# Explaining Cross-Modal Model Behavior with Gradient-Estimation-Based Feature Interaction

**Yi Cai** [*†]
School of Computer Science & Engineering
*Nanyang Technological University*
Singapore
yi.cai@ntu.edu.sg

**Xinpeng Li** [*]
School of Artificial Intelligence
*Nanjing University*
Nanjing, China
lixp@lamda.nju.edu.cn

**Gerhard Wunder**
Dept. of Math. and Comp. Science
*Freie Universität Berlin*
Berlin, Germany
g.wunder@fu-berlin.de

## Abstract

Multi-modal machine learning has become a dominant paradigm for integrating heterogeneous modalities and achieving unprecedented representational capabilities; however, this multi-modal structure poses unique explainability challenges. While standard feature attribution methods characterize individual feature contributions and can be applied to each modality, such first-order explanations are insufficient to capture cross-modal feature interactions, particularly concerning the underlying concept alignment between modalities. To bridge the gap, this paper studies higher-order feature interactions in multi-modal settings and proposes Gradient-Estimation-Based Feature Interaction (GEFI). GEFI is derived from the proxy gradient estimation framework, and we establish its theoretical equivalence to the Shapley Interaction Index at arbitrary orders. Beyond the general theory, we instantiate GEFI-PM for CLIP-like dual-encoder models, enabling efficient estimation of interactions between visual and textual features. Unlike existing cross-modal interaction methods that operate on fixed image patches, GEFI-PM delivers fine-grained interactions at pixel resolution while maintaining computational efficiency, yielding more expressive explanatory insights.

## 1 Introduction

Multi-modal models have achieved remarkable success in integrating heterogeneous information sources. A popular paradigm trains dual encoders to project inputs from different modalities into a shared embedding space, facilitating subsequent processing [1]. Building on these advances, such models have been deployed for complex downstream tasks, such as image-text retrieval [2], visual question answering [3], and embodied AI [4]. Their success is driven by their ability to exploit combined cross-modal evidence rather than treating each modality in isolation; this integration also poses a unique challenge for existing explainability methods [5, 6].

[*]These authors contributed equally to this work.
[†]Yi Cai was affiliated with Freie Universität Berlin when the work was carried out.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Feature attribution is a widely studied explainability task due to its expressive and illustrative result formatting. Given a model $f(\cdot)$, an explicand $\boldsymbol{x} \in \mathbb{R}^d$, and a baseline[1] $\mathring{\boldsymbol{x}}$, feature attribution aims to quantify the contributions $\xi_i$ of individual input features, which can be formally described as:

$$A_f(\boldsymbol{x}, \mathring{\boldsymbol{x}}) \hookrightarrow \{\xi_i\}_{i=1}^d \in \mathbb{R}^d$$

While many existing feature attribution methods can be adapted to multi-modal settings, these *first-order explanations* fall short of capturing cross-modal interactions, which can profoundly influence model decision-making. Specifically, an attribution method can highlight the most relevant input components in each modality to a decision; however, its results do not possess the explanatory power to reveal whether and how these salient features are aligned across observed modalities.

Acknowledging the limitations of first-order explanations, one direction of explainability research emphasizes higher-order feature interactions [7]. Analogous to the definition of the attribution task, the goal of feature interaction can be formalized as:

$$I_f(\boldsymbol{S}, \boldsymbol{x}, \mathring{\boldsymbol{x}}) \hookrightarrow \{\xi_{\boldsymbol{S},i}\}_{i \notin \boldsymbol{S}} \in \mathbb{R}^{d-|\boldsymbol{S}|} \tag{1}$$

where $\boldsymbol{S} \subset \{1, \ldots, d\}$ denotes the feature subset that the interactions anchor on[2]. When $|\boldsymbol{S}| = 1$, Eq. (1) describes the pairwise interaction task, the most common interaction type studied in the literature [5, 6]. Furthermore, the task definition collapses to feature attribution when no anchor feature is selected, *i.e.*, $\boldsymbol{S} = \varnothing$; consequently, feature attribution can be seen as a special case of feature interaction. Compared to first-order explanations, feature interactions provide more detailed insights by highlighting contextual sensitivities through inter-dependencies between features. This form of explanation can be particularly powerful for explaining model functioning in the context of multi-modality, especially for models like Contrastive Language-Image Pre-training (CLIP) [1], where cross-modal interactions are inevitable to ensure image-text alignment.

Following this line of research, we propose *Gradient-Estimation-Based Feature Interaction* (GEFI), a general explanation framework that extends gradient-estimation-based explanation [8, 9] from first-order attribution to interactions of arbitrary orders. We theoretically establish the equivalence between GEFI and the Shapley Interaction Index [10] at any order, grounding the proposed framework on a rigorous foundation within the existing literature. In principle, GEFI[3] supports flexible reuse of one observation set to investigate interactions of arbitrary order across various feature subsets. Building on this framework, we further instantiate GEFI-PM, a computationally efficient variant tailored to the product measure associated with dual-encoder model architectures. This efficiency gain enables pixel-level quantification of feature interactions beyond the fixed image patches considered in existing work, offering more fine-grained insights into cross-modal concept alignments. Through a series of carefully designed experiments dedicated to feature interaction validation, we demonstrate substantial improvements in explanation quality over SOTA competitors, highlighting the benefits of the finer granularity enabled by GEFI-PM.

## 2 Related Work

**Feature attribution** quantifies the marginal contribution of individual features to an inquired model outcome. A popular regime uses gradients to determine attribution scores [11]. Smilkov et al. [12] and Sundararajan et al. [13] improve upon this by mitigating the shattering (via a Gaussian process) and saturation problems (via path integrals), respectively; combining both methods further enhances explanation quality and empirically demonstrates competitive performance [14]. Subsequent gradient-based works study fine-grained problems such as computational efficiency [15], baseline selection [16, 17], and the specification of integral paths [18]. In parallel with white-box solutions strictly depending on internal model access, black-box approaches assess feature importance by probing model responses to input perturbations [19]. A popular trend in this category is to refine sampling and information aggregation strategies [20, 21, 22]. Recent work [8, 9] bridges the white-box and black-box approaches via gradient estimation [23] to derive high-quality explanations

[1] Standard machine learning models cannot process incomplete inputs; a baseline is therefore introduced as a reference point to simulate feature absence, facilitating the computation of feature contributions relative to their presence.

[2] We use the term *anchor* in line with common practice in feature interaction visualization, where each view typically focuses on interaction scores associated with one selected feature or feature subset.

[3] Code is available at: `https://github.com/caiy0220/GEFI`

on a sampling basis. Nevertheless, these attribution methods are limited to providing first-order information, which cannot reveal inter-feature dependencies for context-aware understanding.

**Feature interaction** addresses this limitation of feature attribution by extending the investigation to higher-order dependencies among feature subsets. A prominent line of research formalizes feature interactions through cooperative game theory. Building on the Shapley Interaction Index (SII) [10], several axiomatic formulations extend feature interaction analysis with cooperative game theory [7, 24]. Conceptually, these methods extend the perturbation paradigm of feature attribution to estimate joint effects from feature coalitions. Subsequent work consolidates these approaches within a computational framework [25, 26], and the underlying principles are further extended to analyze multi-modal similarity in Vision-Language Models [6]. Other black-box methods estimate interactions via surrogate modeling; for example, Joukovsky et al. [27] uses a bilinear surrogate to approximate observed model behavior, and Butler et al. [28] improves query efficiency through sparse Fourier transforms. On the other hand, white-box approaches exploit gradient information to capture higher-order effects, such as integrated second-order partial derivatives for pairwise interactions [29] and integrated Jacobians specified for dual-encoder architectures [5]. Despite these advancements, notable challenges persist in feature interaction research. We highlight in particular computational complexity, which remains a bottleneck for deriving fine-grained interaction results. Most existing solutions, if not all, operate on fixed image patches as a compromise due to the exponentially growing interaction space. While image patches are natural input components of vision transformers [30], treating them as indivisible elements ignores how pixel-level changes affect patch representations, thereby limiting explanation expressiveness. We show in this paper that GEFI overcomes this limitation and provides pixel-level cross-modal interactions, significantly improving explanation quality without incurring prohibitive computational overhead.

## 3 Gradient-Estimation-Based Feature Interaction

### 3.1 Preliminary

Let $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ be the target model function, and let $\boldsymbol{x}, \mathring{\boldsymbol{x}} \in \mathbb{R}^d$ denote the explicand and the baseline, respectively. This paper considers a black-box setting, assuming that the model is accessible only through its input and output interfaces. Gradient-estimation-based explanation aims to approximate gradient information from observations collected via model queries. For this purpose, gradient estimation introduces a search distribution $\boldsymbol{\pi}(\cdot|\boldsymbol{x})$ conditioned on the explicand and defines the smoothed objective:

$$J(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{z}|\boldsymbol{x})}[f(\boldsymbol{z})] = \int f(\boldsymbol{z})\boldsymbol{\pi}(\boldsymbol{z}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{z} \tag{2}$$

The score-function gradient estimator can be derived under the mild assumption of continuous differentiability of both $f(\cdot)$ and $\boldsymbol{\pi}(\cdot|\boldsymbol{x})$ [23]:

$$\boldsymbol{\eta}_{\boldsymbol{x}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} J(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{z}|\boldsymbol{x})}[f(\boldsymbol{z}) \cdot \nabla_{\boldsymbol{x}} \log \boldsymbol{\pi}(\boldsymbol{z}|\boldsymbol{x})] \tag{3}$$

Cai et al. [9] generalizes this idea by constructing the search distribution with proxy variables $\boldsymbol{\alpha} \in [0, 1]^d$, which resolves the continuity assumption in the feature space. With each element of $\boldsymbol{\alpha}$ modeling the presence probability of the corresponding feature, these proxies convert the sampling process from continuous perturbations to a series of binary sampling under $\mathrm{Bernoulli}(\boldsymbol{\alpha})$, considering only feature presence (taking the value of $\boldsymbol{x}$) and absence (taking the value of $\mathring{\boldsymbol{x}}$). With the search distribution defined, the feature attribution method GEFA [9] is derived from the proxy gradient estimator through the path integral along the diagonal of the proxy space[4]:

$$\boldsymbol{\xi}^{\text{GEFA}} := \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}|\alpha)} \left[ f(\boldsymbol{z}) \cdot \left( \frac{\boldsymbol{\epsilon} - \alpha}{\alpha(1-\alpha)} \right) \right] \mathrm{d}\alpha \tag{4}$$

where $\boldsymbol{z} = \boldsymbol{x} \odot \boldsymbol{\epsilon} + \mathring{\boldsymbol{x}} \odot \bar{\boldsymbol{\epsilon}}$ is a shorthand for the sampling process through a binary mask $\boldsymbol{\epsilon}$, with $\odot$ denoting element-wise multiplication and $\bar{\boldsymbol{\epsilon}} = \mathbb{1}_d - \boldsymbol{\epsilon}$ denoting the complement of $\boldsymbol{\epsilon}$. Each mask element $\epsilon_i$ is sampled from a Bernoulli distribution configured by the proxy path indicator $\alpha$. As such, proxy gradient estimation shifts the sampling process from a concrete feature space to the proxy space, enhancing its applicability to broader input modalities, including discrete features such as text.

[4]The proxy variables $\boldsymbol{\alpha}$ simplify to a scalar $\alpha$, as all proxies share the same value along the diagonal of the proxy space.

### 3.2 Derivation of GEFI

Next, we demonstrate how the proxy gradient estimator can be extended to quantify higher-order interactions that highlight model activations arising from feature inter-dependencies. Specifically, the second-order feature interaction between a feature pair $(i, j)$ can be derived by simultaneously applying the proxy gradient estimator with respect to both $x_i$ and $x_j$ based on Eq. (4), yielding *Gradient-Estimation-Based Feature Interaction* (GEFI):

$$\xi_{i,j}^{\text{GEFI}} := \underbrace{\int_0^1 \overbrace{\mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}|\alpha)}\Big[f(\boldsymbol{z}) \cdot \frac{(\epsilon_i - \alpha)(\epsilon_j - \alpha)}{\alpha^2(1-\alpha)^2}\Big]}^{\text{Second-order Proxy Gradient Estimator}} \, \mathrm{d}\alpha}_{\text{Proxy Path Integral}} \tag{5}$$

**Theorem 3.1.** *The feature interaction score defined in Eq.* (5) *is equivalent to the second-order SII.*

Following this observation, we find that GEFI can be further generalized to a $k$th-order form that quantifies interaction effects among a given feature subset $\boldsymbol{S} \cup \{j\}$:

$$\xi_{\boldsymbol{S},j}^{\text{GEFI}} := \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}|\alpha)}\Big[f(\boldsymbol{z}) \cdot \frac{\prod_{l \in \boldsymbol{S} \cup \{j\}}(\epsilon_l - \alpha)}{\alpha^k(1-\alpha)^k}\Big] \, \mathrm{d}\alpha \tag{6}$$

where the interaction order $k$ is implicitly specified by the cardinality of the interacting feature subset, namely $|\boldsymbol{S} \cup \{j\}| = k$ as denoted by the subscript.

**Theorem 3.2.** *The general form of GEFI defined in Eq.* (6) *is equivalent to the $k$th-order SII; furthermore, in the special case where $\boldsymbol{S} = \varnothing$, it reduces to the Shapley value for feature attribution.*

The proofs for Theorem 3.1 and Theorem 3.2 are provided in Appendix A.1. According to Eq. (6), interactions at different orders can be estimated by reusing one observation set, since the required queries are sampled from the same distribution. Therefore, the proposed approach incurs no additional query budget when the analysis is extended to different feature subsets or higher orders. Given this paper's primary focus on pairwise feature interaction, the subsequent discussion of GEFI centers on second-order interactions unless explicitly specified otherwise.

## 4 Cross-Modal Feature Interaction - A Case Study on CLIP

### 4.1 Interaction under Product Measure

While GEFI is readily applicable to quantifying feature interactions, the number of queries required to obtain reliable estimates in high-dimensional spaces remains a bottleneck to its practical applicability. To this end, this section considers dual-encoder models such as CLIP and develop a query-efficient variant GEFI-PM by explicitly leveraging their product-measure structure. CLIP-like models output cosine similarities between image and text embeddings produced by the respective encoders. Given a query budget $n = n_{\text{img}} + n_{\text{txt}}$ allocated to the image and text encoders, the product measure yields an observation matrix $\boldsymbol{V} \in \mathbb{R}^{n_{\text{img}} \times n_{\text{txt}}}$ by evaluating all image-text combinations. However, the standard GEFI formulation is restricted to utilize paired observations corresponding to the main diagonal of $\boldsymbol{V}$, leaving the remaining combinations unexploited.

To fully leverage the available observations, we refine Eq. (5) by decoupling the proxy path integrals in the image and text spaces, which facilitates aggregation of information represented by the full observation matrix. We denote the decoupled *product-measure* version as GEFI-PM. Let $\boldsymbol{\epsilon}_{\text{img}}$ and $\boldsymbol{\epsilon}_{\text{txt}}$ denote image and text masks sampled from the distributions determined by the decoupled proxy path indicators $\alpha$ and $\beta$ for each space. The cross-modal interaction determined by GEFI-PM can be formulated as follows considering an image-text feature pair $(i, j)$:

$$\xi_{i,j}^{\text{GEFI-PM}} := \int_0^1 \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)}\left[\mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)}\Big[f(\boldsymbol{z}) \cdot \frac{(\epsilon_{\text{img},i} - \alpha)(\epsilon_{\text{txt},j} - \beta)}{\alpha(1-\alpha)\beta(1-\beta)}\Big]\right] \mathrm{d}\alpha \, \mathrm{d}\beta \tag{7}$$

where a query $\boldsymbol{z}$ is generated from a mask that concatenates the samples prepared for each modality, *i.e.*, $\boldsymbol{\epsilon} = \text{concat}(\boldsymbol{\epsilon}_{\text{img}}, \boldsymbol{\epsilon}_{\text{txt}})$.

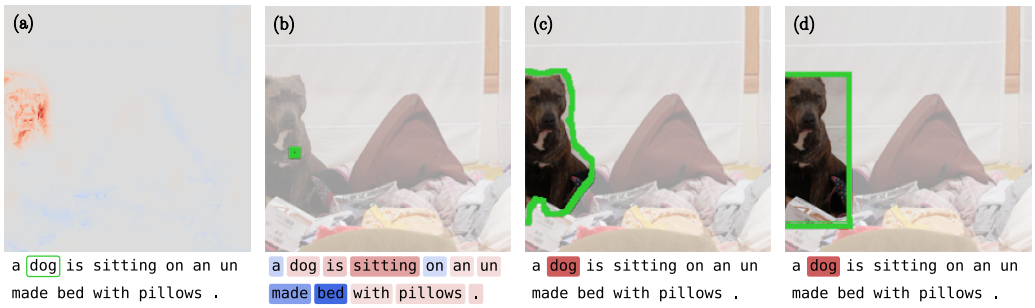



Figure 1: Illustration of the asymmetric semantic power of individual features in cross-modal interaction. The anchor feature is highlighted in green, interaction scores are indicated with red/blue for positive/negative. (a) Interactions between the selected token and pixels highlight the concept of "dog" in the image space. (b) The semantic power of one individual pixel (middle left) is much weaker than a token, yielding less comprehensible results. (c, d) Interactions anchored on visual feature subsets (selected by segmentation or bounding box) improve the expressiveness of the results, demonstrating the aligned concept also from the interactions anchored on image components.

**Proposition 4.1.** *GEFI-PM formulates a probabilistic interaction index induced by a product coalition measure across modalities.*

**Remark 4.2.** *The exact SII can be recovered from Eq.* (7) *by applying a rescaling factor determined by the coalition cardinality of each modality, at the cost of reduced effective sample size.*

Remark 4.2 is a direct consequence of Proposition 4.1, as both GEFI-PM and SII belong to the family of probabilistic interaction indices [31]. This enables recovery of SII by reweighting the probability distribution toward synchronized coalitions. That said, the correction will re-concentrate mass to the main diagonal of the observation matrix, which may cause performance degradation due to the reduced effective sample size. We provide further discussion regarding this point in Appendix A.2.

**Proposition 4.3.** *GEFI-PM implements a uniform probability distribution over modality-wise cardinality pairs and satisfies the linearity, symmetry, and dummy properties.*

Compared with the general GEFI framework, GEFI-PM adopts an alternative proxy path to accommodate practical computational constraints. Although this modification introduces bias relative to SII, it substantially reduces estimation variance with more exhaustive exploitation of available observations, resulting in improved empirical performance through a principled finite-budget bias-variance trade-off.

### 4.2 Interaction Anchored on Feature Subsets

So far, we have presented a solution that is efficient for computing cross-modal interactions in any CLIP-like dual-encoder architecture. Following Eq. (7), it is straightforward to determine interactions anchored on text components, as individual tokens possess clear semantic meaning and therefore yield perceptible interaction results. However, this intuition does not transfer directly to the image space, where individual pixels play substantially weaker semantic roles. A single pixel carries very little semantic information, posing a challenge for extracting meaningful insights from the quantitative results, as illustrated in Figure 1.

To address this challenge in image space, we extend the definition of image anchors from individual pixels to pixel subsets. Here, a subset $\boldsymbol{S}$ is a user-defined collection of pixels in the input image, such as a region representing a concrete object. Recalling that the proxy variable specifies the present probability of a feature, we can extend this probabilistic interpretation to feature subsets by considering the proportion of active features within the subset, namely $p_{\boldsymbol{S}} = \sum_{i \in \boldsymbol{S}} \epsilon_i / |\boldsymbol{S}|$. We update the GEFI-PM formulation accordingly as follows:

$$\xi^{\text{GEFI-PM}}_{\boldsymbol{S},j} := \int_0^1 \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)} \left[ \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)} \left[ f(\boldsymbol{z}) \cdot \frac{(p_{\boldsymbol{S}} - \alpha)(\epsilon_{\text{txt},j} - \beta)}{\alpha(1-\alpha)\beta(1-\beta)} \right] \right] \mathrm{d}\alpha \, \mathrm{d}\beta \tag{8}$$

This definition collapses to the standard binary case where $\epsilon_i \in \{0, 1\}$ when the subset is a singleton $\boldsymbol{S} = \{i\}$, consistent with the cross-modal interaction described in Eq. (7). Compared with the fixed

image patches commonly considered in the feature interaction literature [6, 26, 28], $\boldsymbol{S}$ offers a more flexible definition. Beyond standard image patches, it can also represent arbitrary pixel combinations[5], such as regions specified by bounding boxes or segmentation masks, thereby enabling finer semantic granularity. Analogously, the feature subset concept extends naturally in the text space for interactions anchored on token spans; we defer this formulation to Appendix C for brevity.

It is noteworthy that this subset-anchored variation still represents pairwise feature interactions, since the subset $\boldsymbol{S}$ is treated as a single unit. This is distinct from the higher-order form in Eq. (6), where the elements of $\boldsymbol{S}$ are considered individually and aggregated through a product operator.

