GRADIENT DIFFERENCE ENABLES EFFICIENT SECOND-ORDER VISION-LANGUAGE INTERACTION EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language encoders such as CLIP excel at connecting visual and linguistic representations, yet interpreting their predictions remains challenging. Existing first-order explanation methods identify important image regions or text tokens individually, but fail to capture cross-modal feature interactions necessary for understanding dual-encoder predictions. Recent studies have therefore moved beyond first-order attribution toward second-order interaction explanations. However, the current frameworks face high computational or memory costs as text lengths and candidate interactions grow. To address these limitations, we propose a lightweight gradient-difference framework that derives second-order interaction explanations directly from existing first-order gradient-based attribution methods. By masking a target textual concept or visual region and measuring the resulting change in matching score gradients, our approach isolates phrase-conditioned visual explanations and region-conditioned textual explanations without introducing specialized interaction machinery or expensive sampling. Applied to standard methods (yielding Grad-CAM+, GAME+, and Grad-ECLIP+), our framework achieves performance comparable to state-of-the-art second-order approaches while significantly reducing computational overhead. The quantitative evaluations and human studies further demonstrate that our approach yields both faithful and user-informative vision-language correspondences.

1 INTRODUCTION

Vision-language encoders such as CLIP (Radford et al., 2021) have become fundamental building blocks for connecting visual and linguistic representations. By independently encoding an image and a text into a shared embedding space and measuring their similarity, CLIP enables a wide range of applications, including zero-shot recognition, image-text retrieval, and multimodal understanding. Despite their success, however, the mechanisms by which vision-language encoders compare visual and text inputs remain difficult to interpret. Understanding which image regions match a particular word or phrase is especially important for assessing whether a model relies on meaningful vision-language correspondences or spurious associations.

A common approach to interpreting CLIP is to assign importance scores to individual image patches or text tokens by employing first-order attribution methods, such as attention-based Rollout (Abnar & Zuidema, 2020), similarity map CLIPSurgery (Li et al., 2025), and gradient-related methods Grad-CAM (Selvaraju et al., 2017), GAME (Chefer et al., 2021), and Grad-ECLIP (Zhao et al., 2026). These methods produce visual saliency maps that reveal which image regions and text tokens contribute to an image-text similarity score. However, these methods fundamentally provide unimodal explanations: they identify regions important to the overall prediction, without revealing the interactions between image regions and textual concepts. This limitation is especially consequential for vision-language models, whose predictions depend on relationships between two modalities. For example, given an image containing a dog and a car and a caption describing both, such as the example in Fig. 1, a first-order heatmap may highlight both objects (Fig. 1left), but cannot indicate whether a highlighted region is important because it corresponds to the word "dog," "car," or another part of the caption.

Recent studies have therefore moved beyond first-order attribution toward second-order interaction explanations, which explicitly characterize dependencies between image regions and text to-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

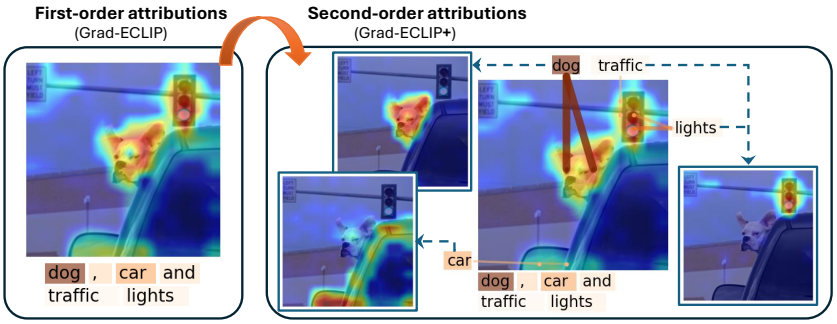


Figure 1: Visual explanations for the input image and sentence "dog, car and traffic lights." **Left:** The first-order attributions obtained from Grad-ECLIP Zhao et al. (2026). **Right:** The corresponding second-order attributions generated by the our gradient-difference method, denoted as Grad-ECLIP+. The overlaid heatmap and colors indicate the importance of specific image regions and words. Connecting lines present the high interaction responses between image regions and words, with darker colors and thicker lines signifying greater importance. The three sub-heatmaps pointed by dashed-arrows correspond to the "car", "dog", and "traffic lights" in the sentence.

kens. exCLIP (Moeller et al., 2025) derives a second-order attribution formulation and produces an image-text interaction tensor whose elements quantify the contribution of individual patch-token pairs to the similarity prediction. FIxLIP (Baniecki et al., 2025) further formulates interaction explanation from a game-theoretic perspective using weighted Banzhaf interactions (Tsai et al., 2023). While these works reveal the image-text relationships underlying the similarity predictions, which first-order saliency methods alone can obscure, they also introduce methodological machinery beyond conventional first-order explanation methods and large computational costs. As the number of patches and tokens increases, the resulting computation of exCLIP become memory-intensive when deriving interactions through their Integrated Hessians methodology (Janizek et al., 2021) along paths from reference inputs, while FIxLIP requires a substantial number of model inferences with the exponentially growing number of possible masked input combinations. Thus, exCLIP is constrained by memory consumption as the input sequence grows, while FIxLIP faces increasing computational cost as the number of players and sampled coalitions grows.

Considering the practicality and scalability issues of previous works, we propose a simple yet effective gradient-difference framework to obtain the rich interaction information required by second-order explanations directly from existing first-order gradient-based explanations, without introducing an expensive interaction-specific attribution framework. Specifically, given a target word or phrase in the text, we mask the corresponding text tokens and measure the resulting change in the gradients of the original and masked matching scores. This gradient difference captures how the contribution of each image region changes when the target textual concept is removed, thereby revealing the interaction between the visual region and the selected word or phrase. Symmetrically, given a target image region, we mask the corresponding image patches, and the resulting gradient difference identifies the words or phrases whose contributions change most when the selected visual region is removed.

This bidirectional formulation provides a natural bridge between first- and second-order explanations, enabling a consistent suite of explanations. When no target word, phrase, or image region is specified, the original first-order method is recovered and used to explain the overall image–text similarity. Consequently, the same underlying first-order attribution mechanism can be used to obtain conventional visual saliency as well as image-text interaction explanations. Specifically, we apply our gradient difference framework on Grad-CAM, GAME, and Grad-ECLIP, which enables the second-order versions, denoted as Grad-CAM+, GAME+, and Grad-ECLIP+.

Finally, comparing against representative second-order explanation methods, we evaluate our methods on CLIP models on three datasets: ImageNet (Russakovsky et al., 2015), MS COCO (Lin et al., 2014), and Flickr30k (Plummer et al., 2015). To evaluate the faithfulness of fine-grained associations in second-order explanations, we extend standard deletion/insertion metrics (Moeller et al., 2025; Zhao et al., 2026) to a *phrase-targeted* deletion/insertion protocol, which tests whether the resulting explanations correctly associate image regions with the corresponding words or phrases. Furthermore, we conduct human studies to assess both the faithfulness and usefulness of the ex-

planations. On the one hand, faithfulness to humans complements deletion/insertion metrics by assessing whether the associations identified by an explanation method align with human judgments of which image regions correspond to particular words. On the other hand, usefulness tests whether the visual explanations can improve a user's understanding of the model's prediction. The comprehensive experiment results show that the gradient-difference-enabled second-order explanation methods, especially Grad-ECLIP+, have comparable performance to the SOTA methods, while significantly reducing the implementation costs.

2 Related works

2.1 Vision-Language Models

Vision-language models (VLMs) (Radford et al., 2021; Tschannen et al., 2025; Wan et al., 2024; Zhai et al., 2023) learn representations that connect visual and language information, and have become fundamental to tasks such as image-text retrieval (Wang et al., 2019), image captioning Xu et al. (2015), visual question answering Antol et al. (2015), etc. Among them, the dual-encoder architecture as used in CLIP (Radford et al., 2021) employs separate image and text encoders that map the two modalities into a shared embedding space, which are trained with a contrastive objective. Given an image-text pair, the semantic correspondence is measured through the similarity between the two embeddings. This architecture has demonstrated strong zero-shot transferability and has become a widely used foundation for vision-language representation learning.

