EasyChair Preprint

№ XXX

Concept Attribution and Dual Explainability in Vision-Language Models

Laura-Luisa Voicu, Sebastian-Antonio Toma, Vlad Andrei Negru, Camelia Lemnaru and Rodica Potolea

October 17, 2025

Concept Attribution and Dual Explainability in Vision-Language Models

1st Laura-Luisa Voicu
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
lauraluisavoicu@gmail.com

2nd Sebastian-Antonio Toma
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
sebastianantoniotoma@gmail.com

3rd Vlad Andrei Negru
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Vlad.Negru@campus.utcluj.ro

4th Lemnaru Camelia
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Camelia.Lemnaru@campus.utcluj.ro

5th Rodica Potolea
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Rodica.Potolea@campus.utcluj.ro

***Abstract*—In this paper, we investigate methods for applying explainability to Vision-Language Models (VLMs). The main problem is given by the absence of a common representation between text and image, thus it is necessary to determine a modality for aligning the two information streams. Our research is particularized on the Contrastive Language-Image Pretraining (CLIP) model whose architecture is based on two independent encoders, both using Transformers. We propose a Concept Attribution method that addresses the problem of fusing visual encoder signals with textual gradients. This method is optimized by hybridizing the current model with a Large Language Model (LLM), which provides a developed linguistic context and implicitly an increase in the precision of explanations for multimodal reasoning. The fusion process is controlled through parameters that balance the textual and visual influence on explainability. Furthermore, for an objective validation of the results, we integrated Faithfulness Metrics, responsible for both analyzing the visual response and refining its visual representation through heatmaps. In addition, we demonstrated the utility of the method by applying it in a complex Dual Explainability system, which validates the coherence between the prompt given by the user and the model response.**

***Index Terms*—Explainability, VLM, MLLM, CLIP, Transformers**

I. Introduction

This paper aims to investigate explainability methods for vision language models. Explainability (XAI) [7] is needed especially for models whose backbone is based on visual transformers [5], as the complexity of their internal architectures increases, hindering understanding of model predictions. This leads to a lack of transparency and trust among users in critical domains such as Autonomous Driving [8] [9]. Although there are a considerable number of XAI methods, most of them are specialized in single-input rather than multimodal settings [10]. The goal of this research is to develop a method for Explainability that is compatible with multimodal Vision Language Models (VLM) [6].

We designed Concept Attribution to provide visual explanations of decisions made by VLMs. The work is customized for the CLIP (Contrastive Language-Image Pre-training) [3] model with Visual Transformer as the backbone for the visual encoder, but also for hybrid extensions between CLIP and LLM, replicating the architecture of the Dolphins [2] model.

In addition, to demonstrate the functionality of the method, we also developed a Dual Explainability approach, which consists of analyzing the relevance and accuracy of the model's responses in relation to specific prompts addressed by the user.

II. Related work

A. Vision-Language Models

Vision language models (VLMs) used for tasks such as Visual Question Answering (VQA) aim to test a system's ability to understand and interpret the surrounding worlds using natural languages [20]. As a general structure, such models integrate visual and textual encoders that transform inputs into vectors.

This representation creates a space in which the image can interact with the text. Visual encoders, typically trained on manually annotated datasets, introduce limitations in terms of generalization and adaptation to new concepts. This fact justifies the need to improve the visual encoder's operation so that it can learn from large and diverse datasets, and is not constrained by a fixed number of labels.

B. CLIP

The Contrastive Language-Image Pretraining model solves this problem by its ability to learn visual concepts through linguistic supervision. The zero-shot capabilities make it suitable for classification and segmentation tasks, which is essential for its integration into complex vision-language architectures.

CLIP consists of two independent encoders: one with a Transformer backbone for text, and another for image that can have either a Vision Transformer (ViT) or Convolutional Neural Network (CNN) backbone [18]. These encoders process the embeddings separately and then project them into

a shared vision language latent space, as is called into [1], where they compute the similarity between them using cosine similarity [16].

C. Forward Propagation

Forward propagation can be seen as an initial association between image and text. At this early stage, we do not know whether the association is correct or not. The image is divided into patches, each of them converted into a vector embedding. The text is also split into tokens and converted into embeddings. Self-attention is calculated in forward pass, meaning that CLIP can now create connections between each patch. At a conceptual level, these connections can be visualized as a graph: patches with similar elements are closer, while the different ones are farther apart.

