Seeing Through VisualBERT: A Causal Adventure on Memetic Landscapes

Dibyanayan Bandyopadhyay
Indian Institute of Technology Patna
Patna, Bihar, India
`dibyanayan_2321cs14@iitp.ac.in`

Mohammed Hasanuzzaman
EEECS, Queen's University Belfast
Belfast, BT7 1NN, UK
`m.hasanuzzaman@qub.ac.uk`

Asif Ekbal
Indian Institute of Technology Jodhpur
Jodhpur, Rajasthan, India
`asif@iitp.ac.in`

Abstract

Detecting offensive memes is crucial, yet standard deep neural network systems often remain opaque. Various input attribution-based methods attempt to interpret their behavior, but they face challenges with implicitly offensive memes and non-causal attributions. To address these issues, we propose a framework based on a Structural Causal Model (SCM). In this framework, VisualBERT is trained to predict the class of an input meme based on both meme input and causal concepts, allowing for transparent interpretation. Our qualitative evaluation demonstrates the framework's effectiveness in understanding model behavior, particularly in determining whether the model was right due to the right reason, and in identifying reasons behind misclassification. Additionally, quantitative analysis assesses the significance of proposed modelling choices, such as de-confounding, adversarial learning, and dynamic routing, and compares them with input attribution methods. Surprisingly, we find that input attribution methods do not guarantee causality within our framework, raising questions about their reliability in safety-critical applications[1]. The project page is at: `https://newcodevelop.github.io/causality_adventure/`

1 Introduction

Memes have evolved from spreading humor to being used for disseminating offensive content, necessitating the development of neural multimodal systems to detect such content (Kiela et al., 2021). However, these systems often lack transparency, undermining public trust in real-world applications and highlighting the need for interpretability and trustworthiness.

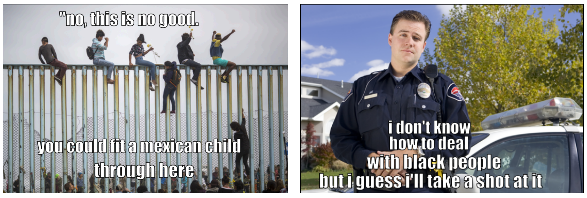

"no, this is no good.
you could fit a mexican child through here
i don't know how to deal with black people but i guess i'll take a shot at it

Figure 1: The Underlying notion of these memes is *racism*, which is implicit. None of the input attribution methods could decipher the notion of *racism* solely through input attribution.

While Large Language Models (LLMs) and Vision Language Models (VLMs) could predict offensive memes and provide self-explanations, these explanations are not always faithful to model behavior (Madsen et al., 2024; Agarwal et al., 2024). Our self-consistency checks (§Appendix B) confirm this issue for offensive meme detection, prompting us to focus on enhancing the reliability of existing multimodal classifiers. Current interpretability techniques, such as input attributions (e.g., Integrated Gradient (Sundararajan et al., 2017)), struggle with implicit content and causality. For implicitly offensive memes, attribution methods fail to capture underlying concepts like 'racism' (refer to Figure 1). Also, they only detect influential features without considering their causal impact (Chattopadhyay et al., 2019).

Causality-based techniques like CausaLM (Feder et al., 2021) and Amnesic Probing (Elazar et al., 2021) offer solutions but face scalability issues and focus on global rather than local explanations like: 'Is adjective important for sentiment analysis' or 'Is part-of-speech information crucial for word prediction?' We

---

[1] This paper contains various racist and offensive memes and keywords which do not reflect authors' beliefs.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10715–10731
November 12-16, 2024 ©2024 Association for Computational Linguistics

address these limitations with a novel causal framework integrating VisualBERT (Li et al., 2019) with causal concepts. Our approach extracts implicit context from the meme as a set of causal concepts and uses dynamic routing and adversarial learning to predict meme offensiveness based on both meme content and contribution from causal concepts.

Quantitative analysis (Section 6) shows that traditional interpretability techniques, which rely on correlation, do not always align with causality, emphasizing that *correlation does not imply causation*. Through ablation studies, we demonstrate that our framework bases its predictions on relevant causal concepts, enhancing trustworthiness. Qualitative analysis (Section 7) indicates whether the model's decisions are justifiable and explains its error cases.

Our proposed framework is novel, model-agnostic, and acts as a proof-of-concept which demonstrates the potential of using causal analysis to elucidate the decision-making process of multimodal classifiers.

2 Related Work

Causal Interpretability: Causal interpretability aims to understand how counterfactuals cause model outputs to change, thus estimating the causal effect of inputs (Feder et al., 2022). A subfield, causal mediation analysis, explores the mechanisms behind these effects (Geiger et al., 2021; Vig et al., 2020; Meng et al., 2023). Generating exact counterfactuals is challenging (Abraham et al., 2022; Calderon et al., 2022), so recent work focuses on approximations (Geiger et al., 2021) or counterfactual representation (Feder et al., 2021; Elazar et al., 2021; Ravfogel et al., 2021). Our current research concentrates on counterfactual representation. Most of the existing works target single modality (e.g. text or vision) (Feder et al., 2021; Goyal et al., 2020) and answer global questions about feature importance (Elazar et al., 2021). We propose a method for answering local questions about specific concepts (e.g., 'Is the meme offensive due to the presence of *Holocaust* as a concept?') while addressing scalability issues of prior methods (Feder et al., 2022). Our framework incorporates concept annotations and integrates with VisualBERT for trustworthy local causal interpretability.

Multimodal Interpretability. Recently, there has been a surge in multimodal models for various tasks (Ding et al., 2021; Du et al., 2022; Li et al., 2023; Liu et al., 2023; Zhu et al., 2023), yet research on generating explanations for their predictions remains limited. Researchers primarily rely on interpretability techniques like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017a) and various input attribution methods (Sundararajan et al., 2017; Lundberg and Lee, 2017b; Shrikumar et al., 2017). However, recently, there has been a shift towards generating natural language explanations, bridging the gap between text-only and multimodal systems. Methods like NLX-GPT (Sammani et al., 2022) and Semantify (Bandyopadhyay et al., 2024) offer solutions but fail to fully capture implicit causal meanings or the causal impact of input features (Chattopadhyay et al., 2019). This gap motivated us to develop a framework that enables causal interpretations of implicit inputs.

3 Causal Process

3.1 Causal Diagram

Our framework is based on a Structural Causal Model (SCM) that integrates both the causal explanation process and multimodal classification objectives, drawing inspiration from Geiger et al. (2021). We assume an exogenous variable E_1 that generates causal concepts c_1, ... to c_n. Another exogenous variable E_2 controls meme text t and image v representation. The collection of the concepts $\{c_i\}_{i=1}^n$ controls the latent representation L. (t, v) *along with* L *controls the intermediate representation* I*, which further controls the output* y *of the model.* Figure 2 represents this in details.

Counterfactual Representation. Inspired by CausaLM (Feder et al., 2021), we want to intervene I such that the generated counterfactual representation I^{CF_i} is insensitive to concept c_i and similar to I for all the other concepts except ${c_i}^2$.

To achieve this, we observe the SCM, depicted in Figure 2, where $I = f(L, (t, v))$ and L represents the latent, a weighted sum of concept representations c_i, given by $L = \sum_{i=1}^n w_i \cdot c_i$. Creating a counterfactual latent (L^{-i}) unaffected by a specific concept c_i is straightforward to achieve by setting $w_i = 0$. This explicit modelling offers simple implementation and ensures I^{CF_i} represents a counterfactual unaffected by c_i, while still influenced by other concepts.

Need for De-confounding. In this formulation

---

[2] Note a slight abuse of notation here, $c_i \in \mathbb{R}^{1\times 768}$ refers to concept representation instead of 'textual' concepts.

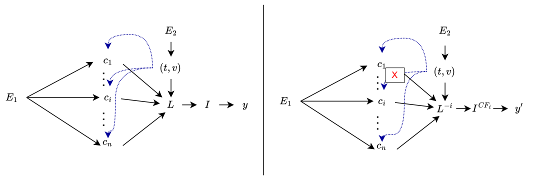


Figure 2: **Left:** The causal process is illustrated by a SCM. **Right:** Causal intervention selectively intervenes on a concept c_i to nullify its effect on the model. This generates the intermediate counterfactual representation I^{CF_i}. To measure the causal effect of concept c_i, we take the Individual treatment effect (ITE) as $|y' - y|$. Dotted blue lines denote that meme content representations (t,v) (along with E_1) generate causal concepts c_i. X demonstrates causal intervention, i.e. breaking the link between (t,v) and c_i, realized by setting $w_i = 0$.

of obtaining the counterfactual representation, a challenge arises because concept c_i might be correlated with other concepts $\{c_j\}$ (e.g. 'holocaust' is correlated to 'nazism'), such that L^{-i} can be retrieved from $\{c_j\}$. We term these influencing concepts like $\{c_j\}$ as 'control concepts' and c_i as the 'treated concept', following the terminology in Feder et al. (2021). To address this challenge, we propose a novel step called de-confounding (refer to Section 4.3). Here, we constrain our framework to prevent the recovery of a treated concept c_i from the counterfactual latent representation L^{-i}, which is essentially a linear combination of potential control concepts. The intuition is that once a concept is removed from the representation containing it, it cannot be recovered.