With the dual-encoder formulation, the similarity score of a vision-language encoder depends jointly on the visual and textual inputs. Consequently, understanding the prediction requires not only identifying important image regions or words individually, but also characterizing how features from the two modalities interact. This observation motivates recent efforts to develop interaction-based second-order explanations for CLIP and other vision-language encoders.

2.2 First-Order Explanations for VLMs

Many explainability methods attribute a model prediction to individual input features; for vision-language models, this means identifying image regions associated with a text query or word importances associated with an image. Attention Rollout (Abnar & Zuidema, 2020) propagates attention information through Transformer layers, while relevance-based approaches such as GAME (Chefer et al., 2021) use attention or relevance propagation with gradients to obtain prediction-specific explanations. However, attention maps do not necessarily correspond directly to the features driving the final matching score, and sparse attention patterns in CLIP can yield unsatisfactory explanations. Other approaches construct image-text similarity maps directly from local visual features and text representations. For example, CLIPSurgery (Li et al., 2025) and MaskCLIP (Zhou et al., 2022) use forward similarity information to localize textual concepts in images. Although these methods can produce interpretable spatial maps, they rely primarily on the forward computation and therefore do not necessarily measure the contribution of features to the final prediction. Gradient-based approaches instead use the sensitivity of the matching score to intermediate features. Grad-CAM (Selvaraju et al., 2017) obtains feature importance from gradients of the prediction, while Grad-ECLIP (Zhao et al., 2026) adapts this principle specifically to CLIP and the Transformer architecture. These first-order approaches provide useful information about which image regions or text tokens contribute to the *overall* similarity. However, they assign importance to features individually and therefore cannot explicitly reveal the relationships between features across modalities (e.g., which image region corresponds to which text token). This limitation is particularly important for vision-language encoders, whose predictions fundamentally depend on interactions between their two inputs.

2.3 Second-Order Explanations for VLMs

To overcome the limitation of first-order explanations, recent work has investigated second-order explanations that can explicitly characterize cross-modal correspondences between image regions and text tokens. exCLIP (Moeller et al., 2025) introduced a general second-order attribution framework for differentiable dual encoders, deriving interaction attribution from integrated mixed derivatives between the two inputs. More recently, FIxLIP (Baniecki et al., 2025) formulates explanation of vision-language encoders as a cooperative game. It treats image patches and text tokens as players, and the similarity of masked image-text pairs defines the game value. FIxLIP uses weighted Banzhaf interactions (Tsai et al., 2023) to decompose the similarity prediction into individual contributions

and pairwise interactions. Despite these advances, exCLIP relies on integrated second-order derivatives and can produce memory-costly, high-dimensional interaction representations, while FIxLIP requires repeated model evaluations over masked subsets and faces quadratic growth in the number of interaction terms; both pose practical challenges when the input size (e.g., text length) increases.

Rather than designing a separate interaction model or relying on extensive masked inferences, our work investigates whether existing first-order explanation methods can be systematically extended to capture cross-modal interactions. By taking the difference between gradients computed under different textual or visual conditions, we obtain a quantity that captures how the importance of one modality changes as a function of the other modality, bridging the gap between efficient first-order gradient attribution and interaction-based second-order explanation.

3 Gradient difference-enabled second-order explanations

3.1 Preliminaries

Vision-language encoders such as CLIP consist of an image encoder and a text encoder that independently map an image and a text description into a shared embedding space. Given an image I and a text input T, $\mathbf{F}_I = E_I(I) \in \mathbb{R}^D, \mathbf{F}_T = E_T(T) \in \mathbb{R}^D$ denote their corresponding image and text embeddings. The model computes the cosine similarity as

$$S(I,T) = \cos(E_I(I), E_T(T)) = \frac{\mathbf{F}_I^\top \mathbf{F}_T}{||\mathbf{F}_I||_2 ||\mathbf{F}_T||_2}, \tag{1}$$

which is the predicted matching score to be explained. First-order explanations generate visualizations that highlight important regions in the image and important word in the sentence, related to producing the matching score. We first briefly introduce the gradient-based first-order methods, including Grad-CAM, GAME, and Grad-ECLIP, from the image viewpoint, as in (Chefer et al., 2021; Zhao et al., 2026), where image patch importances are calculated with the given text prompt T. The word importances generated for the text prompt T, given the input image I, can be obtained by analogous calculation from the text encoder.

Let $\mathbf{A} \in \mathbb{R}^{H \times W \times C}$ denote an intermediate feature representation in the image encoder, with spatial locations indexed by i and feature channels indexed by c. Gradient-based explanation methods obtain evidence of the importance of the intermediate features from the gradient of the matching score with respect to $\mathbf{A}$:

$$\mathbf{G} = \frac{\partial S(I,T)}{\partial \mathbf{A}}. \tag{2}$$

The resulting gradient is then transformed into a spatial explanation map according to the following specific explanation method:

- **Grad-CAM** (Selvaraju et al., 2017) computes the channel importance weights by globally averaging the gradient over spatial locations, and the spatial explanation is then obtained by weighting and aggregating the feature maps.
- **GAME** (Chefer et al., 2021) is a gradient- and relevance-based Transformer explanation method. It propagates prediction-specific relevance through the Transformer and combines the relevance with the attention structure to produce a spatial explanation.
- **Grad-ECLIP** (Zhao et al., 2026) is a state-of-the-art first-order attribution method for explaining image-text matching in VLMs. It uses the intermediate feature values together with prediction-related channel and spatial importance weights. In particular, the channel importance is obtained from gradients of the image-text matching score with respect to the final attention layer output, while a loosened attention map is used to determine spatial importance. The final heatmap is obtained by weighting and aggregating the intermediate feature values using these channel and spatial weights.

Overall, we denote the complete transformation from the target gradient and intermediate features to a spatial heatmap by a function $\mathcal{M}(\cdot)$. Thus, the gradient-based first-order explanation methods can be abstractly written as

$$H = \mathcal{M}(\mathbf{A}, \mathbf{G}). \tag{3}$$

This notation allows us to focus on the common gradient signal shared by gradient-based explanation methods. Despite their different feature aggregation mechanisms, these first-order methods share a common structure: a gradient of the matching score provides prediction-specific information, which is subsequently transformed into a spatial explanation.

3.2 Gradient Difference for Cross-Modal Interaction

First-order explanations describe the importance of image regions for the complete sentence. However, a sentence may contain multiple concepts, and the resulting heatmap does not explicitly identify which region is associated with a particular word or phrase.

To obtain phrase-specific explanations, we consider the change in the matching score when we remove a target part. Let P denote the target phrase and let $T_{\setminus P}$ denote the text obtained by masking the tokens corresponding to P with zeros. We define the phrase-induced score difference as

$$\Delta S_P = S(I, T) - S(I, T_{\setminus P}). \tag{4}$$

A positive ΔS_P indicates that the target phrase contributes positively to image-text matching, while a negative one indicates the opposite. Then, the gradient of this score difference with respect to the image representation can be obtained directly from two first-order gradients. Specifically,

$$\Delta \mathbf{G}_P = \frac{\partial \Delta S_P}{\partial \mathbf{A}} = \frac{\partial S(I,T)}{\partial \mathbf{A}} - \frac{\partial S(I, T_{\setminus P})}{\partial \mathbf{A}} = \mathbf{G}_{\text{orig}} - \mathbf{G}_{\text{mask}}. \tag{5}$$

We refer to $\Delta \mathbf{G}_P$ as the gradient difference. The original gradient $\mathbf{G}_{\text{orig}}$ measures the sensitivity of the matching score to image features when the complete sentence is present, whereas $\mathbf{G}_{\text{mask}}$ measures the corresponding sensitivity after removing the target phrase. Their difference therefore measures how image sensitivity changes with the presence of the phrase.

This change in sensitivity characterizes the interaction between the image representation and the target phrase. In this sense, the gradient difference provides a finite-difference approximation to a cross-modal second-order interaction. If the text representation perturbation induced by masking P is viewed as $\Delta \mathbf{z}_P$, then we can derive

$$\Delta \mathbf{G}_P \approx \frac{\partial^2 S}{\partial \mathbf{A}\, \partial \mathbf{z}_P} \Delta \mathbf{z}_P, \tag{6}$$

for which we put the details in the Appendix. Therefore, phrase masking and gradient differences provide a practical way to obtain cross-modal interactions from two ordinary backward passes instead of explicitly computing a Hessian.