D. Backward Propagation

Gradients are computed during backward propagation. Starting from forward pass, where the image and text are combined as a first step in the fusion process, the cosine similarity is computed between image-text embeddings. The Gradient signal is produced by maximizing the similarity score between image and text [13] [19].

$$\text{sim}(I, T) = \frac{E_{\text{image}}(I) \cdot E_{\text{text}}(T)}{\|E_{\text{image}}(I)\| \|E_{\text{text}}(T)\|}$$

The goal of the gradients is to indicate how much a patch in the image should change to be aligned with the text. In other words, the gradient indicates how much a function F needs to be modified to maximize its value.

E. Attention Heads Selection

Attention heads are components in Transformers architecture [4] that are responsible for the fusion between text tokens and image patches.

For CLIP with ViT backbone, whose architecture is based on Multi-Head Self-Attention [14], heads are specialized in certain semantic properties, as demonstrated in [1]. The authors show that even without human intervention, heads in the last 4 layers of CLIP have acquired specialized roles, such as detecting visual attributes for certain features, heads specialized in localization or counting, etc.

III. Proposed Solution

Concept Attribution is a method that controls the influence of textual data in VLMs based on Self-Attention architectures such as Contrastive Language-Image Models (CLIP).

The two main elements Concept Attribution is build on are Self-Attention Maps and Gradient Signals, computed during Forward and Backward propagation.

Concept Attribution method has two purposes:

- Assigning textual concepts to specific Visual representations.
- Combining the two sources of information: self-attention, computed in Forward Pass (which contains information about the structure of the image) and gradients, computed in Backward Pass (which contain information about the prompt text).

A. Image-Text Fusion

$$\text{Weighted-Attn} = (1 - \alpha) \cdot \tilde{\text{SA}} + \alpha \cdot \tilde{\text{G}}$$

Where:

- (α) is Text Strength parameter.
- $\tilde{\text{SA}}$ is the normalized Self-Attention, computed as:

$$\tilde{\text{SA}} = \frac{\text{SA} - \min(\text{SA})}{\max(\text{SA}) - \min(\text{SA}) + \epsilon}$$

- $\tilde{\text{G}}$ is the normalized Gradient, computed as:

$$\tilde{\text{G}} = \frac{\text{G} - \min(\text{G})}{\max(\text{G}) - \min(\text{G}) + \epsilon}$$

- ϵ is a small constant to prevent disivion by zero.

The fusion formula is applied to each patch in the image, and it combines the two sources of information:

- Self-attention, which captures structural relationship among patches.
- Text gradients, which reflect the influence of the entire prompt on each patch.

The text is tokenized by CLIP and then used to compute the gradient used in formula. As a result, each patch has a weight that represents its relevance to the given token.

Before self-attention and gradients are integrated in the final formula, they need to be normalized to reduce them to the same value range. This step is necessary because self-attention in Visual-Transformers (ViT) can have different values in interval $[0, 1]$, whereas gradients may have a wider range. Additionally, the value $\epsilon = 1e-8$ is added to prevent division by zero.

Another important aspect targeted by normalization is preserving the meaning of Text Strength (α) parameter, so that when $\alpha = 0.5$, the formula will determinate an equal contribution from Self-Attention and Gradients. In other words, the value $\alpha = 0.5$ represents equal influence from both image and gradients.

B. Text Strength parameter (α)

Text Strength (α) parameter controls the balance between internal structure of the image and the influence of text. Concept Attribution's goal is to accurately highlight regions in the image related to the text meaning. Therefore, the Text Strength value should be as close to 1 as possible if we want to get visual representations of text influence on images.

Another important aspect of the Text Strength parameter is that it can also be used across different layers in hybrid architectures, such as CLIP + LLM which is used by Dolphins model.

C. Text Boosting Factor (β)

Text Boosting Factor amplifies weak signals from textual gradients to compensate the excessive attention of Visual Transformers to patch structures. It also highlights subtle but significant regions for the text (e.g. textures, edges).

Why is Text Strength not enough?

Text Strength parameter controls only the image-text balance, not the amplification of individual components.

For example, if α = 0.9 and $\beta \to 0$ and the text highlights certain subtle parts of the image, the influence of α will be insignificant in the visual representation, in contradiction with the explanations from III-B.