Causal Effect of Concept c_i. Drawing inspiration from existing literature (Feder et al., 2021), we formulate the causal effect of concept c_i (for a specific input) as:

$$\begin{aligned}\widehat{RITE}_i &=< \phi(I_{DC}^{CF_i}) - \phi(I_{DC}) > \\ &=< \phi(f((t,v), L_{DC}^{-i})) - \phi(f((t,v), L_{DC})) >\end{aligned} \tag{1}$$

Here, $\widehat{RITE}_i$ represents the *"Representation-based Individual Treatment Effect"*, for c_i, with DC as subscript representing the de-confounding objective. The function f, modelled as VisualBERT in this paper, takes the input meme as text and visual representation ((t, v)) and latent L_{DC}, outputting a representation I_{DC}. This intermediate representation is then passed through a classifier ϕ. *Essentially, $\widehat{RITE}_i$ denotes the absolute change in predicted class probability due to the absence of concept c_i and could be used to measure its causal effect on the model.*

3.2 Concept Annotation

Figure 2 illustrates the causal process, which relies on the integration of concepts alongside meme inputs to facilitate model predictions. The selection of these concepts (which form the 'concept set') is pivotal, aiming to i) encapsulate the breadth of themes present within the training dataset while ii) minimizing redundancy.

Scalability. To make the annotation process efficient and scalable, we use the following approach. Starting empty, the concept set expands as new concepts are introduced. For example, if the first meme includes 'terrorism' and 'holocaust', these are added to the set. If the second meme includes 'terrorism' and 'racism', the set becomes 'terrorism', 'holocaust', 'racism'. Therefore, we ***only*** *append new concepts to the set if they are absent, to minimize redundancy and counter overlap between similar concepts.*

Annotation Process. We enlisted three annotators, all postgraduate students aged 25-27 with expertise in multimodal machine learning and prior experience curating datasets published in reputable venues, to annotate the concept set. We ensured ethics and took active steps to ensure their well-being, as detailed in Appendix Section G. Let us denote the three annotated concept sets as $\{c_1\}$, $\{c_2\}$, and $\{c_3\}$. We then calculate concept representation $r_t(c_1^i)$, where c_1^i denotes ith concept from set $\{c_1\}$. Similarly, let us denote meme representation as $r_t(T^j) \odot r_v(V^j)$, where T^j, V^j are the text and image of the jth meme from the test set, and $\odot$ illustrates element-wise multiplication. Here, $r_t(T^j) \in \mathbb{R}^{1\times768}$ and $r_v(V^j) \in \mathbb{R}^{1\times768}$ represent CLIP (Radford et al., 2021) text and vision encodings respectively for the jth meme. We then calculate the **total similarity** of the set $\{c_1\}$ to the memes in the training set as $\sum_{j=0}^{N} \sum_{i=0}^{n} r_t(c_1^i)^T \cdot (r_t(T^j) \odot r_v(V^j))$, where N is the number of memes in the training set and n is the number of concepts in the set $\{c_1\}$. Similarly, this total similarity is calculated for sets $\{c_2\}$, and $\{c_3\}$. We observe that the total similarity of set $\{c_2\}$ is the highest which leads us to choose this as the final set. The concepts in $\{c_2\}$ are shown in Table 1.

Ambiguity Resolution. Annotators were instructed to maintain precision and leverage existing annotated concepts when annotating new memes.

Despite these guidelines, ambiguities often emerged, especially when new meme concepts were semantically similar to already annotated concepts. In these instances, annotators were encouraged to engage in discussions to reach a consensus if they encountered difficulties in identifying or formulating a concept. Additionally, they were advised to utilize shared resources, such as dictionaries, to identify semantically related terms.

4 Methodology

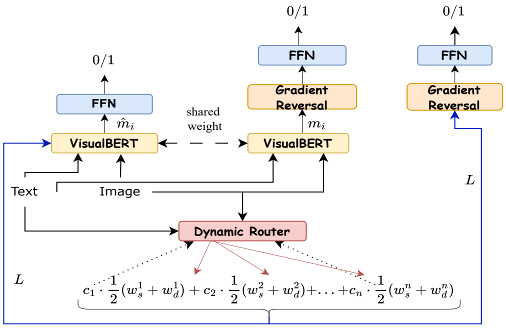


Figure 3: Model architecture comprising of VisualBERT, a dynamic routing layer and a gradient reversal layer. 0/1: non-offensive/offensive.

Our method relies on a three-step process, namely i) Dynamic Routing, ii) Adversarial learning and iii) De-confounding.

Model Inputs. A meme consists of text T and image I. We extract image features using FasterRCNN (Ren et al., 2016), yielding dimensions $v \in \mathbb{R}^{\mathbb{B}\times\mathbb{N}\times 768}$. Text tokens pass through the model embedding layer to generate text features with dimensions $t \in \mathbb{R}^{\mathbb{B}\times\mathbb{M}\times 768}$. Concatenating these text and image features (t, v) results in input dimensions of $\mathbb{R}^{\mathbb{B}\times(\mathbb{M}+\mathbb{N})\times 768}$ for VisualBERT. Additionally, we introduce a latent representation L, with dimensions $\mathbb{R}^{\mathbb{B}\times 1\times 768}$, obtained by weighted summation of concept features c_i. Formally, $L = \sum_{i=0}^{n}(w_d^i + w_s^i)c_i$, where w_d^i is a dynamic weight, $w_d^i = \rho(c_i, T)$, dependent on the ith concept c_i and the meme text T. This functional relationship is learned through dynamic routing. Similarly, w_s^i are static weights, $w_s^i = \tau(c_i)$, dependent only on the concept c_i.

4.1 Dynamic Routing.

Need for Dynamic Routing. The key idea behind dynamic routing is to learn dynamic weights that determine the importance of each concept for a prediction based on both the meme input and the concept itself. These weights are functions of the meme's text, and concept, serving as learnable parameters that control each concept's influence on the prediction. Without dynamic routing, static weights are used, treating all concepts equally, regardless of the specific meme input.

How is it achieved?

To model the interaction between the M text features $\{t_j\}_{j=0}^{M}$ produced by VisualBERT for a given input T and the concept feature c_i, we learn a weight W, which modifies t_j as $t_{ij} = W_{ij} \cdot t_j$. Further, the interaction between c_i and t_j can be modelled by taking a dot product between them. $p_{ij} = t_j^T \cdot c_i$ demonstrates this operation, where p_{ij} is a scalar. To normalize p_{ij} between 0 and 1, we use the softmax function: $b_{ij} = \frac{exp(p_{ij})}{\sum_{k=0}^{n} exp(p_{ik})}$, where n is the number of concepts. To measure the effect of all text inputs on concept i, we calculate the weighted mean: $s_i = \frac{1}{M}\sum_{j=0}^{M} b_{ij} \cdot t_{ij}$, where M is the number of text tokens input to the model. Here, s_i shows the cumulative effect of all the text inputs on concept c_i. To model the interaction between c_i and T, we want the length of s_i to approximate their interaction. Specifically, longer vectors should have a unit length, whereas shorter vectors should have a length close to zero. This is achieved using the following squashing function (Sabour et al., 2017):

$$v_i = \text{squash}(s_i) = \frac{||s_i||^2}{1 + ||s_i||^2} \cdot \frac{s_i}{||s_i||} \qquad (2)$$

The length of v_i acts as the dynamic weight between c_i and T_i, such that $w_d^i = \rho(c_i, T) = ||v_i||$.

4.2 Adversarial Learning

Notation. Let m_i be a $\mathbb{R}^{1\times 768}$ dimensional vector output corresponding to the '[CLS]' token when VisualBERT processes the text T and image I. Similarly, let $\hat{m_i}$ be the output vector when VisualBERT processes the text T, image I, and latent L (by concatenating with image representation). For classifying an input meme, we utilize a feed-forward neural network (FFN) on the representation $\hat{m_i}$.