### 4.3 Computational Consideration

With pairwise interactions defined, we now turn to computational efficiency. Under a limited query budget $n$, Eq. (5) can be approximated by a Monte-Carlo estimator, starting with initializing a set of $n$ query masks from $\boldsymbol{\epsilon} \sim \text{Bernoulli}(\alpha)$ [9]. For these masks, a weighting matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$ is constructed, where each element denotes the per-observation weighting factor $\frac{(\epsilon_i - \alpha)}{\alpha(1-\alpha)}$ for a feature. Each factor is determined by the corresponding mask value and the path indicator that specifies the sampling distribution. Each row of $\boldsymbol{Q}$ corresponds to a single mask $\boldsymbol{\epsilon}$ to generate one query $\boldsymbol{z}$, together reflecting the query budget $n$. Each column contains the weights for a specific feature over the sampling path $\alpha \sim \mathcal{U}_{[0,1]}$, with the number of columns matching the feature space dimensionality. Since this initialization is independent of specific explicands, it can be treated as a one-time offline cost and reused across different explicands. Additionally, the offline mask initialization allows the incorporation of more complicated sampling strategies, *e.g.*, antithetic sampling [32] and mask resizing [21], to further improve query effectiveness. We defer this discussion to Appendix G.1 as an orthogonal research direction since it deviates from the main focus of this paper.

Let $d_{\text{img}}$ and $d_{\text{txt}}$ denote the dimensionalities of the image and text spaces, respectively, with $d_{\text{img}} + d_{\text{txt}} = d$. The aforementioned initialization scheme can be adapted to Eq. (7), where the weighting terms are split into two matrices, $\boldsymbol{Q}_{\text{img}} \in \mathbb{R}^{n_{\text{img}} \times d_{\text{img}}}$ and $\boldsymbol{Q}_{\text{txt}} \in \mathbb{R}^{n_{\text{txt}} \times d_{\text{txt}}}$, to account for the decoupled sampling across modalities. Computing all cross-modal interactions directly would incur a time complexity of $\mathcal{O}(n_{\text{img}} n_{\text{txt}} \cdot d_{\text{img}} d_{\text{txt}})$, as it requires iterating over all mask pairs and the corresponding entries in the observation matrix $\boldsymbol{V}$. However, this cost can be significantly reduced by reorganizing the information aggregation procedure. Since each row of $\boldsymbol{V}$ corresponds to a text mask and each column to an image mask, the cross-modal interaction matrix can be estimated as:

$$\boldsymbol{\xi}^{\text{GEFI-PM}} \approx (\boldsymbol{V} \times \boldsymbol{Q}_{\text{img}})^{\top} \times \boldsymbol{Q}_{\text{txt}} \in \mathbb{R}^{d_{\text{img}} \times d_{\text{txt}}} \tag{9}$$

where each element $\xi_{i,j}^{\text{GEFI-PM}}$ represents the cross-modal interaction between the image-text feature pair indexed by $(i, j)$. This structured aggregation reduces the time complexity to $\mathcal{O}(n_{\text{txt}} d_{\text{img}} \cdot (n_{\text{img}} + d_{\text{txt}}))$, relying on matrix products at a complexity comparable to computing the initial cosine similarity scores. Compared to a brute-force implementation, Eq. (9) lowers computational complexity by one order of magnitude and provides a highly parallelizable matrix-product structure for efficient real-time execution on modern GPUs.

## 5 Interaction Evaluation

To quantitatively evaluate the quality of visual feature interaction conditioned on textual components and to facilitate direct comparison with SOTA approaches, we follow common practices in cross-modal explanation evaluation and consider two schemes: *object localization* and *input perturbation*.

**Object Localization** [5] assesses how well the interactions between image features and a selected textual token align with the ground-truth annotation associated with that token. Within the localization scheme, we employ two evaluation metrics[6]: *Point Game Accuracy (PGA)* [6, 34, 35] and *Intersection over Union (IoU)*. PGA is a low-granularity metric that reflects how often an explainer points the weighted center of its most interactive region into the ground-truth mask; IoU provides a more detailed view of the exact spatial overlap between the two regions. Throughout the experiment, the explainer-determined region is defined by the top-$k$ features with the highest intersection scores,

[5] Pixels within $\boldsymbol{S}$ are not necessarily adjacent.

[6] We have also considered *Point Game Energy (PGE)* as a metric; the experimental results are deferred to Appendix E.1 due to observed inconsistencies in metric implementations [5, 33, 34].

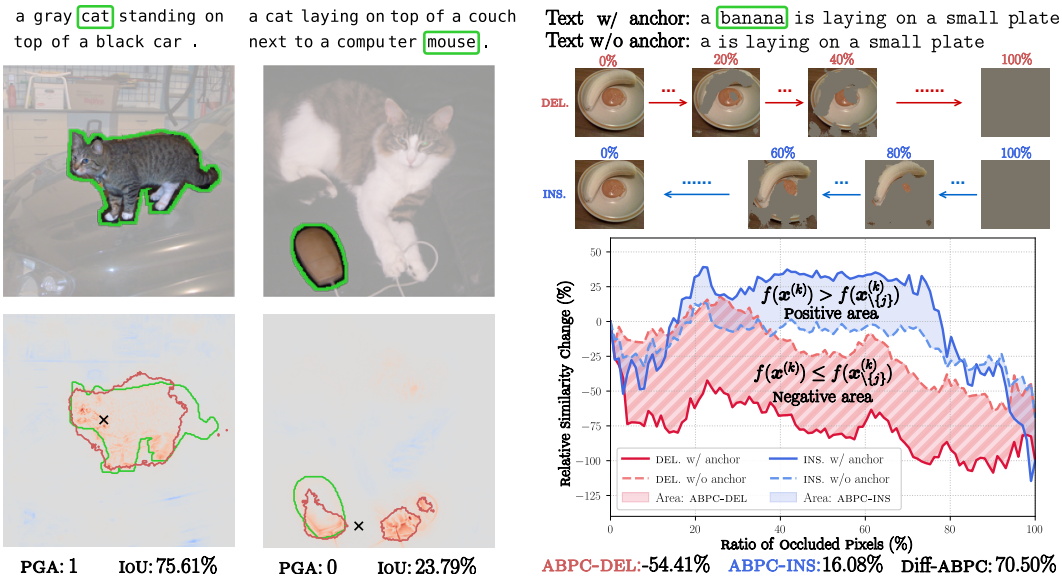



Figure 2: Illustrative examples of the evaluation schemes. **(Left) Object localization** assesses the alignment between the most interactive region and the ground truth for a designated anchor token. PGA measures how often the weighted center of an explainer-determined region falls within the ground truth region; IoU quantifies the overlap between the two regions. **(Right) Input perturbation** reflects explanation quality by the effectiveness of explanation-guided perturbations in affecting model outcomes. Specifically, interaction effects are captured by differences in progressive perturbation trends between the contrastive pair. The deletion and insertion results are aggregated into the second-order difference DIFF-ABPC to provide more comprehensive assessments.

where $k$ is the number of features determined by the cardinality of the ground-truth mask. Please note that the number $k$ is rounded to an integer multiple of the image patch size, facilitating a fair comparison between the proposed solution (pixel-level) and SOTA competitors (image-patch-level). Annotations in both bounding-box and segment forms are considered, with the results reported separately. Figure 2 (left) provides illustrative examples of metrics for the localization scheme.

**Input Perturbation** [36] quantifies explanation quality by measuring the performance degradation following different perturbation priorities [21]. In the context of cross-modal feature interaction, this scheme is adapted with conditional perturbation for a fixed text input [5]. However, the standard perturbation scheme is designed for the attribution task and does not generalize trivially to second-order explanations studied in this paper. To better reveal the impact of interactions, we propose **Input Perturbation on Contrastive Pair** – a second-order variant of the standard scheme. Given interactions anchored on a text component indexed by $j$, the standard scheme is complemented by the same perturbation on a contrastive text part $\boldsymbol{x}_{\setminus\{j\}}$; the interaction power on the anchor $j$ can be quantified by measuring the area between the perturbation curves (ABPC) of the contrastive pair:

$$\text{ABPC} = \frac{1}{d_{\text{img}}} \sum_{k=1}^{d_{\text{img}}} f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k)}_{\setminus\{j\}})$$

where $\boldsymbol{x}^{(k)}$ is a shorthand for the input with the top-$k$ image features removed as determined by the explanation results. Conditioned on the perturbation priority, the metric can be specified as ABPC-DEL for highest-first removal and ABPC-INS for lowest-first. Figure 2 (right) presents a schematic illustration of the perturbation metrics. For more comprehensive reflection on explanation quality, we further define DIFF-ABPC as the signed difference between ABPC-INS and ABPC-DEL, which aggregates the results of both perturbation regimes:

$$\text{DIFF-ABPC} = \text{ABPC-INS} - \text{ABPC-DEL}$$

Of particular note, the proposed DIFF-ABPC on the contrastive pair is formulated as a second-order difference. This inherent alignment with the second-order nature of dual interactions makes it suitable for evaluating interaction results.

Table 1: Average execution time $\pm$ std. (in seconds) of evaluated interaction methods.

| Model | GEFI-PM | exCLIP | Faith-Shap | FIxLIP | ProxySPEX |
|---|---|---|---|---|---|
| ViT-B/16 | **3.95**$_{\pm 0.12}$ | 4.54$_{\pm 0.16}$ | 41.28$_{\pm 1.98}$ | 68.31$_{\pm 1.46}$ | 36.89$_{\pm 1.63}$ |
| ViT-B/32 | **1.51**$_{\pm 0.05}$ | 4.30$_{\pm 0.13}$ | 8.56$_{\pm 0.86}$ | **1.51**$_{\pm 0.29}$ | 18.04$_{\pm 3.28}$ |

Additionally, localization metrics are predicated on an *alignment assumption* between model reasoning and human semantic grounding, under which an ideal explainer is expected to bridge interactions with visual human annotations corresponding to a designated text anchor. This premise, however, is not universally valid. To avoid relying solely on the alignment assumption, perturbation metrics are included to focus on model responses to explanation-guided interventions. Altogether, the selected metrics reflect explanation quality from complementary perspectives.

# 6 Experiments

**Models and Datasets.** To evaluate the effectiveness of the proposed method, we conduct extensive experiments across pre-trained CLIP models [1], including the ViT-B/32 and finer-grained ViT-B/16 variants. These models are tested on two popular open-source vision-language datasets: the validation split of Microsoft's Common Objects in Context (COCO) [37] and the test split [38] of Flickr30k (Flickr) [39, 40]. For *object localization* tests, we use the existing phrase-to-bounding-box annotations on Flickr and apply dictionary-based synonym matching on COCO to associate caption tokens with ground-truth annotations [5]. For *input perturbation* tests, we select five anchor tokens from each caption to evaluate image-space interactions. Specifically, these anchor tokens are selected based on their first-order attributions determined by IG [13], excluding stop words and punctuation. We elaborate further on the experimental settings for the design tests in Appendix D. Beyond CLIP backbones, Appendix F extends the evaluation to other dual-encoder models, showing that GEFI-PM consistently outperforms the competitors across a broader range of experimental settings and further supporting the generalizability of the findings presented in the main text.

**Baseline Methods.** We compare GEFI-PM against representative SOTA competitors dedicated to investigating cross-modal interactions: 1. *exCLIP* [5], a white-box approach aggregating gradient information through the integrated Jacobian; 2. *Faith-Shap* [24], an interaction index extending Shapley values to feature interactions; 3. *FIxLIP* [6], a weighted variant of the Banzhaf interaction index; 4. *ProxySPEX* [28], a surrogate-based approach leveraging sparse Fourier transforms. These choices cover representative categories in interaction research, including white-box, game-theoretic, and surrogate-based approaches. While acknowledging additional established baselines [7, 25, 27], we prioritize recent advances to focus on state-of-the-art comparisons.

**Runtime Costs.** To facilitate a fair comparison, we set the same query budget of $n = 5,000$ for all sampling-based methods, including GEFI-PM. The sole exception is exCLIP, whose budget is defined by the interpolation steps $n = 20$. Table 1 reports the average runtime of each tested explainer on the two model variants. Despite operating directly in pixel space, GEFI-PM achieves the lowest runtime across the reported cases, benefiting from its organized and highly parallelized computation. Although the white-box exCLIP may become more efficient as the model scales further, GEFI-PM retains a clear advantage over the other black-box approaches. Its main efficiency bottleneck is model throughput, whereas other black-box explainers involve substantial post-processing after model queries, inducing rapidly increasing computational costs as the number of features grows.

## 6.1 Objective Localization

Figure 3 presents the performance of the tested feature interaction methods in the object localization task. Since segmentation annotations are only available for COCO, the combination of datasets and ground-truth types yields three test cases, each tested with the two selected CLIP variants. The quantitative performance of each competitor is measured by PGA (bar plots) and IoU (box plots), illustrated separately with the same scales for each metric across different cases. Across all test cases, GEFI-PM consistently outperforms all baselines, which perform comparably to one another. Among them, exCLIP and ProxySPEX generally achieve stronger results, although their relative rankings

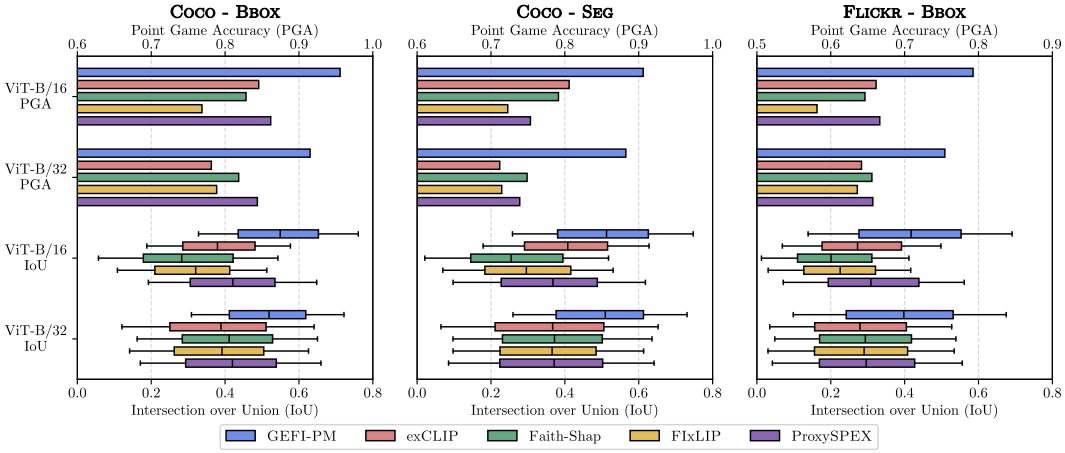



Figure 3: Competitor performance in the object localization tests. Results are grouped by dataset and considered ground truth type (BBOX stands for bounding box and SEG for segmentation), as specified in the subplot titles. Within each group, the upper half reports PGA with *bar plots*, and the lower half reports IoU using *box plots*. The two horizontal bands of each metric correspond to results acquired with ViT-B/16 and ViT-B/32, respectively, as indicated by the y-axis labels.

Table 2: Average Diff-ABPC (normalized) of the competitors in the input perturbation tests.

| Model | Dataset | GEFI-PM | exCLIP | Faith-Shap | FIxLIP | ProxySPEX | Random |
|---|---|---|---|---|---|---|---|
| ViT-B/16 | COCO | **.1764** | .0859 | .0885 | <u>.1328</u> | .1316 | -.0001 |
| | FLICKR | **.1464** | .0704 | .0565 | <u>.1009</u> | .0988 | -.0002 |
| ViT-B/32 | COCO | **.1682** | .0762 | <u>.1475</u> | .1411 | .1384 | .0002 |
| | FLICKR | **.1245** | .0532 | <u>.0958</u> | <u>.0958</u> | .0915 | -.0002 |

A higher Diff-ABPC score is better. The **best** results are shown in boldface and the <u>second-best</u> results are underlined.

vary across different cases. As anticipated, aligning interactions with segmentation masks is more difficult than with bounding boxes. Nevertheless, GEFI-PM demonstrates robust performance in this challenging setting with only marginal degradation, particularly in terms of IoU.

Another noteworthy observation is that GEFI-PM benefits from the increased capability of ViT-B/16 and achieves higher scores than with ViT-B/32. However, Faith-Shap and FIxLIP exhibit performance degradation from ViT-B/32 to ViT-B/16, especially when measured by IoU. We hypothesize that these degradations result from the increasing number of image patches, as the fixed query budget becomes insufficient for the fourfold expansion of the visual player set. In contrast, GEFI-PM operates directly at the pixel level and is therefore invariant to different image-patch partitions.

We attribute the performance gains by GEFI-PM to its fine-grained, pixel-level interaction modeling. For methods operating on fixed image patches, localizing the ground-truth mask becomes challenging when an object lies at the intersection of adjacent patches. Such cases, however, are not inherently different for GEFI-PM. Although this discussion is specific to the localization task, it highlights the greater expressiveness of pixel-level interactions, a lens looking inside predefined image patches.

### 6.2 Input Perturbation

Table 2 reports the performance of the evaluated explainers in the input perturbation tests. Performance is reported on Diff-ABPC, an aggregate measure of the insertion and deletion schemes [41]. To facilitate a dataset-level summary despite scale variations in model outputs across test inputs, we further normalize the Diff-ABPC scores by the corresponding total degradation of each input after complete removal of the image component. The normalized Diff-ABPC scores are generally small since they are second-order differences that capture only the impact of pairwise interactions among many possible forms of contribution. Thus, the values in the table should not be interpreted in isolation, but rather by comparing explainers under the same evaluation setting. Alongside the selected baselines, random removal is included as a sanity check; it shows that perturbations without

explanatory information fail to reveal interaction effects. Appendix E.2 provides the full results on the perturbation tests, including ABPC-DEL, ABPC-INS, and the standard first-order metrics.

Consistent with the localization results, GEFI-PM surpasses the competitors across all test cases. The higher scores indicate that our approach effectively identifies the most interactive features for a selected anchor feature, inducing larger outcome differences between contrastive pairs during progressive feature removal. In contrast, exCLIP, while competitive among the baselines in the localization tests, struggles in the perturbation tests. As acknowledged by its authors [5], exCLIP's results are not strictly tied to input-level spatial features at specific positions, preventing it from fully capturing the overall model behavior. This discrepancy across evaluation schemes also underlines the necessity of the complementary experimental design adopted in this work. In addition to the quantitative results, Appendix H provides a comprehensive set of qualitative examples illustrating the outcomes of the proposed method, which are omitted from the main text due to space constraints.

## 7 Conclusion

This paper presents a feature interaction method with a clean derivation from first-order proxy gradient estimation. To the best of our knowledge, GEFI is the first effort to investigate second-order interactions at the pixel level, while prior work trades explanation expressiveness for computational tractability by focusing on fixed image patches. Through a case study on CLIP-like models, we investigate second-order interactions from both theoretical and empirical perspectives and demonstrate the improved explanation quality provided by GEFI. Our results highlight that unraveling image patches is feasible for second-order explanations and that greater expressiveness is necessary for characterizing cross-modal model behavior.

## Acknowledgement

Yi Cai and Gerhard Wunder were supported by the Federal Ministry of Education and Research of Germany (BMBF) in the program of "Souverän. Digital. Vernetzt.", joint project "AIgenCY: Chances and Risks of Generative AI in Cybersecurity", project identification number 16KIS2013. Gerhard Wunder was also supported by BMBF joint project "6G-RIC: 6G Research and Innovation Cluster", project identification number 16KISK020K. Xinpeng Li would like to thank Prof. Kai Ming Ting and the School of Artificial Intelligence at Nanjing University for providing the computing resources used in this research.

## References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[4] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023.

[5] Lucas Moeller, Pascal Tilli, Thang Vu, and Sebastian Padó. Explaining caption-image interactions in CLIP models with second-order attributions. *Transactions on Machine Learning Research*, 2025.

[6] Hubert Baniecki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, and Przemyslaw Biecek. Explaining similarity in vision-language encoders with weighted Banzhaf interactions. In *Advances in Neural Information Processing Systems*, 2025.

[7] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9259–9268. PMLR, 2020.

[8] Yi Cai and Gerhard Wunder. On gradient-like explanation under a black-box setting: When black-box explanations become as good as white-box. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5360–5382. PMLR, 2024.

[9] Yi Cai, Thibaud Ardoin, and Gerhard Wunder. GEFA: A general feature attribution framework using proxy gradient estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 6165–6192. PMLR, 2025.

[10] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

[11] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

[12] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. In *Proceedings of the ICML Workshop on Visualization for Deep Learning, Sydney, Australia, 10 August 2017*, 2017.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[14] Yi Cai, Thibaud Ardoin, Mayank Gulati, and Gerhard Wunder. Rethinking explanation evaluation under the retraining scheme. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19826–19834, 2026.