D. Attention Heads Selection

Based on [1] study, Concept Attribution comes with an improvement referring to dynamical heads selection in real-time, based on text balance and redundancy penalization.

$$\text{HeadScore} = \alpha \cdot \text{TextRelevancce} - \beta \cdot \text{DiversityPenalty} \quad (1)$$

$$\text{TextRelevance} = \text{ImageEmbeddings} \cdot \text{TextProjection} \quad (2)$$

$$\text{DiversityPenalty} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \quad (3)$$

$$\text{A} = CurrentHead$$
$$\text{B} = PreviousHead$$

- β's goal is to avoid selecting heads with similar specializations, ensuring that the selected heads cover as many text features as possible.

E. Head Selection Strategies

Approaching the Head Selection problem, three strategies have been discussed:

1) Head selection is made only in the last four layers of CLIP-ViT, without applying dynamic Head Selection. As suggested in [1], the last four layers retain most of the direct features of text's influence on the image.
2) With this strategy, we select the ten best heads from last layer of CLIP-ViT, based on the score computed with formula (1). In this manner, the selected heads are those whose specializations best match and respond optimally to the current prompt.
3) This last strategy combines the two above, meaning we dynamically select the top ten most suitable heads from the last four layers. This strategy may be suitable when the prompt is ambiguous because it may need to expand the search for more general features in early layers.

F. Contrastive Texts

We define contrastive texts as concepts opposite to the initial prompt. Introducing them into the Concept Attribution method strengthens prompt-specific regions by eliminating activations for elements irrelevant to the contrastive text.

Automating Contrastive Concept Selection

Since manual selection of contrastive concept made by users can be often erroneous, either due to subjectivity or the difficulty of understanding the notion of contrastivity, we integrated an automatic selection mechanism for contrastive context into the Concept Attribution method. This mechanism is based on zero-shot capabilities of CLIP which is already part of its architecture, making the system more efficient, robust, and user-friendly.

The automatic selection process involves:

- Identifying the main concepts in the image through CLIP's visual encoder. This step will result in a list of concepts present in the image, obtained by computing the similarity between the embeddings in the image and a predefined list of candidate labels.
- Selection of the concept that results in the lowest degree of similarity to the concept deduced from the input.

An advantage of this method is its computational efficiency, as it leverages CLIP. The model identifies key concepts within the image that can be then used as contrastive concepts.

A disadvantage of this method is that the model does not understand the complex relationships between objects (for example, between "pedestrian" and "street" an elaborate distinction cannot be made, both of which can be categorized as traffic elements). However, the contrastive concept must be as different as possible, opposite (but not syntactically) to the original one, so the more it belongs to a semantic field that has no links with that of the original concept, the better the results.

G. Concept Attribution in Hybrid Systems

The need to extend the CLIP-ViT model arises because it relies on Self-Attention which, as observed in previous sections, captures structural relationships between image patches without considering the influence of the given prompt text [12]. Since the final goal of the work is to apply an explainability method to a Multimodal Large Language Model (MLLM), Concept Attribution must support such systems. Therefore, we extend CLIP-ViT with an LLM, following the same architecture as Dolphins [2]. In classic CLIP-ViT, Concept Attribution relies strictly on self-attention from the visual encoder and gradients calculated as cosine similarity between image-text embeddings.

However, the hybrid system brings the following improvements:

- Provides linguistic context projected into CLIP space and used in the text relevance computation. CLIP's tokenizer does not understand the word itself (for example, for "vehicle" does not understand that it can refer to any

of "car", "bus", "truck"), while the integrated LLM has contextual understanding capacity and thus can project detailed concepts and the generated heatmaps are much more accurate. For example, simple CLIP would activate the same areas regardless of the context provided in the prompt.
- Gradients computed during the LLM's response generation reflects the influence of the patches. It is also possible to determine the regions in the image that determined certain descriptions of the response.

H. Faithfulness metrics

To validate Concept Attribution, beyond the visual representations through heatmaps, we also integrated Faithfulness Metrics for an objective evaluation.

These metrics involve measuring the correlation between explanations generated and represented through heatmaps with the internal logic of CLIP-ViT. They aim to objectively measure the model's performance, which translates into how well the attention map identifies the relevant areas in the image for the model to understand the textual content. In other words, they measure how much the model's decision is affected when certain regions marked in attention map are disturbed [15].

Finally, we can consider that attention map is correct if the mask applied to the most relevant regions highlighted by it degrades the model's performance on the given task [**?**].