Need for Adversarial Learning. Typically, input text and images contain sufficient information to classify a meme into offensive or non-offensive classes, rendering the latent representation less effective compared to text and image inputs alone. To enhance the effectiveness of the latent representation to match that of text and image inputs, we

0. holocaust (535)	1. nazism (510)	2. genocide (292)	3. funny (3000)	4. anti-muslim (345)	5. terrorism (276)
6. violence (360)	7. politics (122)	8. racism (405)	9. international-relation (88)	10. adult (172)	11. gore (178)
12. misogynistic (381)	13. immigration (430)	14. extremism (431)	15. immoral (151)	16. white supremacy (205)	17. indecency (831)

Table 1: Annotated concepts and the number of memes that have this concept in brackets. Note that the total number of concepts may exceed the number of memes as a meme can have multiple concepts. Annotators could see the train labels while annotating and were instructed to label non-offensive memes as funny.

employ adversarial learning. The objective is to make both m_i and L invariant to the output class, while their combined representation $\hat{m_i}$ should retain discriminatory information for classifying memes. The aim is for L to be as effective as the combined text and image inputs.

How is it achieved? To achieve this, we utilize a Gradient Reversal Layer (Ganin and Lempitsky, 2015) before passing L and m_i to two separate classifiers for the offensiveness detection task. These classifiers help learn class-invariant L and m_i, while the classifier utilizing $\hat{m_i}$ learns class-dependent representation.

4.3 De-confounding

The premise of de-confounding and its necessity is described in Section 3.1. Before de-confounding, we assume, that there exists some $\overline{W}$, which can project each L^{-i} to c_i. More formally, $\overline{W} \cdot L^{CF} = C$, where $L^{CF} = [-L^{-1}-, -L^{-2}-, ..., -L^{-n}-] \in \mathbb{R}^{768\times n}$, and $C = [-c_1-, -c_2-, ..., -c_n-] \in \mathbb{R}^{768\times n}$, where $L^{-i}, c_i \in \mathbb{R}^{1\times 768}$ and there are n concepts.

For De-confounding, we aim to learn a projection matrix P (a nullspace projection matrix of $\overline{W}$) such that projecting each c_i through it yields counterfactual latent $L^{CF'}$ incapable of reconstructing C. The following theorem illustrates this.

Theorem. If P is a nullspace projection matrix of $\overline{W}$ and $C' = P \cdot C$ *for all* i, then $\overline{W} \cdot L^{CF'} = 0$ *for all* i. For *Proof*, refer to Appendix Section §C. After de-confounding, we denote the latent L with a subscript, i.e., L_{DC}.

5 Experimental setups

The experimental setups and dataset details are elaborated in the Appendix Section F due to space constraints. Here we introduce various metrics and baselines.

5.1 Simulating Model Outcome

Definition. Simulatability, as defined in Hase et al. (2020), refers to how well explanations from model M help an observer (e.g. another simpler model, called a simulator model) predict outputs of M. Intuitively, better simulatability would reflect faithful explanation as there is a pattern between explanation as input and model prediction as output for the simulator to learn. In our proposed framework, explanation would mean causally sorted concepts in descending order. Sorting is necessary to preserve the causal importance order between the concepts.

Notation. Let us assume an ordered set $X^j_{cau} = \{x^j_i\}^n_{i=1}$ contains n concepts sorted by their causal attribution score (measured by $\widehat{RITE}$ score) in a decreasing manner for the jth meme in the dataset. Similarly, let us assume an ordered set X^j_{attr} refers to n concepts sorted by their input attribution score by an input attribution method like Integrated Gradient.

Rank Correlation. A positive correlation between these sets X^j_{cau} and X^j_{attr} indicates that highly attributed concepts also influence model outcomes (causality), while a negative correlation suggests the opposite. We measure correlation by both Kendal's (KENDALL, 1938) and Spearman's rank correlation (Spearman, 1904) criteria.

Accuracy Metrics. Denoting $\hat{x}^j_i$ as the text representation of x^j_i, the rank-adjusted[3] BoW representation $\hat{X}^j$ (note that we drop cau or $attr$ as subscript to show a generic case) of the set X^j would be $\frac{1}{n}\sum_{i=1}^{n}\gamma^i\hat{x}^j_i$, where $\gamma = 0.9$ is a positive non-zero constant. We train a simulator (a support vector machine-SVM model (Cortes and Vapnik, 1995)) on: i) $\hat{X}^j$, ii) concatenation of $\hat{X}^j$ and $\hat{m}^j_i$, denoted by $[\hat{X}^j; \hat{m}^j_i]$, and iii) $\hat{m}^j_i$, to predict the original model prediction $\hat{y}^j$. The jth superscript reflects the jth meme. Intuitively, the SVM simulates the original model based on the provided information (either one of cases (i), (ii), or (iii)). The performance of the simulator can be seen in Table 2. *F1 w/ exp* denotes the simulator performance in case (i) when only the ranked concepts X_{cau} or X_{attr} were used as simulator input. Similarly, *F1 w/ inp* denotes case (iii), where the multimodal represen-

---

[3] rank adjustment is necessary for maintaining the order.

tation ($\hat{m}_i^j$) is used for model input. Lastly, *F1 w/ both* denotes case (ii).

Comprehensiveness and Sufficiency. To measure the impact of set X_{cau} or X_{attr} on simulator performance, we employ two metrics: i) Comprehensiveness and ii) Sufficiency (DeYoung et al., 2020). Comprehensiveness quantifies the reduction in simulator model confidence when $\hat{X}^j$ replaces $[\hat{X}^j; \hat{m}_i^j]$ as simulator input. Denoting the simulator by S, comprehensiveness is $S([\hat{X}^j; \hat{m}_i^j])_k - S(\hat{m}_i^j)_k$ for predicted class k. A higher comprehensiveness score indicates more importance of attribution set X^j for the simulator. Sufficiency is defined as $S([\hat{X}^j; \hat{m}_i^j])_k - S(\hat{X}^j)_k$. It requires a higher average comprehensiveness and lower average sufficiency score for X^j to be considered more simulatable.

5.2 Quantifying Trustworthiness

To measure the trustworthiness of a model, understanding whether its predictions originate from relevant concepts within the input is crucial, akin to assessing if the model is 'right for the right reasons.' To ascertain this, we annotate offensive memes from the test set with relevant concepts from a predefined set of 18 concepts. Subsequently, we employ averaged Precision@5 (P@5), Recall@5 (R@5), and Mean Average Precision (MAP@5) to assess the relevance of the top five concepts (w.r.t. the annotation) from both the X_{cau} and the attribution set X_{attr} obtained through various input attribution methods. Technical specifics are detailed in the Appendix Section E. Better scores in all these metrics reflect that the top attributed concepts align with human judgement, thus essentially making the model more trustworthy in return.

5.3 Baselines

We employ several standard input attribution methods to calculate attribution scores, dividing them into two groups based on their underlying mechanisms:

Path-integral based Gradient Attribution:

- Integrated Gradients (IG) (Sundararajan et al., 2017): IG attributes the importance of features by integrating gradients along the path from a baseline input to the actual input. It ensures that attribution is distributed across all input features in a manner that satisfies the completeness property (§Appendix A).
- DeepLIFT (Shrikumar et al., 2019): This method compares the activation of each neuron to a reference, and assigns importance to the neuron based on the comparison score.
- DeepLIFTSHAP (Lundberg and Lee, 2017a): A variant of DeepLIFT that aligns with SHAP values, combining both methods to compute attribution based on a cooperative game-theory approach.
- GradientSHAP (Ancona et al., 2018): GradientSHAP samples a point between an input-baseline pair and computes the mean gradients with respect to an output class across all such pairs.

Simple Gradient Attribution:

- Saliency (Simonyan et al., 2014): It Identifies key input features by computing output gradients, highlighting features most influential to the model's prediction.
- Input × Gradient (Shrikumar et al., 2017): This method computes the element-wise product of the input and its gradient to measure each feature's contribution to the prediction.

We use the Captum library (`https://captum.ai/`) to calculate attributions for each method and apply these attributions to three matrices: i) the concept matrix ($C \in \mathbb{R}^{18\times768}$), ii) the textual embedding ($t \in \mathbb{R}^{\mathbb{B}\times\mathbb{M}\times768}$), and iii) the image embedding ($v \in \mathbb{R}^{\mathbb{B}\times\mathbb{N}\times768}$).

Comparison with Our Methodology. Traditional input attribution methods outlined above rank features (or concepts) by assigning scores based on their importance to the model's output. In contrast, our approach adopts a causal perspective, using the $\widehat{RITE}$ score to measure how much each concept causally influences the prediction.

While our causal framework differs from standard attribution techniques, we explore an intriguing link between the completeness property of path-integral methods and causality, which we discuss in more detail in Appendix A.