[15] Adrian Hill, Neal McKee, Johannes Maeß, Stefan Bluecher, and Klaus Robert Muller. Smoothed differentiation efficiently mitigates shattered gradients in explanations. In *Advances in Neural Information Processing Systems*, 2025.

[16] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[17] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, 2021.

[18] Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 58090–58104. PMLR, 2024.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[20] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4768–4777, 2017.

[21] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceeedings of the British Machine Vision Conference 2018*. British Machine Vision Association, 2018.

[22] Xinpeng Li and Kai Ming Ting. Distribution-based feature attribution for explaining the predictions of any classifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23221–23229, 2026.

[23] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(1):5183–5244, 2020.

[24] 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.

[25] Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *Advances in Neural Information Processing Systems*, volume 36, pages 11515–11551, 2023.

[26] Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 130324–130357, 2024.

[27] Boris Joukovsky, Fawaz Sammani, and Nikos Deligiannis. Model-agnostic visual explanations via approximate bilinear models. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1770–1774. IEEE, 2023.

[28] Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Advances in Neural Information Processing Systems*, 2025.

[29] 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.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[31] Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

[32] Dimo Brockhoff, Anne Auger, Nikolaus Hansen, Dirk V Arnold, and Tim Hohm. Mirrored sampling and sequential selection for evolution strategies. In *International Conference on Parallel Problem Solving from Nature*, pages 11–21. Springer, 2010.

[33] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[34] Chenyang Zhao, Janet H Hsiao, and Antoni B Chan. Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):5967–5985, 2024.

[35] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[36] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2016.

[37] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[38] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[39] 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.

[40] 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. *International Journal of Computer Vision*, 123(1):74–93, 2017.

[41] Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. *Transactions on Machine Learning Research*, 2024.

[42] Luca Martino, Víctor Elvira, and Francisco Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[43] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

[44] 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.

# A Mathematical Proofs

## A.1 Proof of GEFI's equivalence to Shapley Interaction Index

This section presents the proofs of Theorem 3.1 and Theorem 3.2. We first prove the equivalence of the second-order interaction, and then generalize this observation to the higher-order form.

***Proof of Theorem 3.1 (Equivalence to Second-order SII).*** Let $\Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}})$ denote the difference operator in the second-order SII, as defined in [10]:

$$\Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}}) := f(\boldsymbol{z}_{\boldsymbol{T}}) - f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}) - f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}}) + f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i,j\}})$$

The goal of the proof can be formally presented as:

$$\xi_{i,j}^{\text{GEFI}} \stackrel{?}{=} \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{i,j\}} \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} \Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}})$$

The right-hand side presents the second-order SII, where $i$ and $j$ are the feature indices referring to the target features, between which the interaction is investigated, and $\boldsymbol{T} \subseteq \{1,\dots,d\}\setminus\{i,j\}$ is a coalition that excludes the target feature pair. The notation $\boldsymbol{z}_{\boldsymbol{T}}$ is a shorthand for a sample with only features indexed by $\boldsymbol{T}$ being present, namely:

$$\boldsymbol{z}_{\boldsymbol{T}} := \boldsymbol{x} \odot \boldsymbol{\epsilon} + \mathring{\boldsymbol{x}} \odot \bar{\boldsymbol{\epsilon}}, \text{ where } \epsilon_l = \begin{cases} 1, \text{ if } l \in \boldsymbol{T} \\ 0, \text{ if } l \notin \boldsymbol{T} \end{cases}$$

Focusing on the sample $\boldsymbol{z}_{\boldsymbol{T}}$ determined by such a coalition $\boldsymbol{T}$, its contribution $w_{i,j}^{\boldsymbol{T}}$ to the result $\xi_{i,j}^{\text{GEFI}}$ can be computed as follows according to Eq. (5):

$$\begin{aligned}
w_{i,j}^{\boldsymbol{T}} &:= \int_0^1 \alpha^{|\boldsymbol{T}|}(1-\alpha)^{d-|\boldsymbol{T}|} \cdot f(\boldsymbol{z}_{\boldsymbol{T}}) \cdot \frac{(0-\alpha)(0-\alpha)}{\alpha^2(1-\alpha)^2}\,\mathrm{d}\alpha && \text{(Absence of } i \text{ and } j\text{)} \\
&= \int_0^1 \alpha^{|\boldsymbol{T}|}(1-\alpha)^{d-|\boldsymbol{T}|-2} \cdot f(\boldsymbol{z}_{\boldsymbol{T}})\,\mathrm{d}\alpha \\
&= \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} \cdot f(\boldsymbol{z}_{\boldsymbol{T}}) && \text{(Beta-function)}
\end{aligned}$$

Here, the weight of a sample can be interpreted as its aggregated mass over the path integral of proxy gradient estimation. Analogously, the observations at $\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}$, $\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}}$, and $\boldsymbol{z}_{\boldsymbol{T}\cup\{i,j\}}$ are weighted by:

$$\begin{aligned}
w_{i,j}^{\boldsymbol{T}\cup\{i\}} &= \int_0^1 \alpha^{|\boldsymbol{T}|+1}(1-\alpha)^{d-|\boldsymbol{T}|-1} \cdot f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}) \cdot \frac{(1-\alpha)(0-\alpha)}{\alpha^2(1-\alpha)^2}\,\mathrm{d}\alpha \\
&= -\frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}) \\
w_{i,j}^{\boldsymbol{T}\cup\{j\}} &= \int_0^1 \alpha^{|\boldsymbol{T}|+1}(1-\alpha)^{d-|\boldsymbol{T}|-1} \cdot f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}}) \cdot \frac{(0-\alpha)(1-\alpha)}{\alpha^2(1-\alpha)^2}\,\mathrm{d}\alpha \\
&= -\frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}}) \\
w_{i,j}^{\boldsymbol{T}\cup\{i,j\}} &= \int_0^1 \alpha^{|\boldsymbol{T}|+2}(1-\alpha)^{d-|\boldsymbol{T}|-2} \cdot f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i,j\}}) \cdot \frac{(1-\alpha)(1-\alpha)}{\alpha^2(1-\alpha)^2}\,\mathrm{d}\alpha \\
&= \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i,j\}})
\end{aligned}$$

Summing over the coalition quadruples for all possible $\boldsymbol{T}$ concludes the proof:

$$\begin{aligned}
\xi_{i,j}^{\text{GEFI}} &= \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{i,j\}} w_{i,j}^{\boldsymbol{T}} + w_{i,j}^{\boldsymbol{T}\cup\{i\}} + w_{i,j}^{\boldsymbol{T}\cup\{j\}} + w_{i,j}^{\boldsymbol{T}\cup\{i,j\}} \\
&= \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{i,j\}} \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-2)!}{(d-1)!} \Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}}) \qquad (10)
\end{aligned}$$

□

Next, we generalize the proof to the more general $k$th-order form as claimed by Theorem 3.2.

***Proof of Theorem 3.2 (Equivalence to $k$th-order SII).*** To facilitate the proof, we use a shorthand notation $\boldsymbol{S}_{+j}$ for the interacting feature set $\boldsymbol{S}_{+j} = \boldsymbol{S} \cup \{j\}$, where $|\boldsymbol{S}_{+j}| = k$, and introduce one additional notation $\boldsymbol{V} \subseteq \boldsymbol{S}_{+j}$ representing the subset of the interaction set, which is involved in the computation of the corresponding discrete difference operator. Please note that the second-order interaction shown above is a special case when $\boldsymbol{S} = \{i\}$.

Following Eq. (6), the weight of the sample $\boldsymbol{z}_{\boldsymbol{T}\cup\boldsymbol{V}}$ can be derived as:

$$\begin{aligned} w_{\boldsymbol{S},j}^{\boldsymbol{T}\cup\boldsymbol{V}} &= \int_0^1 \alpha^{|\boldsymbol{T}|+|\boldsymbol{V}|}(1-\alpha)^{d-|\boldsymbol{T}|-|\boldsymbol{V}|} \cdot f(\boldsymbol{z}_{\boldsymbol{T}\cup\boldsymbol{V}}) \cdot \frac{\prod_{\boldsymbol{S}_{+j}\setminus\boldsymbol{V}}(0-\alpha)\cdot\prod_{\boldsymbol{V}}(1-\alpha)}{\alpha^k(1-\alpha)^k}\,\mathrm{d}\alpha \\ &= \int_0^1 \alpha^{|\boldsymbol{T}|+|\boldsymbol{V}|}(1-\alpha)^{d-|\boldsymbol{T}|-|\boldsymbol{V}|} \cdot \alpha^{-|\boldsymbol{V}|}(1-\alpha)^{|\boldsymbol{V}|-k} \cdot (-1)^{k-|\boldsymbol{V}|} f(\boldsymbol{z}_{\boldsymbol{T}\cup\boldsymbol{V}})\,\mathrm{d}\alpha \\ &= \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-k)!}{(d-k+1)!} \cdot (-1)^{k-|\boldsymbol{V}|} f(\boldsymbol{z}_{\boldsymbol{T}\cup\boldsymbol{V}}) \end{aligned}$$

The previous conclusion with second-order interaction can then be easily generalized to $k$th-order by collecting all weighted observations:

$$\begin{aligned} \xi_{\boldsymbol{S},j}^{\text{GEFI}} &= \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\boldsymbol{S}_{+j}} \sum_{\boldsymbol{V}\subseteq\boldsymbol{S}_{+j}} w_{\boldsymbol{S},j}^{\boldsymbol{T}\cup\boldsymbol{V}} \\ &= \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\boldsymbol{S}_{+j}} \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-k)!}{(d-k+1)!} \cdot \underbrace{\sum_{\boldsymbol{V}\subseteq\boldsymbol{S}_{+j}} (-1)^{k-|\boldsymbol{V}|} f(\boldsymbol{z}_{\boldsymbol{T}\cup\boldsymbol{V}})}_{\text{Möbius transform}} \end{aligned}$$

The special case holds when setting $k = 0$, namely:

$$\xi_{\varnothing,j}^{\text{GEFI}} = \xi_j^{\text{GEFA}} = \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{j\}} \frac{|\boldsymbol{T}|!(d-|\boldsymbol{T}|-1)!}{d!} \cdot (f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}}))$$

which is exactly the Shapley value, with only a few differences in notation. □

### A.2 GEFI-PM – A Probabilistic Interaction Index

This section expands the discussion of Proposition 4.1 and Remark 4.2 by showing the corresponding proofs and specifying the rescaling factors to recover SII from GEFI-PM.

***Proof of Proposition 4.1 (A Probabilistic Interaction Index).*** To prove that GEFI-PM belongs to the family of probabilistic interaction indices, the goal is to show the following:

$$\xi_{i,j}^{\text{GEFI-PM}} \overset{?}{=} \sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{i,j\}} \mathbb{P}(\boldsymbol{T}) \cdot \Delta_{ij}(f(\boldsymbol{z}_{\boldsymbol{T}}))$$

where $\mathbb{P}(\cdot)$ denotes the probability distribution of possible coalitions, which should sum up to one, *i.e.*, $\sum_{\boldsymbol{T}\subseteq\{1,\dots,d\}\setminus\{i,j\}} \mathbb{P}(\boldsymbol{T}) = 1$. According to the definition of GEFI-PM, we can first determine the expected contribution[7] $v_{i,j}^{\boldsymbol{T}}$ from a coalition $\boldsymbol{T}$ to $\xi_{i,j}^{\text{GEFI-PM}}$:

$$\begin{aligned} v_{i,j}^{\boldsymbol{T}} &:= \int_0^1 \beta^{|\boldsymbol{T}_{\text{txt}}|}(1-\beta)^{d_{\text{txt}}-|\boldsymbol{T}_{\text{txt}}|} \int_0^1 \alpha^{|\boldsymbol{T}_{\text{img}}|}(1-\alpha)^{d_{\text{img}}-|\boldsymbol{T}_{\text{img}}|} \frac{f(\boldsymbol{z}_{\boldsymbol{T}})\cdot(0-\alpha)(0-\beta)}{\alpha(1-\alpha)\beta(1-\beta)}\,\mathrm{d}\alpha\,\mathrm{d}\beta \\ &= \frac{|\boldsymbol{T}_{\text{img}}|!(d_{\text{img}}-|\boldsymbol{T}_{\text{img}}|-1)!}{d_{\text{img}}!} \cdot \frac{|\boldsymbol{T}_{\text{txt}}|!(d_{\text{txt}}-|\boldsymbol{T}_{\text{txt}}|-1)!}{d_{\text{txt}}!} f(\boldsymbol{z}_{\boldsymbol{T}}) \\ &= \frac{1}{d_{\text{img}}}\binom{d_{\text{img}}-1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}}\binom{d_{\text{txt}}-1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \cdot f(\boldsymbol{z}_{\boldsymbol{T}}) \end{aligned}$$

[7] To avoid overloading $w_{i,j}^{\boldsymbol{T}}$ used for GEFI, we denote the analogous coalition contribution by $v_{i,j}^{\boldsymbol{T}}$ for GEFI-PM.

where $\boldsymbol{T}_{\text{img}}$ and $\boldsymbol{T}_{\text{txt}}$ denote the present feature sets in the image and text spaces, respectively, forming the coalition $\boldsymbol{T} = \boldsymbol{T}_{\text{img}} + \boldsymbol{T}_{\text{txt}}$. Collecting the coalition quadruples that compose the difference operator for all possible $\boldsymbol{T}$ yields:

$$\xi_{i,j}^{\text{GEFI-PM}} = \sum_{\boldsymbol{T} \subseteq \{1,\dots,d\} \setminus \{i,j\}} \frac{1}{d_{\text{img}}} \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}} \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \cdot \Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}}) \tag{11}$$

which constitutes the form of probabilistic interaction indices. Next, we show that the weights of each difference operator sum up to one. Recalling the decoupled sampling of the two modalities, the weighting terms of coalitions can be reorganized as:

$$\begin{aligned}
&\sum_{\boldsymbol{T} \subseteq \{1,\dots,d\} \setminus \{i,j\}} \frac{1}{d_{\text{img}}} \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}} \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \\
=&\sum_{\boldsymbol{T}_{\text{img}} \subseteq \{1,\dots,d_{\text{img}}\} \setminus \{i\}} \frac{1}{d_{\text{img}}} \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \sum_{\boldsymbol{T}_{\text{txt}} \subseteq \{1,\dots,d_{\text{txt}}\} \setminus \{j\}} \frac{1}{d_{\text{txt}}} \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \\
=&\sum_{|\boldsymbol{T}_{\text{img}}|=0}^{d_{\text{img}}-1} \frac{1}{d_{\text{img}}} \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|} \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \sum_{|\boldsymbol{T}_{\text{txt}}|=0}^{d_{\text{txt}}-1} \frac{1}{d_{\text{txt}}} \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|} \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \\
=&\underbrace{\sum_{|\boldsymbol{T}_{\text{img}}|=0}^{d_{\text{img}}-1} \frac{1}{d_{\text{img}}}}_{d_{\text{img}} \text{ terms}} \cdot \underbrace{\sum_{|\boldsymbol{T}_{\text{txt}}|=0}^{d_{\text{txt}}-1} \frac{1}{d_{\text{txt}}}}_{d_{\text{txt}} \text{ terms}} \\
=&1
\end{aligned}$$

Putting together, the weighting terms of difference operators comply with the probability interpretation; therefore, GEFI-PM belongs to the probabilistic-interaction family. □

By showing that both GEFI (unbiased to SII) and GEFI-PM share the form of probabilistic interaction index, the statement in Remark 4.2 becomes evident, subject to a rescaling factor determined by the cardinality of coalitions for each modality:

$$s(|\boldsymbol{T}_{\text{img}}|, |\boldsymbol{T}_{\text{txt}}|) = \frac{d_{\text{img}} d_{\text{txt}}}{d-1} \binom{d-2}{|\boldsymbol{T}|}^{-1} \cdot \binom{d_{\text{img}} - 1}{|\boldsymbol{T}_{\text{img}}|} \cdot \binom{d_{\text{txt}} - 1}{|\boldsymbol{T}_{\text{txt}}|}$$

In practice, a total number of $n_{\text{img}} \cdot n_{\text{txt}}$ rescaling factors can be prepared once the query masks are initialized, and then applied to the observation matrix through a dot-wise product. Although the recovery of unbiasedness to SII is possible and computationally manageable, the rescaling will shift the focus back on the main diagonal of the observation matrix, negatively affecting explanation quality, with the information content of the remaining measurements almost ignored. The information lost due to the rescaling, which can be considered as importance sampling, can be formally measured by the effective sample size [42] defined as follows:

$$\text{ESS} = \frac{1}{\sum_{i=0}^{n} s_i^2}$$

where $s_i$ is the normalized weighting term of each element in the observation matrix, which sum up to $\sum_{i=0}^{n} s_i = 1$. Above definition of ESS is in range of $[1, n]$, which can be further normalized to $(0, 1]$ through the total number of observations and adapted to the product-measure context of cross-modal interaction:

$$\hat{\text{ESS}} = \frac{1}{|\boldsymbol{V}| \sum_{v \in \boldsymbol{V}} s_v^2}$$

It is easy to show that $\hat{\text{ESS}} \leq 1$ using the Cauchy-Schwarz inequality. The equality is only achieved when all samples received the same weight, which is exactly the case of GEFI-PM that all observations are processed equally:

$$\hat{\text{ESS}}^{\text{GEFI-PM}} = \frac{1}{|\boldsymbol{V}| \sum_{v \in \boldsymbol{V}} \left(\frac{1}{n_{\text{img}} \cdot n_{\text{txt}}}\right)^2} = \frac{1}{\left(\sum_{v \in \boldsymbol{V}} \frac{1}{n_{\text{img}} \cdot n_{\text{txt}}}\right)^2} = 1$$



(a) (b) (c)

Figure 4: Importance sampling of the observation matrix by different GEFI variants; (a) The vanilla version of GEFI considers only the synchronized coalitions on the main diagonal; (b) GEFI-PM leverages all observations uniformly; (c) Rescaling on GEFI-PM shifts the focus back onto the diagonal.

Any rescaling terms, including $s(|\boldsymbol{T}_{\text{img}}|, |\boldsymbol{T}_{\text{txt}}|)$, that distort the uniform sampling weights will degrade $\hat{\text{ESS}}$. While analytically quantifying the degradation is difficult due to the factorial terms, we complement the claim of information waste with empirical evidence. Following the setting in the main text, where $n_{\text{img}} = 224 \times 224$ and $n_{\text{txt}} = 77$, we can compute the $\hat{\text{ESS}}$ using the definition of the rescaling term $s(|\boldsymbol{T}_{\text{img}}|, |\boldsymbol{T}_{\text{txt}}|)$ for budgets evenly allocated for the two modalities, which yields:

$$\hat{\text{ESS}}^{\text{RESCALE}} \approx 0.129$$

As comparison, the standard GEFI considers only the main diagonal, and its best case $\hat{\text{ESS}}$, where $n' = n_{\text{img}} = n_{\text{txt}}$, of GEFI is:

$$\hat{\text{ESS}}^{\text{GEFI}} = \frac{1}{|\boldsymbol{V}| \sum_{n'} \frac{1}{n'}^2} = \frac{1}{n'}$$

The concrete value will be $0.02$ when the budget in one dimensionality is $n' = 500$ and can further decrease as the number of observations increases.

Figure 4 illustrates the weighting matrices of the three GEFI variants. It can be seen that rescaling GEFI-PM has a broader coverage of the observation matrix as compared to the standard GEFI, thereby more effectively leverages the collected information while both versions preserve unbiasedness to SII, aligning with the results of the case study that $0.129 \gg 0.02$. However, this value is still considerably low as compared to GEFI-PM due to its narrowed focus, leading to noisy and less expressive results; therefore, we do not consider applying the rescaling factors for the designed experiments.

### A.3 Further properties of GEFI-PM

This section discusses the properties of GEFI-PM as stated in Proposition 4.3, which contains the four properties: 1. Cardinality-Balanced, 2. Linearity, 3. Symmetry, and 4. Dummy. We provide formal definition of these properties before starting the proofs.

