Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis: a Case Study on Hateful Memes

Yosuke Miyanishi[1], **Minh Le Nguyen**[1]
[1]Japan Advanced Institute of Science and Technology
1-1 Asahi-dai, Nomi-shi, Ishikawa, Japan
{yosuke.miyanishi, nguyenml}@jaist.ac.jp

Abstract

Amidst the rapid expansion of Machine Learning (ML) and Large Language Models (LLMs), understanding the semantics within their mechanisms is vital. Causal analyses define semantics, while gradient-based methods are essential to eXplainable AI (XAI), interpreting the model's 'black box'. Integrating these, we investigate how a model's mechanisms reveal its causal effect on evidence-based decision-making. Research indicates intersectionality - the combined impact of an individual's demographics - can be framed as an Average Treatment Effect (ATE). This paper demonstrates that hateful meme detection can be viewed as an ATE estimation using intersectionality principles, and summarized gradient-based attention scores highlight distinct behaviors of three Transformer models. We further reveal that LLM Llama-2 can discern the intersectional aspects of the detection through in-context learning and that the learning process could be explained via meta-gradient, a secondary form of gradient. In conclusion, this work furthers the dialogue on Causality and XAI. Our code is available online (see *External Resources* section).

Keywords: intersectionality, multimodal, XAI, causal, treatment effect, hateful memes, LLM, context

1. Introduction

The domain of causality offers profound insights into the data generation processes, revealing the intricate architecture of the problems at hand. A meticulous examination of these generative processes is indispensable for deep comprehension of phenomena with significant social implications. This paper is dedicated to conducting a rigorous case study in this vein, bridging the gap between the foundational principles of science and the cutting-edge capabilities of Machine Learning (ML) technologies.

EXplainable Artificial Intelligence (XAI) emerges as a critical paradigm in shedding light on ML models' often opaque inner workings. While previous research has ventured into various domains, the application of XAI principles to causal analysis remains scarcely explored. By integrating causality and XAI, this study aims to enrich our understanding of social phenomena and how they are reflected in state-of-the-art (SOTA) ML models facing the representation of the phenomena.

The rise of online hate speech, especially hateful memes (Fig. 1, top) —comprising both text and image, has prompted significant research. While multimodal ML algorithms have seen substantial improvements, efforts focus more on benchmarking and maximizing performance, including the Hateful Memes Challenge competition (Kiela et al., 2020), rather than applying XAI methods. Existing approaches also lack a focus on causal architecture. This study defines hateful meme detection as an Average Treatment Effect (ATE) estimation problem for input data modalities (image and text) and examines the effects through the prism of XAI.

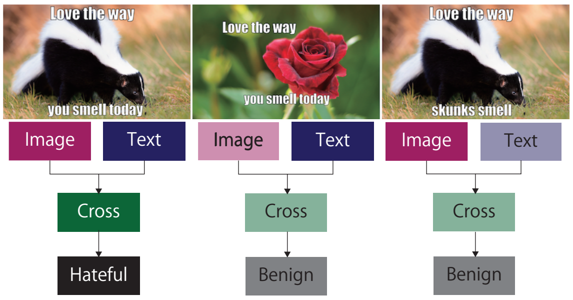

Figure 1: Visualization of a hateful meme and its corresponding confounders. (top) Meme samples and (bottom) their directed acyclic graph representation. (left) A hateful meme highlights cross-modal interactions between its image and text components that contribute to its hatefulness. (middle) The image benign confounder showcases original text and a non-hateful image, resulting in reduced cross-modal interactions and a benign classification. (right) The text benign confounder comprises an original image and non-hateful text. Note: The samples depicted are illustrative and do not exist in the dataset. ©Getty Images

Intersectionality[1], or *the network of connections between social categories such as race, class, and gender, especially when this may result in addi-*

[1] Oxford Dictionary

tional disadvantage or discrimination, acts as a bridge between ML and social science. Though broadly applied in social science and used for debiasing in ML, its wider applications are limited. How can we use this concept as a *generalized* tool for broader problems? Motivated by this question, this paper proposes reframed intersectionality, explores whether causally formalized intersectionality can address a broader range of problems, and evaluates *inductive* bias in ML models.

Furthermore, the excellence of Large Language Models (LLMs) across various benchmarks has been showcased, particularly in few-shot learning. The concept of in-context learning presents a promising avenue, but its formal causal evaluation is limited. Here, we address this problem.

Our contribution could be summarized as:

- Formalization of hateful meme detection as an intersectional causal effect estimation problem, allowing performance assessment based on data generation process.

- Introduction of reframed causal intersectionality to evaluate inductive bias, marking a step towards broader applications, including demographics-nationality intersectionality, financial inclusion, and clinical diagnosis.

