Multi-Modal Information Bottleneck Attribution with Cross-Attention Guidance

Pauline Bourigault[1,3]
p.bourigault22@imperial.ac.uk
Emmanuelle Bourigault[2]
emmanuelle@robots.ox.ac.uk
Danilo P. Mandic[3]
d.mandic@imperial.ac.uk

[1] Department of Computing,
Imperial College London
[2] Department of Engineering,
University of Oxford
[3] Department of Electrical and
Electronic Engineering,
Imperial College London

Abstract

For the progression of interpretable machine learning, particularly in the intersection of vision and language, ensuring transparency and comprehensibility in model decisions is crucial. This work introduces an enhancement to the Multi-modal Information Bottleneck attribution method by integrating cross-attention mechanisms. This targets the core challenge of improving the interpretability of vision-language pretrained models, such as CLIP, by fostering more discerning and relevant latent representations. The proposed method filters and retains essential information across modalities, leveraging cross-attention to dynamically focus on pertinent visual and textual features for any given context. Through evaluations using CLIP as an example, we demonstrate improvements in attribution accuracy and interpretability over existing attribution methods, including gradient-based, perturbation-based, attention-based, and information-theoretic methods. By providing a more nuanced understanding of model decisions, this work contributes to offer a promising avenue for deploying vision-language models in critical domains such as healthcare.

1 Introduction

Vision-Language Pretrained Models (VLPMs) have become essential tools for a variety of vision-language tasks, leveraging vast multimodal datasets to learn intricate image-text associations [10]. For instance, the CLIP model [14] excels in tasks such as Visual Question Answering thanks to its training on 400 million image-text pairs. Yet, VLPMs like Vision Transformers (ViTs) [5], which form the backbone of models such as CLIP, are complex and opaque. This causes challenges for interpretation, which is a critical drawback for applications in sensitive areas such as healthcare or assistive technologies. Enhancing the interpretability of VLPMs is crucial for their safe, reliable, and trustworthy application in these fields. To address this challenge, attribution methods offer post-hoc explanations by assigning significance scores to input features, aiding in understanding a model's decision-making process. These methods generate visual heatmaps for vision models and

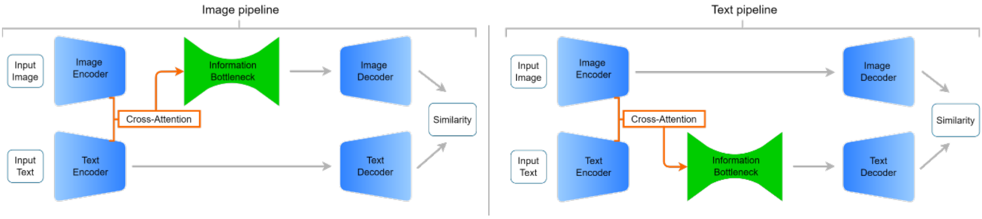

Figure 1: The CA-M2IB framework, illustrating the image pipeline (left) and text pipeline (right).

score input tokens for language models, enhancing model transparency across various tasks and architectures. Current methods are gradient-based [16, 19], perturbation-based [11], attention-based [4, 13], or information-theoretic [8, 15, 23]. Moreover, in contrast to existing attribution methods focusing primarily on unimodal models, Wang *et al.* [23] recently introduced a multi-modal attribution approach (M2IB) that leverages the information bottleneck principle [21], without needing ground-truth labels, and focuses on maximizing feature relevance across modalities. In this work, we introduce the M2IB attribution method with cross-attention guidance to dynamically modulate the information flow between modalities. This approach, termed CA-M2IB, leverages the intuition that the relevance of features in one modality can be enhanced by the contextual information provided by another modality, thereby enriching the latent representations for attribution (see Fig. 1). The cross-attention mechanism acts as a bridge that adaptively filters and aligns modality-specific embeddings based on their mutual relevance. Our evaluations demonstrate the effectiveness of CA-M2IB in pinpointing significant features for both modalities (e.g., image and text), providing a novel tool for enhancing the interpretability and trustworthiness of VLPMs in safety-critical settings.

2 Preliminaries

In this section, we delve into a simple multi-modal adaptation of the information bottleneck concept, and focus on its extension in the multi-modal context.