Concept Drop measures how much the image-text similarity score drops when regions deemed relevant by the Heatmap are removed [11]. A score greater than 0 indicates that the identified areas are critical for the given concept.

$$\text{ConceptDrop} = \frac{\text{Sim}(I,T) - \text{Sim}(I \odot (1-M), T)}{\text{Sim}(I,T)} \times 100\% \tag{4}$$

- Sim(I, T) represents the similarity score given by CLIP based on an image I and a text T.
- $\text{Sim}(I \odot (1-M), T)$ represents the similarity score over which the mask M is applied, where (1-M) represents the inverted heatmap to highlight the behavior of the CA method when relevant elements in the image are removed.
- Sim(I, T) normalization is needed to transform the Concept Drop score into a perecentage indicating how much of the original relevance was lost.

Interpretation

The problem of determining the formula starts from an absurd assumption: "If I remove all the areas that the Heat Map considers important, how much does the prediction degrade?". Thus, if certain areas chosen by Concept Attribution are eliminated and a degradation of the prediction is observed, it means that those were truly relevant areas. Therefore, it can be considered that the Concept Attribution method correctly identified the parts that contributed to the deduction.

Negative Suppression measures how well the heatmap avoids irrelevant regions or even regions relevant to the contrastive concept. It measures the difference between the similarity score of the image with the positive text and the similarity score of the image with the contrastive text. If the obtained value exceeds a threshold, then the model was able to distinguish the real from the contrastive one, even after eliminating the regions, meaning that the generated Heat Map did not suppress essential discriminative information.

$$\text{NS} = \frac{\text{Sim}(I, T^+)}{\text{Sim}(I, T^-) + \epsilon} \tag{5}$$

- $\text{Sim}(I, T^+)$ represents the similarity between the image and the given prompt.
- $\text{Sim}(I, T^-)$ represents the similarity between the image and the contrastive (opposite) prompt.

Interpretation

$\text{Sim}(I, T^-)$ measures the activation on the wrong concepts. If these values are high, the generated Heat Map is not correct, activating on irrelevant concepts, the model not being able to distinguish the concepts. Negative Suppression detects heatmaps that do not focus on concrete concepts, being too general. It also measures the specificity of the explanations given by the model.

Optimal Values:

- $NS > 5$: The heatmap avoids activating regions specific to the contrastive text.
- $NS < 2$: The heatmap activates regions specific to the contrastive text.

Integrating Faithfulness Metrics in Concept Attribution

Faithfulness Metrics also contribute to dynamic adjustments of the final heatmap, as they are integrated into the Concept Attribution method. Improvements brought by the integration of the metrics:

- We refine relevant areas by using Concept Drop metric. The resulting score can be seen as a feedback from the model, which tells how critical the Attention-Map is for the given concept. If the score is high and the identified areas are relevant, in the final representation given by the heatmap the contribution of those areas is amplified to be more visible.
- We suppress noise and false activations by using the Negative Suppression metric and the contrastive concept. If the Negative Suppression score is low, it means that the heatmap cannot sufficiently distinguish the input concept and the contrastive concept. It is necessary to reduce the intensities in those regions, thus teaching the model to penalize and suppress them.
- We optimize explanations by improving heatmap generation process. Faithfulness metrics integrated in Concept Attribution can be considered as a post-process that validates the quality of explanations. The improved

heatmap not only indicates where CLIP looks to make the final decision, but also why those regions are considered important by the model.

I. Background Suppression

Background suppression is an enhancement to Concept Attribution that filters out weak activations from heatmaps, keeping only the areas that are significant for the prompt. It acts like a high-pass filter for model accuracy, removing noise and improving accuracy.

$$H_f(i,j) = \begin{cases} H(i,j), & \text{if } H(i,j) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$\text{where } \tau = \alpha \cdot \max(H) \quad \text{, with } \alpha \in (0,1) \tag{7}$$

Optimal Values:

- For small $\alpha \in (0.1, 0.3)$: Subtle, less obvious details in the image are preserved.
- For large $\alpha \in (0.4, 0.6)$: More aggressive background removal.