- Demonstration that attention attribution scores (Hao et al., 2021) divided by modality interaction describe the causal effect accurately, unlike non-divided scores. This finding opens doors for causal explainability in multimodal settings (Liu et al., 2022).

- Pioneering formal and empirical analysis of LLM's meta-optimization process in the multimodal in-context setting.

2. Related Work

2.1. Causal ML and XAI

Causal Inference (CI) occupies a pivotal role in the elucidation of social phenomena and the interpretation of intervention outcomes. It bifurcates into two primary methodologies: the graphical and structural schemas for modeling reality (Pearl, 2001), alongside the framework for potential outcome prediction (Rubin, 2008). CI's utility spans a diverse array of sectors, including but not limited to, medicine (Vlontzos et al., 2022), manufacturing (Vuković and Thalmann, 2022), and the social sciences (Sengupta and Srivastava, 2021), guiding the interpretation of data within these fields. Furthermore, CI principles have been applied within Machine Learning (ML) and its allied disciplines, giving rise to the subfield of Causal ML. Causal ML encompasses research into natural language processing (Yang et al., 2022), hate speech detection (Chakraborty and Masud, 2022), and the study of image-text multimodality (Sanchez et al., 2022). Notably, the theoretical underpinnings of hateful memes, as a convergence of these interests, remain underexplored. Our research attempts to map out graphical and formal representations of the causal structures underlying hateful memes.

In conjunction, EXplainable Artificial Intelligence (XAI) (Speith, 2022; Joshi et al., 2021; Barredo Arrieta et al., 2020) has sought to demystify the internal mechanisms of ML models. XAI's domain of inquiry extends across various fields, including medicine (Holzinger, 2021) and energy (Machlev et al., 2022), with a particular focus on both model-agnostic (Sundararajan et al., 2017; Gaur et al., 2021; Marcos et al., 2019) and model-specific (Hao et al., 2021; Holzinger et al., 2021) evaluations. However, the intersection of causality with XAI remains nascent. This study investigates XAI's capability in assessing attributions to causality metrics, emphasizing gradient-based interpretations as central to XAI endeavors.

Since ML models typically minimize the gradient for optimization, components with steep gradients toward the model's decision-making are considered crucial. The gradient-based XAI approach (Selvaraju et al., 2017), often model-specific, finds pertinent application in the analysis of Transformers (Vaswani et al., 2017), which underpin most SOTA models in natural language processing (NLP). Here, quantifying the attribution of attention matrix weights via the gradient emerges as a direct method (Hao et al., 2021). Our research proposes both theoretical and empirical advancements in causal analysis, leveraging this gradient-based methodology to enhance understanding and interpretation within the causality domain.

2.2. Intersectionality

Intersectionality, a bias indicator of multiple demographics within various domains, has inspired a few causal analyses (Yang et al., 2021; Bright et al., 2016). While XAI techniques have been used to alleviate its negative impact in ML literature (Lalor et al., 2022; Simatele and Kabange, 2022), our work redefines intersectionality for broader problems, and pioneers the quantification of inductive intersectional bias.

2.3. Hate Speech and Hateful Memes

Hate speech and hateful memes have attracted substantial ML research attention, involving various models (Das et al., 2020; Lippe et al., 2020) and datasets (Kiela et al., 2020; de Gibert et al., 2018; Davidson et al., 2017; Sabat et al., 2019; Suryawanshi et al., 2020). Previous analytical works have focused on racial bias (Sengupta and Srivastava, 2021; Sharma et al., 2022), virality (Ling et al., 2021; Chakraborty and Masud, 2022), and propaganda techniques (Dimitrov et al., 2021), and a few have applied XAI methods (Cao et al., 2021; Hee et al., 2022; Deshpande and Mani, 2021). This study builds upon these by formalizing hateful meme detection as a causal effect estimation problem and emphasizing the importance of modality interaction.

2.4. LLM

Large Language Models (LLMs), known for their powerful in-context few-shot learning capabilities (Brown et al., 2020) in various NLP and multimodal tasks, are emerging as significant tools (Zhao et al., 2023). To understand their inner workings, meta-gradient, or the update of the attention weights as a secondary form of the gradient, explains in-context learning empirically (Coda-Forno et al., 2023; Chen et al., 2022) and theoretically (Dai et al., 2023; von Oswald et al., 2023).
However, understanding LLM's causal power remains a complex and emerging area of study (Ban et al., 2023; Kıcıman et al., 2023). Moreover, unlike the traditional classifier-attached-encoder model (e.g. BERT (Devlin et al., 2019)) with *predicted probability*, how to estimate the causal effect of chatbot-style LLM and analyze its inner workings quantitatively remains elusive. This study demonstrates that a dedicated task design could be used to estimate LLM's causal effect and that a meta-gradient could explain its inner workings concerning that effect.