**Definition A.1** (Cardinality-Balanced)**.** *Let $t_{\text{img}}$ and $t_{\text{txt}}$ denote the cardinality of coalitions in the image and text space, and define the aggregated cardinality mass as:*

$$\mathbb{P}_{t_{\text{img}}, t_{\text{txt}}} := \sum_{\substack{|\boldsymbol{T}_{\text{img}}| = t_{\text{img}} \\ |\boldsymbol{T}_{\text{txt}}| = t_{\text{txt}}}} \mathbb{P}(\boldsymbol{T}_{\text{img}}, \boldsymbol{T}_{\text{txt}})$$

*A probabilistic interaction index is cardinality-balanced if satisfying:*

$$\mathbb{P}_{t_{\text{img}}, t_{\text{txt}}} = \frac{1}{d_{\text{img}} \cdot d_{\text{txt}}}$$

**Definition A.2** (Linearity)**.** *Linearity requires the interaction scores for a linear composition of two functions should equal the weighted sum of the separate interaction scores for each function, formally:*

$$\xi_{i,j}^{(af_1 + bf_2)} = a\xi_{i,j}^{(f_1)} + b\xi_{i,j}^{(f_2)}, \ \forall i, j \in \{1, \cdots, d\}$$

**Definition A.3** (Symmetry)**.** *Let $f(\cdot)$ be a function that is symmetric in two variables $x_i$ and $x_j$, i.e., $f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}) = f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}})$, $\forall \boldsymbol{T} \subseteq \{1,\cdots,d\}\backslash\{i,j\}$. Symmetry requires the equality to always hold for interaction scores of the two features anchored on an arbitrary third feature $x_k$, namely, $\xi_{i,k} = \xi_{j,k}$.*

**Definition A.4** (Dummy)**.** *Dummy requires the interaction scores anchored on a feature $x_i$ to always be zero if $f(\cdot)$ is not functionally dependent on that feature.*

***Proof of Proposition 4.3 (Cardinality-Balanced).*** We have previously shown that GEFI-PM belongs to the probabilistic-interaction family as formalized by Eq. (11), where the probability of each coalition is:

$$\mathbb{P}(\boldsymbol{T}) = \mathbb{P}(\boldsymbol{T}_{\text{img}}, \boldsymbol{T}_{\text{txt}}) = \frac{1}{d_{\text{img}}}\binom{d_{\text{img}}-1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}}\binom{d_{\text{txt}}-1}{|\boldsymbol{T}_{\text{txt}}|}^{-1}$$

The aggregated cardinality mass for GEFI-PM can then be determined as follows, where the combinatorial numbers cancel out:

$$\begin{aligned}
\mathbb{P}_{t_{\text{img}},t_{\text{txt}}} &= \sum_{\substack{|\boldsymbol{T}_{\text{img}}|=t_{\text{img}} \\ |\boldsymbol{T}_{\text{txt}}|=t_{\text{txt}}}} \frac{1}{d_{\text{img}}}\binom{d_{\text{img}}-1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}}\binom{d_{\text{txt}}-1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \\
&= \binom{d_{\text{img}}-1}{|\boldsymbol{T}_{\text{img}}|} \cdot \binom{d_{\text{txt}}-1}{|\boldsymbol{T}_{\text{txt}}|} \cdot \frac{1}{d_{\text{img}}}\binom{d_{\text{img}}-1}{|\boldsymbol{T}_{\text{img}}|}^{-1} \cdot \frac{1}{d_{\text{txt}}}\binom{d_{\text{txt}}-1}{|\boldsymbol{T}_{\text{txt}}|}^{-1} \\
&= \frac{1}{d_{\text{img}} \cdot d_{\text{txt}}}
\end{aligned}$$

□

***Proof of Proposition 4.3 (Linearity).*** GEFI-PM clearly satisfies *Linearity* by its definition in Eq. (7). Since it is clear from the context, we omit the superscript denoting method names for clarity and utilize the space to note different functions as referred to by the linearity property. For an arbitrary image-text feature pair, GEFI-PM satisfies:

$$\begin{aligned}
\xi_{i,j}^{(af_1+bf_2)} &= \int_0^1\int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)}\left[\mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)}\left[[af_1(\boldsymbol{z}) + bf_2(\boldsymbol{z})] \cdot \frac{(\epsilon_{\text{img},i}-\alpha)(\epsilon_{\text{txt},j}-\beta)}{\alpha(1-\alpha)\beta(1-\beta)}\right]\right] \mathrm{d}\alpha\mathrm{d}\beta \\
&= \int_0^1\int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)}\left[\mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)}\left[af_1(\boldsymbol{z}) \cdot \frac{(\epsilon_{\text{img},i}-\alpha)(\epsilon_{\text{txt},j}-\beta)}{\alpha(1-\alpha)\beta(1-\beta)}\right]\right] \mathrm{d}\alpha\mathrm{d}\beta \\
&\quad + \int_0^1\int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)}\left[\mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)}\left[bf_2(\boldsymbol{z}) \cdot \frac{(\epsilon_{\text{img},i}-\alpha)(\epsilon_{\text{txt},j}-\beta)}{\alpha(1-\alpha)\beta(1-\beta)}\right]\right] \mathrm{d}\alpha\mathrm{d}\beta \\
&= \xi_{i,j}^{(af_1)} + \xi_{i,j}^{(bf_2)}
\end{aligned}$$

□

***Proof of Proposition 4.3 (Symmetry).*** Let $x_i$ and $x_j$ be two symmetric features on $f(\cdot)$, and $k$ be an arbitrary third feature. We leverage the conclusion of Proposition 4.1 to facilitate the proof of *Symmetry*, which aims to show $\xi_{i,k}^{\text{GEFI-PM}} = \xi_{j,k}^{\text{GEFI-PM}}$. According to the definition of probabilistic interaction index, the interaction score determined by GEFI-PM can be reformulated as:

$$\begin{aligned}
\xi_{i,k}^{\text{GEFI-PM}} &= \sum_{\boldsymbol{T}\subseteq\{1,\ldots,d\}\backslash\{i,k\}} \mathbb{P}(\boldsymbol{T}) \cdot \Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}}) \\
&= \sum_{\boldsymbol{T}'\subseteq\{1,\ldots,d\}\backslash\{i,j,k\}} \left[\mathbb{P}(\boldsymbol{T}') \cdot \Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'}) + \mathbb{P}(\boldsymbol{T}'_{+j})\Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'_{+j}})\right]
\end{aligned}$$

where $\boldsymbol{T}'_{+j}$ is a shorthand notation for $\boldsymbol{T}' \cup \{j\}$. This reformulation separates the information aggregation process into two conditions: measured coalition with or without the symmetric feature $x_j$ of $x_i$, which also applies to the interaction score between the feature pair $(j,k)$:

$$\xi_{j,k}^{\text{GEFI-PM}} \sum_{\boldsymbol{T}'\subseteq\{1,\ldots,d\}\backslash\{i,j,k\}} \left[\mathbb{P}(\boldsymbol{T}') \cdot \Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'}) + \mathbb{P}(\boldsymbol{T}'_{+i})\Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'_{+i}})\right]$$

The probability terms $\mathbb{P}(\boldsymbol{T}')$ are exactly the same for the first component; the equality $\mathbb{P}(\boldsymbol{T}'_{+i}) = \mathbb{P}(\boldsymbol{T}'_{+j})$ also holds given the same cardinality of the coalition. As for the difference operator, the equality of the first term is obvious following the symmetric roles between $x_i$ and $x_j$:

$$\begin{aligned}\Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'}) &= f(\boldsymbol{z}_{\boldsymbol{T}'}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{k\}}) + f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,k\}}) \\ &= f(\boldsymbol{z}_{\boldsymbol{T}'}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{k\}}) + f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{j,k\}}) \\ &= \Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'})\end{aligned}$$

where the components in the first line are equal to the corresponding components in the second line, for example, the pair of $f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i\}})$ and $f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{j\}})$. Similarly, the equality of the second difference operator can be shown by unfolding the composing components:

$$\begin{aligned}\Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'_{+j}}) &= f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{j,k\}}) + f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,j,k\}}) \\ &= f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,k\}}) + f(\boldsymbol{z}_{\boldsymbol{T}'\cup\{i,j,k\}}) \\ &= \Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'_{+i}})\end{aligned}$$

Putting the above findings together conclude the proof:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{T}') \cdot \Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'}) + \mathbb{P}(\boldsymbol{T}'_{+j}) \Delta_{ik} f(\boldsymbol{z}_{\boldsymbol{T}'_{+j}}) = \mathbb{P}(\boldsymbol{T}') \cdot \Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'}) + \mathbb{P}(\boldsymbol{T}'_{+i}) \Delta_{jk} f(\boldsymbol{z}_{\boldsymbol{T}'_{+i}}) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\, \boldsymbol{T}' \subseteq \{1, \ldots, d\} \backslash \{i, j, k\} \\ &\Rightarrow \xi^{\text{GEFI-PM}}_{i,k} = \xi^{\text{GEFI-PM}}_{j,k}\end{aligned}$$

□

***Proof of Proposition 4.3 (Dummy)***. The *Dummy* property provides the precondition that $f(\cdot)$ is functionally independent of $x_i$, namely:

$$f(\boldsymbol{z}_{\boldsymbol{T}}) = f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}}),\ \forall\, \boldsymbol{T} \subseteq \{1, \ldots, d\} \backslash \{i\}$$

Considering this precondition and the probabilistic-interaction form of GEFI-PM, it can be easily shown that any interaction anchored on $i$ is zero:

$$\begin{aligned}\xi^{\text{GEFI-PM}}_{i,j} &= \sum_{\boldsymbol{T} \subseteq \{1,\ldots,d\} \backslash \{i,j\}} \mathbb{P}(\boldsymbol{T}) \cdot \Delta_{ij} f(\boldsymbol{z}_{\boldsymbol{T}}) \\ &= \sum_{\boldsymbol{T} \subseteq \{1,\ldots,d\} \backslash \{i,j\}} \mathbb{P}(\boldsymbol{T}) \cdot \underbrace{f(\boldsymbol{z}_{\boldsymbol{T}}) - f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i\}})}_{=0} + \underbrace{f(\boldsymbol{z}_{\boldsymbol{T}\cup\{i,j\}}) - f(\boldsymbol{z}_{\boldsymbol{T}\cup\{j\}})}_{=0} \\ &= 0\end{aligned}$$

which holds for $j \in \{1, \ldots, d\} \backslash \{i\}$. □

# B Implementation Details of GEFI-PM

Algorithm 1 provides a schematic overview to implement GEFI-PM following the description in Section 4. The first two lines (1 and 2) abstract the query construction processes following Bernoullis conditioned on different proxy variables, which are detailed in Algorithm 1. It is noteworthy that the query construction processes are identical for image and text modalities; they are separated in two lines since we apply an additional technique during the sampling in the image space for more effective masking. The query construction function returns the queries and the weighting factors for features according to their presence status in the corresponding query. The generated queries are then process by the model (line 3), which returns the full observation matrix $\boldsymbol{V}$. Specifically, model probing is composed of two stages: the image and text queries are encoded as embeddings by the dual encoders of a CLIP-like model, and the observation matrix is produced through a product measure on the normalized embeddings. Finally, the interaction scores (line 4) are determined by summarizing the collected information following Eq. (9).

**Algorithm 1** GEFI-PM Explanation Scheme

**Input:** $f(\cdot)$: target model; $\boldsymbol{x}$: explicand; $\mathring{\boldsymbol{x}}$: baseline; $n_{\text{img}}, n_{\text{txt}}$: image and text budgets
**Output:** $\boldsymbol{\xi} \in \mathbb{R}^{d_{\text{img}} \times d_{\text{txt}}}$: feature interaction scores;

$\boldsymbol{Z}_{\text{img}}, \boldsymbol{Q}_{\text{img}} = \text{GetImageQueries}(\boldsymbol{x}, \mathring{\boldsymbol{x}}, n_{\text{img}})$ *# Query Construction*
$\boldsymbol{Z}_{\text{txt}}, \boldsymbol{Q}_{\text{txt}} = \text{GetTextQueries}(\boldsymbol{x}, \mathring{\boldsymbol{x}}, n_{\text{txt}})$
$\boldsymbol{V} = f(\boldsymbol{Z}_{\text{img}}, \boldsymbol{Z}_{\text{txt}}) \in \mathbb{R}^{n_{\text{img}} \times n_{\text{txt}}}$ *# Collecting observations*
$\boldsymbol{\xi} = (\boldsymbol{V} \times \boldsymbol{Q}_{\text{img}})^T \times \boldsymbol{Q}_{\text{txt}}$
**return** $\boldsymbol{\xi}$

# C Extended Discussion on Interaction Anchored on Feature Subsets

Section 4.2 introduces the concept of computing pairwise interactions anchored on a feature subset. The discussion in the main text focuses on feature grouping in the image space, which addresses the reduced expressiveness due to the weaker semantic role of individual pixels. As previously mentioned, this concept can be easily extended to the text space, that is, token spans, instead of single tokens, can be considered as an anchor for interaction computation, formally as:

$$\xi^{\text{GEFI-PM}}_{i,\boldsymbol{S}_{\text{txt}}} := \int_0^1 \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)} \left[ \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)} \left[ f(\boldsymbol{z}) \cdot \frac{(\epsilon_{\text{img},i} - \alpha)(p_{\boldsymbol{S}_{\text{txt}}} - \beta)}{\alpha(1-\alpha)\beta(1-\beta)} \right] \right] \mathrm{d}\alpha \, \mathrm{d}\beta$$

To distinguish from a visual feature subset, we denote a subset of tokens as $\boldsymbol{S}_{\text{txt}} \subseteq \{1, \ldots, d_{\text{txt}}\}$; the underlying concept is not affected. Analogously to the image space, the definition of $\boldsymbol{S}_{\text{txt}}$ allows arbitrary token selection, *e.g.*, a continuous text span, a set of non-adjacent tokens, or a mixture of both. Furthermore, the concept of subset anchor can also be applied simultaneously in both image and text spaces, rendering:

$$\xi^{\text{GEFI-PM}}_{\boldsymbol{S}_{\text{img}},\boldsymbol{S}_{\text{txt}}} := \int_0^1 \int_0^1 \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{img}}|\alpha)} \left[ \mathbb{E}_{\boldsymbol{\pi}(\boldsymbol{\epsilon}_{\text{txt}}|\beta)} \left[ f(\boldsymbol{z}) \cdot \frac{(p_{\boldsymbol{S}_{\text{img}}} - \alpha)(p_{\boldsymbol{S}_{\text{txt}}} - \beta)}{\alpha(1-\alpha)\beta(1-\beta)} \right] \right] \mathrm{d}\alpha \, \mathrm{d}\beta$$

While allowing flexible examination of feature interactions in different forms, it should be noted that the practical use of subset anchors would require a proper design of human-machine interfaces to facilitate active user selection/definition of these subsets.

# D Further Details on Experimental Settings

## D.1 Experimental Environment

The proposed method is implemented using Python 3.11.2 with standard packages. The primary packages were *Numpy* 1.26.4, *PyTorch* 2.6.0, and *Torchvision* 0.21.0. The CUDA version was 12.6 for GPU support. All experiments were conducted on a machine operating Debian 12 with the following hardware specifications:

- Processor: Intel i9-10980XE, 18 cores
- Memory: 32GB DDR4
- GPU: NVIDIA RTX A5500, 24GB

## D.2 Detailed Experimental Settings

**Object Localization.** In addition to the annotation-caption matching described in Section 6, two additional restrictions are applied during the preparation of test entries following previous work [5]. First, the test is restricted to annotations associated with objects that are the only entity of their class in the corresponding image; this avoids unnecessary discrepancies when the tested model does not capture all instances from the same class simultaneously. Second, all tested ground-truth annotations should have a bounding box of at least $32 \times 32$ and remain uncropped after the standard CLIP preprocessing pipeline. It should be noted that the second rule, particularly the restriction on bounding box size, clearly favors explainers operating on fixed image patches; the test could

otherwise contain unsolvable entries for such approaches if a small ground-truth mask does not cover any image-patch center.

**Input Perturbation.** We repeat the input perturbation test on interactions determined by different anchor tokens for each image-caption pair. As previously mentioned, five anchor tokens are selected for each entry, prioritized in descending order of their first-order attribution scores. This design allows exploration of interaction quality when anchored on input components of various importance to a specific inference process. While the GEFI framework supports direct estimation of feature attribution, we use IG [13] as the first-order explainer for anchor selection to avoid favoring the proposed approach. The `captum` [43] implementation is adopted for IG. To derive token-level attributions, `captum.attr.LayerIntegratedGradients` computes attributions in the input embedding layer and aggregates them into input tokens by summing attribution scores assigned to each token embedding. Please note that the selected anchor tokens do not necessarily restrict to concrete concepts associated with nouns, and the test is still possible as it relies solely on model probing without leveraging any ground-truth annotations. That said, stop words and punctuation are filtered out to improve the comprehensibility of the results.

### D.3 Implementation of Baseline Methods

To facilitate an effective comparison with SOTA approaches, we leverage the official implementations of the selected baselines, which are listed below:

- *exCLIP* [5]: `https://github.com/lucasmllr/exCLIP`. We refactored the core logic within the `Explainer.explain` method for better runtime memory efficiency. The change avoids instantiating excessively large intermediate tensors while maintaining strict algorithmic equivalence to the original implementation, resolving the out-of-memory issue.
- *Faith-Shap* [24]: `https://github.com/chepingt/Faith-Shap`.
- *FIxLIP* [6]: `https://github.com/hbaniecki/fixlip`.
- *ProxySPEX* [28]: `https://pypi.org/project/spectralexplain/`. The official implementation is published as the `spectralexplain` package, and `version 0.0.4` is used.

All baselines use the hyperparameter configurations recommended by the respective authors throughout the designed experiments.

## E Additional Experimental Results

This section presents the full experimental results, which do not fit into the main text due to space constraints. The full results generally lead to the same conclusion and complement those reported in Section 6 with more fine-grained details, including additional metrics and standard deviation information.

### E.1 Detailed Results of Object Localization

The quantitative results on the object localization tests are reported with two metrics: PGA (in Table 3) and IoU (in Table 4). Consistent with the visualizations in Figure 3, GEFI-PM demonstrates superior performance in all test cases. In addition to the selected baselines, Tables 3 and 4 report a random baseline in the last row of each table, serving as an indicator of task difficulty and aiding better perception of the values. For clarity, the random baseline is omitted from the main results.

We also incorporate the results of *Point Game Energy (PGE)*, another metric considered in the literature [5, 6], in Table 5. Conceptually, PGE quantifies the fraction of the total explanation "*energy*" falling within the ground-truth region relative to the entire image. However, the definition of the so-called energy fraction is inconsistent in the literature: some work restricts the computation to positive attributions/interactions [5], whereas others sum over all scores [33, 34]. The former implementation can lead to unfair comparisons because different proportions of features may be included depending on how an explainer distributes positive and negative scores. Explainers who assign positive scores only to the top features will easily achieve better performance under this definition. As for the latter

Table 3: Object localization results: PGA ↑.

| Method | ViT-B/32 | | | ViT-B/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | Flickr | COCO | | Flickr |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.9151** | **.8826** | **.7545** | **.9557** | **.9061** | **.7926** |
| exCLIP | .7814 | .7118 | .6416 | .8455 | <u>.8058</u> | .6613 |
| Faith-Shap | .8184 | <u>.7489</u> | .6557 | .8284 | .7913 | .6463 |
| FIxLIP | .7886 | .7145 | .6359 | .7687 | .7227 | .5814 |
| ProxySPEX | <u>.8437</u> | .7389 | <u>.6569</u> | <u>.8618</u> | .7534 | <u>.6664</u> |
| Random | .5772 | .4517 | .3657 | .5772 | .4517 | .3657 |

↑: A higher value is better.

Table 4: Object localization results: IoU ($\pm$std.) ↑.

| Method | ViT-B/32 | | | ViT-B/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.4985**$_{\pm.17}$ | **.4794**$_{\pm.19}$ | **.3846**$_{\pm.20}$ | **.5359**$_{\pm.16}$ | **.4948**$_{\pm.18}$ | **.4057**$_{\pm.20}$ |
| exCLIP | .3841$_{\pm.19}$ | .3601$_{\pm.20}$ | .2871$_{\pm.18}$ | .3842$_{\pm.15}$ | <u>.3983</u>$_{\pm.16}$ | .2833$_{\pm.15}$ |
| Faith-Shap | .4004$_{\pm.18}$ | .3618$_{\pm.19}$ | .2950$_{\pm.18}$ | .3114$_{\pm.18}$ | .2832$_{\pm.17}$ | .2301$_{\pm.16}$ |
| FIxLIP | .3830$_{\pm.18}$ | .3504$_{\pm.18}$ | .2860$_{\pm.17}$ | .3200$_{\pm.16}$ | .3046$_{\pm.16}$ | .2330$_{\pm.14}$ |
| ProxySPEX | <u>.4146</u>$_{\pm.18}$ | <u>.3620</u>$_{\pm.19}$ | <u>.2998</u>$_{\pm.18}$ | <u>.4219</u>$_{\pm.17}$ | .3605$_{\pm.18}$ | <u>.3152</u>$_{\pm.18}$ |
| Random | .1495$_{\pm.15}$ | .0924$_{\pm.11}$ | .1142$_{\pm.12}$ | .1495$_{\pm.15}$ | .0924$_{\pm.11}$ | .1142$_{\pm.12}$ |

case, PGE is unbounded and can become badly defined when the denominator, given by the sum of present scores, is zero or negative.

Given the aforementioned limitations, we choose not to present PGEs as our main results, even though GEFI-PM outperforms the others. In our implementation, we aggregate the two definitions and adopt a distribution-aware thresholding strategy. Our computation of PGE restricts consideration to the top $50\%$ most salient features, leading to an entry-specific threshold definition. This definition can also be fine-grained with ground-truth annotations, *e.g.*, by considering the cardinality of ground-truth masks as described for IoU.

