Elsevier, v. 10, n. 6, p. 1331–1354, 2024.

[13] ESTEVA, A. et al. Deep learning-enabled medical computer vision. *NPJ digital medicine*, Nature Publishing Group UK London, v. 4, n. 1, p. 5, 2021.

[14] RICHMOND, K. M. et al. Explainable ai and law: An evidential survey. *Digital Society*, Springer, v. 3, n. 1, p. 1, 2024.

[15] SHAH, N.; BHAGAT, N.; SHAH, M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, Springer, v. 4, n. 1, p. 9, 2021.

[16] ČERNEVIČIENĖ, J.; KABAŠINSKAS, A. Explainable artificial intelligence (xai) in finance: a systematic literature review. *Artificial Intelligence Review*, Springer, v. 57, n. 8, p. 216, 2024.

[17] LIPTON, Z. C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, v. 16, n. 3, p. 31–57, jun. 2018. ISSN 1542-7730, 1542-7749. Available at: <https://dl.acm.org/doi/10.1145/3236386.3241340>.

[18] FAN, F.-L. et al. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, IEEE, v. 5, n. 6, p. 741–760, 2021.

[19] GHORBANI, A.; ABID, A.; ZOU, J. Interpretation of neural networks is fragile. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 3681–3688.

[20] KASHEFI, R. et al. *Explainability of Vision Transformers: A Comprehensive Review and New Perspectives*. arXiv, 2023. Version Number: 1. Available at: <https://arxiv.org/abs/2311.06786>.

[21] YU, J. et al. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[22] ANTOL, S. et al. Vqa: Visual question answering. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015.

[23] LIANG, F. et al. Open-vocabulary semantic segmentation with mask-adapted clip. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2023. p. 7061–7070.

[24] ZHAO, S. et al. Exploiting unlabeled data with vision and language models for object detection. In: SPRINGER. *European conference on computer vision*. [S.l.], 2022. p. 159–175.

[25] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

[27] OORD, V. den. Representation learning with contrastive predictive coding. *arXiv e-prints*, p. arXiv, 2018.

[28] SINGH, A. et al. Flava: A foundational language and vision alignment model. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 15638–15650.

[29] YAO, L. et al. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[30] JANG, J. et al. Unifying vision-language representation space with single-tower transformer. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2023. v. 37, n. 1, p. 980–988.

[31] YAO, L. et al. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, v. 35, p. 9125–9138, 2022.

[32] LUO, H. et al. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2023. p. 23033–23044.

[33] TSCHANNEN, M. et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[34] DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[35] RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. " 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*. [S.l.: s.n.], 2016. p. 1135–1144.

[36] GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.

[37] ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. [S.l.], 2014. p. 818–833.

[38] SELVARAJU, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, Springer, v. 128, p. 336–359, 2020.

[39] CHATTOPADHAY, A. et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE. *2018 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2018. p. 839–847.

[40] PETSIUK, V.; DAS, A.; SAENKO, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[41] LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964.

[42] BACH, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 7, p. e0130140, 2015.

[43] ABNAR, S.; ZUIDEMA, W. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[44] CHEFER, H.; GUR, S.; WOLF, L. Transformer interpretability beyond attention visualization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2021. p. 782–791.

[45] CHEFER, H.; GUR, S.; WOLF, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 397–406.

[46] WANG, Y.; RUDNER, T. G.; WILSON, A. G. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, v. 36, p. 16009–16027, 2023.

[47] LI, Y. et al. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, Elsevier, p. 111409, 2025.

[48] KINDERMANS, P.-J. et al. The (un) reliability of saliency methods. In: *Explainable AI: Interpreting, explaining and visualizing deep learning*. [S.l.]: Springer, 2019. p. 267–280.

[49] BANSAL, N.; AGARWAL, C.; NGUYEN, A. Sam: The sensitivity of attribution methods to hyperparameters. In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 8673–8683.

[50] MISHRA, S. et al. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021.

[51] TU, W.; DENG, W.; GEDEON, T. A closer look at the robustness of contrastive language-image pre-training (clip). *Advances in Neural Information Processing Systems*, v. 36, p. 13678–13691, 2023.

[52] MOAYERI, M. et al. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 19087–19097.

[53] DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 248–255.

[54] KRIZHEVSKY, A.; NAIR, V.; HINTON, G. Cifar-10 (canadian institute for advanced research). 2009. *URL http://www. cs. toronto. edu/kriz/cifar. html*, v. 5, n. 4, p. 1, 2009.

[55] GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.