3. Methodology

3.1. Background

3.1.1. Average Treatment Effect

The *Average Treatment Effect (ATE)* (Rubin, 2008) is a key metric for assessing causal impacts. It represents the average difference in outcomes between treated and untreated groups. This measurement facilitates a standardized evaluation of causal effects across diverse scenarios. For a binary treatment $BT \in (0,1)$ yielding outcome θ_{BT}, ATE is defined as:

$$ATE = \theta_1 - \theta_0 \tag{1}$$

This research repurposes intersectional ATE to evaluate hateful meme detection models.

3.1.2. Causal Intersectionality

Building upon the textual definition, causal intersectionality (Bright et al., 2016) challenges the simplistic aggregation of individual demographic effects. Instead, it highlights the complex, synergistic interactions between multiple demographic factors, acknowledging the nuanced dynamics that influence causal relationships in social studies. Defining causal intersectionality (Eq. 2) involves binary vectors for two demographics (e.g., gender D_1 and color D_2) marked as $D = \{D_1, D_2\}$, and the outcome θ.

$$\theta_D \neq \sum_i \theta_{D_i} \tag{2}$$

Using this causality structure, we assess multimodal models within a causal context.

3.1.3. Attention Attribution Score

Simply put, *attention attribution score* quantifies the contribution of attention weights to the model's decision-making. Initially, a seminal work (Sundararajan et al., 2017) introduced the *integrated gradient* method for quantifying model component's attribution. This method calculates the contribution of specific model components based on the gradient's integral over that component. Upon this work, another study (Hao et al., 2021) reported its applicability to Transformer's attention weights, deriving attention attribution score ($attr$ herein). This approach proves critical for interpreting the Transformer's behavior, especially in understanding how the attention matrix influences model outputs. Given hyperparameter α, $attr$ computes the integrated gradient for attention matrix A relative to a Transformer's output θ.

$$attr = A * \int_{\alpha=0}^{1} \frac{\partial\theta(\alpha A)}{\partial A} d\alpha \tag{3}$$

We employ modality-wise averaged $attr$ differences as causal effect indicators.

3.1.4. Learning Objectives: Binary Classifier vs LLM

For the classifier-attached-encoder model, hateful meme detection aligns with binary classification. In summary, given *both hateful and benign pairs*, the classifier tries to maximize the classification performance. With a ground-truth label y_{gt} and a loss function f_{loss}, the learning objective when training a model θ is:

$$\underset{\theta}{\text{argmin}} -\{y_{gt} f_{loss}(\theta) + (1 - y_{gt}) f_{loss}(1 - \theta)\} \tag{4}$$

On the other hand, in in-context learning, hateful samples and their counterpart confounders are presented to LLM *in parallel*. This differs from the objective above in that the information of hateful samples is not used to handle confounders, and vice versa. For example, facing the hateful sample in the zero-shot setting, the meta-objective it is trying to meta-optimize to is:

$$\underset{\theta}{\text{argmin}} -y_{gt} f_{loss}(\theta) \tag{5}$$

This study delves into the meta-objective for LLM in a few-shot context for the equivalent comparison.

3.1.5. Meta-Optimization in Few Shot Setting

Meta-optimization, in the realm of in-context learning, mirrors gradient descent (Irie et al., 2022; Dai et al., 2023). Given query q as an input, a dual learning process of linear attention θ - fine-tuning and in-context learning with attention update ΔA - is contrasted with A_{ZSL}, the weight in a zero-shot setting.

$$\theta = (A_{ZSL} + \Delta A)q \tag{6}$$

ΔA functions as a gradient variant named meta-gradient. Our extension encapsulates attention attribution and its subsequent in-context learning results.

3.2. Proposed Methodology Overview

Expanding on the original causal intersectionality (Eq. 2), we define intersectionality for text T and image I modalities. With $X_1 = (T_1, I_1)$ indicating hateful content and two types of benign samples $X_0 \in \{(T_1, I_0), (T_0, I_1)\}$, the multimodal intersectionality is:

$$\theta_{X_1} \neq \sum_{X_0} \theta_{X_0} \tag{7}$$

The remainder of this section elaborates on causal intersectionality in meme detection, $attr$-based modality assessment, and LLM evaluation, visualized in Fig. 2.