6 Quantitative Findings

1. Classifier performance. Model performance on the test set is 70.36% as measured by the F1 score. This reflects the VisualBERT in our framework works well and even exceeds some of the benchmark models evaluated as a part of the FB Hateful Meme Dataset (Kiela et al., 2021).

Methods	Causal Rank Correlation		Explainability of Simulator		Performance of Simulator		
	Kendall's Tau	*Spearman's rho*	*Comprehensiveness*	*Sufficiency*	*F1 w/ both*	*F1 w/ inp*	*F1 w/ exp*
Int. Grad.	0.017	0.025	0.030	0.089	0.66	0.62	0.53
Saliency	**0.859**	**0.916**	0.005	0.116	0.63	0.62	0.42
DeepLift	-0.001	-0.002	0.029	0.021	0.67	0.62	0.65
DeepLiftSHAP	-0.003	-0.005	**0.035**	**0.017**	0.67	0.62	**0.68**
GradientSHAP	0.010	0.014	0.008	0.117	0.63	0.62	0.45
Input x Grad	-0.001	-0.002	0.034	0.029	**0.68**	0.62	0.67
Causal	1	1	0.002	0.112	0.63	0.62	0.46

Table 2: Table shows an assessment of input attribution-based methods in two directions: i) Their rank correlation with causality, ii) Explanation capability of the extracted keywords as measured by a simulator. Note that here we do not use de-confounding for input-attribution methods as we want to measure the *overall effect* of these concepts on the model outcome (§list A)

Input Attr/Causal	Full			w/o dyn. routing			w/o adversarial			w/o deconfounding		
	R@5	*P@5*	*MAP@5*	*R@5*	*P@5*	*MAP@5*	*R@5*	*P@5*	*MAP@5*	*R@5*	*P@5*	*MAP@5*
Int. Grad.	0.29	0.20	0.20	0.26	0.17	0.19	0.28	0.19	0.17	0.27	0.18	0.17
Saliency	0.27	0.19	0.18	0.19	0.12	0.16	0.29	0.19	0.16	0.32	0.21	0.25
DeepLIFT	0.30	0.21	0.21	0.27	0.18	0.17	0.26	0.18	0.16	0.26	0.18	0.18
DeepLIFTSHAP	0.31	0.21	0.22	0.31	0.22	0.22	0.30	0.20	0.20	0.29	0.19	0.20
GradSHAP	0.30	0.20	0.20	0.30	0.21	0.21	0.29	0.20	0.20	0.29	0.20	0.20
Input x Grad	0.30	0.21	0.21	0.26	0.17	0.17	0.27	0.19	0.17	0.26	0.18	0.17
Causal	**0.26**	**0.18**	**0.17**	0.21	0.14	0.16	0.20	0.14	0.13	0.30	0.20	0.19

Table 3: Red border denotes scores obtained by *Causal* framework, i.e. sorting keywords with their $\widehat{RITE}$ score. The scores from alternative input-attribution methods are also displayed.

2. Simple attribution methods align better with Causality. Table 2 presents Kendall's tau and Spearman's rho (averaged across all memes in the test set) for correlation comparison. Notably, simpler methods like Saliency show a stronger correlation with the causal set than complex methods like DeepLIFT, suggesting their potential to capture causal relationships. This suggests that simpler methods may offer a clearer and more direct understanding of causality within the model's decision-making process.

3. Complex attribution methods are more simulatable. Consider the *F1 w/ exp* scores for DeepLIFT, DeepLIFTSHAP, and Input x Grad as input attribution methods (§Table 2). The attribution set X_{attr} obtained by these methods achieves $\sim 66\%$ F1 score of the simulator, indicating a high correlation between X_{attr} and the original model prediction $\hat{y}$ (termed as highly simulatable). Also both comprehensiveness and sufficiency scores are higher for these models.

4. Correlation does not imply causation. Note the negative correlation of X_{attr} with the X_{cau} set for complex attribution methods (§Table 2), highlighting the fact that *Correlation (simulation) does not always imply Causation*. Additionally, observe the lower *F1 w/ exp* score obtained by the Saliency attribution set, despite its higher rank correlation to the *Causal* set. This suggests that keywords causally related to model outcome may not always be easily simulatable. This outcome is in line with the study of Bastings et al. (2022), where authors showed that simple attribution methods are more faithful compared to complex methods. *Note that although the findings match, their approach is non-causal and does not overlap with ours at all.*

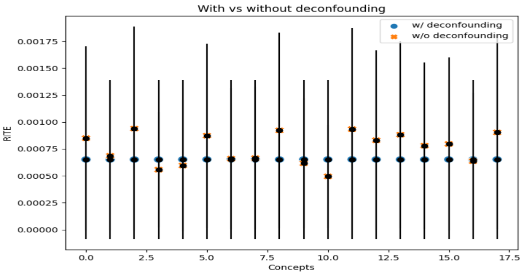


Figure 4: Comparison of mean $\widehat{RITE}$ score between w/ de-confounding and w/o de-confounding strategies

5. Dynamic routing is paramount. Within our framework (§Table 3), without any ablation of modelling components (shown by Full), the X_{cau} set achieves 0.26, 0.18, and 0.17 scores for R@5, P@5, and MAP@5, respectively. Disabling dynamic routing (w/o dyn. routing) results in static

Meme Idx	Top-5 Causal	Actual	Predicted	Possible Reason
32	‘anti muslim’, ‘terrorism’, ‘genocide’, ‘violence’, ‘adult’	Non Offensive	Offensive	Dataset bias
93	‘violence’ ‘racism’, ‘genocide’, ‘anti muslim’, ‘holocaust’	Non Offensive	Offensive	Dataset bias
92	‘violence’, ‘genocide’, ‘terrorism’, ‘extremism’, ‘nazism’	Non Offensive	Offensive	Inadequate context
65	‘indecency’, ‘holocaust’, ‘funny’, ‘immigration’, ‘racism’	Offensive	Non Offensive	Background knowledge
73	‘holocaust’, ‘violence’, ‘gore’, ‘nazism’, ‘funny’	Offensive	Non Offensive	Modality conflict
76	‘misogynistic‘, ‘gore’, ‘holocaust’, ‘extremism’, ‘nazism’	Offensive	Non Offensive	Bias and modality conflict

Table 4: Meme index with corresponding causal keywords and their possible reason behind error cases.

weighting of concepts irrespective of meme input, leading to reduced scores for the X_{cau} set across all metrics compared to the Full framework, indicating diminished trustworthiness.

6. De-confounding is necessary for establishing correlation between causal concepts and model output. Without de-confounding, input attribution-based methods like DeepLIFTSHAP and GradientSHAP do not perform as well as they do with other configurations (e.g. w/o dyn. routing). Refer to Table 3. This suggests that without de-confounding, the model struggles to find a proper correlation between causal concepts and model output. Empirically, in Figure 4, we illustrate the mean $\widehat{RITE}$ scores of the concepts with and without de-confounding. Without de-confounding, due to widely varied mean $\widehat{RITE}$ scores for several concepts, certain concepts dominate the top positions in the causal set, regardless of the input, indicating potential bias in the model's causal attribution and making it difficult for input attribution methods to establish a correlation between model input and output, resulting in lower scores across metrics.

7. Necessity of de-confounding for causality. The observation that the ‘causal’ set performs better without de-confounding may raise questions about its necessity. However, as illustrated in Section 3, de-confounding remains a principally valid choice. Without de-confounding, as discussed in the previous paragraph, certain concepts often dominate the top positions in the causal set, regardless of the input. Conversely, with de-confounding, all concepts have similar $\widehat{RITE}$ values and their standard deviation, ensuring equal representation across test set examples (§Figure 4). Therefore, despite achieving lower scores, the framework with de-confounding is more trustworthy due to its consistent performance, lack of bias towards certain concepts in the causal attribution set, and equal representation of concepts across inputs.

7 Analysing the model through Causal Lens

Is the model always right due to the right reason? No, it is not always the case. We gain insight into this by examining the model causally. In Figure 6, we present two examples of memes from the test set along with the model's predictions, both correctly classified as offensive.

In the first example, the meme's offensiveness is linked to the Holocaust and antisemitic ideas, accurately reflected in the Top-5 causal concepts. Notably, concepts like ‘Violence’, ‘Holocaust’, and ‘Nazism’ from the gold standard set directly relate to the meme's context. With two out of the Top-5 causal keywords aligning with the meme's context, we conclude that the model's prediction was correct for the right reason.