### E.2 Detailed Results of Input Perturbation

Table 6 reports perturbation results for contrastive pairs, measured by ABPC-DEL, ABPC-INS, and DIFF-ABPC. According to these interaction-specific metrics, GEFI-PM outperforms the SOTA baselines for almost all test cases, with particular advantages in ABPC-DEL. The only exception is the ABPC-INS score on ViT-B/32 COCO, where GEFI-PM still retains competitive performance with the best-performing baseline. The values in Table 6 are averaged over all five anchor tokens across all image-caption entries. More detailed results, grouped by the attribution rank of anchor tokens, are presented in Tables 7 – 9. These fine-grained results demonstrate the consistent performance of GEFI-PM across anchor tokens with different attribution magnitudes, and the observed explanation quality aligns with the overall assessments.

Since the proposed input perturbation on contrastive pair inherently contains the standard perturbation scheme, we further report results using metrics originally designed for first-order explanations. Standard input perturbation considers only the original input and quantifies explanation quality by measuring the area under the perturbation curve (AUPC), formally as:

$$\text{AUPC} = \frac{1}{d_{\text{img}}} \sum_{k=1}^{d_{\text{img}}} f(\boldsymbol{x}^{(k)}) \tag{12}$$

Table 5: Object localization results: PGE ($\pm$std.) ↑.

| Method | ViT-B/32 | | | ViT-B/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.6965**$_{\pm.25}$ | **.5854**$_{\pm.25}$ | **.5267**$_{\pm.28}$ | **.7613**$_{\pm.23}$ | **.5958**$_{\pm.24}$ | **.5526**$_{\pm.28}$ |
| exCLIP | .5849$_{\pm.28}$ | .4824$_{\pm.26}$ | .4581$_{\pm.27}$ | .6279$_{\pm.22}$ | .5525$_{\pm.22}$ | .4772$_{\pm.25}$ |
| Faith-Shap | .5571$_{\pm.24}$ | .4502$_{\pm.22}$ | .4567$_{\pm.27}$ | .6466$_{\pm.27}$ | .5631$_{\pm.26}$ | .5057$_{\pm.31}$ |
| FIxLIP | .4834$_{\pm.24}$ | .3843$_{\pm.21}$ | .3715$_{\pm.24}$ | .4267$_{\pm.26}$ | .3415$_{\pm.22}$ | .3247$_{\pm.20}$ |
| ProxySPEX | .5646$_{\pm.23}$ | .4241$_{\pm.21}$ | .4190$_{\pm.25}$ | .6035$_{\pm.22}$ | .4483$_{\pm.21}$ | .4555$_{\pm.25}$ |
| Random | .2346$_{\pm.19}$ | .1552$_{\pm.14}$ | .1885$_{\pm.16}$ | .2346$_{\pm.19}$ | .1552$_{\pm.14}$ | .1885$_{\pm.16}$ |

In the deletion scheme, a more effective explainer should prioritize removing the most relevant features, yielding a downward-bending curve (lower is better); the opposite trend is expected for insertion (higher is better). Table 10 present the overall assessments based on AUPC-DEL, AUPC-INS, and DIFF-AUPC, with more fine-grained results grouped by token ranks reported in Tables 11 – 13.

In the first-order evaluation, GEFI-PM performs favorably under the deletion scheme but is less competitive in insertion tests. The assessments of other baselines also reveal varying degrees of inconsistency between the corresponding AUPC-DEL and AUPC-INS scores. In fact, this disagreement problem is known in the standard input perturbation test. Blücher et al. [41] points out that the informativeness of deletion and insertion depends on the underlying occlusion strategy and its relation to the random baseline; therefore, AUPC-DEL and AUPC-INS do not by default provide fully interchangeable views of explanation quality. A straightforward yet powerful remedy[8] is to aggregate both perturbation directions through DIFF-AUPC = AUPC-INS − AUPC-DEL. The resulting metric provides a more comprehensive assessment of explanation quality and highlights GEFI-PM as the best-performing approach in the last column of Table 10 and in Table 13.

Although the first-order results favor GEFI-PM, we emphasize that the standard input perturbation is not, by itself, a proper test for interaction evaluation. A counterexample supporting this claim is provided by incorporating feature attribution methods for tests covering different anchors. Such first-order methods can easily outperform the interaction approaches discussed in this paper for lower-ranked anchors, despite the fact that they produce identical results across anchors and reveal no interaction information.

Beyond the discussion of the first-order metric, we also note that this disagreement problem is inherently canceled by construction under the definition of ABPC, which can be easily proved by an argument analogous to the analysis in [41].

[8]We refer readers to the original article [41] for further details on the proofs and analysis.

Table 6: Input perturbation results **with contrastive pairs** following different (second-order) metrics.

| Test Case | Method | ABPC-DEL ↓ | ABPC-INS ↑ | Diff-ABPC ↑ |
|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **-.1642**$_{\pm.28}$ | .0040$_{\pm.15}$ | **.1682**$_{\pm.29}$ |
| | exCLIP | -.1047$_{\pm.23}$ | -.0285$_{\pm.16}$ | .0762$_{\pm.16}$ |
| | Faith-Shap | -.1385$_{\pm.26}$ | **.0089**$_{\pm.18}$ | .1475$_{\pm.25}$ |
| | FIxLIP | -.1349$_{\pm.24}$ | .0062$_{\pm.18}$ | .1411$_{\pm.23}$ |
| | ProxySPEX | -.1334$_{\pm.24}$ | .0050$_{\pm.17}$ | .1384$_{\pm.22}$ |
| ViT-B/32 FLICKR | GEFI-PM | **-.1210**$_{\pm.16}$ | **.0035**$_{\pm.09}$ | **.1245**$_{\pm.15}$ |
| | exCLIP | -.0851$_{\pm.13}$ | -.0319$_{\pm.11}$ | .0532$_{\pm.10}$ |
| | Faith-Shap | -.1036$_{\pm.15}$ | -.0077$_{\pm.09}$ | .0958$_{\pm.13}$ |
| | FIxLIP | -.1046$_{\pm.15}$ | -.0087$_{\pm.09}$ | .0958$_{\pm.13}$ |
| | ProxySPEX | -.1010$_{\pm.15}$ | -.0094$_{\pm.09}$ | .0915$_{\pm.13}$ |
| ViT-B/16 COCO | GEFI-PM | **-.1663**$_{\pm.18}$ | **.0102**$_{\pm.11}$ | **.1764**$_{\pm.17}$ |
| | exCLIP | -.1114$_{\pm.16}$ | -.0254$_{\pm.10}$ | .0859$_{\pm.13}$ |
| | Faith-Shap | -.1051$_{\pm.16}$ | -.0166$_{\pm.09}$ | .0885$_{\pm.13}$ |
| | FIxLIP | -.1356$_{\pm.18}$ | -.0028$_{\pm.10}$ | .1328$_{\pm.17}$ |
| | ProxySPEX | -.1248$_{\pm.16}$ | .0069$_{\pm.09}$ | .1316$_{\pm.16}$ |
| ViT-B/16 FLICKR | GEFI-PM | **-.1281**$_{\pm.15}$ | **.0184**$_{\pm.10}$ | **.1464**$_{\pm.17}$ |
| | exCLIP | -.0884$_{\pm.12}$ | -.0181$_{\pm.09}$ | .0704$_{\pm.11}$ |
| | Faith-Shap | -.0750$_{\pm.12}$ | -.0184$_{\pm.08}$ | .0565$_{\pm.09}$ |
| | FIxLIP | -.1038$_{\pm.14}$ | -.0029$_{\pm.09}$ | .1009$_{\pm.14}$ |
| | ProxySPEX | -.0947$_{\pm.13}$ | .0041$_{\pm.08}$ | .0988$_{\pm.14}$ |

↓: A lower value is better.

Table 7: Input perturbation results by attribution ranks of anchor tokens: ABPC-DEL ($\pm$std.) ↓.

| **ABPC-DEL** | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **-.2798**$_{\pm.45}$ | **-.1901**$_{\pm.24}$ | **-.1498**$_{\pm.23}$ | **-.1035**$_{\pm.16}$ | **-.0852**$_{\pm.13}$ |
| | exCLIP | -.2115$_{\pm.38}$ | -.1227$_{\pm.18}$ | -.0897$_{\pm.17}$ | -.0498$_{\pm.13}$ | -.0392$_{\pm.12}$ |
| | Faith-Shap | -.2581$_{\pm.45}$ | -.1624$_{\pm.20}$ | -.1222$_{\pm.18}$ | -.0765$_{\pm.13}$ | -.0609$_{\pm.12}$ |
| | FIxLIP | -.2515$_{\pm.39}$ | -.1571$_{\pm.19}$ | -.1208$_{\pm.18}$ | -.0745$_{\pm.14}$ | -.0585$_{\pm.11}$ |
| | ProxySPEX | -.2534$_{\pm.40}$ | -.1565$_{\pm.19}$ | -.1175$_{\pm.18}$ | -.0707$_{\pm.13}$ | -.0562$_{\pm.12}$ |
| ViT-B/32 FLICKR | GEFI-PM | **-.1705**$_{\pm.18}$ | **-.1349**$_{\pm.20}$ | **-.1123**$_{\pm.13}$ | **-.1027**$_{\pm.14}$ | **-.0838**$_{\pm.14}$ |
| | exCLIP | -.1346$_{\pm.15}$ | -.0928$_{\pm.14}$ | -.0779$_{\pm.12}$ | -.0693$_{\pm.13}$ | -.0503$_{\pm.09}$ |
| | Faith-Shap | -.1592$_{\pm.18}$ | -.1156$_{\pm.18}$ | -.0937$_{\pm.12}$ | -.0856$_{\pm.14}$ | -.0633$_{\pm.09}$ |
| | FIxLIP | -.1593$_{\pm.18}$ | -.1169$_{\pm.19}$ | -.0952$_{\pm.12}$ | -.0870$_{\pm.14}$ | -.0638$_{\pm.10}$ |
| | ProxySPEX | -.1567$_{\pm.18}$ | -.1127$_{\pm.18}$ | -.0918$_{\pm.12}$ | -.0831$_{\pm.13}$ | -.0598$_{\pm.09}$ |
| ViT-B/16 COCO | GEFI-PM | **-.2646**$_{\pm.22}$ | **-.1895**$_{\pm.19}$ | **-.1493**$_{\pm.15}$ | **-.1174**$_{\pm.14}$ | **-.0999**$_{\pm.12}$ |
| | exCLIP | -.2026$_{\pm.21}$ | -.1298$_{\pm.17}$ | -.0938$_{\pm.14}$ | -.0693$_{\pm.11}$ | -.0521$_{\pm.10}$ |
| | Faith-Shap | -.2006$_{\pm.21}$ | -.1201$_{\pm.16}$ | -.0878$_{\pm.13}$ | -.0614$_{\pm.11}$ | -.0461$_{\pm.10}$ |
| | FIxLIP | -.2367$_{\pm.22}$ | -.1533$_{\pm.19}$ | -.1200$_{\pm.15}$ | -.0883$_{\pm.13}$ | -.0692$_{\pm.11}$ |
| | ProxySPEX | -.2192$_{\pm.21}$ | -.1421$_{\pm.17}$ | -.1074$_{\pm.14}$ | -.0815$_{\pm.12}$ | -.0640$_{\pm.10}$ |
| ViT-B/16 FLICKR | GEFI-PM | **-.1749**$_{\pm.17}$ | **-.1377**$_{\pm.14}$ | **-.1204**$_{\pm.13}$ | **-.1103**$_{\pm.12}$ | **-.0967**$_{\pm.19}$ |
| | exCLIP | -.1302$_{\pm.15}$ | -.0970$_{\pm.12}$ | -.0815$_{\pm.11}$ | -.0743$_{\pm.10}$ | -.0586$_{\pm.10}$ |
| | Faith-Shap | -.1163$_{\pm.15}$ | -.0829$_{\pm.11}$ | -.0682$_{\pm.10}$ | -.0595$_{\pm.09}$ | -.0474$_{\pm.10}$ |
| | FIxLIP | -.1554$_{\pm.18}$ | -.1109$_{\pm.13}$ | -.0969$_{\pm.12}$ | -.0867$_{\pm.11}$ | -.0685$_{\pm.13}$ |
| | ProxySPEX | -.1399$_{\pm.16}$ | -.1036$_{\pm.12}$ | -.0888$_{\pm.11}$ | -.0771$_{\pm.10}$ | -.0636$_{\pm.13}$ |

Table 8: Input perturbation results by attribution ranks of anchor tokens: ABPC-INS ($\pm$std.) ↑.

| ABPC-INS | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | -.0205$_{\pm.13}$ | -.0066$_{\pm.18}$ | .0046$_{\pm.17}$ | .0198$_{\pm.16}$ | **.0261**$_{\pm.11}$ |
| | exCLIP | -.0662$_{\pm.21}$ | -.0369$_{\pm.18}$ | -.0299$_{\pm.12}$ | -.0054$_{\pm.16}$ | .0004$_{\pm.11}$ |
| | Faith-Shap | **-.0138**$_{\pm.18}$ | **.0040**$_{\pm.19}$ | **.0092**$_{\pm.20}$ | **.0226**$_{\pm.19}$ | .0253$_{\pm.12}$ |
| | FIxLIP | -.0175$_{\pm.19}$ | .0004$_{\pm.18}$ | .0067$_{\pm.19}$ | .0206$_{\pm.17}$ | .0235$_{\pm.12}$ |
| | ProxySPEX | -.0157$_{\pm.17}$ | .0005$_{\pm.18}$ | .0037$_{\pm.17}$ | .0186$_{\pm.18}$ | .0204$_{\pm.11}$ |
| ViT-B/32 FLICKR | GEFI-PM | **-.0031**$_{\pm.09}$ | **.0012**$_{\pm.11}$ | **.0038**$_{\pm.07}$ | **.0042**$_{\pm.11}$ | **.0115**$_{\pm.08}$ |
| | exCLIP | -.0458$_{\pm.13}$ | -.0373$_{\pm.15}$ | -.0324$_{\pm.09}$ | -.0290$_{\pm.10}$ | -.0148$_{\pm.08}$ |
| | Faith-Shap | -.0149$_{\pm.09}$ | -.0081$_{\pm.10}$ | -.0085$_{\pm.08}$ | -.0078$_{\pm.09}$ | .0007$_{\pm.08}$ |
| | FIxLIP | -.0177$_{\pm.09}$ | -.0089$_{\pm.11}$ | -.0099$_{\pm.08}$ | -.0076$_{\pm.09}$ | .0006$_{\pm.09}$ |
| | ProxySPEX | -.0162$_{\pm.09}$ | -.0102$_{\pm.09}$ | -.0108$_{\pm.08}$ | -.0094$_{\pm.09}$ | -.0005$_{\pm.08}$ |
| ViT-B/16 COCO | GEFI-PM | -.0105$_{\pm.12}$ | .0049$_{\pm.12}$ | **.0091**$_{\pm.11}$ | **.0208**$_{\pm.11}$ | **.0296**$_{\pm.11}$ |
| | exCLIP | -.0475$_{\pm.12}$ | -.0297$_{\pm.11}$ | -.0238$_{\pm.10}$ | -.0171$_{\pm.10}$ | -.0062$_{\pm.09}$ |
| | Faith-Shap | -.0276$_{\pm.09}$ | -.0208$_{\pm.09}$ | -.0162$_{\pm.09}$ | -.0127$_{\pm.09}$ | -.0038$_{\pm.08}$ |
| | FIxLIP | -.0125$_{\pm.11}$ | -.0035$_{\pm.10}$ | -.0017$_{\pm.10}$ | -.0019$_{\pm.11}$ | .0072$_{\pm.10}$ |
| | ProxySPEX | **-.0036**$_{\pm.09}$ | **.0064**$_{\pm.09}$ | .0064$_{\pm.09}$ | .0099$_{\pm.09}$ | .0167$_{\pm.09}$ |
| ViT-B/16 FLICKR | GEFI-PM | **.0159**$_{\pm.11}$ | **.0158**$_{\pm.10}$ | **.0216**$_{\pm.12}$ | **.0171**$_{\pm.08}$ | **.0214**$_{\pm.09}$ |
| | exCLIP | -.0280$_{\pm.11}$ | -.0162$_{\pm.09}$ | -.0174$_{\pm.09}$ | -.0174$_{\pm.08}$ | -.0112$_{\pm.07}$ |
| | Faith-Shap | -.0222$_{\pm.10}$ | -.0160$_{\pm.08}$ | -.0184$_{\pm.08}$ | -.0196$_{\pm.07}$ | -.0157$_{\pm.07}$ |
| | FIxLIP | -.0007$_{\pm.10}$ | -.0017$_{\pm.09}$ | -.0028$_{\pm.11}$ | -.0072$_{\pm.08}$ | -.0019$_{\pm.08}$ |
| | ProxySPEX | .0038$_{\pm.09}$ | .0075$_{\pm.08}$ | .0047$_{\pm.09}$ | .0010$_{\pm.07}$ | .0036$_{\pm.07}$ |

Table 9: Input perturbation results by attribution ranks of anchor tokens: Diff-ABPC ($\pm$std.) ↑.

| Diff-ABPC | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **.2593**$_{\pm.45}$ | **.1835**$_{\pm.24}$ | **.1544**$_{\pm.32}$ | **.1234**$_{\pm.14}$ | **.1113**$_{\pm.12}$ |
| | exCLIP | .1454$_{\pm.24}$ | .0858$_{\pm.16}$ | .0598$_{\pm.13}$ | .0444$_{\pm.11}$ | .0396$_{\pm.09}$ |
| | Faith-Shap | .2444$_{\pm.36}$ | .1664$_{\pm.24}$ | .1315$_{\pm.26}$ | .0992$_{\pm.16}$ | .0862$_{\pm.11}$ |
| | FIxLIP | .2339$_{\pm.30}$ | .1575$_{\pm.22}$ | .1276$_{\pm.25}$ | .0951$_{\pm.14}$ | .0820$_{\pm.11}$ |
| | ProxySPEX | .2378$_{\pm.32}$ | .1570$_{\pm.22}$ | .1212$_{\pm.21}$ | .0894$_{\pm.15}$ | .0767$_{\pm.10}$ |
| ViT-B/32 FLICKR | GEFI-PM | **.1675**$_{\pm.16}$ | **.1361**$_{\pm.18}$ | **.1161**$_{\pm.13}$ | **.1069**$_{\pm.12}$ | **.0953**$_{\pm.16}$ |
| | exCLIP | .0889$_{\pm.13}$ | .0555$_{\pm.10}$ | .0455$_{\pm.08}$ | .0403$_{\pm.08}$ | .0355$_{\pm.08}$ |
| | Faith-Shap | .1443$_{\pm.16}$ | .1075$_{\pm.14}$ | .0851$_{\pm.10}$ | .0779$_{\pm.11}$ | .0640$_{\pm.10}$ |
| | FIxLIP | .1416$_{\pm.16}$ | .1080$_{\pm.15}$ | .0853$_{\pm.11}$ | .0794$_{\pm.11}$ | .0645$_{\pm.12}$ |
| | ProxySPEX | .1406$_{\pm.16}$ | .1025$_{\pm.13}$ | .0810$_{\pm.10}$ | .0737$_{\pm.10}$ | .0593$_{\pm.10}$ |
| ViT-B/16 COCO | GEFI-PM | **.2542**$_{\pm.20}$ | **.1945**$_{\pm.18}$ | **.1584**$_{\pm.14}$ | **.1381**$_{\pm.15}$ | **.1295**$_{\pm.13}$ |
| | exCLIP | .1550$_{\pm.17}$ | .1001$_{\pm.14}$ | .0699$_{\pm.10}$ | .0522$_{\pm.09}$ | .0459$_{\pm.08}$ |
| | Faith-Shap | .1730$_{\pm.18}$ | .0993$_{\pm.14}$ | .0717$_{\pm.12}$ | .0487$_{\pm.10}$ | .0423$_{\pm.07}$ |
| | FIxLIP | .2241$_{\pm.20}$ | .1498$_{\pm.18}$ | .1183$_{\pm.15}$ | .0864$_{\pm.14}$ | .0764$_{\pm.11}$ |
| | ProxySPEX | .2156$_{\pm.19}$ | .1484$_{\pm.16}$ | .1138$_{\pm.13}$ | .0915$_{\pm.14}$ | .0807$_{\pm.11}$ |
| ViT-B/16 FLICKR | GEFI-PM | **.1908**$_{\pm.16}$ | **.1534**$_{\pm.14}$ | **.1420**$_{\pm.16}$ | **.1274**$_{\pm.12}$ | **.1181**$_{\pm.23}$ |
| | exCLIP | .1023$_{\pm.12}$ | .0808$_{\pm.10}$ | .0641$_{\pm.12}$ | .0570$_{\pm.08}$ | .0474$_{\pm.10}$ |
| | Faith-Shap | .0941$_{\pm.12}$ | .0669$_{\pm.09}$ | .0498$_{\pm.09}$ | .0399$_{\pm.07}$ | .0316$_{\pm.06}$ |
| | FIxLIP | .1548$_{\pm.16}$ | .1091$_{\pm.13}$ | .0941$_{\pm.15}$ | .0796$_{\pm.11}$ | .0666$_{\pm.14}$ |
| | ProxySPEX | .1437$_{\pm.15}$ | .1110$_{\pm.12}$ | .0935$_{\pm.14}$ | .0781$_{\pm.10}$ | .0672$_{\pm.16}$ |