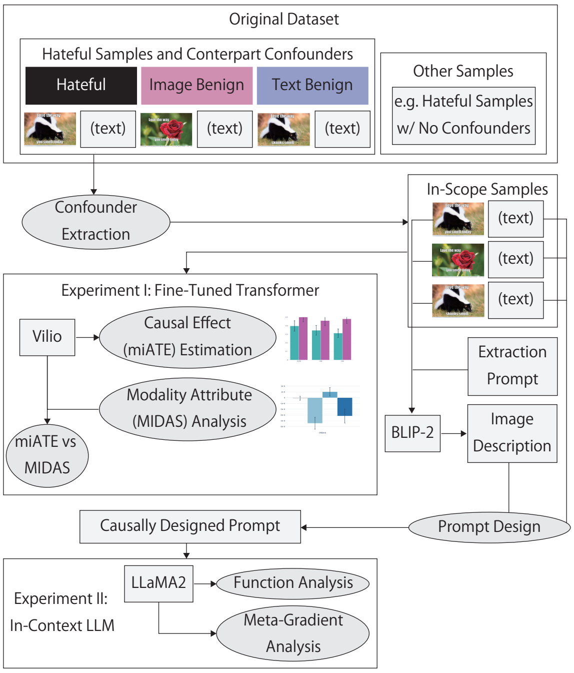

Figure 2: A schematic overview of our proposed methodology. Rectangular boxes denote data or models, while circular shapes represent the processes involved.

3.3. Causal Multimodal Intersectionality

3.3.1. Intersectionality Reframed

We broaden causal intersectionality, positing its utility beyond human demographics to include arbitrary components. This reframed intersectionality assesses interconnections between *arbitrary categories such as user demographics or input modalities*, and how they amplify effects on significant issues - capturing indirect effects[2] in social contexts. This adaptation remains consistent with the original causal formalism (Eq. 2).

3.3.2. Performance Measurement with Causal Intersectionality

In the data generation process of hateful memes, a text and an image, which are benign in isolation, jointly produce hate. Applying demographic intersectionality concepts, we introduce the *multimodal intersectional Average Treatment Effect (miATE)*.

$$miATE = \theta_{X_1} - \sum_{X_0} \theta_{X_0} \tag{8}$$

Model performance is assessed considering differences in each modality. Hateful samples are categorized by original benign text confounders T_0^{org} and image benign confounders I_0^{org}. The *Confounder Extraction* section provides more details.

3.4. Attention Attribution Score by Interaction

We introduce the Modality Interaction Disentangled Attribution Score (MIDAS), which quantifies the contribution from various interaction types $t \in \{within\_image, within\_text, cross\_modal\}$ to a model's decision. Given an input $X_i = (T, I)$, with N denoting the number of elements for interaction type t, $MIDAS$ computes the modality-wise average avg_t of the attention attribution score $attr^{X_i}$, analogous to $miATE$. Following the initial work on attention attribution score (Hao et al., 2021), $MIDAS$ is calculated from the last hidden layer, excluding $[CLS]$ and $[SEP]$ tokens.

$$\begin{aligned} MIDAS &= avg_t\{attr^{X_1} - \sum_{X_0}(attr^{X_0})\} \\ avg_t(X) &= \frac{1}{N}\sum^{t} X \end{aligned} \tag{9}$$

Note that to mitigate the negative impact of class imbalance (Hossain et al., 2022), θ and $attr$ is averaged per sample and confounder category $(T_1, I_1), (T_1, I_0), (T_0, I_1)$.

[2] Defined as an effect of two variables X_1 and X_2 to variable X_3 via another variable Z

3.5. Formal Relation between miATE and MIDAS

We examine the relationship between $miATE$ and $MIDAS$ by proposing that $MIDAS$ can be perceived as an attention attribution to $miATE$. Given $G(A)$ as the one-step gradient for an attention matrix A, we see that $attr$ approximates the product of the gradient and the attention (Eq. 10 - true when $\alpha = 1$). $MIDAS$ is expressed as the difference of that value between hateful and benign content (Eq. 11). Furthermore when $MIDAS$ is aggregated across n samples that are representative of the entire dataset, it depicts the variation in attention expectancy normalized with the function $\mathcal{N}$ across these samples, as shown in Eq. 11. In essence, we propose that $MIDAS$ acts as an attention-focused representation of the model's causal effect, i.e. $miATE$.

$$\begin{aligned} attr^{X_i} &\simeq A^{X_i} * G(A^{X_i}) \\ where\ G(A^{X_i}) &= \frac{\partial \theta(A)}{\partial A^{X_i}} \end{aligned} \tag{10}$$

$$\begin{aligned} MIDAS \simeq avg_t\{&A^{X_1} * G_{norm}(A^{X_1}) \\ &- \sum_{X_0}(A^{X_0} * G_{norm}(A^{X_0})\} \\ \sum_n MIDAS \simeq \mathbb{E}[&A^{X_1}] - \sum_{X_0} \mathbb{E}[A^{X_0}] \end{aligned} \tag{11}$$

$$\begin{aligned} \text{where} \quad & G_{norm}(A) = \mathcal{N}(G(A)) \\ & \mathcal{N}(G(A)) : G(A) \rightarrow (0,1) \end{aligned}$$