2.1 Information Bottleneck Concept

The information bottleneck method offers a theoretical basis for compressing neural network representations [20]. It aims to distill a stochastic latent variable Z, derived from an input X via a parametric encoder $p_{Z|X}(z \mid x;\theta)$, that captures crucial information regarding a specific target Y. This process is achieved by maximizing the mutual information between Z and Y, subject to a limitation on the mutual information between Z and X. Formally, this is defined as the optimization challenge

$$\max_\theta I(Z,Y;\theta) \quad \text{s.t.} \quad I(Z,X;\theta) \leq \bar{I}, \tag{1}$$

where $I(\cdot,\cdot;\theta)$ denotes the mutual information metric and $\bar{I}$ symbolizes a bound on information compression. This dilemma can be reframed as maximizing the function

$$\mathcal{F}(\theta) = I(Z,Y;\theta) - \beta I(Z,X;\theta). \tag{2}$$

Here, β serves as a Lagrange multiplier balancing between acquiring a latent representation highly informative of Y and ensuring the compression of the representation regarding X [1].

2.2 Multi-Modal Information Bottleneck

The information bottleneck loss in (2), which focuses on optimizing latent representations to predict specific targets from inputs, is not directly applicable here for VLPMs due to the absence of or the costly acquisition of task-specific targets Y. Instead, the aim is to define an optimization objective following the ones for self-supervised models, such as CLIP or SLIP [12, 14], that utilize pairs of text and images. The challenge of learning attribution maps in a multi-modal context significantly diverges from that in unimodal tasks. In image-text representation learning scenarios typical of VLPMs, we usually deal with text descriptions rather than explicit labels. They serve as inputs to guide us towards a representation learning objective independent of task-specific labels that relies on the multi-modal nature of the inputs. Acknowledging the intrinsic value of related information across multiple modalities, such as text descriptions pertinent to images, Wang *et al.* [23] hypothesized that an effective image encoding should encapsulate details about its corresponding text and vice versa. This led to propose a multi-modal information bottleneck objective for a modality m from the set $\mathcal{M} = \{\text{modality } 1, \text{modality } 2\}$ as

$$\mathcal{F}_m(\theta_m) = I(Z_m, E_{m'}; \theta_m) - \beta\, I(Z_m, X_m; \theta_m), \tag{3}$$

where $m' = \mathcal{M} \backslash m$ is the complementary modality to m, and $E_{m'}$ denotes the embedding of modality m'.

3 Multi-Modal Information Bottleneck with Cross-Attention Guidance for Attribution

3.1 Attribution via Cross-Attention M2IB

Based on (3), Wang *et al.* [23] introduced a multi-modal information bottleneck principle (M2IB) for attribution of VLPMs. Expanding upon this foundational work, we introduce a multi-modal information bottleneck approach, CA-M2IB, that harnesses cross-attention to refine the attribution process for VLPMs. CA-M2IB is characterized by the integration of a variational approximation, to then derive a tractable optimization objective that is fine-tuned with respect to a set of attribution parameters influenced by cross-attention dynamics. The derivation of the objective closely follows the steps in [1, 23].

Let us denote the cross-attention function that computes attention weights from modality m to m', where Θ_A represents the parameters of the attention network. These attention weights are used to create an attention-modulated embedding $\tilde{E}_{m'}$, which dynamically highlights features in $E_{m'}$ based on their relevance to Z_m. The cross-attention mechanism is formalized as

$$\tilde{E}_{m'} = \text{softmax}\left(\frac{Q(Z_m; \Theta_A) K(E_{m'}; \Theta_A)^T}{\sqrt{d_k}}\right) V(E_{m'}; \Theta_A), \tag{4}$$

where Q, K and V represent the query, key, and value matrices respectively, and d_k the scaling factor used to avoid overly large values of the dot product. We define the attribution