Table 10: Input perturbation results **without contrastive pairs** following different (first-order) metrics.

| Test Case | Method | AUPC-DEL ↓ | AUPC-INS ↑ | Diff-AUPC ↑ |
|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **.4612**$_{\pm.34}$ | .7456$_{\pm.41}$ | **.2842**$_{\pm.42}$ |
| | exCLIP | .6865$_{\pm.46}$ | .6938$_{\pm.50}$ | .0073$_{\pm.32}$ |
| | Faith-Shap | .5591$_{\pm.43}$ | .8208$_{\pm.58}$ | .2615$_{\pm.38}$ |
| | FIxLIP | .5615$_{\pm.47}$ | .8047$_{\pm.59}$ | .2435$_{\pm.43}$ |
| | ProxySPEX | .5894$_{\pm.46}$ | **.8218**$_{\pm.58}$ | .2325$_{\pm.36}$ |
| ViT-B/32 FLICKR | GEFI-PM | **.4158**$_{\pm.24}$ | .6899$_{\pm.20}$ | **.2742**$_{\pm.32}$ |
| | exCLIP | .6011$_{\pm.20}$ | .6055$_{\pm.20}$ | .0046$_{\pm.24}$ |
| | Faith-Shap | .5063$_{\pm.20}$ | **.7119**$_{\pm.20}$ | .2056$_{\pm.25}$ |
| | FIxLIP | .5000$_{\pm.21}$ | .6929$_{\pm.20}$ | .1930$_{\pm.30}$ |
| | ProxySPEX | .5298$_{\pm.19}$ | .7080$_{\pm.19}$ | .1782$_{\pm.24}$ |
| ViT-B/16 COCO | GEFI-PM | **.4885**$_{\pm.25}$ | .7661$_{\pm.26}$ | **.2773**$_{\pm.35}$ |
| | exCLIP | .7109$_{\pm.24}$ | .7339$_{\pm.25}$ | .0229$_{\pm.27}$ |
| | Faith-Shap | .6787$_{\pm.22}$ | .8418$_{\pm.22}$ | .1631$_{\pm.20}$ |
| | FIxLIP | .6055$_{\pm.30}$ | .7988$_{\pm.29}$ | .1931$_{\pm.43}$ |
| | ProxySPEX | .6704$_{\pm.23}$ | **.8760**$_{\pm.24}$ | .2056$_{\pm.27}$ |
| ViT-B/16 FLICKR | GEFI-PM | **.4424**$_{\pm.23}$ | .6992$_{\pm.22}$ | **.2568**$_{\pm.32}$ |
| | exCLIP | .6787$_{\pm.21}$ | .6831$_{\pm.22}$ | .0043$_{\pm.26}$ |
| | Faith-Shap | .6567$_{\pm.19}$ | .7803$_{\pm.20}$ | .1235$_{\pm.17}$ |
| | FIxLIP | .5752$_{\pm.25}$ | .7490$_{\pm.27}$ | .1735$_{\pm.40}$ |
| | ProxySPEX | .6460$_{\pm.20}$ | **.8198**$_{\pm.22}$ | .1741$_{\pm.26}$ |

Please note the difference between the concepts of ABPC and AUPC.

Table 11: Input perturbation results **without contrastive pair**: AUPC-DEL ($\pm$std.) ↓.

| **AUPC-DEL** | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **.3958**$_{\pm.29}$ | **.4470**$_{\pm.33}$ | **.4614**$_{\pm.34}$ | **.4949**$_{\pm.43}$ | **.5161**$_{\pm.32}$ |
| | exCLIP | .6548$_{\pm.46}$ | .6841$_{\pm.50}$ | .6904$_{\pm.44}$ | .7065$_{\pm.46}$ | .6992$_{\pm.44}$ |
| | Faith-Shap | .4966$_{\pm.36}$ | .5356$_{\pm.46}$ | .5679$_{\pm.48}$ | .5962$_{\pm.44}$ | .6074$_{\pm.38}$ |
| | FIxLIP | .5112$_{\pm.45}$ | .5410$_{\pm.45}$ | .5674$_{\pm.48}$ | .5889$_{\pm.47}$ | .6055$_{\pm.48}$ |
| | ProxySPEX | .5205$_{\pm.47}$ | .5645$_{\pm.42}$ | .6030$_{\pm.54}$ | .6274$_{\pm.42}$ | .6396$_{\pm.42}$ |
| ViT-B/32 FLICKR | GEFI-PM | **.3936**$_{\pm.24}$ | **.4009**$_{\pm.29}$ | **.4280**$_{\pm.20}$ | **.4175**$_{\pm.24}$ | **.4390**$_{\pm.24}$ |
| | exCLIP | .6006$_{\pm.18}$ | .6011$_{\pm.18}$ | .5933$_{\pm.23}$ | .6025$_{\pm.23}$ | .6079$_{\pm.17}$ |
| | Faith-Shap | .4751$_{\pm.19}$ | .4951$_{\pm.19}$ | .5117$_{\pm.19}$ | .5151$_{\pm.22}$ | .5342$_{\pm.19}$ |
| | FIxLIP | .4736$_{\pm.21}$ | .4917$_{\pm.19}$ | .5039$_{\pm.20}$ | .5073$_{\pm.21}$ | .5244$_{\pm.24}$ |
| | ProxySPEX | .4900$_{\pm.20}$ | .5166$_{\pm.19}$ | .5381$_{\pm.18}$ | .5435$_{\pm.20}$ | .5605$_{\pm.19}$ |
| ViT-B/16 COCO | GEFI-PM | **.4302**$_{\pm.23}$ | **.4756**$_{\pm.25}$ | **.5044**$_{\pm.25}$ | **.5186**$_{\pm.26}$ | **.5190**$_{\pm.25}$ |
| | exCLIP | .6719$_{\pm.26}$ | .7021$_{\pm.23}$ | .7236$_{\pm.25}$ | .7349$_{\pm.25}$ | .7256$_{\pm.23}$ |
| | Faith-Shap | .6196$_{\pm.24}$ | .6709$_{\pm.23}$ | .6914$_{\pm.22}$ | .7114$_{\pm.21}$ | .7041$_{\pm.20}$ |
| | FIxLIP | .5435$_{\pm.31}$ | .5889$_{\pm.30}$ | .6143$_{\pm.30}$ | .6431$_{\pm.29}$ | .6440$_{\pm.28}$ |
| | ProxySPEX | .6182$_{\pm.25}$ | .6602$_{\pm.23}$ | .6816$_{\pm.23}$ | .6987$_{\pm.23}$ | .6968$_{\pm.21}$ |
| ViT-B/16 FLICKR | GEFI-PM | **.4158**$_{\pm.22}$ | **.4299**$_{\pm.24}$ | **.4458**$_{\pm.23}$ | **.4463**$_{\pm.23}$ | **.4744**$_{\pm.22}$ |
| | exCLIP | .6782$_{\pm.20}$ | .6753$_{\pm.22}$ | .6787$_{\pm.20}$ | .6753$_{\pm.19}$ | .6865$_{\pm.25}$ |
| | Faith-Shap | .6270$_{\pm.19}$ | .6484$_{\pm.20}$ | .6636$_{\pm.17}$ | .6694$_{\pm.18}$ | .6753$_{\pm.18}$ |
| | FIxLIP | .5366$_{\pm.23}$ | .5659$_{\pm.25}$ | .5781$_{\pm.25}$ | .5835$_{\pm.24}$ | .6138$_{\pm.26}$ |
| | ProxySPEX | .6221$_{\pm.20}$ | .6323$_{\pm.20}$ | .6509$_{\pm.19}$ | .6519$_{\pm.18}$ | .6724$_{\pm.21}$ |

Table 12: Input perturbation results **without contrastive pair**: AUPC-INS (±std.) ↑.

| AUPC-INS | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | .7930$_{\pm.44}$ | .7681$_{\pm.41}$ | .7500$_{\pm.42}$ | .7178$_{\pm.34}$ | .6899$_{\pm.41}$ |
| | exCLIP | .7158$_{\pm.49}$ | .7007$_{\pm.46}$ | .6909$_{\pm.49}$ | .6860$_{\pm.52}$ | .6724$_{\pm.52}$ |
| | Faith-Shap | .8564$_{\pm.54}$ | .8403$_{\pm.57}$ | .8193$_{\pm.64}$ | .8022$_{\pm.56}$ | .7793$_{\pm.57}$ |
| | FIxLIP | .8433$_{\pm.52}$ | .8296$_{\pm.61}$ | .8027$_{\pm.63}$ | .7842$_{\pm.58}$ | .7578$_{\pm.62}$ |
| | ProxySPEX | **.8564**$_{\pm.52}$ | **.8403**$_{\pm.62}$ | **.8203**$_{\pm.60}$ | **.8042**$_{\pm.61}$ | **.7822**$_{\pm.53}$ |
| ViT-B/32 FLICKR | GEFI-PM | .7026$_{\pm.20}$ | .6997$_{\pm.21}$ | .6875$_{\pm.19}$ | .6865$_{\pm.20}$ | .6733$_{\pm.21}$ |
| | exCLIP | .6016$_{\pm.23}$ | .6040$_{\pm.18}$ | .6133$_{\pm.19}$ | .6074$_{\pm.17}$ | .6011$_{\pm.21}$ |
| | Faith-Shap | **.7256**$_{\pm.20}$ | **.7212**$_{\pm.20}$ | **.7114**$_{\pm.21}$ | **.7085**$_{\pm.18}$ | .6919$_{\pm.19}$ |
| | FIxLIP | .7139$_{\pm.21}$ | .7041$_{\pm.20}$ | .6924$_{\pm.21}$ | .6895$_{\pm.20}$ | .6655$_{\pm.20}$ |
| | ProxySPEX | .7222$_{\pm.20}$ | .7163$_{\pm.20}$ | .7056$_{\pm.19}$ | .7026$_{\pm.18}$ | **.6919**$_{\pm.19}$ |
| ViT-B/16 COCO | GEFI-PM | .8066$_{\pm.26}$ | .7803$_{\pm.26}$ | .7598$_{\pm.26}$ | .7466$_{\pm.27}$ | .7305$_{\pm.25}$ |
| | exCLIP | .7661$_{\pm.25}$ | .7446$_{\pm.25}$ | .7295$_{\pm.25}$ | .7173$_{\pm.26}$ | .7080$_{\pm.24}$ |
| | Faith-Shap | .8760$_{\pm.22}$ | .8491$_{\pm.22}$ | .8398$_{\pm.23}$ | .8267$_{\pm.22}$ | .8120$_{\pm.20}$ |
| | FIxLIP | .8550$_{\pm.25}$ | .8149$_{\pm.28}$ | .7959$_{\pm.30}$ | .7681$_{\pm.30}$ | .7515$_{\pm.30}$ |
| | ProxySPEX | **.9033**$_{\pm.22}$ | **.8828**$_{\pm.23}$ | **.8765**$_{\pm.25}$ | **.8643**$_{\pm.25}$ | **.8472**$_{\pm.21}$ |
| ViT-B/16 FLICKR | GEFI-PM | .7031$_{\pm.22}$ | .7056$_{\pm.21}$ | .7002$_{\pm.23}$ | .7061$_{\pm.24}$ | .6807$_{\pm.22}$ |
| | exCLIP | .6748$_{\pm.19}$ | .6851$_{\pm.20}$ | .6885$_{\pm.24}$ | .6860$_{\pm.23}$ | .6816$_{\pm.21}$ |
| | Faith-Shap | .7974$_{\pm.21}$ | .7861$_{\pm.19}$ | .7793$_{\pm.21}$ | .7725$_{\pm.20}$ | .7646$_{\pm.20}$ |
| | FIxLIP | .7812$_{\pm.28}$ | .7563$_{\pm.25}$ | .7451$_{\pm.29}$ | .7412$_{\pm.28}$ | .7207$_{\pm.23}$ |
| | ProxySPEX | **.8267**$_{\pm.20}$ | **.8286**$_{\pm.20}$ | **.8203**$_{\pm.24}$ | **.8218**$_{\pm.24}$ | **.8022**$_{\pm.17}$ |

Table 13: Input perturbation results **without contrastive pair**: Diff-AUPC (±std.) ↑.

| Diff-AUPC | Method | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|---|
| ViT-B/32 COCO | GEFI-PM | **.3972**$_{\pm.41}$ | **.3210**$_{\pm.40}$ | **.2886**$_{\pm.43}$ | **.2230**$_{\pm.40}$ | **.1740**$_{\pm.41}$ |
| | exCLIP | .0610$_{\pm.36}$ | .0167$_{\pm.32}$ | .0005$_{\pm.30}$ | -.0207$_{\pm.32}$ | -.0267$_{\pm.28}$ |
| | Faith-Shap | .3601$_{\pm.41}$ | .3042$_{\pm.38}$ | .2517$_{\pm.38}$ | .2057$_{\pm.31}$ | .1720$_{\pm.38}$ |
| | FIxLIP | .3320$_{\pm.40}$ | .2883$_{\pm.44}$ | .2354$_{\pm.41}$ | .1951$_{\pm.44}$ | .1523$_{\pm.43}$ |
| | ProxySPEX | .3359$_{\pm.35}$ | .2759$_{\pm.42}$ | .2172$_{\pm.33}$ | .1770$_{\pm.35}$ | .1423$_{\pm.29}$ |
| ViT-B/32 FLICKR | GEFI-PM | **.3093**$_{\pm.32}$ | **.2986**$_{\pm.31}$ | **.2595**$_{\pm.28}$ | **.2688**$_{\pm.32}$ | **.2344**$_{\pm.35}$ |
| | exCLIP | .0011$_{\pm.26}$ | .0032$_{\pm.21}$ | .0200$_{\pm.26}$ | .0049$_{\pm.25}$ | -.0066$_{\pm.24}$ |
| | Faith-Shap | .2507$_{\pm.27}$ | .2261$_{\pm.26}$ | .1996$_{\pm.23}$ | .1936$_{\pm.28}$ | .1576$_{\pm.20}$ |
| | FIxLIP | .2402$_{\pm.31}$ | .2122$_{\pm.27}$ | .1886$_{\pm.26}$ | .1821$_{\pm.29}$ | .1410$_{\pm.34}$ |
| | ProxySPEX | .2322$_{\pm.27}$ | .1998$_{\pm.24}$ | .1675$_{\pm.23}$ | .1597$_{\pm.23}$ | .1313$_{\pm.20}$ |
| ViT-B/16 COCO | GEFI-PM | **.3765**$_{\pm.31}$ | **.3047**$_{\pm.34}$ | **.2554**$_{\pm.35}$ | **.2281**$_{\pm.36}$ | **.2118**$_{\pm.35}$ |
| | exCLIP | .0947$_{\pm.27}$ | .0425$_{\pm.26}$ | .0059$_{\pm.26}$ | -.0179$_{\pm.29}$ | -.0176$_{\pm.25}$ |
| | Faith-Shap | .2566$_{\pm.22}$ | .1782$_{\pm.20}$ | .1487$_{\pm.19}$ | .1152$_{\pm.17}$ | .1080$_{\pm.17}$ |
| | FIxLIP | .3113$_{\pm.39}$ | .2262$_{\pm.41}$ | .1819$_{\pm.44}$ | .1252$_{\pm.43}$ | .1076$_{\pm.45}$ |
| | ProxySPEX | .2852$_{\pm.27}$ | .2227$_{\pm.26}$ | .1948$_{\pm.26}$ | .1658$_{\pm.28}$ | .1504$_{\pm.24}$ |
| ViT-B/16 FLICKR | GEFI-PM | **.2874**$_{\pm.32}$ | **.2759**$_{\pm.32}$ | **.2544**$_{\pm.33}$ | **.2600**$_{\pm.35}$ | **.2063**$_{\pm.29}$ |
| | exCLIP | -.0033$_{\pm.23}$ | .0096$_{\pm.24}$ | .0099$_{\pm.26}$ | .0104$_{\pm.27}$ | -.0051$_{\pm.28}$ |
| | Faith-Shap | .1704$_{\pm.19}$ | .1378$_{\pm.16}$ | .1161$_{\pm.19}$ | .1029$_{\pm.14}$ | .0898$_{\pm.13}$ |
| | FIxLIP | .2445$_{\pm.39}$ | .1903$_{\pm.39}$ | .1674$_{\pm.42}$ | .1577$_{\pm.40}$ | .1069$_{\pm.37}$ |
| | ProxySPEX | .2043$_{\pm.24}$ | .1964$_{\pm.26}$ | .1694$_{\pm.28}$ | .1697$_{\pm.27}$ | .1301$_{\pm.23}$ |

Table 14: Additional object localization results on ResNet-based CLIP models: PGA ↑.

| Method | RN50 | | | RN50x4 | | |
|---|---|---|---|---|---|---|
| | COCO | | Flickr | COCO | | Flickr |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.9359** | **.8988** | **.7987** | **.9322** | **.8817** | **.7705** |
| exCLIP | .8374 | .8049 | .6905 | .7028 | .6658 | .6030 |
| Faith-Shap | .8519 | .7588 | .6914 | .8609 | .7895 | .6802 |
| FIxLIP | .8509 | .7624 | .6538 | .7525 | .6847 | .3236 |
| ProxySPEX | .8473 | .7425 | .6745 | .8365 | .7642 | .6341 |
| Random | .5763 | .4508 | .3669 | .5736 | .4616 | .4092 |

# F Extended Evaluation on Additional Dual-Encoder Models

Due to space constraints, Section 6 focuses on the ViT-based CLIP backbone given its popularity. To enhance the generalizability of our findings, this section extends the evaluation to a broader set of dual-encoder vision-language models.

**Additional Models.** We carefully select two ResNet-based CLIP models, RN50 and RN50x4 [1], and two SigLIP-2 models, SigLIP2-B/16 and SigLIP2-L/16 [44]. The ResNet-based CLIP models introduce an alternative visual encoder architecture, and SigLIP-2 represents a different contrastive learning regime. Together, these choices provide broader coverage of representative solutions for dual-encoder architectures and training paradigms, enhancing empirical evidence for the main claims of the paper.

Throughout the extended experiments, the evaluation protocols, datasets, and metrics strictly follow those used to produce the main results, while the spatial granularity and input resolutions are adjusted to fit the respective model configurations. Specifically, RN50 and RN50x4 operate with input resolutions of $224 \times 224$ and $288 \times 288$, respectively. SigLIP2-B/16 and SigLIP2-L/16 use ViT encoders with a patch size of 16 and input resolutions of $224 \times 224$ and $256 \times 256$, respectively. Since exCLIP is designed for OpenAI CLIP models and is not directly applicable to SigLIP-2, we report its results as unavailable for the SigLIP-2 experiments.

## F.1 Results of Extended Evaluation

Tables 14–19 report the additional object localization results. On RN50 and RN50x4, GEFI-PM achieves the best performance across all datasets, annotation types, and localization metrics. This suggests that the advantage of GEFI-PM does not depend on the ViT patch-token structure, but rather stems from its ability to estimate fine-grained cross-modal interactions at the pixel level. Similar observations apply to the results on SigLIP2-B/16 and SigLIP2-L/16; GEFI-PM consistently outperforms the competing methods across the reported localization settings. These results indicate that the proposed approach remains effective across model families with different training objectives and architectural choices.

Table 20 further reports input perturbation results with contrastive pairs. For the ResNet-based CLIP models, GEFI-PM obtains the best ABPC-DEL, ABPC-INS, and Diff-ABPC scores in every test case. For SigLIP-2, GEFI-PM achieves the strongest ABPC-DEL and Diff-ABPC scores across both datasets and model scales, although ABPC-INS alone is less conclusive. The previous discussion in Appendix E.2 provides an interpretation of this observation: deletion and insertion can emphasize different aspects of perturbation-based evaluation, and Diff-ABPC provides a more comprehensive interaction-specific assessment by aggregating both directions.

These additional results align with the main findings in Section 6, providing empirical evidence for GEFI-PM's generalizability when explaining different models. This conclusion generally aligns with our theoretical foundation, which does not assume a predefined patch-based input structure.