3.6. LLM

3.6.1. Causal Objective: Implicit miATE maximization

Before discussing LLM, we show that training a classifier implicitly addresses the $miATE$ maximization problem. Specifically, the objective (Eq. 6) over a hateful-confounder pair could be written as:

$$\underset{\theta}{\text{argmin}} -\{f_{loss}(\theta_{X_1}) + \sum_{X_0} f_{loss}(1 - \theta_{X_0})\} \tag{12}$$

Here, we see that the first term maximizes the first term of Eq. 8 and the second term minimizes the second term of Eq. 8. In contrast, zero-shot LLM only addresses the first term of Eq. 12. In the next section, we show how we design the task for LLM to aim for the same goal.

3.6.2. Causal Task Design

To meta-optimize to the causal task, the hateful-confounder pair should be given *to an identical, not separate, meta-optimizer*, or the optimizer cannot have any information about intersectional causality (second term of Eq. 12). Table 1 shows the causality-oriented design of a representative prompt for hateful meme detection.

User: Out of image-caption pairs #0 to #2, select the most likely hateful or sarcastic pair with a potential label (hateful or sarcastic). If all pairs are benign, please say so.
#0: image: 'Skunk', caption: 'Love the way you smell today'
#1: image: 'Rose flower', caption: 'Love the way you smell today'
#2: image: 'Skunk', caption: 'Love the way skunks smell'
System:

Table 1: An illustrative prompt for the causal objective in the zero-shot scenario. In the few-shot context, answers are delivered succinctly (e.g., *#0 is hateful.*).

Note that this task design inherently counteracts sample imbalance since it simultaneously represents these hateful, original benign, and picked benign samples.

3.6.3. Meta-Optimization for Causal Objective

Meta-optimization for the causal objective poses challenges (Niu et al., 2021) like complicated instruction and varied available labels. The optimization process consists of:

1. *Task Type Classification (TTC)*. LLM recognizes the task as a binary classification.

2. *Label Identification (LI)*. LLM provides probable labels, e.g., *hateful*, *sarcastic* (Chauhan, 2020), and *benign*. Note that including the *sarcastic* label addresses its nuanced overlap with hatefulness (Sundaram et al., 2022), capitalizing on LLM's comprehension of complex social phenomena embedded in training corpora. We still regard this task as a binary classification of *all-pair-benign* vs *one-pair-hateful*, with a subtask of *hateful-sarcastic* classification.

See Table 2 for a set of examples.

Subtask	Label	Response
TTC	Negative	Sorry, I couldn't understand your instructions.
TTC	Positive	#1 could be sarcastic.
LI	Negative	#1 could be sarcastic.
LI	Positive	#0 could be hateful.

Table 2: Synthesized responses and subtask labels for in-context learning. Refer to Table 1 for the corresponding instruction prompt.

The meta-optimization process is segmented into these subtasks, with the understanding that LI follows a successful TTC, denoted as $TTC = 1$. The output of meta-optimized Transformer block θ could be formalized as:

$$\theta^{SubTask} = (A^{SubTask} + \Delta A^{SubTask})q$$
$$\theta = \begin{cases} \theta^{TTC} + \theta^{LI} & (TTC = 1) \\ \theta^{TTC} & otherwise \end{cases} \quad (13)$$
$$where\ SubTask \in \{TTC, LI\}$$

4. Experimental Settings

4.1. Data Preparation

4.1.1. Hateful Memes Dataset

Our study utilizes the Hateful Memes Challenge dataset (Kiela et al., 2020) and focuses primarily on the *dev_seen* subset. Unimodal hateful samples (Das et al., 2020; Lippe et al., 2020) are omitted from our study.

4.1.2. Confounder Extraction

From the dataset, 162 pairs of hateful (T_1^{org}, I_1^{org}) and benign samples (T_1^{org}, I_0^{org}) or (T_0^{org}, I_1^{org}) are identified. Since most of the pairs have either one of the text or image confounders, not both, three random inputs with the missing modality (I_0^{picked} or T_0^{picked}) are concatenated with the other modality to accommodate the requirements of Eq. 8, resulting in a uniquely crafted subset $(T_1^{org}, I_0^{picked})$ and $(T_0^{picked}, I_1^{org})$. The structure of this subset is summarized in Table 3.

Sample Category	Number of Samples
Hateful	162
Image Benign	78
Text Benign	84
Picked Image Benign	234
Picked Text Benign	252

Table 3: Samples utilized in our analysis. This table categorizes samples as Hateful (T_1^{org}, I_1^{org}), Image Benign (T_1^{org}, I_0^{org}), Text Benign (T_0^{org}, I_1^{org}), Picked Image Benign $(T_1^{org}, I_0^{picked})$, and Picked Text Benign $(T_0^{picked}, I_1^{org})$.