J. Dual Explainability

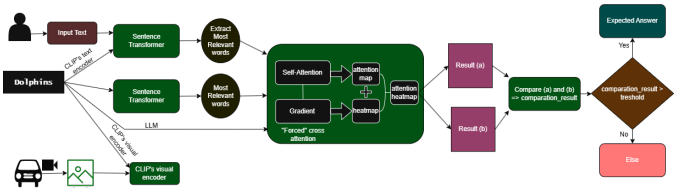

Fig. 1. Dual Explainability Architecture

Dual explainability, whose architecture is presented in Figure 1, is a method that verifies how concrete the answer given by the model is in relation to the user prompt. It starts from the assumption that a correct answer from model must activate regions at least similar to the regions activated by the user prompt, and therefore the generated heatmap must be at least similar.

There are two exceptional cases useful for understanding how the model works:

- The model guesses the answer, using the biases from the training data and without relying on the image
- The model focuses on irrelevant details in the image, even if the answer is correct.

IV. Experimental work

This section presents the evolution of the Concept Attribution method, demonstrating the improvements made to the explainability of Vision Language Models.

A. Choosing parameters for Text Strength and Text Boosting Factor

Choosing optimal value for Text Strength

- $\alpha \to 0$: Used for early layers to detect simple elements, such as pixels, edges, textures.
- $\alpha \to 1$: Used for late layers to integrate semantic information.

We tested different values for Text Strength (Figure 2): 0, 0.5, 0.8, 1. The results show that the alpha=0.8 ensure a balance between image structure and text influence. Lower values retain too much structural information and the heatmaps are less specific (all of them seem to point out same regions regardless of the text).

Choosing optimal value for Text Boosting Factor

- For $\beta < 1$, tests have shown that Gradients tend to be noisier that Self-Attention. Values lower than 1 for Text Boosting Factor act like a high-pass filter, keeping only strong Gradient signals and remove noise.
- For $\beta > 10$, oversaturated gradients create a balance between all patches, making it impossible to distinguish the critical regions from the important ones. As a result, patches lose granularity and heatmaps become overloaded with maximum values almost everywhere.

For Text Boosting Factor, $\beta = 0.5$ shows the best results compared to weaker and stronger boosting (Figure 3). The relevant regions in the image will be subtly amplified without damaging structural information. Grater values will cause oversaturation because all patches will become equally imporant.

B. Choosing the strategy for Head Selection

We chose the Head Selection strategy by analysing the Concept Attribution's behaviour on a large amount of tests, varying both image-text pairs, as well as Text Strength and Text Boosting Factor parameters (Figure 4).

TABLE I
Comparison of Head Selection Strategies

Strategy	Advantages	Limitations	Use Cases
Layers	Easy,fast; Good for capturing global structure.	Ignores individual head specializations.	Generic classification, prioritize visual context.
Heads	High precision for specific concepts.	Potential risk of overgitting to specific prompts.	Visual segmentation based on textual concepts.
Both	Balances global understanding with local relevance specific to the prompt.	Higher computational cost due to searching in multiple layers.	Complex or ambiguous prompts require a layered understanding of the image.

Conclusion reached from the experiments and comparisons from Table I:

- **Heads** strategy is suitable for clear concepts.