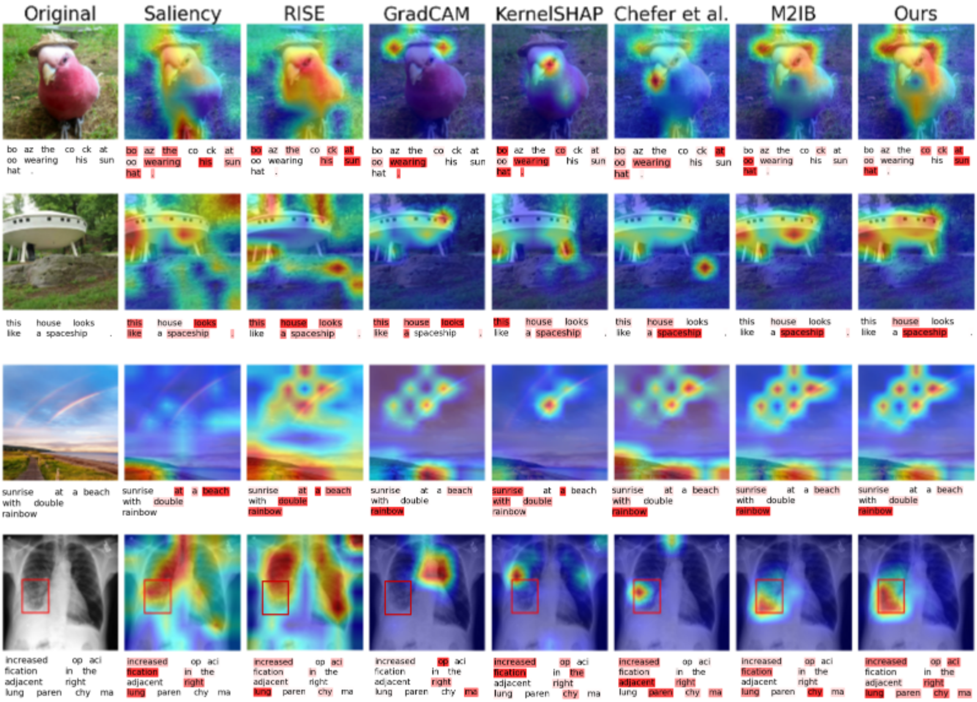
Original
Saliency
RISE
GradCAM
KernelSHAP
Chefer et al.
M2IB
Ours
bo az the co ck at oo wearing his sun hat .
this house looks like a spaceship .
sunrise at a beach with double rainbow
increased op aci fication in the adjacent right lung paren chy ma

Figure 2: Attribution maps for image-text pair examples. Ground-truth bounding boxes [2] on the fourth row related to the text are indicated with red rectangles.

objectives for each modality (*e.g.*, image and text) as

$$\mathcal{F}_{\text{image}}(\theta_{\text{image}}, \Theta_A) = \gamma_{\text{image}} I(Z_{\text{image}}, \tilde{E}_{\text{text}}; \theta_{\text{image}}, \Theta_A) - \beta_{\text{image}} I(Z_{\text{image}}, X_{\text{text}}; \theta_{\text{image}}),$$
$$\mathcal{F}_{\text{text}}(\theta_{\text{text}}, \Theta_A) = \gamma_{\text{text}} I(Z_{\text{text}}, \tilde{E}_{\text{image}}; \theta_{\text{text}}, \Theta_A) - \beta_{\text{text}} I(Z_{\text{text}}, X_{\text{image}}; \theta_{\text{text}}),$$

where $\theta_{\text{image}} = \{\lambda_{\text{image}}, \sigma_{\text{image}}, \ell_{\text{image}}\}$ and $\theta_{\text{text}} = \{\lambda_{\text{text}}, \sigma_{\text{text}}, \ell_{\text{text}}\}$ are modality-specific sets of parameters, and γ scales the influence of cross-attention. We then define the latent representation encoding process to incorporate attention-modulated embeddings. This involves adjusting the parametric encoders to utilize $\tilde{E}_{m'}$, thereby integrating cross-modal contextual information into the latent representations. Under the assumption of dimensional independence within the latent space, the encoded latent representation Z_m, given an input x_m and conditioned on both θ_m and Θ_A, is expressed as

$$Z_m \mid x_m; \theta_m, \Theta_A = h_m(x_m; \lambda_m) \odot f_m^{\ell_m}(\tilde{E}_{m'}) + \sigma_m(\mathbf{1}_J - h_m(x_m; \lambda_m)) \odot \varepsilon, \quad (5)$$

where $h_m(x_m; \lambda_m) \in \mathbb{R}^J$ denotes a modality-specific mapping function parameterized by λ_m, responsible for determining the degree to which each dimension of $f_m^{\ell_m}(\tilde{E}_{m'})$ is retained in the latent representation. $f_m^{\ell_m}(\tilde{E}_{m'}) \in \mathbb{R}^J$ represents the output of the ℓ_mth layer of a neural network embedding function f_m, now functionally dependent on the attention-augmented embedding $\tilde{E}_{m'}$ rather than the original input x_m alone. $\sigma_m^2 \in \mathbb{R}_{>0}$ is a hyperparameter controlling the variance of the noise introduced into the latent space. $\mathbf{1}_J \in \mathbb{R}^J$ is a vector of ones, facilitating the modulation of noise across the dimensions of Z_m. $\varepsilon \sim \mathcal{N}(0, I_J)$ is a noise vector drawn from a standard multivariate normal distribution. $\odot$ denotes the Hadamard product, enabling the selective blending of signal and noise within the latent representation. (5) implies that for elements where $[h_m(x_m;\lambda_m)]_i$ is close to 1, the corresponding dimensions of Z_m are predominantly influenced by the contextual embeddings $\tilde{E}_{m'}$, reflecting direct modulation by cross-attention. Conversely, dimensions where $[h_m(x_m;\lambda_m)]_i$ is near 0 become predominantly noise, indicating suppression of those features in the latent representation. This approach allows to dynamically adjust the flow of information through the latent space, to balance between preserving relevant cross-modal interactions and minimizing irrelevant information.