Our formulation is complementary to existing explicit second-order attribution approaches. For example, exCLIP derives cross-modal interactions using integrated Hessians, while FIxLIP estimates interactions using weighted Banzhaf indices. Our approach instead asks a more specific question: "how does the image attribution change when a particular phrase is removed?" This enables existing gradient-based explanation mechanisms to be directly extended to phrase-region interaction explanations.

3.3 Gradient difference-enabled second-order explanations

The gradient-difference formulation is method-agnostic. The gradient from the prediction w.r.t. the feature maps used in first-order methods measures how the features affect the matching score, while in the second-order method, we measure the sensitivity related to the score change when the given target phrase P is removed, by introducing perturbation on the text representation, as shown in (6). Therefore, given any gradient-based first-order explanation method $\mathcal{M}$, we obtain its second-order counterpart by replacing the original gradient with the gradient difference,

$$H^{+}_{\mathcal{M}}(P) = \mathcal{M}\left(\mathbf{A}, \Delta \mathbf{G}_P\right) = \mathcal{M}\left(\mathbf{A}, \mathbf{G}_{\text{orig}} - \mathbf{G}_{\text{mask}}\right). \tag{7}$$

The resulting heatmap is conditioned on the target phrase P and therefore represents the image regions whose attribution changes most when the phrase is removed. Symmetrically, with analogous derivation of (6) from the view of text encoder, the word importance correspond to a specific image region can be obtained by checking the gradient changes on the text token features when parts of the image is masked and cause the image representation perturbation (see the example in Figure 4(a)).

We instantiate this general formulation with three representative gradient-based explanation methods, Grad-CAM, GAME, and Grad-ECLIP, which leads to **Grad-CAM+**, **GAME+**, and **Grad-ECLIP+** by replacing the image gradient w.r.t. the original text with the corresponding gradient difference. The three proposed methods therefore share the same gradient-difference principle while preserving the respective designs of their first-order counterparts.

4 EXPERIMENTS

In the experiments, we empirically validate the performance of the gradient-difference-enabled **Grad-CAM+**, **GAME+**, and **Grad-ECLIP+** with evaluations on explanation faithfulness to model (Sec. 4.1), and the faithfulness and usefulness to humans by user studies (Sec. 4.2), and compare with the SOTA methods. We also provide the evaluation of localization ability of the explanation maps in Sec. A.1. For the explanation faithfulness to model, we propose the phrase-target deletion and insertion protocol, aiming at evaluating the patch-word association that constitutes the second-order explanation. We mainly use the CLIP dual encoders with the OpenAI pretrained models of two sizes: ViT-B/32 with 7×7 image patches and ViT-B/16 with 14×14. The experiments are all run on a single 48G NVIDIA RTX 6000.

Datasets We conduct our evaluations on three commonly used datasets: the Flickr30k dataset (Young et al., 2014) with collected entity annotations (Plummer et al., 2015), the MS COCO Caption2017 (Lin et al., 2014), and ImageNet (Russakovsky et al., 2015). In Flickr30k, there are 5000 image-caption pairs with object phrase annotations in the validation set, including the first word location in the sentence and the corresponding bounding box location on the image. Finally, we average the results for each method on the same subset of 4438 images, where images with long captions were removed due to out-of-memory (OOM) errors from exCLIP. For the MS COCO validation set, which contains 5 captions for one image, we detect all phrases related to the objects in the captions, and keep the ones that correspond to a single instance with the bounding box annotation. This results in 11706 object phrases from 3046 images. For ImageNet, we construct 5000 samples by randomly selecting 4 distinct categories from the 1000 ImageNet classes and drawing a single validation image from each. We stitch these four images into a single composite layout paired with a caption listing the 4 category names separated by commas (e.g., "tow truck, sturgeon, sax, wool"), as illustrated in Figure 2.

Baselines We compare our Grad-CAM+, GAME+, and Grad-ECLIP+ against three baselines: ITSM follows the approaches from Li et al. (2025); Zhou et al. (2022) using forward similarity between image patch features and token features by applying the final projection layer to the individual hidden representations; the recent exCLIP Moeller et al. (2025) derives the image patch and word interactions from the integrated mixed derivatives between the two inputs; the state-of-the-art second-order method FIxLIP Baniecki et al. (2025) is a perturbation-based approach, which solves the vision-language cooperative game with weighted Banzhaf interactions Tsai et al. (2023). Each baseline is built with different properties. For FIxLIP, we adopt a middle value for p in the weighted Banzhaf interactions as $p = 0.5$, and 2^{19} sampled masks (budget) in their implementation. A comparison of the computational cost in memory and processing time is listed in Table 1. With the ITSM only requiring a forward inference as the baseline, the gradient-based methods Grad-CAM+, GAME+, and Grad-ECLIP+ remain as computationally lightweight as the first-order versions. Due to the huge time cost with ViT-B/16, we only conduct the quantitative evaluations for FIxLIP with the ViT-B/32.

Table 1: Computational cost (memory/processing time) comparison on **one single image-text pair** on ViT-B/16 with 196 image patches and different number of text tokens N.

Method	$N = 10$		$N = 20$		$N = 30$	
ITSM	0.8G	0.02s	0.8G	0.02s	0.8G	0.02s
FIxLIP	1.9G	**31m**	1.9G	**34m**	1.9G	**40m**
exCLIP	**16G**	9s	**27G**	11s	**39G**	14s
Grad-CAM+	1G	0.3s	1G	0.6s	1G	0.9s
GAME+	1.4G	0.4s	1.4G	0.8s	1.4G	1.2s
Grad-ECLIP+	1G	0.3s	1G	0.6s	1G	0.9s

4.1 PHRASE-TARGETED DELETION AND INSERTION

In this section, we evaluate the faithfulness of explanations through perturbation deletion and insertion experiments. Previous second-order methods were evaluated in different ways. exCLIP adopts the conditional perturbation (Sammani et al., 2024), which keeps the sentence and deletes/inserts image patches based on the heatmap aggregated over all words, resulting in a sentence-level evaluation. Meanwhile, FIxLIP deletes/inserts either image patches or text tokens in each step based on the attribution values, which evaluates the importance of patches and tokens to the whole image-text matching rather than the correlation between image regions and parts of the sentence. Neither of these methods test cross-modal attribution of partial inputs (i.e., image regions and individual words). Therefore, we propose phrase-targeted deletion and insertion, which tests whether 2nd-order attribution maps correctly associate image regions with the corresponding words or phrases.

Metric: Concretely, as shown in Fig. 2, the heatmaps corresponding to the annotated phrases in each dataset are generated by each method. For the *deletion process*, with the importance in

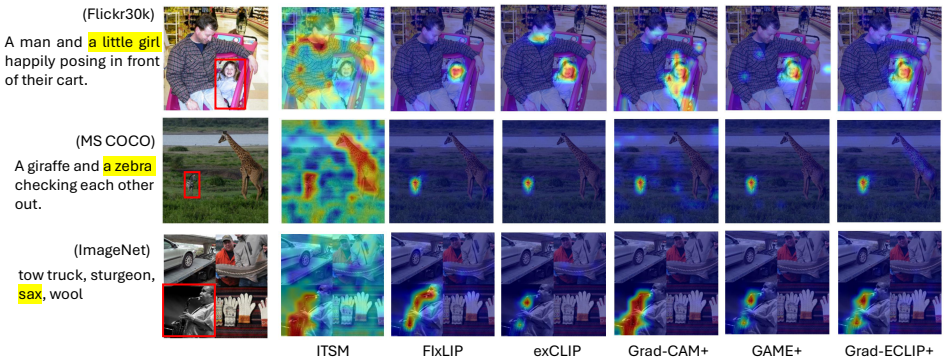


Figure 2: The image-caption examples from Flickr30k (Plummer et al., 2015), MS COCO (Lin et al., 2014), and ImageNet (Russakovsky et al., 2015), and the attributions for the selected object phrase in the captions (marked in yellow) generated by each method, with corresponding bounding boxes (marked in red) on the image.

Table 2: Explanation faithfulness evaluation using phrase-targeted deletion (PD) and insertion (PI). The top-2 results are bolded and underlined.