Table 15: Additional object localization results on SigLIP-2 models: PGA ↑.

| Method | SigLIP2-B/16 | | | SigLIP2-L/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | Flickr | COCO | | Flickr |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.8277** | **.7842** | **.6260** | **.7531** | **.6714** | **.5245** |
| exCLIP | – | – | – | – | – | – |
| Faith-Shap | .6261 | .5524 | .4441 | .3988 | .3091 | .3213 |
| FIxLIP | <u>.6448</u> | <u>.5773</u> | <u>.4700</u> | .5329 | <u>.4654</u> | <u>.4155</u> |
| ProxySPEX | .6217 | .5409 | .3850 | .4307 | .3455 | .3426 |
| Random | .5826 | .4556 | .3564 | <u>.5808</u> | .4547 | .3629 |

Table 16: Additional object localization results on ResNet-based CLIP models: IoU ($\pm$std.) ↑.

| Method | RN50 | | | RN50x4 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.5306**$_{\pm.17}$ | **.4942**$_{\pm.19}$ | **.4145**$_{\pm.20}$ | **.4911**$_{\pm.17}$ | **.4708**$_{\pm.18}$ | **.3861**$_{\pm.19}$ |
| exCLIP | .3373$_{\pm.16}$ | .3259$_{\pm.16}$ | .2826$_{\pm.15}$ | .3279$_{\pm.15}$ | .3072$_{\pm.15}$ | .2604$_{\pm.14}$ |
| Faith-Shap | <u>.4258</u>$_{\pm.17}$ | <u>.3733</u>$_{\pm.17}$ | .3145$_{\pm.18}$ | <u>.3949</u>$_{\pm.17}$ | <u>.3598</u>$_{\pm.17}$ | <u>.2747</u>$_{\pm.17}$ |
| FIxLIP | .4194$_{\pm.17}$ | .3672$_{\pm.17}$ | .3006$_{\pm.18}$ | .3472$_{\pm.18}$ | .3075$_{\pm.18}$ | .1695$_{\pm.16}$ |
| ProxySPEX | .4251$_{\pm.17}$ | .3725$_{\pm.17}$ | <u>.3157</u>$_{\pm.18}$ | .3850$_{\pm.17}$ | .3482$_{\pm.17}$ | .2692$_{\pm.17}$ |
| Random | .1495$_{\pm.15}$ | .0923$_{\pm.11}$ | .1142$_{\pm.12}$ | .1493$_{\pm.15}$ | .0922$_{\pm.11}$ | .1142$_{\pm.12}$ |

Table 17: Additional object localization results on SigLIP-2 models: IoU ($\pm$std.) ↑.

| Method | SigLIP2-B/16 | | | SigLIP2-L/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.4226**$_{\pm.20}$ | **.3957**$_{\pm.21}$ | **.3014**$_{\pm.20}$ | **.3629**$_{\pm.21}$ | **.3255**$_{\pm.21}$ | **.2649**$_{\pm.20}$ |
| exCLIP | – | – | – | – | – | – |
| Faith-Shap | <u>.2613</u>$_{\pm.19}$ | .1997$_{\pm.15}$ | <u>.1933</u>$_{\pm.16}$ | <u>.2371</u>$_{\pm.19}$ | .1593$_{\pm.15}$ | .1827$_{\pm.16}$ |
| FIxLIP | .2465$_{\pm.16}$ | <u>.2142</u>$_{\pm.15}$ | .1746$_{\pm.13}$ | .2121$_{\pm.16}$ | <u>.1730</u>$_{\pm.14}$ | .1617$_{\pm.13}$ |
| ProxySPEX | .2512$_{\pm.18}$ | .1862$_{\pm.15}$ | .1914$_{\pm.16}$ | .2362$_{\pm.19}$ | .1602$_{\pm.14}$ | <u>.1861</u>$_{\pm.16}$ |
| Random | .1514$_{\pm.16}$ | .0934$_{\pm.11}$ | .1130$_{\pm.12}$ | .1512$_{\pm.16}$ | .0931$_{\pm.11}$ | .1129$_{\pm.12}$ |

Table 18: Additional object localization results on ResNet-based CLIP models: PGE ($\pm$std.) ↑.

| Method | RN50 | | | RN50x4 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.7360**$_{\pm.25}$ | **.6120**$_{\pm.25}$ | **.5703**$_{\pm.28}$ | **.7019**$_{\pm.24}$ | **.5908**$_{\pm.24}$ | **.5421**$_{\pm.28}$ |
| exCLIP | .5724$_{\pm.27}$ | <u>.4901</u>$_{\pm.25}$ | .4558$_{\pm.43}$ | .4932$_{\pm.22}$ | .4106$_{\pm.20}$ | .3925$_{\pm.21}$ |
| Faith-Shap | <u>.5880</u>$_{\pm.23}$ | .4707$_{\pm.22}$ | <u>.4869</u>$_{\pm.26}$ | <u>.6547</u>$_{\pm.23}$ | <u>.5399</u>$_{\pm.22}$ | <u>.5068</u>$_{\pm.29}$ |
| FIxLIP | .5123$_{\pm.22}$ | .3989$_{\pm.20}$ | .3824$_{\pm.22}$ | .4516$_{\pm.22}$ | .3520$_{\pm.20}$ | .2531$_{\pm.26}$ |
| ProxySPEX | .5635$_{\pm.23}$ | .4450$_{\pm.21}$ | .4473$_{\pm.25}$ | .5967$_{\pm.23}$ | .4851$_{\pm.22}$ | .4354$_{\pm.26}$ |
| Random | .2345$_{\pm.19}$ | .1551$_{\pm.14}$ | .1884$_{\pm.16}$ | .2344$_{\pm.19}$ | .1551$_{\pm.14}$ | .1884$_{\pm.16}$ |

Table 19: Additional object localization results on SigLIP-2 models: PGE ($\pm$std.) ↑.

| Method | SigLIP2-B/16 | | | SigLIP2-L/16 | | |
|---|---|---|---|---|---|---|
| | COCO | | FLICKR | COCO | | FLICKR |
| | BBOX | SEG | BBOX | BBOX | SEG | BBOX |
| GEFI-PM | **.5836**$_{\pm.28}$ | **.4780**$_{\pm.27}$ | **.4045**$_{\pm.28}$ | **.4791**$_{\pm.25}$ | **.3746**$_{\pm.23}$ | **.3596**$_{\pm.26}$ |
| exCLIP | – | – | – | – | – | – |
| Faith-Shap | .4147$_{\pm.31}$ | .3379$_{\pm.28}$ | .2811$_{\pm.30}$ | .2346$_{\pm.25}$ | .1735$_{\pm.21}$ | .1918$_{\pm.26}$ |
| FIxLIP | .3995$_{\pm.97}$ | .3032$_{\pm.79}$ | .2658$_{\pm.36}$ | .3113$_{\pm.21}$ | .2282$_{\pm.17}$ | .2364$_{\pm.18}$ |
| ProxySPEX | .4827$_{\pm.29}$ | .3978$_{\pm.27}$ | .3101$_{\pm.28}$ | .3192$_{\pm.34}$ | .2469$_{\pm.30}$ | .2074$_{\pm.32}$ |
| Random | .2372$_{\pm.19}$ | .1568$_{\pm.14}$ | .1868$_{\pm.16}$ | .2371$_{\pm.19}$ | .1567$_{\pm.14}$ | .1867$_{\pm.16}$ |

Table 20: Additional input perturbation results with contrastive pairs following second-order metrics.

| Test Case | Method | ABPC-DEL ↓ | ABPC-INS ↑ | Diff-ABPC ↑ |
|---|---|---|---|---|
| RN50 COCO | GEFI-PM | **-.1727**$_{\pm.25}$ | **.0054**$_{\pm.20}$ | **.1781**$_{\pm.32}$ |
| | exCLIP | -.1379$_{\pm.24}$ | -.0601$_{\pm.17}$ | .0778$_{\pm.19}$ |
| | Faith-Shap | -.1444$_{\pm.23}$ | -.0032$_{\pm.15}$ | .1412$_{\pm.24}$ |
| | FIxLIP | -.1439$_{\pm.24}$ | -.0048$_{\pm.16}$ | .1392$_{\pm.26}$ |
| | ProxySPEX | -.1368$_{\pm.23}$ | -.0042$_{\pm.15}$ | .1327$_{\pm.24}$ |
| RN50 FLICKR | GEFI-PM | **-.1310**$_{\pm.14}$ | **.0104**$_{\pm.09}$ | **.1415**$_{\pm.13}$ |
| | exCLIP | -.1069$_{\pm.13}$ | -.0397$_{\pm.11}$ | .0672$_{\pm.10}$ |
| | Faith-Shap | -.1093$_{\pm.13}$ | -.0070$_{\pm.09}$ | .1023$_{\pm.12}$ |
| | FIxLIP | -.1110$_{\pm.13}$ | -.0081$_{\pm.09}$ | .1030$_{\pm.12}$ |
| | ProxySPEX | -.1052$_{\pm.13}$ | -.0074$_{\pm.09}$ | .0978$_{\pm.12}$ |
| RN50x4 COCO | GEFI-PM | **-.1747**$_{\pm.25}$ | **.0479**$_{\pm.18}$ | **.2227**$_{\pm.25}$ |
| | exCLIP | -.1249$_{\pm.24}$ | -.0267$_{\pm.18}$ | .0981$_{\pm.17}$ |
| | Faith-Shap | -.1277$_{\pm.23}$ | .0188$_{\pm.15}$ | .1465$_{\pm.21}$ |
| | FIxLIP | -.1271$_{\pm.23}$ | .0041$_{\pm.17}$ | .1312$_{\pm.20}$ |
| | ProxySPEX | -.1196$_{\pm.23}$ | .0161$_{\pm.15}$ | .1357$_{\pm.20}$ |
| RN50x4 FLICKR | GEFI-PM | **-.1465**$_{\pm.21}$ | **.0506**$_{\pm.15}$ | **.1971**$_{\pm.20}$ |
| | exCLIP | -.1030$_{\pm.19}$ | -.0174$_{\pm.16}$ | .0856$_{\pm.13}$ |
| | Faith-Shap | -.0928$_{\pm.19}$ | .0088$_{\pm.14}$ | .1016$_{\pm.15}$ |
| | FIxLIP | -.0741$_{\pm.18}$ | -.0265$_{\pm.16}$ | .0476$_{\pm.14}$ |
| | ProxySPEX | -.0897$_{\pm.18}$ | .0066$_{\pm.14}$ | .0963$_{\pm.15}$ |
| SigLIP2-B/16 COCO | GEFI-PM | **-.0977**$_{\pm.29}$ | .0177$_{\pm.26}$ | **.1153**$_{\pm.25}$ |
| | exCLIP | – | – | – |
| | Faith-Shap | -.0403$_{\pm.22}$ | .0178$_{\pm.19}$ | .0581$_{\pm.14}$ |
| | FIxLIP | -.0677$_{\pm.25}$ | **.0320**$_{\pm.25}$ | .0997$_{\pm.27}$ |
| | ProxySPEX | -.0396$_{\pm.23}$ | .0086$_{\pm.19}$ | .0483$_{\pm.12}$ |
| SigLIP2-B/16 FLICKR | GEFI-PM | **-.1259**$_{\pm.41}$ | .0081$_{\pm.34}$ | **.1341**$_{\pm.36}$ |
| | exCLIP | – | – | – |
| | Faith-Shap | -.0468$_{\pm.30}$ | .0033$_{\pm.26}$ | .0502$_{\pm.19}$ |
| | FIxLIP | -.0854$_{\pm.38}$ | **.0219**$_{\pm.32}$ | .1072$_{\pm.38}$ |
| | ProxySPEX | -.0450$_{\pm.31}$ | -.0031$_{\pm.27}$ | .0419$_{\pm.18}$ |
| SigLIP2-L/16 COCO | GEFI-PM | **-.1232**$_{\pm.51}$ | -.0225$_{\pm.51}$ | **.1007**$_{\pm.45}$ |
| | exCLIP | – | – | – |
| | Faith-Shap | -.0465$_{\pm.37}$ | -.0137$_{\pm.36}$ | .0328$_{\pm.16}$ |
| | FIxLIP | -.0789$_{\pm.40}$ | **.0062**$_{\pm.40}$ | .0851$_{\pm.45}$ |
| | ProxySPEX | -.0383$_{\pm.37}$ | -.0205$_{\pm.36}$ | .0178$_{\pm.12}$ |
| SigLIP2-L/16 FLICKR | GEFI-PM | **-.2117**$_{\pm.58}$ | -.0658$_{\pm.57}$ | **.1459**$_{\pm.55}$ |
| | exCLIP | – | – | – |
| | Faith-Shap | -.1161$_{\pm.41}$ | -.0844$_{\pm.39}$ | .0317$_{\pm.16}$ |
| | FIxLIP | -.1430$_{\pm.47}$ | **-.0281**$_{\pm.47}$ | .1149$_{\pm.60}$ |
| | ProxySPEX | -.0926$_{\pm.40}$ | -.0803$_{\pm.39}$ | .0123$_{\pm.12}$ |

# G Implementation Details of GEFI-PM

This section presents in-depth discussions of implementation details of GEFI-PM, including sampling strategies and the impacts of hyperparameters. Unless specified otherwise, the results reported in this selection are collected from tests with the ViT-B/32 model on the COCO dataset, with the total query budget fixed at $5,000$.

## G.1 Mask Sampling Strategies

According to Eq. (5), GEFI describes an independent sampling process in contrast to the standard Shapley measure, which relies strictly on the quadruples that compose the second-order difference operator. This flexible process enables the incorporation of many existing sampling strategies. During our implementation of GEFI-PM, we first consider antithetic sampling [32], a powerful variance-reduction technique broadly recognized in the gradient-estimation literature [23]. The implementation of antithetic sampling under the framework of GEFI-PM is straightforward: for each constructed mask and its corresponding path indicator $(\boldsymbol{\epsilon}, \alpha)$, a mirroring pair will be added to the total mask set, namely $(\bar{\boldsymbol{\epsilon}}, 1-\alpha)$. The deployment of antithetic sampling halves the initial mask set size, since it is doubled after mirroring. The mirroring process is applied to both image and text spaces.

In addition to antithetic sampling, we further consider the dimensionality issue in the image space. Since GEFI-PM operates on the pixel level, reliable interaction estimation is unrealistic given a budget of $5,000$, which is much smaller than the feature space dimensionality. Implementing GEFI-PM without additional sampling refinement generally leads to poor visual results. To this end, we consider two existing mask sampling strategies from the literature: mask smoothing [9] and mask resizing [21]. Despite the different implementation choices, the underlying concepts of the two strategies are similar: they softly group pixels by assigning similar mask values to adjacent features. As a result, the sampling process will tend to create masks that cover larger regions rather than injecting isolated noises. This will largely enhance the effectiveness of queries to a model by highlighting its sensitivities to the absence of certain input components. While these sampling strategies also reduce the effective feature space, they are fundamentally different from fixed image patches, since they allow interaction patterns to be captured in arbitrary spatial forms and preserve pixel-level presentation. In the following, we elaborate on each sampling strategy with details.

**Pixel (Naïve Sampling)** Independent pixel-wise sampling can be applied as the default choice for GEFI-PM. To sample $\boldsymbol{\epsilon} \sim \mathrm{Bernoulli}(\alpha)$, one can start by initializing random mask values $\boldsymbol{\varepsilon}$ from a uniform distribution, which is followed by a binarization process once the path indicator $\alpha \sim \mathcal{U}_{[0,1]}$ is sampled, *i.e.*, $\boldsymbol{\epsilon} = \boldsymbol{\varepsilon} > \boldsymbol{\alpha}$. As previously mentioned, the performance of this default option will collapse as the dimensionality increases.

**Smooth (Mask Smoothing)** Mask smoothing was first described in [8] and further adopted by [9]. This can be seen as a post-processing of the initialized continuous-value masks $\boldsymbol{\varepsilon}$ described in the naïve approach. Specifically, a Gaussian kernel will be applied to the raw masks to assign similar mask values to adjacent features. Since the Gaussian process will also change the underlying distribution of each mask value, the path indicator $\alpha$ will follow a quantile interpretation during binarization rather than specifying a concrete threshold. Compared to isolated pixel-level sampling, mask smoothing tends to generate masks that cover broader regions, which can be of arbitrary shapes.

**Resize (Mask Resizing)** Mask resizing is similar to the previous alternative in concept but follows a different implementation. It defines an initial mask size for $\boldsymbol{\varepsilon}$, which is then resized to the target shape by bilinear interpolation after instantiation. To ensure equivalent sampling distributions over individual features, a spatial shift is applied along both the x- and y-axes, with the shift step randomly between 0 and the rounded ratio of the target size to the initial mask size. The binarization process for mask resizing is identical to the smoothing alternative following a quantile interpretation of $\alpha$. Among the three alternatives, mask resizing generates masks that span the largest regions. Although the resulting mask shapes are biased towards squares, this design still has an advantage over the others as it aligns with the square partitioning of image patches in ViT.

For the experimental results reported in the main text, we adopt *mask resizing* as the default sampling strategy with an initial mask size of $7 \times 7$, corresponding to a partition of the image space into patches of size 32.

Table 21: Quantitative evaluation of GEFI-PM variants across different image mask sampling strategies.

| Evaluation | | Metrics | GEFI-PM Variants | | |
|---|---|---|---|---|---|
| | | | Pixel | Smooth | Resize |
| Object Localization | BBOX | PGA ↑ | .5926 | .8591 | **.9151** |
| | | IoU ↑ | $.1511_{\pm.15}$ | $.3270_{\pm.15}$ | $\mathbf{.4985}_{\pm.17}$ |
| | | PGE ↑ | $.2369_{\pm.19}$ | $.4601_{\pm.20}$ | $\mathbf{.6965}_{\pm.25}$ |
| | SEG | PGA ↑ | .4724 | .7868 | **.8826** |
| | | IoU ↑ | $.0946_{\pm.11}$ | $.3252_{\pm.15}$ | $\mathbf{.4794}_{\pm.19}$ |
| | | PGE ↑ | $.1576_{\pm.14}$ | $.3843_{\pm.18}$ | $\mathbf{.5854}_{\pm.25}$ |
| Input Perturbation (Normalized) | ABPC-DEL↓ | | $-.1407_{\pm.23}$ | $\mathbf{-.1741}_{\pm.29}$ | $-.1642_{\pm.28}$ |
| | ABPC-INS ↑ | | $-.0331_{\pm.17}$ | $-.0102_{\pm.15}$ | $\mathbf{-.0040}_{\pm.15}$ |
| | Diff-ABPC ↑ | | $.1077_{\pm.17}$ | $.1638_{\pm.27}$ | $\mathbf{.1682}_{\pm.29}$ |

### G.1.1 Empirical Comparison of Different Sampling Strategies

To complement the discussion of different sampling strategies, Table 21 offers a brief comparison among the three alternatives described in the previous section. For mask smoothing, a Gaussian kernel with width 31 and standard deviation 5 is applied; the initial mask size for resizing is $7 \times 7$ as used in the main tests. The performance of the alternatives is reported with the six metrics concerning the two design tests.

In line with the previous discussion, resizing generally achieves the best performance among the alternatives, given its explicit alignment with the prior knowledge about the square patch partitioning. The naïve pixel-level sampling struggles in all tests. As anticipated, the exponentially large search space renders reliable proxy gradient estimation nearly impossible, resulting in noisy visual results and humble quantitative performance. Mask smoothing offers greater sampling flexibility and also demonstrates competitive performance compared to resizing in the input perturbation tests; however, its results in object localization tests are considerably lower. This discrepancy likely occurs because spatial smoothing exploits no inductive bias about feature organization, which is insufficient to suppress the high-frequency noise that disrupts precise spatial grounding.

The reported small-scale empirical results support the analytical discussion on the sampling strategies, which lead to our choice of mask resizing for implementing GEFI-PM.

## G.2 Ablation on Hyperparameters

After fixing the sampling strategy to mask resizing, GEFI-PM now has two primary hyperparameters that configure its deployment: 1. the initial mask size, 2. the image mask ratio. In the following discussion, we refer to the initial mask size by the corresponding patch size to align with the common practice in the literature. Specifically, a patch size of $32 \times 32$ refers to an initial mask resolution of $7 \times 7$, and similarly, $16 \times 16$ corresponds to a $14 \times 14$ mask size. Please note that the initial mask size is flexible, as it does not need to be an integer divisor of the target image shape. As for the image mask ratio, it is another aspect of configuration, allocating the query budget $n = n_{\text{img}} + n_{\text{txt}}$ to optimize explanation quality. Given the significantly larger dimensionality of the image space, it is natural to narrow its potential value range to $[0.5, 1.0]$. While a greedy strategy would be to allocate the most budget to $n_{\text{img}}$, this leads to a cost of reduced size of the observation matrix, which is given by $n_{\text{img}} \cdot n_{\text{txt}}$. Solving this optimization task appears non-trivial and, very likely, is task-dependent. In our main results, we empirically set the image budget ratio to $0.8$. To better understand the effects of these two hyperparameters, we further extend our experiments to test GEFI-PM's performance under different configurations, aiming to empirically reveal their joint effects.