3.2 Formulating the Variational Objective

The variational objective for modality m can be formulated as

$$\mathcal{F}_m(\theta_m,\Theta_A) = \gamma_m \mathcal{F}_m^{\text{fit}}(\theta_m,\Theta_A) - \beta_m \mathcal{F}_m^{\text{compress}}(\theta_m), \tag{6}$$

where $\mathcal{F}_m^{\text{fit}}(\theta_m,\Theta_A)$ represents the cross-attention enhanced fitting term, designed to maximize the mutual information between the latent representation Z_m and the embedding $\tilde{E}_{m'}$. The term $\mathcal{F}_m^{\text{compress}}(\theta_m)$ is designed to quantify the amount of compression of the latent representation Z_m with respect to the input X_m. This term essentially measures how much information about X_m is being retained in Z_m, with the goal of minimizing unnecessary information to ensure that Z_m captures only the most relevant features, and is expressed as

$$\mathcal{F}_m^{\text{compress}}(\theta_m) = \mathbb{E}_{p_{X_m}}\left[\mathbb{D}_{KL}\left(p_{Z_m|X_m}(\cdot \mid X_m;\theta_m)\,||\,q_{Z_m}(\cdot)\right)\right]. \tag{7}$$

Here, $p_{Z_m|X_m}(\cdot \mid X_m;\theta_m)$ is the conditional distribution of the latent representation Z_m given X_m, parameterized by θ_m, $q_{Z_m}(\cdot)$ is a predefined prior distribution over Z_m, typically chosen to be a simple distribution such as the standard Gaussian $N(0,I)$ to encourage the model to learn efficient and generalizable representations. The compression term thereby acts as a regularizer that mitigates overfitting and promotes the learning of generalizable features. The fitting term, which includes cross-attention, aims at maximizing the alignment between Z_m and $\tilde{E}_{m'}$, and can be defined using a variational approximation for mutual information as

$$\mathcal{F}_m^{\text{fit}}(\theta_m,\Theta_A) = \int p(x_m)\,p(\tilde{e}_{m'} \mid x_m)\,p(z_m \mid x_m;\theta_m)\log q(\tilde{e}_{m'} \mid z_m;\theta_m,\Theta_A)\,dx_m\,d\tilde{e}_{m'}\,dz_m\,. \tag{8}$$

The term $q(\tilde{e}_{m'} \mid z_m;\theta_m,\Theta_A)$ is the variational approximation that models the conditional distribution of $\tilde{E}_{m'}$ given Z_m. The balance between fitting and compression is maintained through the hyperparameter β_m, allowing for the adjustment of their relative importance in the overall objective $\mathcal{F}_m$. Given (8), to define a tractable variational optimization objective, we can propose the following expression for the empirical distribution from which to sample X_m and $\tilde{E}_{m'}$, that is

$$\hat{p}(x_m,\tilde{e}_{m'}) = \frac{1}{N}\sum_{n=1}^{N}\left[\delta_0\left(x_m - x_m^{(n)}\right) \times \delta_0\left(\tilde{e}_{m'} - \tilde{f}_{m'}\left(x_{m'}^{(n)};\Theta_A\right)\right)\right], \tag{9}$$

where $\delta_0(\cdot)$ denotes the Dirac delta function, ensuring that $\hat{p}(x_m,\tilde{e}_{m'})$ assigns probability mass only to the observed pairs of inputs and their corresponding attention-augmented embeddings. The function $\tilde{f}_{m'}(x_{m'}^{(n)};\Theta_A)$ represents the process of obtaining the attention-augmented embedding $\tilde{E}_{m'}$ for input $x_{m'}^{(n)}$ under the VLPM, parameterized by Θ_A. The empirical distribution $\hat{p}(x_m,\tilde{e}_{m'})$ incorporates the impact of cross-attention on the embeddings and allows for the tractable estimation of the variational optimization objective $\hat{\mathcal{F}}_m^{\text{emp}}(\theta_m,\Theta_A)$ by sampling from the observed data and its cross-attention augmented embeddings. We express the empirical estimation of the variational objective as

$$\hat{\mathcal{F}}_m^{\text{emp}}(\theta_m,\Theta_A) = \frac{1}{N}\sum_{n=1}^{N}\Big[\gamma_m \int p(z_m \mid x_m^{(n)};\theta_m)\log q(\tilde{e}_{m'} \mid z_m;\theta_m,\Theta_A)\,dz_m \\ -\beta_m \mathbb{D}_{KL}\left(p_{Z_m|X_m}(\cdot \mid x_m^{(n)};\theta_m)\,||\,q_{Z_m}(\cdot)\right)\Big]. \tag{10}$$