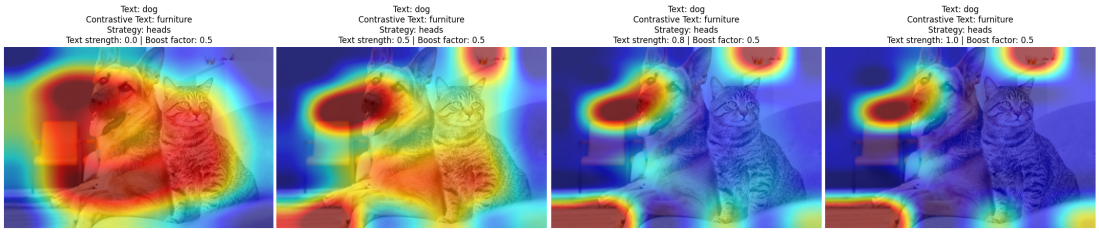

Fig. 2. Tests on Text Strategy

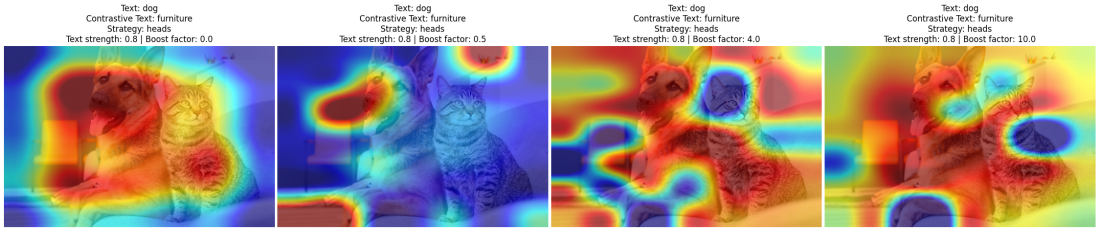

Fig. 3. Tests on Text Boosting Factor with different values

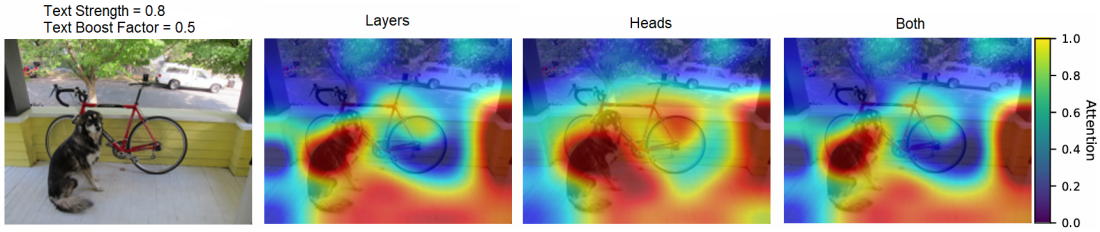

Fig. 4. Tests on Head Selection strategies

- **Both** strategy is better for complex and ambiguous prompts, and also when subtle features might require multi-layer understanding.

C. *Faithfulness metrics*

Performance validation with Concept Drop and Negative Suppression

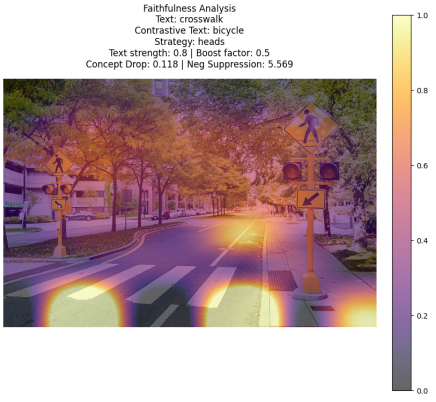

Fig. 5. Faithfulness metrics for found object

1) Results obtained for correct image-text correspondence:

- Concept Drop = 0.118: Positive value indicates that removing highlighted regions will cause the image-text similarity to drop down by 11.8 %, so the heatmap identifies the relevant crosswalk regions correctly.
- Negative suppression = 5.569: Value greater than 5 meaning that heatmap suppress activations for contrastive concept = ”human”, the model can tell the difference between ”crosswalk” and irrelevant ”human” regions.

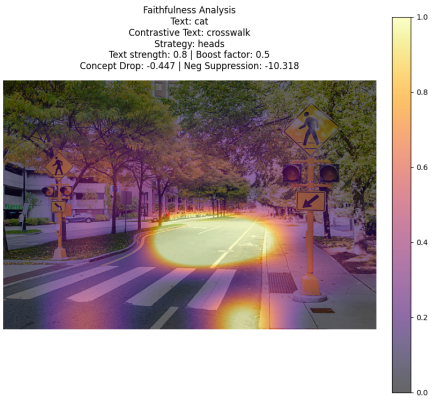

Fig. 6. Faithfulness metrics for not found object

2) Results obtained for the wrong image-text correspondence:

- Concept Drop = -0.447: Negative values indicates that highlighted regions increase similarity with contrastive word ”cat”, meaning that the heatmap is incorrect, missing the ”cat” entirely. This is a good result since there is no ”cat” in the image.
- Negative suppression = -10.318: Negative values indicates that the heatmap activates more for the contrastive concept ”crosswalk”, so the model fails to distinguish ”cat” from irrelevant ”crosswalk”.

Faithfulness metrics worked as intended and exposed False positive results. The heatmap only highlights noise.