In the second example, despite the meme being antisemitic, the identified causal concepts are irrelevant to antisemitism, although highly offensive. However, certain keywords (such as ‘Holocaust’ and ‘Genocide’) identified by GradientSHAP are attributed. Three of these (‘Racism’, ‘Holocaust’, and ‘Genocide’) directly relate to the meme's context. Initially, it may seem the model was right for the right reason. However, the absence of overlap between the set of causal concepts and the gold standard concept set indicates that the model classified the meme correctly but with erroneous causal attribution. This discrepancy suggests that although accurate input attributions exist, the model may base its decision on different causal concepts. Such analyses can help a user to trust the model predictions.

8 Error Analysis

In this section, we analyze the model's performance using causal concepts. For example, in the first row of Table 4 (meme index 32), the model identifies ‘anti-muslim’ among the top 5 causal keywords, even though they are irrelevant. The presence of ‘Mohammed’ alone leads the model to classify the

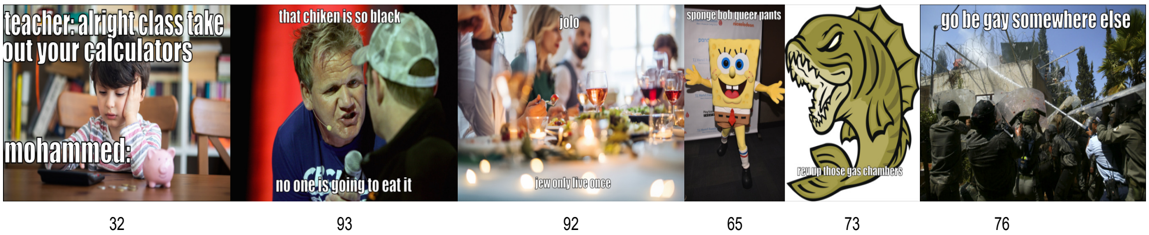


Figure 5: Memes of Table 4.

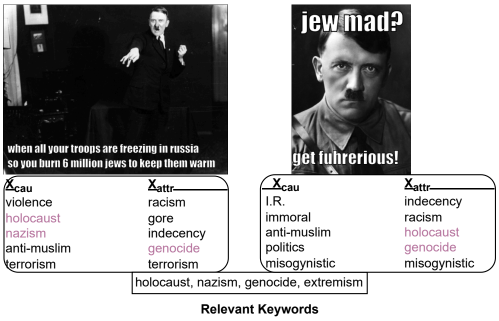


Figure 6: Evaluating trustworthiness of the model

meme as offensive, indicating a dataset-wide bias towards words like 'Mohammed'. Similarly for the second meme, the mere presence of the concept of 'blackness' was sufficient for classifying the meme as offensive despite having nothing related to racism in it. This also shows dataset bias where lots of racists ($\sim 47.1\%$) and religious memes ($\sim 39.3\%$) are present.

Similarly, in the third meme, the word 'jew' prompts the model to associate the meme with concepts like 'violence', and 'genocide' possibly due to insufficient visual context and association of the word jew in antisemitic offensive memes.[4]

In the fourth meme, although offensive, the model finds it humorous due to a lack of background knowledge, especially regarding the wordplay on SpongeBob, and the smiling SpongeBob face may further contribute to misclassification.

The fifth meme's misclassification can be attributed to a humorous background image unrelated to the meme text, creating a *modality conflict*. Note that the fourth and fifth memes which are classified as non-offensive have the 'funny' keyword as a part of the Top 5 causal keywords.

The sixth meme shows a sexually offensive remark but due to conflicting visual modality (showing violence), this meme got misclassified as non offensive. The keywords generated can also be attributed to dataset bias as shown in the first two memes.

By examining causal keywords alongside model inputs, this type of error analysis offers insights into why the model made mistakes.

9 Conclusion

In this paper, we introduce a multimodal causal framework aimed at transparently analyzing VisualBERT predictions. Guided by an SCM, the framework compels VisualBERT to base its decisions on interpretable human-understandable concepts. Evaluation on the Facebook hateful meme dataset reveals key insights: i) Input attribution methods may lack causal underpinning, ii) Modelling choices significantly influence relevant causal attributions, enhancing model trustworthiness. The qualitative analysis delves into whether the model is 'right for the right' reasons and uncovers causal factors behind misclassifications.

The simplicity and versatility of our framework (i.e. the underlying Structural Causal Model and its translation to modelling choices) allow its application across various tasks and multimodal models. Although we show the importance of our architecture on meme offensive detection tasks as a testbed, its application may be important in medicine where the need for trustworthy systems is paramount.

Limitations

While our approach demonstrates promising results, there are some limitations to consider. Firstly, the reliance on a specific dataset, such as the Facebook Hateful Meme dataset, and a specific model, like VisualBERT, may limit the generalizability of our findings to other datasets and models.

Secondly, the concept annotation process introduces challenges as it relies on human annotators

---

[4] The biases identified in the model's behavior do not reflect the views or biases of the authors. The analysis aims to highlight the model's limitations and the importance of addressing bias in machine learning systems.

to define and refine the concept set. This process may introduce subjectivity and biases. To address this challenge, employing more robust annotation guidelines, inter-annotator agreement assessments, and sensitivity analyses can enhance the reliability of the concept annotation process.

At the outset, the concept of the paper acts as a seed or proof of concept, further generalizability of which is to be explored through a chain of related future studies. Specifically, exploring potential applications of the framework in other domains beyond meme classification would be valuable. The framework could be applied in areas such as content moderation, sentiment analysis, and trend analysis in social media, news media, marketing, and public opinion research and medicine.

Ethical Declaration

We acknowledge the potential for misuse of our annotated concepts, which could be employed to filter memes based on racial prejudices. To mitigate this risk, human moderation and intervention are crucial. The purpose of annotating concepts is to facilitate research into the analysis and undertstanding of offensive memes on the internet. When used appropriately, we believe it serves as a valuable resource.

Further the involvement of annotators to annotate potentially triggering meme may seem problematic. On the other hand, we completely ensure annotators' well-being by making voluntary free session with institutional counsellors available at any time. Also, participation in this process was purely at their own wish and they have been warned on exposing themselves to various offensive and trigerring contents which were marked as a disclaimer. We followed four broad ethical principles during the annotation process: i) Annotators were fully briefed on the nature of the task and provided informed consent to participate. ii) Annotators had access to psychological support via our institutional counseling system. iii) Annotators were compensated fairly in line with institute regulations. iv) The privacy and confidentiality of student participants were strictly protected throughout the study.

By adhering to these protocols, we ensured that the ethical concerns associated with using students to label offensive memes were adequately addressed, prioritizing their well-being and ethical treatment. Our study underwent evaluation and approval by our Institutional Review Board (IRB) before proceeding for either annotating offensive memes in the first place or using students to annotate these memes.

Acknowledgements

Dibyanayan Bandyopadhyay acknowledges Prime Minister's Research Fellowship (PMRF). The authors further acknowledge anonymous reviewers for their useful feedback.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Preprint*, arXiv:2205.14140.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. *Preprint*, arXiv:1711.06104.

Dibyanayan Bandyopadhyay, Asmit Ganguly, Baban Gain, and Asif Ekbal. 2024. Semantify: Unveiling memes with robust interpretability beyond input attribution. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6189–6197. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland. Association for Computational Linguistics.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. 2019. Neural network attributions: A causal perspective. *Preprint*, arXiv:1902.02302.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Preprint*, arXiv:2006.00995.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. *Preprint*, arXiv:1409.7495.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Preprint*, arXiv:2106.02997.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2020. Explaining classifiers with causal concept effect (cace). *Preprint*, arXiv:1907.07165.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

M. G. KENDALL. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Scott Lundberg and Su-In Lee. 2017a. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Scott M. Lundberg and Su-In Lee. 2017b. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? *Preprint*, arXiv:2401.07927.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *Preprint*, arXiv:1506.01497.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *Preprint*, arXiv:1710.09829.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. *Preprint*, arXiv:2203.05081.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences. *Preprint*, arXiv:1704.02685.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not just a black box: Learning important features through propagating activation differences. *Preprint*, arXiv:1605.01713.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint*, arXiv:1312.6034.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Completeness vs. Causality

Completeness is a property inherent to path-integral-based methods. Consider a concept defined by c_i, where $C = \{c_i\}_{i=1}^{n}$ and $n = 18$ represents the number of concepts. The completeness property asserts that the attribution for c_i, denoted as $Attr_i$, when calculated separately from other concepts, is equal to $F(c_i, (t, v)) - F(c_i', (t, v))$. Here, F is the VisualBERT combined with the classifier, (t, v) represents other unchanged model inputs (meme content), and c_i' denotes the baseline, often chosen as a zero vector.

$$Attr_i = F(c_i, o) - F(c_i', o) \tag{3}$$

In this context, $F(x) = (\phi \circ f)(x)$. Choosing c_i as the zero vector makes the attribution exactly equal to the $\widehat{RITE}_i$ score for concept c_i (see Equation 1).