As previously mentioned, we set the image budget ratio to $0.8$ and the spatial patch size to $32$ as the default configuration. To investigate the robustness of GEFI-PM to these parameters, we employ a two-step search. We first perform an initial exploration to delineate the viable search space for each hyperparameter. During this phase, we isolate the two hyperparameters, fixing one at its default value while varying the other over a candidate range. This allows us to identify the appropriate bounds for

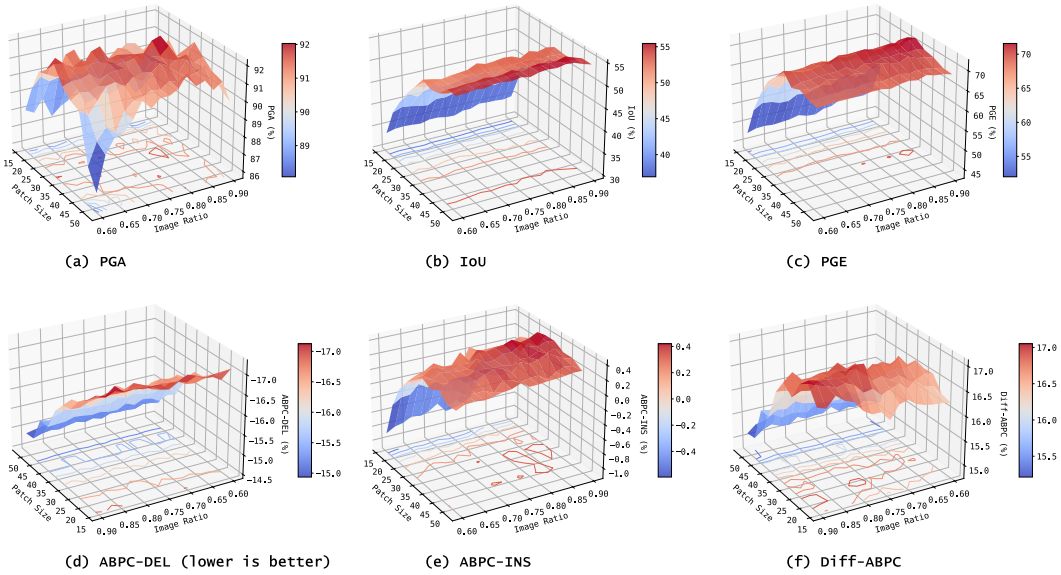



Figure 5: The performance of GEFI-PM across primary hyperparameter configurations.

larger-scale tests. Subsequently, we perform a fine-grained grid search within the identified search ranges to obtain the final reported metrics.

Figure 5 illustrates the results of the fine-grained hyperparameter search across the object localization and input perturbation benchmarks, using bounding boxes as ground-truth masks. Our analysis reveals several key insights regarding the sensitivity of GEFI-PM to its core hyperparameter. For object localization, we observe that intermediate-to-large patch sizes are generally advantageous. This performance gain is likely attributed to the increased capacity of larger patches to capture the full spatial extent of the target object. Additionally, a larger patch size will encourage the explainer to concentrate the mass on the most significant region; therefore, it is beneficial and likely biased for the signal object tests. As for image budget, a higher ratio is generally preferable, consistent with the fact that the significantly higher dimensionality of the image space than that of the text space. Conversely, input perturbation metrics benefit from small-to-medium patch sizes, while maintaining a preference for higher image-budget ratios. This suggests that small patches enable GEFI-PM to sharply identify subtle visual nuances, which are critical to the inference by CLIP. Again, the discrepancies in the two sets of tests underline the complementary nature of object localization and input perturbation, as they focus on different aspects of explanation quality. Difficulties to simultaneously optimize for all metrics also highlight the consistently superior performance of GEFI-PM reported in the main results, which is a consequence of the foundational difference and improvement this paper contributes.

On the other hand, the empirical results demonstrate the robustness of GEFI-PM under different hyperparameter choices, achieving a near-optimal performance across a broad configuration space. Based on these empirical findings, a spatial patch size between 32 and 48, coupled with an image budget allocation ratio between $0.75$ and $0.9$, generally produces desirable outcomes.

## H Qualitative Evaluation

To complement the quantitative results, we provide a set of illustrative examples, which showcase the quality of cross-model interactions generated by the evaluated methods. These examples are selected from the COCO dataset, specifically targeting images containing at least two distinct semantic entities. This selection facilitates a clear demonstration of how interaction outcomes shift when conditioned on different object anchors.

For each image-text pair, we employ two evaluation protocols to better demonstrate the functioning of GEFI-PM in revealing cross-modal interactions. The first protocol quantifies the interaction scores of textual tokens anchored on a visual feature subset (specifically, the set of pixels defined by the ground-truth object segmentation). The second protocol evaluates the spatial distribution of pixels' interaction

scores anchored on the corresponding textual token of the object. Note that baseline methods operate on fixed image patches and cannot process fine-grained object segmentations. Therefore, for the baselines in the first protocol, we identify the ensemble of all image patches with at least 50% spatial overlap with the ground-truth segmentation. The final interaction scores for the textual tokens are derived by averaging the interactions across this ensemble of image patches.

Figures 6 – 9 visualize the interaction results generated by the evaluated methods across various image-text pairs. To enhance visual clarity, the interactions for the visual features are demonstrated after a point-wise multiplication of the original image and the interaction scores, which suppresses irrelevant background regions and highlights the spatial grounding of the model. Additional qualitative examples concentrating solely on the results by GEFI-PM are shown in Figures 10 and 11.

A notable observation is that GEFI-PM is the sole explainer capable of providing fine-grained, pixel-level interactions, with its results consistently demonstrating precise cross-modal alignment. For instance, in Figure 6, GEFI-PM successfully assigns the highest interaction scores to features corresponding to the "cow" and the "man" under both evaluation protocols, highlighting its ability to unravel the detailed inference logic. Interestingly, GEFI-PM assigns a strong negative interaction score to the textual token "dogs" when conditioned on the cow's spatial segmentation in the image, indicating that CLIP decisively discriminates the segmentation region as belonging to a cow rather than dogs; however, when anchored on the token "cow", GEFI-PM highlights pixels corresponding to both the dogs and the cow. This suggests a potential model bias, where CLIP conflates the dogs in the image with the cow due to shared visual patterns such as their black-and-white colors. This observation is shared by most of the tested explainers to some extent, although the highly contrastive interaction scores in the text space pose a challenge for comprehending their results. In general, the tested baseline methods are restricted to coarse, grid-aligned patch-level interactions. These results frequently lack clear semantic coherence and, in some cases, erroneously assign excessively low scores to the relevant objects in the other modality.

In summary, the qualitative evaluations provide empirical validation for our proposed GEFI-PM, demonstrating its capacity to generate highly expressive attributions with remarkable fidelity to the underlying cross-modal logic. By resolving interactions at the pixel level, GEFI-PM transcends the resolution constraints of contemporary baselines, offering an explanatory depth that elucidates the complex dependencies within the model.

# I Broader Impacts

By facilitating granular, pixel-level feature interactions, GEFI enhances the transparency of multi-modal models, a foundational requirement for trustworthiness and accountability during model deployment, particularly in high-stakes domains. We anticipate that the family of proxy-gradient-estimation approaches will equip practitioners and researchers with a toolkit to demystify the internal mechanisms of cross-modal model functioning. With its improved efficiency and flexible feature-subset definition, the greater potential of GEFI can be further realized through an interactive human-machine interface, enabling active and real-time exploration of heterogeneous interaction patterns by reusing the collected observation matrix. This future direction of interactive design can help address the inherent difficulty of representing high-dimensional information and further enhance multi-modal explainability through human-in-the-loop exploration, enabling practical use cases such as model debugging and supervision.

# J Limitations

In this work, we show that GEFI produces fine-grained pixel-level interactions with manageable computational overhead through simultaneous sampling and structured highly parallel interaction estimation. The other side of the coin is that simultaneous estimation under limited budgets, which are often much smaller than the feature space, let alone the coalition space, can render noise into the scores of inactive features. Ideally, these features should receive close-to-zero scores because they do not interact with other features or, more generally, do not contribute to the final decision. However, they may still receive spurious credits due to co-occurrence with impactful or interactive features; this phenomenon can be evident when the tested model is robust against input perturbations. This is, in fact, a common limitation shared by sampling-based approaches. That said, this problem

can be mitigated by further refining the sampling process or by leveraging more advanced variance-reduction techniques, *e.g.*, case-specific design of control variates [9]. Overall, we believe that the aforementioned limitation is addressable, and our discussion points to future directions for enhancing the quality of explanations within the proxy-gradient-estimation framework.




Figure 6: Qualitative comparison of cross-modal interactions. The specific object anchor features are highlighted in green, while the resulting interactions denote positive/negative interaction scores in red/blue, respectively.




Figure 7: Qualitative comparison of cross-modal interactions.

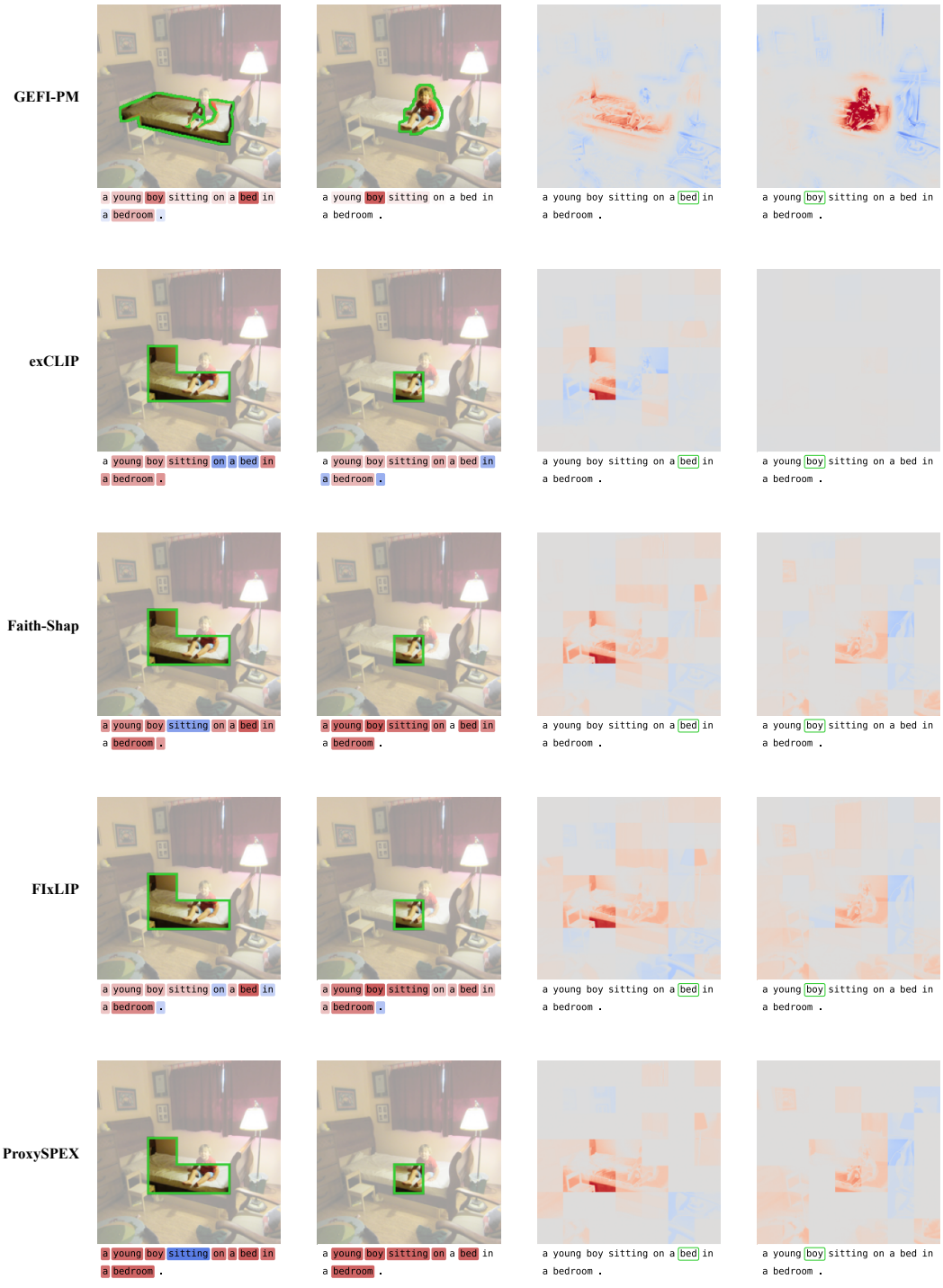



Figure 8: Qualitative comparison of cross-modal interactions.

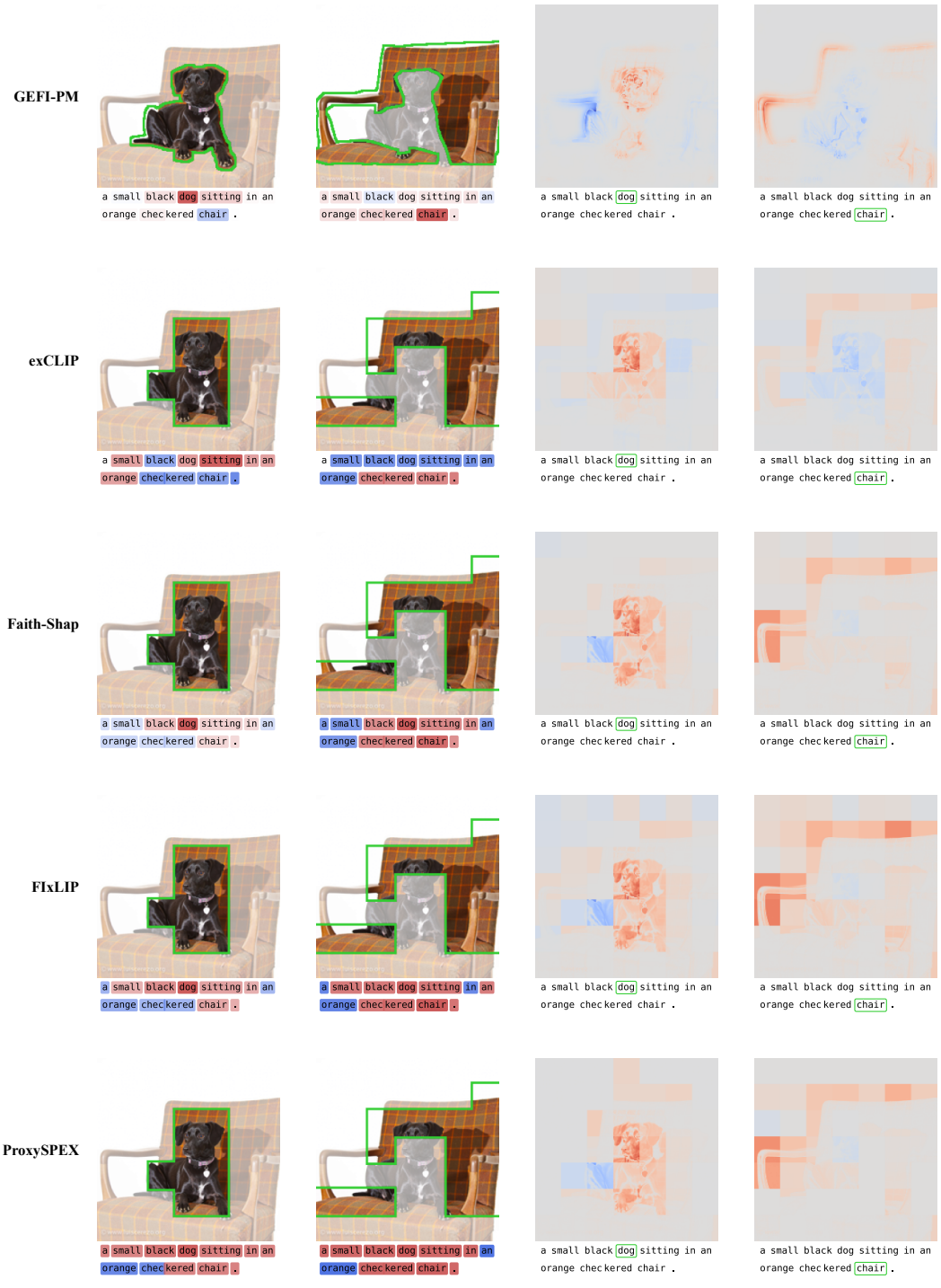



Figure 9: Qualitative comparison of cross-modal interactions.




Figure 10: Additional GEFI-PM interaction visualizations.

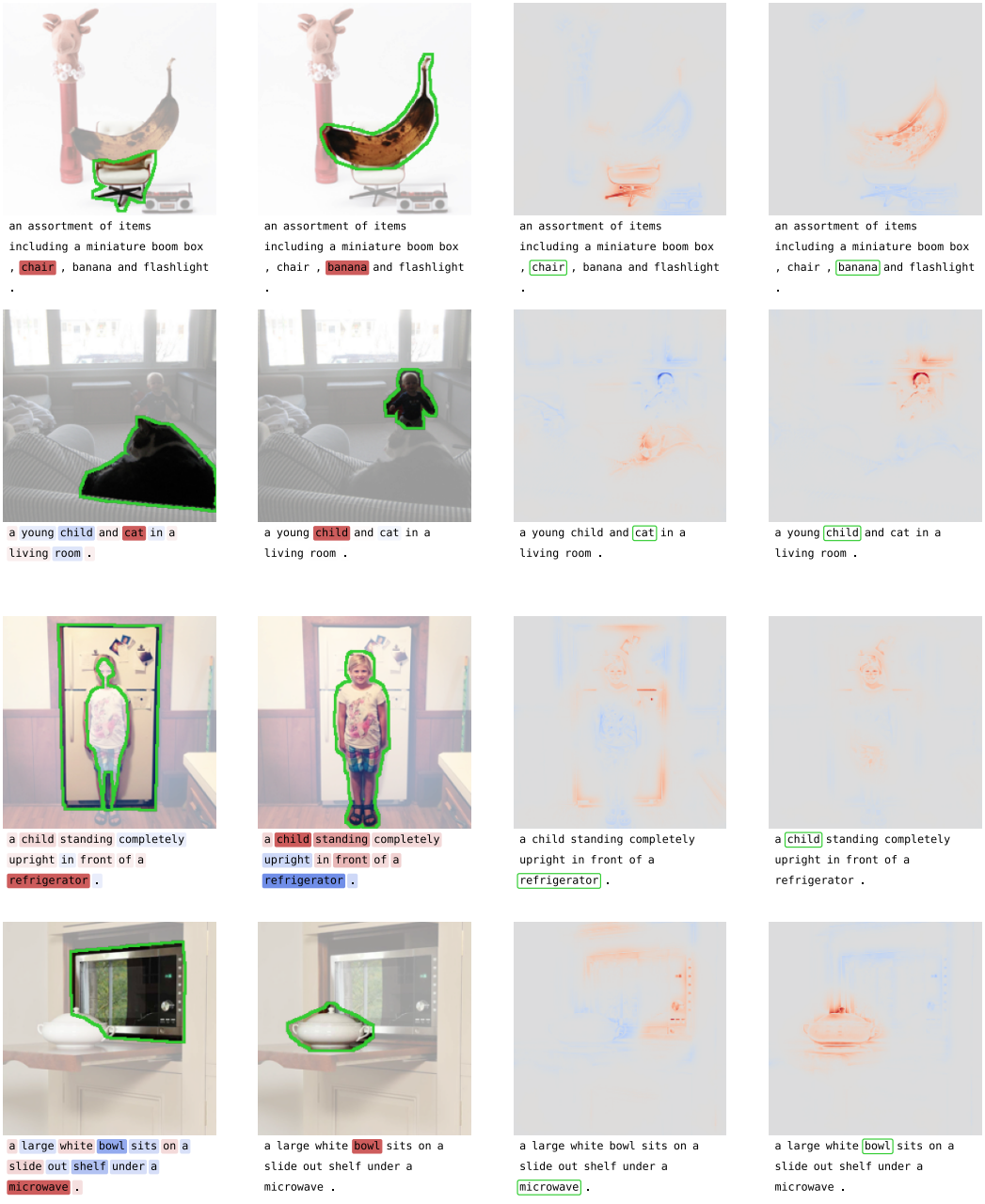



Figure 11: Additional GEFI-PM interaction visualizations.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper studies cross-modal feature interactions as reflected in the abstract and introduction. The relevant contents are presented in Sections 3, 4, and 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the presented work are discussed in Appendix J.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Formal proofs of the main claims are provided in the appendix and correctly associated in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental designs are described with necessary details in Sections 5 and 6, complemented by implementation details in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The full code implementing the proposed method will be released upon the acceptance of the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: A clear description of data splits and pre-processing is provided in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviations and distributional information (through box plots) of the results to facilitate a more comprehensive understanding.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The exact resources for the experiments are provided in Appendix D.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Authors acknowledge and closely follow the NeurIPS code of ethics during the preparation of this paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential impact is discussed with a dedicated section in Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets and codes used during experiments are correctly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.