Dynamic Adjustment of Heatmaps Based on Faithfulness Metrics

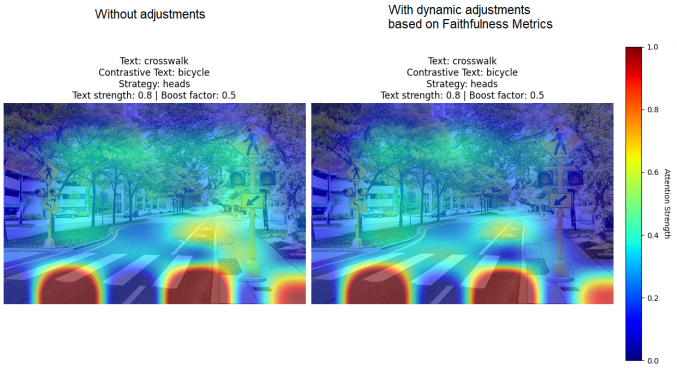

Fig. 7. Comparation between no adjustments and dynamic adjustments on Visual Representation based on Faithfulness Metrics.

D. Dual Explainability

Case I: Divergent but Valid Interpretation

Scenario:

Input prompt ”Can you see any pedestrians in this image?” as shown in Figure 8. Extracted words from prompt (”pedestrians”) and model response (”street”); both concepts are present in the image, but the model chooses a different and still valid one.

Analysis:

- For ”pedestrians”: Concept Drop = 0.696 indicates that the heatmap identifies regions associated with the concept, Negative Suppression = 2.248 indicates that the heatmap partially avoids irrelevant activations from contrastive concepts.
- For ”street”: Concept Drop = -0.265 indicates that the heatmap does not identify regions as well as the previous concept, Negative Suppression = 1.657 indicates that the heatmap is too general and might activate irrelevant regions.
- Conclusion: From a visual perspective, we can conclude that the heatmaps are not identical, but they are alike. Concepts are textually correct for the image, even though there is no pedestrian in the image, this concept is related to ”crosswalk”. The model focuses on contextual inference.

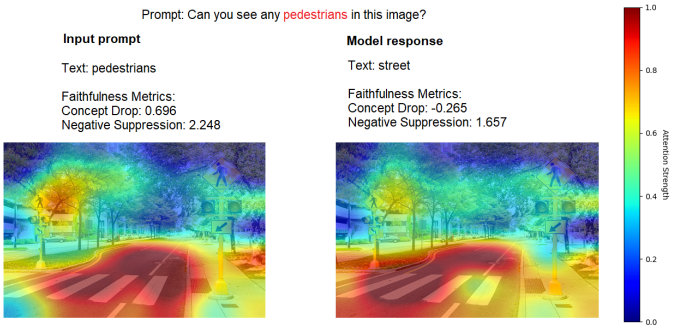

Fig. 8. Dual Explainability: Case I. Divergent but Valid Interpretation.

Case II: Generic prompt

Scenario:

The user provides a generic prompt as the one from Figure 9 (”Describe this image.”), but the model generates a specific response because it is able to idenfity relevant elements in the image.

Analysis:

Both heatmaps have poor metric quality. As expected, the heatmap for the prompt does not focus on specific elements. The heatmap generated using model response seems to indicate the object it is reffering to, but it is not supported by faithfulness metrics.

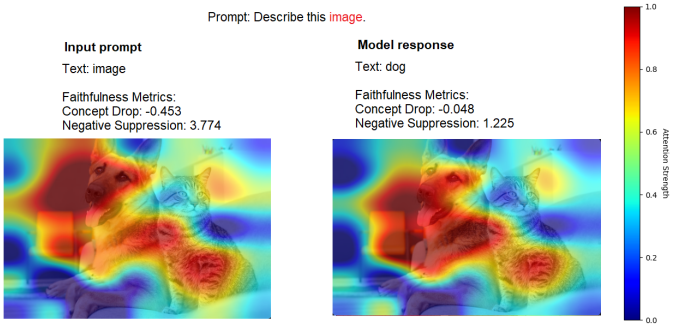

Fig. 9. Dual Explainability: Case I. Divergent but Valid Interpretation.

Case III: Ambiguous prompt

Scenario:

The user provides an ambiguous prompt as ”I don’t like winter.” (Figure 10). In this case, the model offers a general response. **Analysis:** This case addresses a limitation of Dual Explainability: when the prompt is too ambiguous it cannot produce coherent heatmaps, Faithfulness Metrics confirming that.

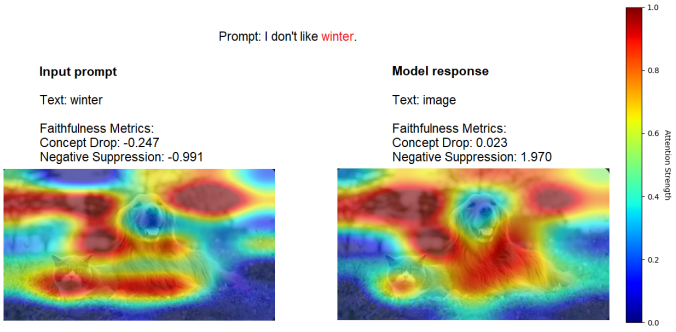

Fig. 10. Dual Explainability: Case I. Divergent but Valid Interpretation.