Using approximations (e.g., Gauss-Legendre approximation) in calculating the path integral results in $Attr_i \approx \widehat{RITE}_i$, implying that $X_{cau} = X_{attr}$. This has an interesting implication:

1. $X_{cau} = X_{attr}$ suggests that path-integral-based attribution methods exhibit causal behavior, which is a highly desirable property.

However, achieving this property requires certain assumptions and modeling choices, outlined below:

1. The background SCM must be chosen as a modeling choice, resulting in the same outcome as using a concept as a zero vector and setting its corresponding static and dynamic weights to zero.

2. We must assume that individual concept attributions are calculated separately, independent of other concepts, with separate zero-baselines for each concept. This is not the usual practice, as attributions are typically calculated with text, vision, and concepts considered as three singular inputs to the model with random baselines. Consequently, all concept embeddings ($C = \{c_i\}_{i=1}^{n}$) are treated as a single input for attribution calculation.

B LLMs/VLMs Self-consistency Checks

Self-consistency checks. Self-consistency checks (Madsen et al., 2024) are methods used to verify the faithfulness of a model's explanation by ensuring that the model's behavior aligns with expected outcomes. In the context of counterfactual explanations, these checks involve modifying the input so that the model predicts the opposite label. By re-evaluating the prediction with this modified input, we can confirm whether the model produces the expected opposite outcome. If it does, the counterfactual explanation is considered faithful. This process is crucial when using an instruction-tuned language model in a conversational setting, where re-evaluation should occur in a new chat session to avoid bias from prior prompts.

LLM self-consistency checks. We define two self-consistency checks for LLMs as illustrated in Figure 8 and Figure 9. In the first session, referred to as Session 1, we query the LLM with the following template: "I have a meme which can be described as caption. The meme text reads: meme text. Is this meme offensive? Only answer Yes/No." The LLM responds with either Yes or No. We then ask the LLM to explain why the meme is offensive or non-offensive. Next, we prompt the LLM to create a counterfactual version of the meme text that would yield the opposite prediction about whether the meme is offensive or non-offensive. This counterfactual text is then embedded into the template and sent back to the LLM as a counterfactual query. The LLM should respond with the opposite label from its initial prediction, as it is a counterfactual according to the model. Figures 8 and 9 show two failure cases in these counterfactual self-consistency checks for LLMs, indicating that LLM explanations are not faithful and should not be trusted to reveal the model's inner workings, despite their plausible appearance.

VLM self-consistency checks. Similar to language models, Visual Language Models (VLMs) can be assessed for self-consistency. Figure 7 presents two instances of self-consistency failures in VLMs using counterfactual explanations. Initially, the VLM is prompted to label a meme as offensive or non-offensive. Subsequently, the same VLM is asked to generate a counterfactual version of the meme text, which is then in-painted back into the meme to create a counterfactual sample. This counterfactual sample is used as a query in the following session. As a counterfactual, the original label should be reversed. However, as demonstrated in Figure 7, this label reversal does not occur in either case, indicating a failure in self-consistency checks for the VLM.

C Proof

The latent representation L' after projecting c_i with P is given by

$$L' = \sum_{i=0}^{n} w_i P \cdot c_i = P \cdot \sum_{i=0}^{n} w_i c_i = P \cdot L \quad (4)$$

From Equation 4, $\overline{W} \cdot L' = \overline{W} \cdot (P \cdot L) = 0$, as $\overline{W} \cdot (P \cdot x) = 0$ for all x.

Furthermore, $w_i \neq 0$, and $w_i \overline{W} \cdot (P \cdot c_i) = 0$. Therefore, $\overline{W} \cdot L^{CF'} = \overline{W} \cdot L' - w_i \overline{W} \cdot (P \cdot c_i) = 0$ for all i.

Following this theorem and its proof, we construct a nullspace projection P of $\overline{W}$. $\overline{W}$ is learned by training a neural network to reconstruct C from L^{CF} for all i. To de-confound the system, we project the concept representations via P, updating C as $C \leftarrow P \cdot C$. This approach is inspired by Iterative Nullspace Projection (INLP)(Ravfogel et al., 2020). After this operation, we denote the latent L with a subscript, i.e., L_{DC}.

D FAQ

1. Question: *What is the need of the framework?*

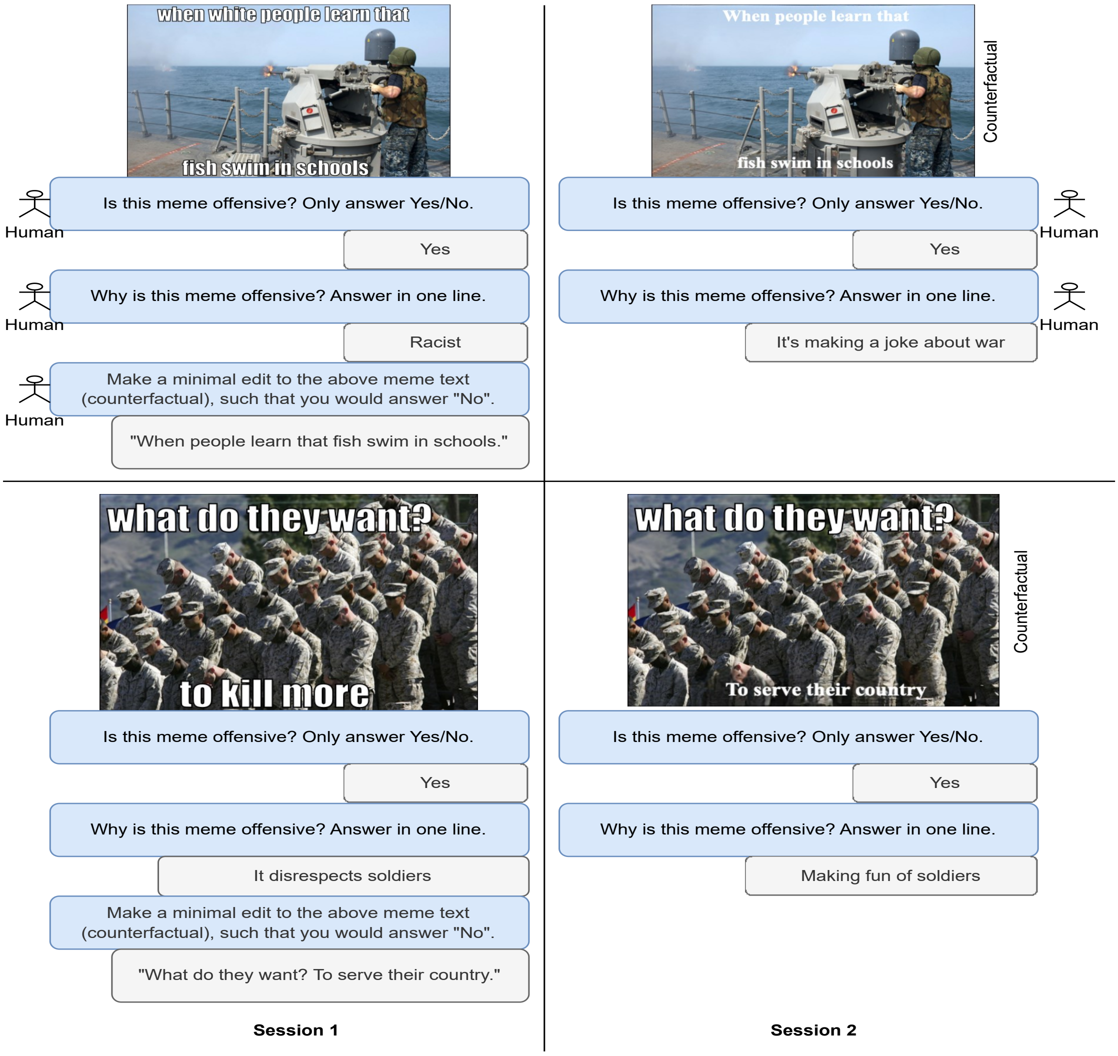
Ans: The proposed framework incorporates a VisualBERT and forces it to base its prediction on causal concepts. This makes the VisualBERT transparent because now we can base its decision on the external causal concepts by looking at their $\widehat{RITE}$ scores. As proposed in the Introduction, input memes contain a lot of implicit contexts, which cannot be attributed by Input attribution methods as they are missing from the meme input itself.