	ViT-B/32						ViT-B/16					
	COCO		Flickr30k		ImageNet		COCO		Flickr30k		ImageNet	
Method	PD↓	PI↑	PD↓	PI↑	PD↓	PI↑	PD↓	PI↑	PD↓	PI↑	PD↓	PI↑
FIxLIP	**0.23**	**1.04**	**0.24**	**1.05**	**-0.09**	**1.28**	-	-	-	-	-	-
ITSM	0.58	0.71	0.58	0.73	0.34	0.82	0.64	1.05	0.64	1.08	-0.07	1.41
exCLIP	0.38	0.89	0.42	0.87	0.07	1.13	0.55	1.10	0.56	1.12	-0.13	**1.59**
Grad-CAM+	0.47	0.80	0.47	0.81	0.05	1.13	0.57	1.08	0.59	1.10	-0.14	1.54
GAME+	0.39	0.87	0.39	0.87	0.08	1.06	0.60	1.06	0.60	1.07	-0.01	1.40
Grad-ECLIP+	0.32	0.89	0.37	0.89	0.05	1.10	**0.48**	**1.11**	**0.49**	**1.13**	**-0.17**	1.51

descending order based on the heatmap, the image patches are deleted from the whole image (I_{whole}, step $t = 0$) to a blank image (I_{blank}, step $t = N_{patch}$). The I_{blank} and the deleted pixels are filled with the image pre-processing mean values. We calculate the cosine similarity $S(I_t, P)$ of the modified image I_t at each step t with the text P *only containing phrase tokens*, where other tokens are replaced by zero. To eliminate the influence of different similarity score ranges with different phrases, for each targeted phrase, the score $S(I_t, P)$ at each step is normalized by:

$$S(I_t, P) = \frac{S(I_t, P) - S(I_{blank}, P)}{S(I_{whole}, P) - S(I_{blank}, P)}, \tag{8}$$

Finally, the similarities at each step are averaged over all phrases. The *insertion process* follows a similar approach, except that the image patches are inserted, according to the descending importance, from I_{blank} to I_{whole} step by step to obtain I_t and the corresponding score.

We plot the normalized similarity score at each step in the deletion/insertion process (See the Appendix Fig. 5), and report the area under the curve (AUC) as the performance of phrase-targeted deletion (PD) and insertion (PI), where a lower PD and higher PI represent a steeper similarity score drop and increase, which indicates a better explanation faithfulness to the model.

Results: The results are shown in Table 2. FIxLIP achieves the best overall performance on the ViT-B/32 due to the use of game theory to solve the attribution assignment by exhaustively considering the cases where each patch and token is masked or not. However, FIxLIP requires an extremely high computational time, rendering it intractable for more patches, such as with ViT-B/16 ($>$ 30 min for a single image-text pair), when processing all samples in the evaluation datasets. Among other methods, Grad-CAM+ presents better performance on stitched images in ImageNet with ViT-B/32, while Grad-ECLIP+ consistently outperforms ITSM, exCLIP, and other gradient-based variances across MS COCO and Flickr30k. Furthermore, when scaled to the finer-grained ViT-B/16 architecture, Grad-ECLIP+ maintains its clear advantage, achieving the lowest PD scores on COCO (0.48), Flickr30k (0.49), and ImageNet (-0.17) while obtaining the highest PI scores (e.g., 1.11 on COCO and 1.13 on Flickr30k). Moreover, all the gradient-difference-enabled methods show clearly better explanation faithfulness than ITSM. Note that a negative PD indicates that the cosine similarity between the target phrase and a deleted image is even lower than with a blank image, while a PI

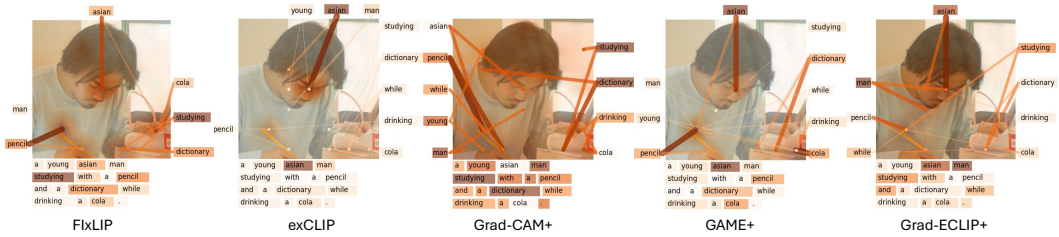


Figure 3: Visualizations of second-order explanations from our framework (Grad-CAM+, GAME+, Grad-ECLIP+) and the state-of-the-art methods FIxLIP and exCLIP. More important image regions and words are marked with darker color. More important associations between image patches and text tokens are indicated with thicker lines and darker color.

over 1.0 represents that the cosine similarity between the target phrase and the inserted part image is higher than with the whole image.

These results demonstrate that integrating our gradient-difference mechanism into first-order attribution methods successfully enables their ability to associate image regions with the corresponding phrases in the caption. In particular, applying on the SOTA first-order Grad-ECLIP, Grad-ECLIP+ shows better explanation faithfulness than the recent representative second-order method exCLIP.

4.2 User Studies

Although quantitative evaluation metrics can provide evidence that the explanation captures information that is associated with the model's prediction and related to the target, they evaluate explanations from the perspective of the model and do not directly assess whether the information revealed is faithful and useful to humans. Therefore, we complement the quantitative faithfulness evaluation with human-subject studies. For second-order explanations, which aim to expose interactions between image regions and text words, we visualize the associations as in the examples in Fig. 3. We normalize the interaction values such that the maximum value is 1. Then we plot the associations that are larger than 0.4, and up to the 2 strongest associations for a word are shown. Note that ITSM is excluded from the user studies since it only obtains fine-grained feature similarities rather than visual explanations.

Plausibility Study: It can be seen from the comparisons in Fig. 3 that the locations of the top-2 patches for the same concept and the strengths are different in each method. Therefore, in the first user study, we assess faithfulness to humans (i.e. plausibility) by investigating which method's results best align with human judgments of which image regions correspond to particular words. Specifically, we generate the association maps for 20 image-text pairs, and participants are asked to choose one from the five randomly placed results with the question "Which option best explains the regions, words, and matching relationships the model focuses on when calculating the similarity of the image and the text" (see the Appendix Fig. 6 for the questionnaire interface). From the 320 responses collected from 16 participants, Grad-ECLIP+ and FIxLIP were selected much more frequently than others, with 38.8% and 31.3% respectively, followed by Grad-CAM+ 14.4%, GAME+ 11.9% and exCLIP 3.8% (χ^2 test, $p<.001$). It demonstrated that the results from Grad-ECLIP+ and FIxLIP obtain significantly higher user trust, indicating they are more interpretable to humans than the other three methods.

Table 3: Selection rates of different methods.

Method	Rate
FIxLIP	31.3%
exCLIP	3.8%
Grad-CAM+	14.4%
GAME+	11.9%
Grad-ECLIP+	**38.8%**

Simulation Study: We further conduct the second user study to investigate whether the second-order explanations can help to improve a user's understanding of the model's prediction by adopting the *forward simulation task* (Hase & Bansal, 2020). Similar to Zhao et al. (2026), the participants are given an image and two captions and asked to judge which caption will receive a higher matching score from the model under two conditions: with or without corresponding patch-word association explanations provided. If the users' judgment accuracy improves when provided with the explanations, then it suggests that the explanations allow the users to better understand the model's predictions. Concretely, the questionnaire contains two parts: Part I showing 10 images with two caption options, followed by Part II showing another 10 images with explanations for each caption option. Considering the faithfulness results (§4.1) and the performance in the Plausibility Study (above), we adopt the two best methods, FIxLIP and Grad-ECLIP+, to generate the association explanations for