Final Variational Optimization Objective We posit that the function $h_m(x_m^{(n)};\lambda_m)$ defines a unique set of parameters $\lambda_m^{(n)}$ for each input $x_m^{(n)}$. This is critical for accurately capturing the unique characteristics of each instance's contribution to the mutual information between modalities. Within this framework, g_m denotes the transformation enacted by the layers subsequent to the information bottleneck in a VLPM for modality m. This transformation includes normalization steps to ensure that the final embeddings for each modality are normalized across their respective dimensions. Such normalization is necessary as it aligns the embeddings. With the normalized embeddings $g_m(z_m)$ and $\tilde{E}_{m'}$, the logarithm of the Gaussian probability density function $q(\tilde{E}_{m'}|g_m(z_m))$ becomes directly proportional to the cosine similarity between $\tilde{E}_{m'}$ and $g_m(z_m)$. This relationship forms the backbone of the final variational optimization objective as

$$\hat{\mathcal{F}}_m^{\text{emp}}(\theta_m,\Theta_A) = \frac{1}{N}\sum_{n=1}^{N}\Big[\gamma_m \int p(z_m \mid x_m^{(n)};\theta_m)\,S_{\text{cosine}}\left(\tilde{e}_{m'},g_m(z_m;\Theta_A)\right)dz_m \\ -\beta_m \mathbb{D}_{KL}\left(p_{Z_m|X_m}(\cdot \mid x_m^{(n)};\theta_m)\,||\,q_{Z_m}(\cdot)\right)\Big], \tag{11}$$

where $S_{\text{cosine}}(\cdot,\cdot)$ is the cosine similarity function. The optimization involves adjusting both θ_m and Θ_A to maximize this empirical objective, using gradient estimation techniques such as Monte Carlo estimation and the reparameterization trick for stochastic gradients. The objective is to enhance the mutual information between the latent representations and the cross-attention modulated embeddings across modalities, fostering the learning of representations that are both informative and reflective of the intrinsic correlations between the modalities.

4 Experiments

We assess our attribution method using CLIP [14] on the Conceptual Captions dataset [18], that encompasses multiple images and captions from the web, and on the Local Alignment Chest X-ray dataset MS-CXR [2], that includes chest X-rays and texts describing radiological findings and that complements the MIMIC-CXR dataset [9] by improving the quality of captions and bounding boxes. In our experiments, we employ a pretrained CLIP model with a ViT-B/32 [5] image encoder and a 12-layer self-attention transformer as text encoder. For Conceptual Captions, we use the pretrained weights of openai/clip-vit-base-patch32. For the MS-CXR dataset, we use a CLIP variant fine-tuned on radiology datasets, CXR-RePaiR [6]. Based on [23], our method integrates an information bottleneck at specific

Table 1: Quantitative Results. Means and standard errors computed over ten random seeds.