2. Question: *What is the need for Adversarial Learning?*

Adversarial Learning ensures the model utilizes the concept alongside the meme inputs. Normally, the meme inputs contain enough information to classify the input meme and simply adding concept representation on top of it does not make the multimodal information any richer. Adversarial learning ensures we learn a non-trivial interaction between multimodal representation and concept representation such that both play an equal role in the final classification task.

3. Question: *With De-confounding, the possibility of top-5 causal concepts being relevant decreases than without it. Then what is its use?*

De-confounding is principally valid because from the counterfactual representation of concept c_i, we should not be able to recover c_i (i.e. its representation). Otherwise, it is not even a counterfactual. Our main goal is to estimate an unbiased value of $\widehat{RITE}$ score, which is obtained after deconfounding. So, although w/o de-confounding scores are higher, they may not reflect true causality as without de-confounding, true counterfactual representations are not generated. Also, an empirical argument is validated in the Section 6 and Figure 4.



Is this meme offensive? Only answer Yes/No.

Yes

Why is this meme offensive? Answer in one line.

Racist

Make a minimal edit to the above meme text (counterfactual), such that you would answer "No".

"When people learn that fish swim in schools."

Is this meme offensive? Only answer Yes/No.

Yes

Why is this meme offensive? Answer in one line.

It's making a joke about war

Is this meme offensive? Only answer Yes/No.

Yes

Why is this meme offensive? Answer in one line.

It disrespects soldiers

Make a minimal edit to the above meme text (counterfactual), such that you would answer "No".

"What do they want? To serve their country."

Is this meme offensive? Only answer Yes/No.

Yes

Why is this meme offensive? Answer in one line.

Making fun of soldiers

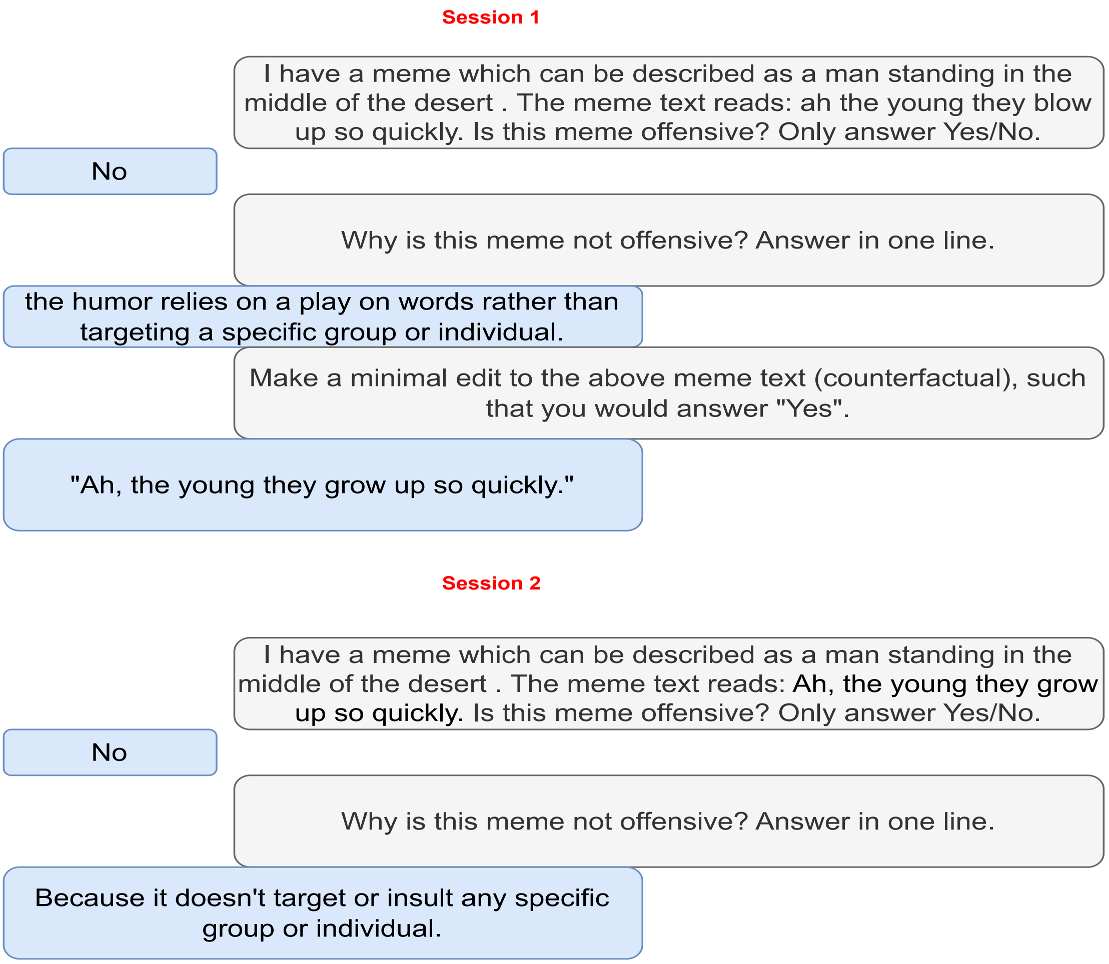
Figure 7: Counterfactual self-consistency checks for VLM, LLaVA-1.6-34B (`https://github.com/haotian-liu/LLaVA`).

4. Question: *Why do you call your framework causal when the only thing 'causal' is how you select the concepts based on their $\widehat{RITE}$ score?*

We assumed that a meme is offensive due to the meme content itself and some contribution from a fixed collection of concepts. That is illustrated as a structural causal model (SCM) in Figure 3.1. Note that every modelling choice is based on the SCM itself. Essentially the framework encompasses an SCM which implements itself as an addition to VisualBERT in the form of i) Adversarial Learning, ii) Dynamic Routing and iii) De-confounding. This forces the SCM to base its decision on meme content as well as on interpretable concepts. Essentially the concepts are called "causal" because they are being modelled as causal in the SCM and the framework implements the SCM on top of VisualBERT.

5. Question: *What is the generality of the proposed framework?*

The framework is overall generic. The framework which implements the SCM is general for every task where implicit domain knowledge is necessary. One such task is meme offensiveness detection. Although the scope of this framework is general, we use it on top of meme offensive detection tasks because of its i) generality of using nontrivial implicit context which is unavailable in input space, and ii) The task has social good applications, and therefore is important to solve. We leave it in the hands of future researchers to use this framework as a proof of concept to describe its



Session 1

I have a meme which can be described as a man standing in the middle of the desert . The meme text reads: ah the young they blow up so quickly. Is this meme offensive? Only answer Yes/No.

No

Why is this meme not offensive? Answer in one line.

the humor relies on a play on words rather than targeting a specific group or individual.

Make a minimal edit to the above meme text (counterfactual), such that you would answer "Yes".

"Ah, the young they grow up so quickly."

Session 2

I have a meme which can be described as a man standing in the middle of the desert . The meme text reads: Ah, the young they grow up so quickly. Is this meme offensive? Only answer Yes/No.

No

Why is this meme not offensive? Answer in one line.

Because it doesn't target or insult any specific group or individual.

Figure 8: An instance where GPT-3.5 offers an counterfactual self-explanation and employs self-consistency verification to assess its faithfulness.

effect in critical domains, such as medicine, where the trustworthiness of machine learning (ML) systems is of utmost importance.

So in summary, the proposed framework is generic whose applicability is shown in one interesting task. The applicability of it to other domains remains as an avenue for further research.