4.2. Experiment I: Fine-Tuned Transformer

4.2.1. Analysis Type

Assuming the predominant contribution of original inputs over the picked ones, in respect of the authors' effort of making the task challenging, we divided the analysis into that of $\{(T_1^{org}, I_1^{org}), (T_1^{org}, I_0^{org}), (T_0^{picked}, I_1^{org})\}$ (denoted as *org. text*), and of $\{(T_1^{org}, I_1^{org}), (T_0^{org}, I_1^{org}), (T_1^{org}, I_0^{picked})\}$ (*org. image*).

4.2.2. Models

We employ author implementation of the SOTA (Kiela et al., 2021) *Vilio* framework (Muennighoff, 2020) for its superior capabilities and adaptable framework, focusing on its three main models: Oscar (Li et al., 2020b), UNITER (Chen et al., 2020), and VisualBERT (Li et al., 2020a), summarized in Table 4. Each model type has three submodels (training corpora or random seed variants), all included in our analysis but the results shown here are from selected one submodel (preliminary analysis shows all submodels exhibit similar trend).

Type	Encoder	Pretraining Task
O	BERT (base)	1) Object tag (or *anchor*) detection 2) Text-image contrastive learning
U	BERT (base)	1) Masked language modeling 2) Masked image modeling 3) Image-text matching 4) Word-region alignment via optimal transport
V	BERT (base)	1) Masked language modeling 2) Image captioning

Table 4: A categorization of Vilio's submodels leveraged in our research. The models are classified into three groups: Oscar (O), UNITER (U), and VisualBERT (V).

4.3. Experiment II: LLM

4.3.1. Models

HuggingFace *Llama-2-13b-chat-hf* (Touvron et al., 2023) is our language model backbone, optimized for chat-style interactions. To convert the image into its textual description, we utilize the BLIP-2 (Li et al., 2023) model with a *FlanT5-XXL* (Chung et al., 2022) backbone.

4.3.2. In-Context Learning

We study Llama-2's behavior on image caption in the original dataset and image description generated by BLIP-2. For in-context learning, the number of samples is limited due to memory restriction. After the response is generated, one of the authors conducts manual labeling since the number of samples is limited (available at our GitHub repository). We gauge performance through accuracy.

4.3.3. Meta-Gradient Evaluation

During our evaluations, we mask redundant subtext (e.g., *#0: image:* and *caption:*) in input prompts.

4.4. Shared Settings

4.4.1. Probing

We employ a probing (Alain and Bengio, 2017) approach with LightGBM (Ke et al., 2017) to explore the impact of modality interaction and in-context learning on causal effect. Responses are split into training (56%), validation (14%), and test (30%) sets. We achieve hyperparameter tuning using Optuna (Akiba et al., 2019). To assess the effects of interaction type t (Experiment I, II) and model type (Experiment I), corresponding categorical variables and interaction terms with $MIDAS$ (Experiment I) or summed attention weights (Experiment II) are added to our analysis. We determine significance using a t-test ($p < 0.05$).

4.4.2. Text-Only Pretrained BERT

For VisualBERT encoder replacement (Experiment I) and BLIP-2-fused-BERT (Experiment II), we use HuggingFace *bert-base-uncased*. In Experiment II, the last four layers of BERT and a linear classifier are trained for 100 epochs with an Adam optimizer (learning rate 5e-5), evaluating its performance across different seeds.

4.4.3. External Resources

All code and experiments are accessible at `https://github.com/HireTheHero/CausalIntersectionalityDualGradient`. Experiments are conducted on a single NVIDIA A100 GPU, either through Google Colaboratory Pro+ or locally.

5. Results & Discussion

5.1. Experiment I: Fine-Tuned Transformer

5.1.1. miATE

First, we assessed each model's performance with $miATE$ (Fig. 3). VisualBERT exhibited the highest disparity, highlighting its bias for text-based tasks (Table 4).

5.1.2. MIDAS Global Analysis

Next, we assessed the model's inner workings (Fig. 4 and 5). $MIDAS$ of Oscar and UNITER (Fig. 4, first and second row) showed predictable trends of attending to one modality while the other is the same. In contrast, VisualBERT's behavior of attending to text-related interactions (third row) mirrored its pretraining tendencies biased towards text (Table 4). Furthermore, replacing VisualBERT's encoder with the one pretrained only with text enhanced the bias (fourth row), which supports the presence of pretraining bias represented in $MIDAS$. We observed no significant model differentiation with the original $attr$ (Fig. 4, left column of each graph), suggesting a simple yet important contribution of modality-wise split.