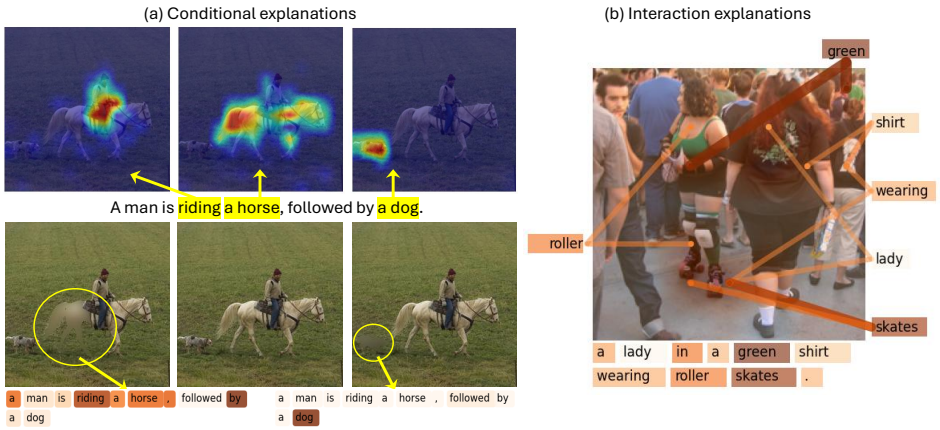


Figure 4: The various approaches to model understanding by gradient-difference-enabled Grad-ECLIP+. (a) Conditional (phrase- or region-targeted) explanations, obtained by removing the target phrase or blurring the object pixels obtained from SAM (Kirillov et al., 2023) (zoom in the yellow circled regions for better view). (b) Interaction explanations. Here, it can be seen that the CLIP model tends to match words independently rather than capture the sentence-level semantics, as evidenced by strong associations for words such as "green" and "shirt" that correspond to different people rather than the specific person described in the sentence.

each image. The explanations from FIxLIP and Grad-ECLIP+ appear alternately in Part II, and the assignment of images is counterbalanced so that a given image appears the same number of times with FIxLIP or Grad-ECLIP+, without an explanation (in Part I) or with an explanation (in Part II), over all participants (see more details in the Appendix Sec. C).

We collected responses from 64 participants in total, and the results are shown in Tab. 4. When provided with the explanations, users' accuracy for judging the higher-scoring caption increased significantly to 66.9% from 57.0% when no explanation is provided (paired t-test, t(63)=-3.58, p<.001). Comparing the two explanation methods, the users had 4.4% higher accuracy when given the explanation maps from Grad-ECLIP+. The user study demonstrates that the gradient-difference-enabled Grad-ECLIP+ displays comparable ability to FIxLIP in faithfulness and usefulness for humans; both can effectively help users better understand the matching score prediction related to image-text interactions. However, Grad-ECLIP+ is more computationally efficient, making it amenable to real-time or interactive explanations.

Table 4: User's accuracy in identifying the better-matched image-caption pairs.

w/o maps	w/ maps	Paired t-test
57.0%	**66.9%**	p<.001
FIxLIP	Grad-ECLIP+	**Paired t-test**
64.7%	**69.1%**	p=0.281

5 Discussion and Conclusion

In this paper, we introduced a lightweight gradient-difference framework that efficiently derives second-order vision-language interaction explanations directly from existing first-order gradient-based attribution methods. By evaluating the gradient changes resulting from targeted concept or region masking, our framework delivers flexible model interpretability as shown in Fig. 4, ranging from targeted conditional heatmaps to comprehensive cross-modal interaction maps. Both quantitative evaluations and user studies demonstrate that our methods achieve performance comparable to state-of-the-art second-order approaches while drastically reducing computational overhead.

Limitations and broader impact. Our framework inherently inherits the limitations of the underlying first-order attribution methods upon which it is built. If a particular vision-language architecture or encoder style is incompatible with or poorly explained by a given first-order gradient method, our gradient-difference framework cannot be applied to that model either. However, at the same time, this dependency provides the advantage in terms of extensibility. Gradient-based first-order attribution methods have been widely adopted for interpreting deep vision and vision-language models, and new methods continue to be developed and improved to address the characteristics of different architectures and representation mechanisms. Our framework can naturally build upon these developments, which provides a general and efficient mechanism for extending existing first-order attribution techniques to capture cross-modal interactions.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We have not used generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, or proving mathematical claims, and generating synthetic datasets is not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involved human participants and was conducted in accordance with the ethical principles for responsible research. Participants were informed about the purpose and procedure of the study before participation and provided informed consent. Participants were compensated for their participation in accordance with the study protocol.

REPRODUCIBILITY STATEMENT

In the Appendix Sec. A, a detailed interpretation of gradient difference is provided. We promise that the code, including the implementation of Grad-CAM+, GAME+, Grad-ECLIP+, phrase-targeted insertion and deletion evaluation, pointing game evaluation, and dataset pre-processing for MS COCO, Flickr30k, and ImageNet, will be open-sourced.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, pp. 4190–4197, 2020.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

Hubert Baniecki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, and Przemyslaw Biecek. Explaining similarity in vision-language encoders with weighted banzhaf interactions. In *NeurIPS*, 2025.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pp. 397–406, 2021.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5540–5552, 2020.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Lucas Moeller, Pascal Tilli, Thang Vu, and Sebastian Padó. Explaining caption-image interactions in CLIP models with second-order attributions. *TMLR*, 2025.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

Fawaz Sammani, Boris Joukovsky, and Nikos Deligiannis. Visualizing and understanding contrastive learning. *IEEE Transactions on Image Processing*, 33:541–555, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *preprint arXiv:2303.15389*, 2023.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André S Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *NeurIPS*, 37:116355–116387, 2024.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, pp. 5764–5773, 2019.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pp. 2048–2057, 2015.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11941–11952, 2023.

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. Gradient-based visual explanation for transformer-based clip. In *ICML*, pp. 61072–61091, 2024.

Chenyang Zhao, Kun Wang, Janet H Hsiao, and Antoni B Chan. Grad-eclip: Gradient-based visual and textual explanations for clip. *TPAMI*, 2026.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pp. 696–712. Springer, 2022.

A INTERPRETATION OF GRADIENT DIFFERENCE

Let $S(\mathbf{x}, \mathbf{z})$ denote the image–text matching score of a VLM, where $\mathbf{x}$ represents the image-side feature and $\mathbf{z}$ represents the text-side representation. For a target phrase P, let $\mathbf{z}$ denote the representation of the original text and let $\mathbf{z}_{\setminus P}$ denote the representation after masking P. The effect of masking the phrase can therefore be represented as a finite perturbation

$$\Delta \mathbf{z}_P = \mathbf{z} - \mathbf{z}_{\setminus P}. \tag{9}$$

Our goal is to characterize how the contribution of an image region changes as a consequence of the presence of P. For an image feature $\mathbf{x}$, this quantity is naturally related to the mixed second-order derivative

$$\frac{\partial^2 S}{\partial \mathbf{x}\, \partial \mathbf{z}}, \tag{10}$$

which measures how the sensitivity of the matching score to the image feature changes with respect to the text representation. Consider the image gradient

$$\mathbf{G}(\mathbf{z}) = \frac{\partial S(\mathbf{x}, \mathbf{z})}{\partial \mathbf{x}}. \tag{11}$$

A first-order Taylor expansion of $\mathbf{G}$ around the original text representation $\mathbf{z}$ gives

$$\mathbf{G}(\mathbf{z}_{\setminus P}) \approx \mathbf{G}(\mathbf{z}) + \frac{\partial \mathbf{G}(\mathbf{z})}{\partial \mathbf{z}} \left(\mathbf{z}_{\setminus P} - \mathbf{z}\right). \tag{12}$$

Since

$$\frac{\partial \mathbf{G}(\mathbf{z})}{\partial \mathbf{z}} = \frac{\partial^2 S}{\partial \mathbf{x}\, \partial \mathbf{z}}, \tag{13}$$

we obtain

$$\mathbf{G}(\mathbf{z}) - \mathbf{G}(\mathbf{z}_{\setminus P}) \approx \frac{\partial^2 S}{\partial \mathbf{x}\, \partial \mathbf{z}} \left(\mathbf{z} - \mathbf{z}_{\setminus P}\right). \tag{14}$$

Consequently, the gradient difference

$$\Delta \mathbf{G}_P = \frac{\partial S(\mathbf{x}, \mathbf{z})}{\partial \mathbf{x}} - \frac{\partial S(\mathbf{x}, \mathbf{z}_{\setminus P})}{\partial \mathbf{x}} \tag{15}$$

is a finite-difference approximation to the action of the mixed second-order derivative on the finite text perturbation $\Delta \mathbf{z}_P$:

$$\Delta \mathbf{G}_P \approx \frac{\partial^2 S}{\partial \mathbf{x}\, \partial \mathbf{z}} \Delta \mathbf{z}_P. \tag{16}$$

This provides a formal interpretation of phrase-image interaction without explicitly constructing the full cross-modal Hessian. $\Delta \mathbf{G}_P$ is the finite-difference response of the image gradient to the phrase perturbation, and therefore captures the component of the second-order interaction associated with removing P.