6. Question: *In the proposed SCM (Equation 1), when a concept is intervened or masked, its information in the text and image remains, potentially affecting the reliability of the causal effects. How do you address this concern?*

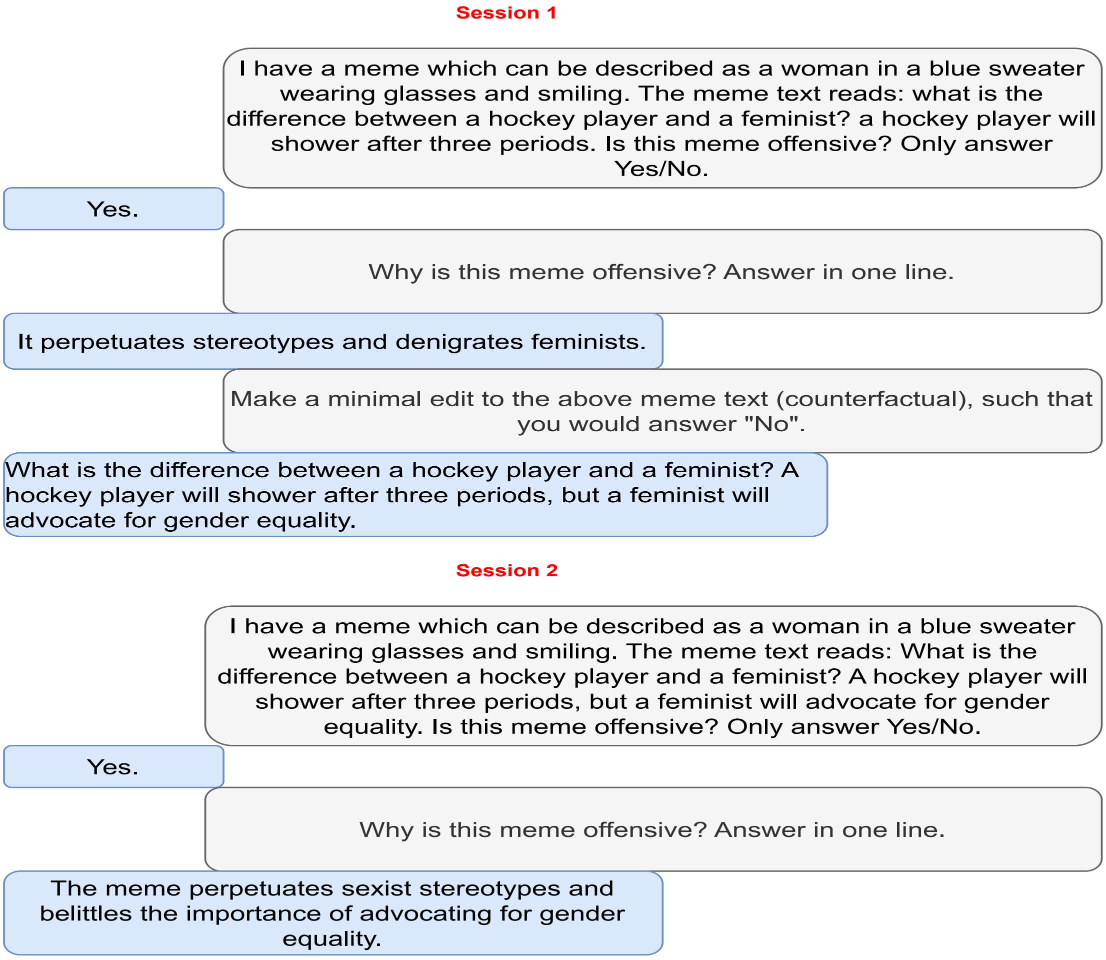
We address this issue in two ways: a) **Gradient Reversal (GR) Layer:** It ensures that the text and image alone, without the explicit presence of the concept, cannot classify the meme as offensive or non-offensive. It forces the model to rely on the explicit presence of the concept for classification.

b) **De-confounding Strategy:** It nullifies the effects of any concept on other interrelated concepts and provides a clearer analysis of each concept's individual impact on predictions.

These mechanisms ensure the reliability of the causal effects derived from equation (1).

7. Question: *The performance of X_{cau} in "explainability of simulator" and "performance of simulator" is not good compared to X_{attr}. Is that problematic?*

It's important to note that evaluating the causal framework based on simulator scores can be misleading. We are using the causal framework to judge how causally important the input attribution based methods are, not the other way around. Better simulator performance of X_{attr} indicates that it has better recognizable patterns, suggesting it can explain the model's decision well. Lower scores



Session 1

I have a meme which can be described as a woman in a blue sweater wearing glasses and smiling. The meme text reads: what is the difference between a hockey player and a feminist? a hockey player will shower after three periods. Is this meme offensive? Only answer Yes/No.

Yes.

Why is this meme offensive? Answer in one line.

It perpetuates stereotypes and denigrates feminists.

Make a minimal edit to the above meme text (counterfactual), such that you would answer "No".

What is the difference between a hockey player and a feminist? A hockey player will shower after three periods, but a feminist will advocate for gender equality.

Session 2

I have a meme which can be described as a woman in a blue sweater wearing glasses and smiling. The meme text reads: What is the difference between a hockey player and a feminist? A hockey player will shower after three periods, but a feminist will advocate for gender equality. Is this meme offensive? Only answer Yes/No.

Yes.

Why is this meme offensive? Answer in one line.

The meme perpetuates sexist stereotypes and belittles the importance of advocating for gender equality.

Figure 9: Another instance where GPT-3.5 offers an counterfactual self-explanation and employs self-consistency verification to assess its faithfulness.

only on simulatability metrics indicate that X_{cau} does not exhibit patterns like X_{attr}. On the flip side, despite high explainability, input attribution methods do not necessarily capture the underlying causality as indicated by lower rank correlation among various highly simulatable input attribution methods (e.g. DeepLIFT, DeepLIFTSHAP). As noted in this paper, “Correlation does not imply causation."

8. Question: *Why most of the examples in the paper are related to racism? Is it because offense related to racism are easy to see?*

This is not the case. Most of the examples are racist or religious because they incorporate most of the memes in the dataset, comprising of 47.1% and 39.3% of memes respectively.

E Evaluation Metrics

P@5: This is defined as among the top 5 causal keywords, how many are also relevant. This can be mathematically formulated as $\frac{n(A \cap B)}{n(B)}$, where A is the set of relevant gold standard keywords, whereas B is the set of top5 causal keywords, and $n(A)$ shows the number of elements in A. $n(B) = 5$, as we are considering the Top 5 keywords. Consider the first example in Figure 6. $A = \{\text{holocaust, nazism, genocide, extremism}\}$ and $B = \{\text{violence, holocaust, nazism, anti-muslim, terrorism}\}$. $A \cap B = \{\text{holocaust, nazism}\}$, which entails $n(A \cap B) = 2$ and $n(B) = 5$. So the P@5 score is $2/5 = 0.4$. We report the averaged P@5 scores across all test set examples in Tables.

R@5: This is defined as among the gold standard relevant keywords, how many are chosen from the top5 causal set. Following the previous notation, this can be defined as $\frac{n(A \cap B)}{n(A)}$. For the previous example, $n(A \cap B) = 2$ and $n(A) = 4$, which entails R@5 is $2/4 = 0.5$.

MAP@5: This metric is defined as a mean of P@k where k ranges from 1 to 5. P@k values are calculated following the previous paragraph. Mathematically, MAP@5 = (P@1+P@2+P@3+P@4+P@5)/5.

F Experimental Setup

Our proposed model was implemented using PyTorch, a Python-based deep-learning library. All experiments (excluding the baselines) were performed on a single Nvidia P100 16GB GPU on CuDA driver version 12.4. The baselines were run on a different CuDA version compared to the baselines (12.1 vs 12.4). We employed the Adam optimizer (Kingma and Ba, 2017) with a learning rate set to $5e-5$ for optimization. The experiments were conducted using the Facebook Hateful Meme dataset (Kiela et al., 2021). The whole set of 8500 memes is broken down into 7840, 660 examples as the train-test split using a random seed 42. The total number of parameters of the model is 350M and to train a single epoch, 30 mins of continuous GPU execution is needed with a fixed batch size of 32.

To mitigate the impact of non-deterministic GPU operations, we used a fixed random seed of 42 across all the experiments.

Note that many attribution methods, such as saliency, rely on gradient backpropagation (particularly the chain rule) to compute attribution scores. Small precision differences caused by changes in GPU hardware can lead to significant variations in these scores.

G Addressing Ethical Concerns Related to Annotation of Offensive Memes

The involvement of annotators in labeling potentially triggering memes may seem problematic. However, we have implemented several measures to ensure the well-being of the annotators:

Voluntary Participation and Informed Consent: Participation in the annotation process was entirely voluntary. Annotators were fully informed about the nature of the task and the potential exposure to offensive and triggering content through a detailed disclaimer. They provided their informed consent before participating.

Psychological Support: We ensured that annotators had access to free, voluntary sessions with institutional counselors at any time. This support system was established to address any psychological distress that might arise from the task.

Fair Compensation: Annotators were fairly compensated for their work, receiving rates according to institutional policies. This compensation was provided in addition to the fellowships that students receive, in accordance with institute regulations.

Confidentiality: We maintained strict confidentiality and privacy for all student participants throughout the study.

Ethical Review and Approval: Our study, including the use of students to annotate offensive memes, underwent evaluation and received approval from our Institutional Review Board (IRB). The IRB reviewed our study design, the nature of the content, and the involvement of student participants to ensure compliance with ethical standards.

Use of Publicly Available Dataset: The memes included in our study were selected from the well-known, publicly available FB Hateful Meme Dataset. This dataset represents a wide range of themes found on the internet, including some offensive and extreme content. Our goal was to analyze how different types of content are perceived and categorized by labeling them with keywords, necessitating the inclusion of a broad spectrum of examples.