	Metric	GradCAM [16]	Saliency [19]	K-SHAP [11]	RISE [13]	Chefer et al. [4]	M2IB [23]	Ours
MSCXR Image	% Conf. Drop (↓)	2.70 ± 0.03	0.76 ± 0.01	2.30 ± 0.04	3.85 ± 0.03	1.82 ± 0.02	0.52 ± 0.01	**0.46 ± 0.01**
	% Conf. Incr. (↑)	12.91 ± 0.47	35.68 ± 0.45	10.45 ± 0.69	7.42 ± 0.45	21.81 ± 0.47	46.56 ± 0.71	**48.93 ± 0.73**
	% ROAR+ (↑)	3.61 ± 0.82	25.97 ± 1.37	12.92 ± 1.03	17.11 ± 0.77	24.84 ± 1.21	39.47 ± 0.88	**43.25 ± 0.90**
	% Localization (↑)	5.66 ± 0.13	22.01 ± 0.17	7.92 ± 0.14	11.19 ± 0.24	21.98 ± 0.26	23.04 ± 0.15	**25.21 ± 0.16**
MSCXR Text	% Conf. Drop (↓)	2.22 ± 0.04	3.28 ± 0.03	2.35 ± 0.05	**1.13 ± 0.02**	2.88 ± 0.03	2.25 ± 0.04	2.02 ± 0.03
	% Conf. Incr. (↑)	36.92 ± 0.55	19.27 ± 0.55	34.68 ± 0.78	**58.34 ± 0.66**	28.56 ± 0.35	36.09 ± 0.70	39.89 ± 0.72
	% ROAR+ (↑)	11.31 ± 0.63	16.11 ± 0.93	14.53 ± 1.11	12.33 ± 1.54	9.29 ± 0.61	16.72 ± 0.76	**18.51 ± 0.79**
CC Image	% Conf. Drop (↓)	4.81 ± 0.01	1.93 ± 0.01	1.88 ± 0.01	1.09 ± 0.01	1.58 ± 0.01	1.08 ± 0.01	**1.01 ± 0.01**
	% Conf. Incr. (↑)	18.18 ± 0.08	23.47 ± 0.12	25.69 ± 0.29	36.43 ± 0.14	38.13 ± 0.12	42.22 ± 0.19	**46.35 ± 0.20**
	% ROAR+ (↑)	2.35 ± 0.42	7.01 ± 0.90	1.59 ± 0.90	3.22 ± 0.99	7.79 ± 0.56	10.82 ± 0.86	**12.37 ± 0.90**
CC Text	% Conf. Drop (↓)	2.14 ± 0.01	1.73 ± 0.01	1.66 ± 0.01	1.27 ± 0.01	1.03 ± 0.01	1.03 ± 0.01	**0.94 ± 0.01**
	% Conf. Incr. (↑)	30.34 ± 0.19	39.51 ± 0.15	**47.65 ± 0.22**	39.05 ± 0.49	39.17 ± 0.11	39.30 ± 0.20	42.46 ± 0.21
	% ROAR+ (↑)	44.32 ± 0.67	44.83 ± 0.66	48.56 ± 3.65	50.14 ± 1.14	54.68 ± 1.27	61.53 ± 1.13	**64.60 ± 1.19**

layers within both the image and text encoders of CLIP. Training the bottleneck follows the setup of the original IBA [15], that duplicates a sample for 10 times for training stabilization and runs 10 iterations with a learning rate of 1 and the Adam optimizer. See Sec. 6.1 in the Supplementary Material for hyperparameter tuning details. Source code is available at https://github.com/PaulineB201/CAM2IB.

4.1 Qualitative results

We qualitatively compare our method with five widely used attribution methods and the recent M2IB framework, as shown in Fig. 2. Our method is able to capture all relevant objects appearing in both modalities, surpassing M2IB, while other methods tend to focus on one major object.

4.2 Localization Test

We assess the efficiency of CA-M2IB by measuring its precision for zero-shot detection in images. This involves converting the saliency map into a binary format, where areas above a 75% score threshold are marked as 1 and the rest as 0, creating a binary prediction map (M_{pred}). Similarly, a ground-truth binary map (M_{gt}) is created based on bounding box data from MS-CXR [2], marking inside-the-box regions as 1 and outside as 0. Samples with several bounding boxes allow to evaluate the method's capability in multi-occurrence identification. The IoU between M_{pred} and M_{gt} is calculated for images of size $n \times m$ as

$$\text{Localization} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{M_{\text{pred}}^{ij} \wedge M_{\text{gt}}^{ij}}}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{M_{\text{pred}}^{ij} \vee M_{\text{gt}}^{ij}}}, \tag{12}$$

where $\mathbb{1}$ is the indicator function, with overlap the logical AND and union the logical OR operations. CA-M2IB achieved an average IoU of 25.21%, surpassing all compared baseline models. Although the IoU score is modest, two main reasons can justify this: (i) M2IB produces segmentation maps rather than bounding boxes, potentially leading to undervalued scores when assessed via bounding boxes, and (ii) the tested VLPM model, not optimized for detection, might capture only a broad link between X-rays and medical captions, affecting precision.