The masking operation provides a finite perturbation. Specifically, the original text and the masked text constitute two evaluations of the VLM at two different points in the text-input space. Instead of evaluating the derivative of the image gradient with respect to the text representation analytically, we measure the change in the image gradient from the point $\mathbf{z}$ to another point $\mathbf{z}_{\setminus P}$. Thus, phrase masking converts a discrete linguistic intervention into a finite difference along the text dimension. When the perturbation is sufficiently local, the gradient change approximates the corresponding directional derivative; when the perturbation is not infinitesimal, it should more precisely be understood as a finite-difference estimate of the accumulated mixed interaction along the path between the two text representations. Hence, phrase masking provides a practical way to access second-order cross-modal information using only two first-order gradient evaluations.

A first-order visual explanation uses

$$\mathbf{G}_{\text{orig}} = \frac{\partial S(I, T)}{\partial \mathbf{A}} \tag{17}$$

to characterize image regions that affect the image–text matching score. We instead compute

$$\Delta \mathbf{G}_P = \frac{\partial S(I, T)}{\partial \mathbf{A}} - \frac{\partial S(I, T_{\setminus P})}{\partial \mathbf{A}}, \tag{18}$$

Table 5: Evaluation of localization ability using Pointing Game Accuracy and Energy (PGA and PGE) for phrase-targeted heat maps. The top-2 results are bolded and underlined.

	ViT-B/32						ViT-B/16					
	COCO		Flickr30k		ImageNet		COCO		Flickr30k		ImageNet	
Method	PGA	PGE	PGA	PGE	PGA	PGE	PGA	PGE	PGA	PGE	PGA	PGE
FIxLIP	81.28	57.04	77.17	56.05	92.21	68.78	-	-	-	-	-	-
ITSM	42.50	37.03	46.81	39.24	60.88	27.38	70.47	38.11	67.12	39.94	80.15	29.50
exCLIP	<u>85.82</u>	**69.41**	<u>78.49</u>	**64.46**	94.07	<u>81.62</u>	<u>86.06</u>	<u>63.53</u>	<u>78.30</u>	58.85	92.16	73.40
Grad-CAM+	64.88	58.45	66.05	59.51	<u>95.00</u>	**82.70**	69.62	61.75	69.82	<u>59.48</u>	**94.77**	**82.66**
GAME+	63.61	54.59	64.87	54.67	92.09	60.78	78.28	56.09	75.33	56.45	<u>94.50</u>	62.04
Grad-ECLIP+	**91.20**	<u>65.40</u>	**83.91**	<u>62.22</u>	**95.64**	67.83	**86.89**	**71.25**	**80.56**	**68.30**	94.37	<u>78.41</u>

which characterizes how the visual sensitivity changes when the phrase P is removed. Applying the original explanation operator to this gradient difference therefore transforms a first-order visual attribution into a phrase-conditioned interaction attribution:

$$H_P^+ = \mathcal{M}\left(A, \Delta\mathbf{G}_P\right), \tag{19}$$

where $\mathcal{M}$ denotes the procedure of an existing first-order method from gradient to heatmap. This leads directly to Grad-CAM+, GAME+, and Grad-ECLIP+ by replacing the original image gradient with the corresponding gradient difference.

A.1 Pointing Game

Following the pointing game test metric adopted in previous visual explanation works (Zhao et al., 2024; Moeller et al., 2025), we evaluate the localization ability of the phrase-targeted explanation maps for each method. Specifically, the ground truth bounding boxes corresponding to the object phrases in the captions are available through the annotations in the MS COCO and Flickr30k datasets, while for ImageNet, we regard the whole sub-image as the ground-truth region, as the red boxes shown on the images in Fig. 2. The Pointing Game Accuracy (PGA) counts a hit score if the location with the largest value in the visual explanation heat map lies within the object region defined by the bounding boxes, and Pointing Game Energy (PGE) calculates the proportion of heat map energy within the ground-truth mask versus the whole map.

The results for localization are shown in Tab. 5. Grad-CAM+ shows an advantage on the ImageNet dataset for both two models, exceeding exCLIP and FIxLIP, while for the more complicated image-caption cases in MS COCO and Flickr30k, Grad-ECLIP+ consistently achieves superior localization performance. Grad-ECLIP+ has the highest PGA on COCO (91.20%), Flickr30k (83.91%), and second-best PGE with ViT-B/32, and leads all evaluated methods with top PGA and PGE performance with ViT-B/16.

B Deletion and Insertion Curves

In this section, we display the deletion and insertion curves in the phrase-targeted explanation faithfulness evaluation, which supports the results in Sec. 4.1.