V. Conclusions

This paper proposes a method for integrating explainability into complex Vision-Language models using a method that addresses the problem of fusion for visual and textual information through Self-Attention signals and Gradients. Concept Attribution optimizes the image-text interaction through parameters dedicated to the text influence, while producing visual representations of explainability that are easy to understand and interpret for users.

The improvements brought to the Concept Attribution method are the Faithfulness Metrics used to validate the model's decision and refine the visual representation of that decision, but also the process of automating the contrastive concepts that highlight the real textual concepts. Dual Explainability also validates the coherence of the model for a certain category of prompts addressed by the user.

As further developments, the system can be extended to process more types of inputs (video, audio) and also, the contrastive concept automation mechanism can be improved with an Image Captioning Model to determine more specific contrastive concepts.

References

[1] Yossi Gandelsman and Alexei A. Efros and Jacob Steinhardt, "Interpreting CLIP's Image Representation via Text-Based Decomposition"

[2] Yingzi Ma and Yulong Cao and Jiachen Sun and Marco Pavone and Chaowei Xiao, "Dolphins: Multimodal Language Model for Driving"

[3] Alec Radford and Jong Wook Kim and Chris Hallacy and Aditya Ramesh and Gabriel Goh and Sandhini Agarwal and Girish Sastry and Amanda Askell and Pamela Mishkin and Jack Clark and Gretchen Krueger and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision"

[4] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, "Attention Is All You Need"

[5] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

[6] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, Jieping Ye, "From Redundancy to Relevance: Enhancing Explainability in Multimodal Large Language Models"

[7] Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez and Javier Del Ser and Adrien Bennetot and Siham Tabik and Alberto Barbado and Salvador García and Sergio Gil-López and Daniel Molina and Richard Benjamins and Raja Chatila and Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI"

[8] Omeiza, Daniel and Webb, Helena and Jirotka, Marina and Kunze, Lars, "Explanations in Autonomous Driving: A Survey"

[9] Shahin Atakishiyev and Mohammad Salameh and Hengshuai Yao and Randy Goebel, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions"

[10] Yunkai Dang and Kaichen Huang and Jiahao Huo and Yibo Yan and Sirui Huang and Dongrui Liu and Mengxi Gao and Jie Zhang and Chen Qian and Kun Wang and Yong Liu and Jing Shao and Hui Xiong and Xuming Hu "Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey"

[11] Julius Adebayo and Justin Gilmer and Michael Muelly and Ian Goodfellow and Moritz Hardt and Been Kim, "Sanity Checks for Saliency Maps"

[12] Sarthak Jain and Byron C. Wallace, "Attention is not Explanation"

[13] Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization"

[14] Peng Jin and Bo Zhu and Li Yuan and Shuicheng Yan, "MoH: Multi-Head Attention as Mixture-of-Head Attention"

[15] Yair Gat and Nitay Calderon and Amir Feder and Alexander Chapanin and Amit Sharma and Roi Reichart, "Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals"

[16] Rémi Kazmierczak and Eloïse Berthier and Goran Frehse and Gianni Franchi, "CLIP-QDA: An Explainable Concept Bottleneck Model"

[17] Weiyan Xie and Xiao-Hui Li and Caleb Chen Cao and Nevin L. Zhang, "ViT-CX: Causal Explanation of Vision Transformers"

[18] Sheng Shen and Liunian Harold Li and Hao Tan and Mohit Bansal and Anna Rohrbach and Kai-Wei Chang and Zhewei Yao and Kurt Keutzer, "How Much Can CLIP Benefit Vision-and-Language Tasks?"

[19] Ravidu Suien Rammuni Silva and Jordan J. Bird, "FM-G-CAM: A Holistic Approach for Explainable AI in Computer Vision"

[20] Chen, Fei-Long and Zhang, Du-Zhen and Han, Ming-Lun and Chen, Xiu-Yi and Shi, Jing and Xu, Shuang and Xu, Bo, "VLP: A Survey on Vision-language Pre-training"