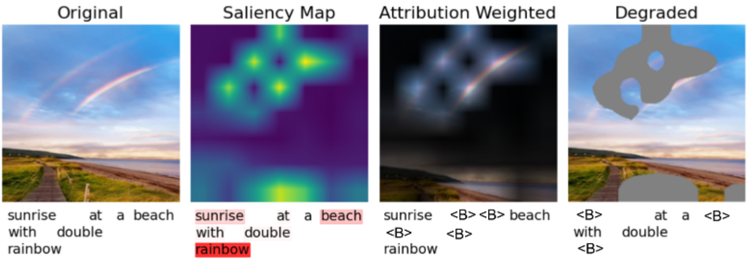
Original
Saliency Map
Attribution Weighted
Degraded
sunrise at a beach with double rainbow
sunrise at a beach with double rainbow
sunrise <B> <B> beach <B> <B> rainbow
<B> at a <B> with double <B>

Figure 3: Visualisation of degradation for the tests performed. In the third column, the saliency maps are combined with the original images through a component-wise multiplication, effectively highlighting regions of interest. Concurrently, text elements scoring below the 50th percentile in terms of attribution are obscured using a placeholder token "<B>". This method underpins both the Increase in Confidence and Drop in Confidence metrics. In contrast, the fourth column visualizations pertain to the augmented training set, ROAR+. Here, regions within images that possess attribution scores exceeding the 75th percentile are averaged to their respective channel means, and text tokens exceeding the 50th percentile attribution threshold are substituted with the placeholder "<B>". Results in Tab. 1 are derived using this blank token "<B>" as a padding element.

4.3 Degradation Tests

To gain a deeper understanding of the effectiveness of CA-M2IB, we employ three metrics for a more nuanced comparison with other leading methods. These metrics are based on the premise that eliminating high attribution features by the model should worsen performance, whereas removing low-scored features might enhance it by reducing noise. We conducted ten tests per metric (five for ROAR+) on 1,000 and 2,000 randomly sampled image-text pairs from MS-CXR and Conceptual Captions, respectively (Tab. 1).

The **Confidence Drop** [3] metric evaluates an attribution method's accuracy by focusing on the importance of features. Lower values indicate better performance. Ideally, removing high-scoring features should not degrade performance significantly. For images, this involves applying the saliency map to the image through point-wise multiplication. For text, only the top 50% of tokens by attribution score are retained [22] (see Fig. 3). The score is

$$\text{Confidence Drop} = \frac{1}{N}\sum_{i=1}^{N} \max(0, o_i - s_i), \tag{13}$$

where o_i is the cosine similarity of features of original images and texts, and s_i is the new cosine similarity when one modality is distilled based on the attribution.[1]

The **Increase in Confidence** [3] metric assesses how well an attribution method identifies and removes non-essential information. By discarding features deemed less important, the model's confidence in its inputs might improve. The higher this metric, the better the attribution method is. It is defined as

$$\text{Confidence Increase} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(o_i < s_i). \tag{14}$$

[1] Drop and Increase in Confidence are implemented in the pytorch-gradcam repository.

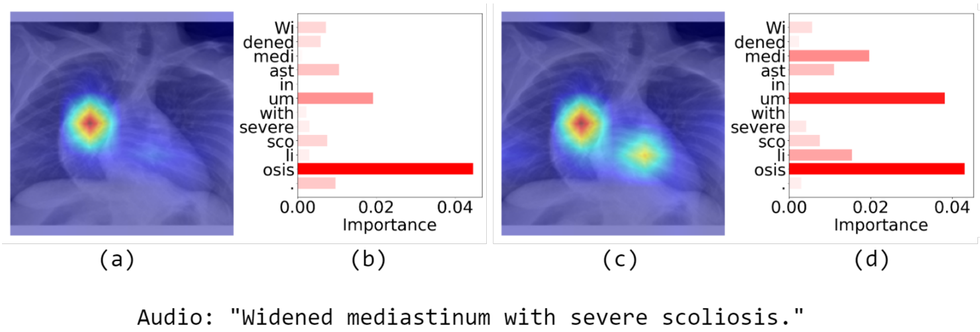

Figure 4: Attribution map on image and text when the other modality is audio with M2IB ((a) and (b)), and CA-M2IB ((c) and (d)).

Remove and Retrain+ (**ROAR+**) [23] involves finetuning the base model using altered images and texts, where key features are substituted with non-informative elements (i.e., channel means of images or padding tokens for texts, see Fig. 3) and tested on a validation set of original inputs. A significant decrease in performance would confirm accuracy of the attribution method as the most useful features to learn for the model are removed. The test dataset was split into 80% for training and 20% for validation, applying the same contrastive loss metric as used for CLIP pretraining. ROAR+ is calculated as $(l_c - l_o)/l_o$, where l_o and l_c are the validation losses of retraining with the original data and degraded data, respectively.

The results are summarized in Tab. 1. CA-M2IB attribution outperforms baseline models in almost all numerical metrics, except for perturbation-based methods, which show improved Increase/Drop in Confidence scores for texts. Perturbation-based methods show better results for short text as they can scan all possible binary masks on the text and then select the best one with the highest confidence score. However, this type of approach is very computationally expensive. We observe that removing tokens or pixels with lower attribution scores with CA-M2IB attribution tends to increase the mutual information with the other modality, while masking by our attribution map tends to decreases the relevance with the other modality. The model also performs worse when retraining on the corrupted data. This strengthens the idea that CA-M2IB in line with M2IB is able to generate useful attribution maps.