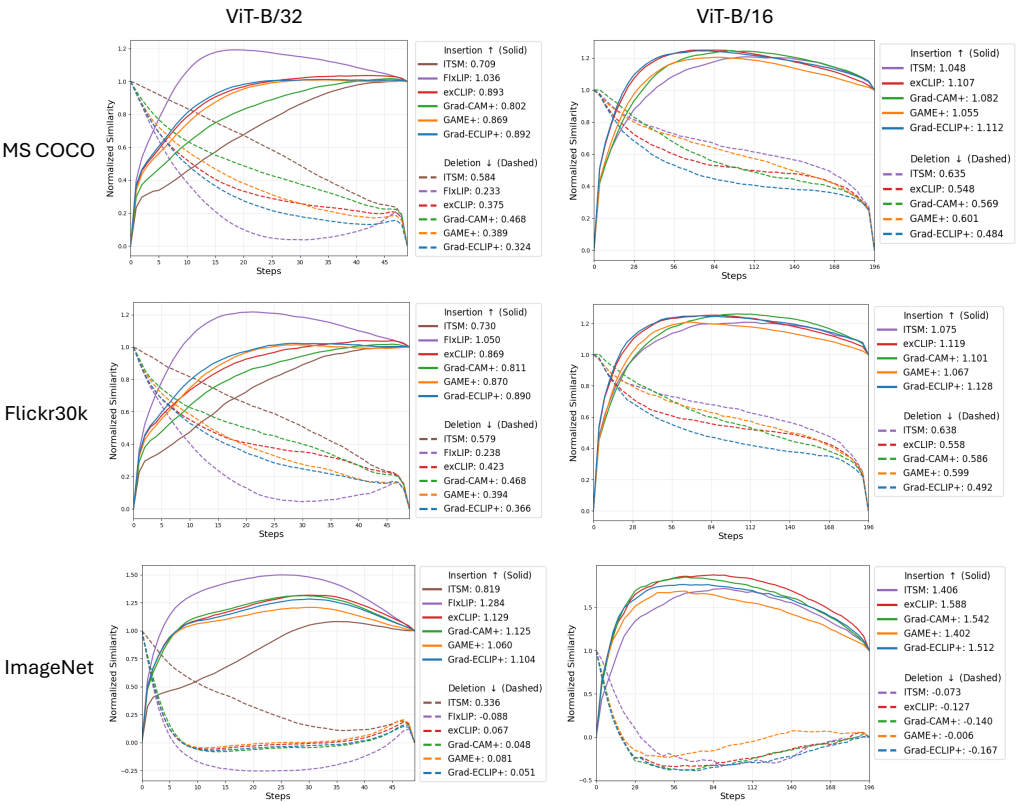


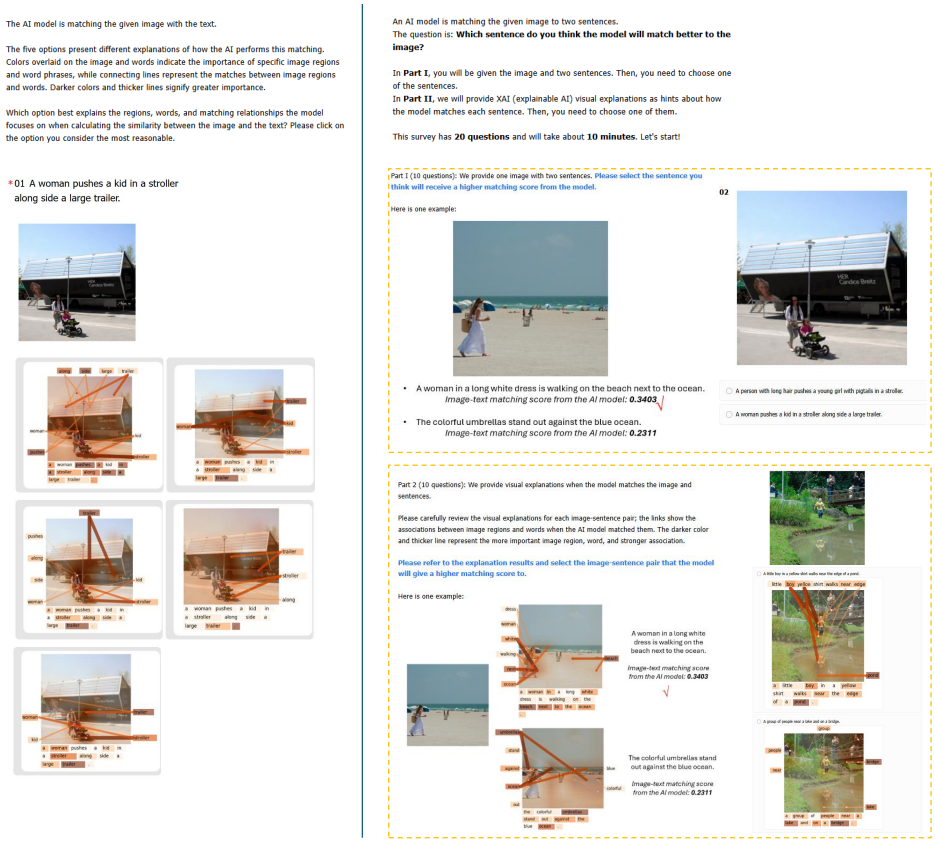
Figure 5: Deletion and insertion curves on MS COCO, Flickr30k, and ImageNet, with ViT-B/32 and ViT-B/16 CLIP encoders.

C More Details about User Studies

In this section, we display more details about the user studies. Fig. 6 gives the screenshots from the questionnaires we used for the plausibility study and simulation study, respectively.

For the plausibility study as shown in Fig. 6(a), the association explanation maps are generated by the five different methods, including FIxLIP, exCLIP, Grad-CAM+, GAME+, and Grad-ECLIP+. We normalize the interaction values for each method so the maximum is 1, then show only associations greater than 0.4, up to 2 strongest associations per word. The questionnaire contains 20 questions, and responses are collected from 16 participants in total.

For the simulation study as shown in Fig. 6(b), the 10 images in the condition without explanation maps (Part I) are displayed before the 10 samples in the condition with explanation maps (Part II), in case users may obtain insights about the model by viewing the explanation samples. To compare the associations between the image regions and words from two captions, we normalize the interaction values from the explanations of the two captions together by making the maximum 1. As the illustration shown in Fig. 7, we build four questionnaires in total to ensure the data assignment counterbalance. We prepare two questionnaires (Q1 and Q2) by swapping the images in Part I and Part II. Since we also compare the two explanation methods, the explanations for the 10 images in Part II of each questionnaire are alternately generated by FIxLIP and Grad-ECLIP+, then another two questionnaires (Q3 and Q4) are prepared to swap the method used on each image in Q1 and Q2. For example, the explanation for the 11th image in Q1 is from FIxLIP, then it is from Grad-ECLIP+ in Q3; The explanation for the next (12th) image is changed to Grad-ECLIP+ in Q1, and the corresponding one in Q3 is from FIxLIP. Overall, the first 10 images and the second 10 images are swapped between Q1 and Q2, and between Q3 and Q4; the adopted explanation methods are swapped between Q1 and Q3, and between Q2 and Q4. Finally, we collect the responses from 16



(a) The Plausibility Study (b) The Simulation Study

Figure 6: The interface for (a) the plausibility study and (b) the simulation study. We give an introduction to the explanation maps for the model and state the questions and required operations in each questionnaire. In (b) the simulation study, there are two parts; we give an example at the beginning of each part to clarify the requirements for the users.

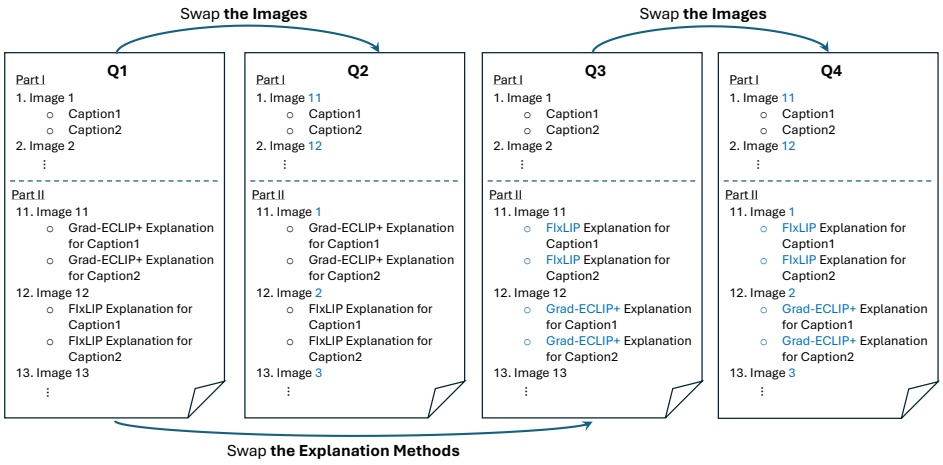


Figure 7: The illustration of counterbalanced assignment in the user study.

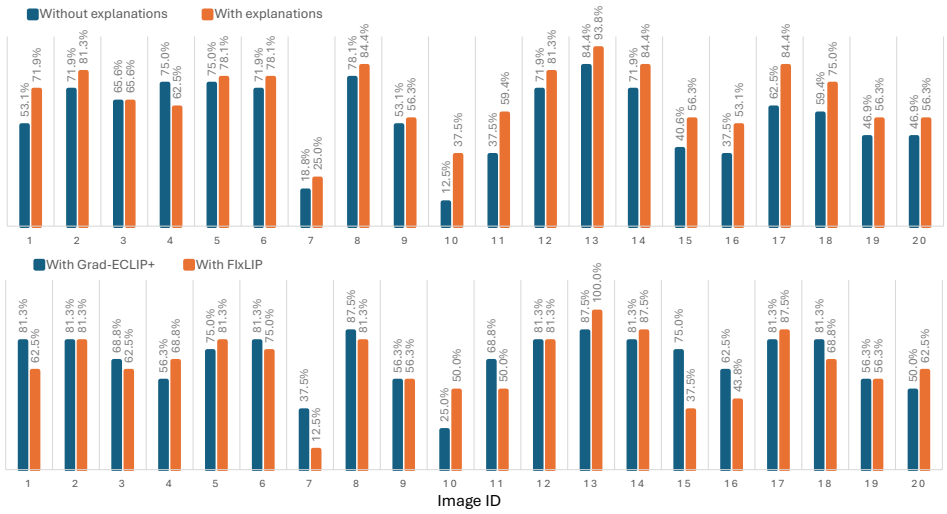
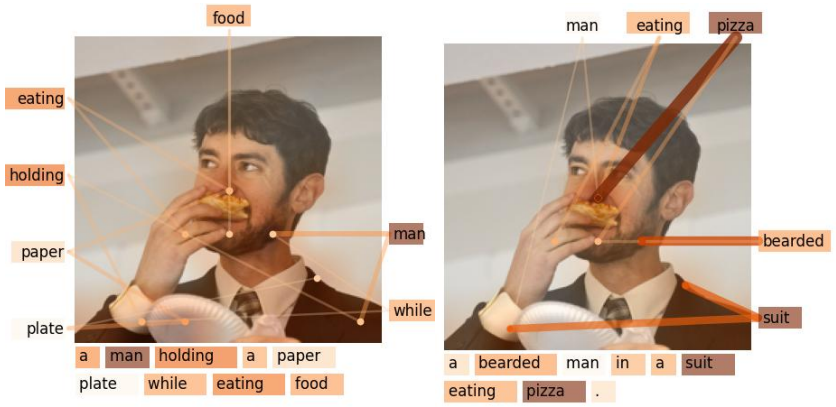


Figure 8: The users' accuracies for each image, comparing (top) without explanations (mean 56.7%) and with explanations (mean 67.0%); (bottom) with Grad-ECLIP+ (mean 68.8%) and with FIxLIP (65.3%).