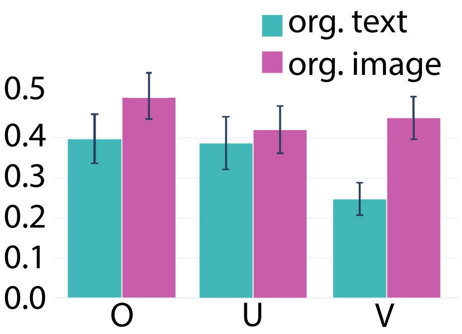

Figure 3: Multimodal Intersectional Average Treatment Effect ($miATE$) across Oscar (O: left), UNITER (U; middle), and VisualBERT (V; right) models, contrasting the samples with original image confounders (*org. image*, cyan) and those with original text confounders (*org. text*, magenta).

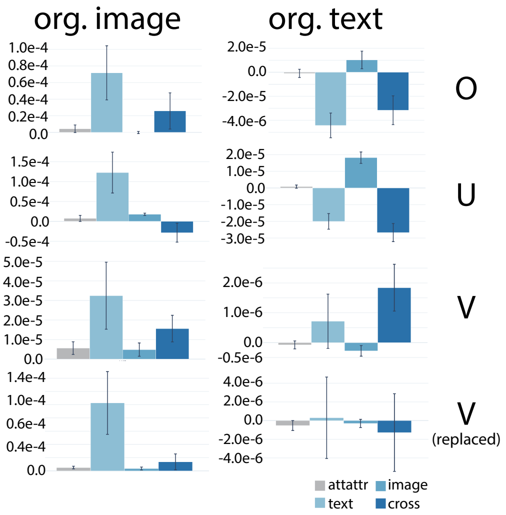

Figure 4: $MIDAS$ for *org. image* (left) and *org. text* (right) samples featuring Oscar (top), UNITER (second row), VisualBERT (third row), and VisualBERT with text-only-pretrained encoder (bottom). From left to right, each graph displays $attr$ with no modality division, $MIDAS_{within\_text}$, $MIDAS_{within\_image}$, and $MIDAS_{cross\_modal}$.

5.1.3. MIDAS Local Analysis

To see if we can interpret the single hateful-benign pair, we extracted local explanation (Chai et al., 2021; Hee et al., 2022). A representative pair (Fig. 5) illustrates that UNITER captures the contrast between a woman and cargo in image confounder analysis (first and third row), and the model similarly attended to the words *dishwasher* and *driving* for text analysis (first and second row).

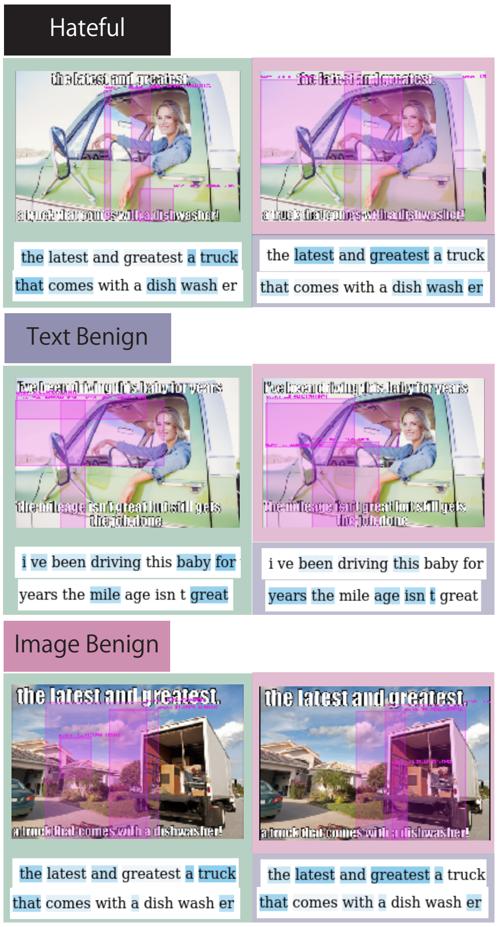
Hateful
the latest and greatest a truck
that comes with a dish wash er
the latest and greatest a truck
that comes with a dish wash er
Text Benign
i ve been driving this baby for
years the mile age isn t great
i ve been driving this baby for
years the mile age isn t great
Image Benign
the latest and greatest.
a truck that comes with a dishwasher
the latest and greatest a truck
that comes with a dish wash er
the latest and greatest a truck
that comes with a dish wash er

Figure 5: Conceptual portrayal of hateful, text benign, and image benign samples derived from UNITER. $MIDAS$ reflects heightened $attr_{cross\_modal}$ (green), $attr_{within\_image}$ (red), or $attr_{within\_text}$ (blue) values. Both image and text inputs spotlight top-scored ROIs and tokens. The text is abbreviated for clarity.