5 Discussion

We have shown that integrating cross-attention within the M2IB framework can further refine the attribution process. It allows for a more nuanced, fine-grained attribution of features between modalities, potentially improving the interpretability of how specific elements of one modality (e.g., words in a sentence) are related to elements in another (e.g., regions in an image).

The approach can be adapted for learning representations of various modalities beyond images and text with models that map the features of these modalities into a common embedding space. For instance, ImageBind [7] facilitates the alignment of different modality embeddings, such as audio, depth sensors (3D), thermal (infrared), and inertial measurement units (IMU), with image embeddings through contrastive learning. We utilized ImageBind as an illustrative example of how CA-M2IB can manage the interpretation of audio-image and audio-text learning representations (Fig. 4).

We note that CA-M2IB seems to generally better perform for specialised vocabulary than M2IB (Figs. 2, 4). For common vocabulary, M2IB performs well compared to other attribution methods, and the addition of the cross-attention mechanism may not influence the overall performance (Fig. 2, row 3). Otherwise, we generally observed a slightly increased or decreased improvement with CA-M2IB compared to M2IB for common words, suggesting that M2IB may be a favourable choice in this context. Indeed, in the experiments with the two datasets Conceptual Captions [18] and MS-CXR [2], the most useful layers for model explanation for image and text selected for CA-M2IB were in line with the ones reported for the M2IB method [23] (see Appendix, Tab. 2). In terms of complexity, CA-M2IB increases the number of parameters by 2% compared to M2IB using CLIP. Experiments on additional datasets would be beneficial, yet we note that the Local Alignment Chest X-ray dataset MS-CXR [2], an improved version of the the MIMIC-CXR dataset [9], is the only dataset available with images and corresponding text captions as well as bounding boxes which allows to perform the localization test (Tab. 1). Besides, existing attribution methods focus primarily on unimodal models hence their inclusion in the comparison with the multimodal M2IB and CA-M2IB methods.

Cross-attention across modalities can help reduce or detect bias in vision-language models (VLMs) by enhancing the interpretability and alignment of features between different modalities, such as images and text. The cross-attention mechanism can dynamically focus on relevant features across modalities. This can contribute to identify biased associations by highlighting which parts of the input data (e.g., specific words or image regions) are influencing model predictions. By integrating cross-attention, CA-M2IB can provide a more nuanced understanding of how different modalities contribute to decisions. This can help in identifying and understanding biases that may arise from the interaction between modalities.

Moreover, cross-attention can be used in frameworks designed to debias VLMs. For instance, methods such as DeAR (Debiasing with Additive Residuals) [17] use cross-attention to adjust representations and reduce biases related to specific identity groups. This involves learning residual representations that offset biased original representations. Cross-attention can also be modified in a disentangled manner to address intersectional biases in models, as demonstrated by approaches like MIST [24]. This approach allows for the simultaneous mitigation of multiple biases (e.g., gender, race, age) without affecting related concepts, thereby reducing compounded biases.

Cross-attention furthermore acts as a bridge that adaptively filters and aligns modality-specific embeddings based on their mutual relevance. This ensures that CA-M2IB focuses on the most informative and unbiased features across modalities. The use of cross-attention within an information bottleneck framework can act as a form of regularization, encouraging the model to rely on robust, generalizable features across modalities. This can encourage to mitigate overfitting and improve model generalization for reducing bias.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *Proceedings of International Conference on Learning Representations*, 2016.

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21, 2022.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision*, pages 839–847, 2018.

[4] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

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

[6] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219, 2021.

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[8] Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. Inserting information bottlenecks for attribution in transformers. *arXiv preprint arXiv:2012.13838*, 2020.

[9] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), 2019.

[10] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. *arXiv preprint arXiv:2107.06383*, 2022.

[11] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of Advances in neural information processing systems*, volume 30, 2017.

[12] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

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

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*, pages 8748–8763, 2021.

[15] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of IEEE international conference on computer vision*, pages 618–626, 2017.

[17] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2023.

[18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics.

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of Workshop at International Conference on Learning Representations*, 2014.

[20] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

[21] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[22] 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 IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[23] Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[24] Hidir Yesiltepe, Kiymet Akdemir, and Pinar Yanardag. Mist: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models. *arXiv preprint arXiv:2403.19738*, 2024.