(a) Match score: **0.304** (b) Match score: **0.351**

Figure 9: An example in the questionnaire where the users' ratio of correctly choosing (b) significantly increases when showing the Grad-ECLIP+ explanation map (from 59.4% to 81.3%).

different participants for each questionnaire, for a total of 64 participants for the simulation study. We make sure the assignment is counterbalanced so that each image is shown to the same number of users with the explanation from FIxLIP and Grad-ECLIP+ under both conditions in Part I and Part II.

In the Sec. 4.2 Simulation Study, we present the averaged accuracy per user (over 64 participants) under the conditions of being given explanation maps or not being given the maps. In this section, we also provide the statistical results from the perspective of users' accuracy with each image. The detailed accuracies for each image under different conditions are shown in Fig. 8, from which we can see that with the explanation maps from Grad-ECLIP+ and FIxLIP can significantly help users to recognize the better-matched caption. In Fig. 9, we show an example in the questionnaire with the Grad-ECLIP+ explanation maps for matching two captions with the same image. It can be seen that the associations between image region with several words in Fig. 9(b) are significantly higher, such as "pizza", "bearded", and "suit", which give insights to users to correctly choose the caption with a higher match score. For this question, 81.3% of participants correctly select the right caption in (b) when given the displayed explanation maps, compared with 59.4% when not showing the maps.

D VISUALIZATIONS ON VIT-B/32 (CLIP) AND VIT-B/16 (CLIP AND EVA-CLIP)

We apply Grad-ECLIP+ on ViT-B/32 from CLIP Radford et al. (2021), and ViT-B/16 from both CLIP and EVA-CLIP Sun et al. (2023). Fig. 10 shows some visualizations. We can see that more detailed information in the sentence can be matched with models of ViT-B/16 than ViT-B/32, such as the "sandals" in the last row.

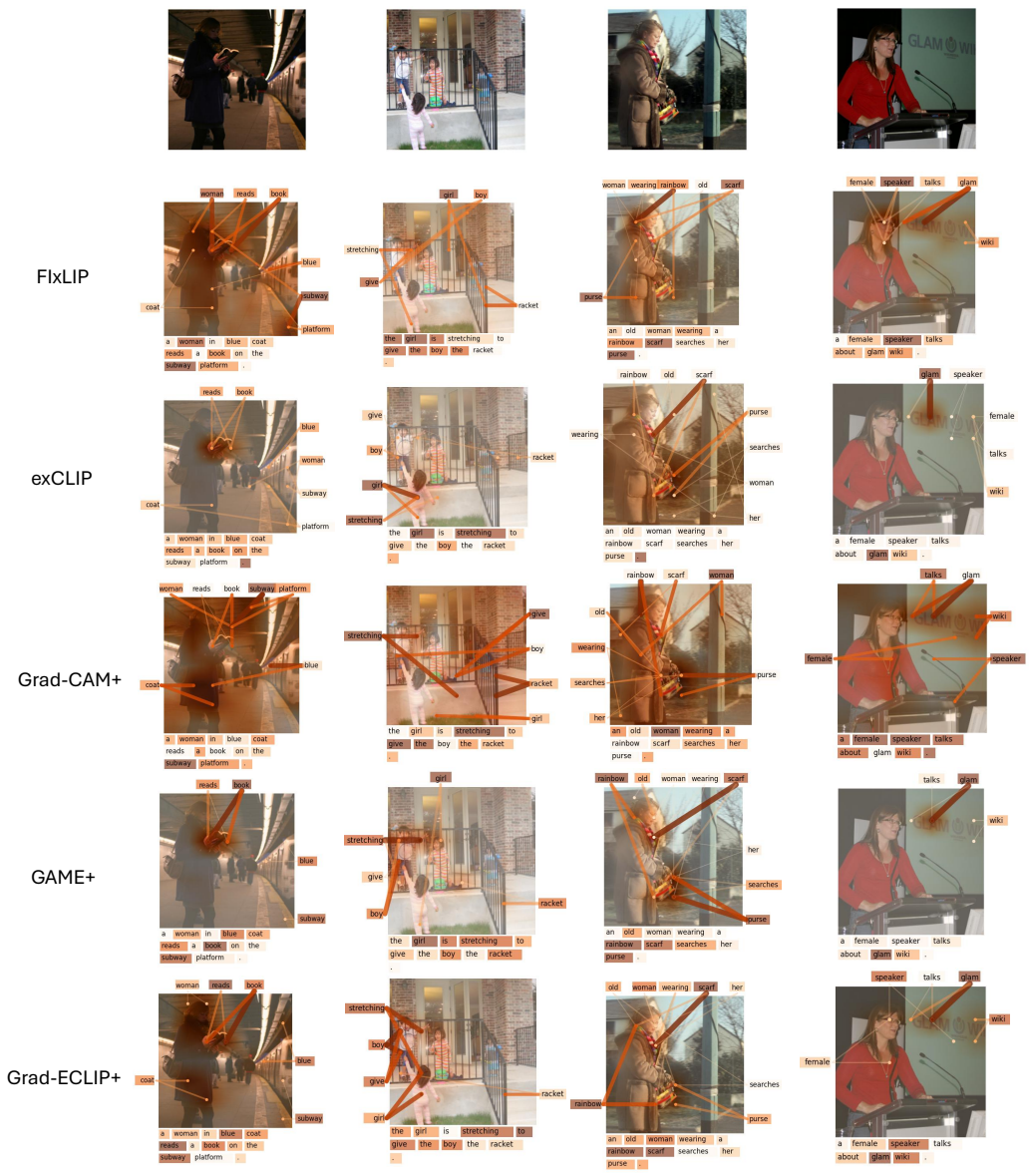


Figure 10: Visualization examples of association explanation using Grad-ECLIP+ on ViT-B/32 from CLIP, and ViT-B/16 from both CLIP and EVA-CLIP.

E MORE VISUALIZATION COMPARISONS WITH THE STATE-OF-THE-ART

Fig. 11 shows more association explanation map comparisons between FIxCLIP, exCLIP, and the methods enabled by our gradient difference framework, including Grad-CAM+, GAME+, and Grad-ECLIP+.

Although the different explanation methods produce distinct attribution patterns, several consistent and interesting behaviors can be observed for the CLIP model. For example, in (a), the word "blue" is not primarily associated with the woman's coat as expected, but instead with the blue top of the train, suggesting that the model may rely on visually salient word–region associations rather than the intended semantic context. In (b), "racket" is repeatedly associated with the stair handrail rather than the tennis racket held by the boy, indicating a mismatch between the textual concept and the relevant object. In (c), both the "scarf" and "purse" are associated with the rainbow-colored regions, suggesting that the model may confuse visually similar attributes across different objects. Finally, in (d), the text "glamwiki" on the presentation slide attracts stronger attention than the female speaker



(a) A woman in blue coat reads a book on the sub way platform.

(b) The girl is stretching to give the boy the racket.

(c) An old woman wearing a rainbow scarf searches her purse.

(d) A female speaker talks about glamwiki.

Figure 11: More visualization comparisons between FIxCLIP, exCLIP, and the methods enabled by our gradient difference framework, including Grad-CAM+, GAME+, and Grad-ECLIP+ .

herself, demonstrating that prominent textual content in the image can distract the model from the semantically relevant person.

Overall, our gradient-difference framework provides a natural bridge between existing first-order gradient-based explanation methods and second-order interaction explanations. By applying our efficient framework, the common patterns revealed by the explanations can provide complementary evidence for understanding and analyzing model behavior, thereby helping users and developers better identify potential weaknesses and interpret the model's decision-making process.