5.1.4. Empirical Relation between MIDAS & miATE

To probe the relationship between $MIDAS$ and $miATE$, we first modeled the entire $(MIDAS, miATE)$ pairs of all the models by a single probe, resulting in a moderate AUC of 75.6 ± 4.20. Next, to see the effectiveness of probing for each model, we applied one probe for one model, resulting in the highest AUC for V (AUC 94.1 ± 3.01), while low-to-moderate for O (60.8 ± 5.18) and U (74.3 ± 2.39). In addition, to see if the probes reflect the findings on $MIDAS$, we analyzed the feature importance (Table 5). Consistent with previous findings, VisualBERT and within-text appear with the highest frequency among model type and interaction type, respectively. In summary, our findings showed a moderate correlation between $MIDAS$ and $miATE$ for BERT-based models, with a particularly robust link for VisualBERT, echoing its distinct model nature.

Interaction Type	O	U	V
within-text	35±25	30±26	253±66
cross-modal	19±14	35±22	223±75
within-image	30±23	54±32	169±74

Table 5: LightGBM probe's feature importance between $MIDAS$ and $miATE$. Values represent frequency counts.

# Few-Shot Samples	Accuracy		
	All(S-)	TTC(S-)	TTC(S+)
0	46.2	61.6	62.6
1	62.9	62.9	60.6
2	62.5	62.5	61.0
3	59.2	59.2	57.6
4	64.3	64.3	71.4

Table 6: Zero-shot (first row) and few-shot (second to fifth row) Llama-2 performance. *All* signifies cumulative sample results, while *TTC* relates to correct TTC samples. Parentheticals (*S+* or *S-*) denote the inclusion of the *sarcastic* label in either positive or negative samples.

5.2. Experiment II: LLM

5.2.1. Effectiveness of BLIP-2 information retrieval

To assess the merit of BLIP-2 information retrieval, we utilized its image description and the original captions to fine-tune BERT pretrained only with text. The resultant enhanced performance (Accuracy 66.9 ± 0.84, AUC: 71.2 ± 1.55) to unimodal BERT benchmarks (Kiela et al., 2020) underlines BLIP-2's effective image information extraction capabilities.

5.2.2. In-Context LLM Performance

Our evaluation with Llama-2 shows that all-sample accuracy improved after one sample (Table 6, left). Interestingly, after just one in-context example, the model achieved exactly the same performance for all samples and $TTC = 1$ samples, meaning impeccable TTC Recall. These results suggest the critical role of in-context examples in task comprehension when the task is challenging in zero-shot settings. Marking the sarcastic label as positive led to better performance at the zero-shot setting but dropped after one example (Table 6, middle and right), implying uncertainty in the decision-making for this label.

5.2.3. Meta-Optimization Evaluation

To gauge the influence of meta-optimization, we applied a probe model to examine the relationship

Interaction Type	A	ΔA
within-text	65±41	24±17
cross-modal	58±33	31±21
within-image	43±25	14±8

Table 7: LightGBM probe's feature importance in Experiment II. Features of zero-shot weights (A) or their few-shot updates (ΔA) are divided by interaction types.

between summed attention weights $(A, \Delta A)$ and TTC label, revealing a moderate AUC of 82.3 ± 6.16. Furthermore, a detailed extraction of feature importance (Table 7) from the probe model allowed us to determine how A and ΔA impact the probe model (left and right). Our findings suggest that while A carries substantial weight in decision-making, its meta-optimized counterpart ΔA also plays a vital role. Regarding the effects of modality interaction (top and bottom), captured by interaction-type-divided weights $A + \Delta A$, each interaction type did contribute to TTC. When examining the differential impacts of each type, however, no significant disparities in their contributions were identified. Addressing the challenge of discerning between them will be a part of the future work.

5.3. Discussion

The primary goal of this paper is to assess models based on the data generation process and its underlying concepts - *hatefulness* in the case of hateful memes. While this deviates from standard ML evaluations focusing on performance metrics like accuracy, it is scientifically valid and relevant to ML problems, like Rubin started his line of causal inference works to analyze the impact of nulled variables (Rubin, 2008). In our study, we demonstrate that the generation of hateful memes embodies multimodal intersectionality, and the SOTA Transformer models effectively capture this nature of the data but are biased by pretraining datasets. In the future, we hope to apply our method to other multimodal problems like the missing modality problem (Ma et al., 2021; Wang et al., 2023), an inherently close one to nulled variable evaluation.

Our study carves a niche by reconceptualizing hateful meme detection through the lens of modality interaction and causal effect. Compared to the seminal work on causal intersectionality (Bright et al., 2016), beyond mere technical insights, we proffer a paradigm shift in causal intersectionality. Our method's key advantage is its unique capability for modality-wise causal analysis, a novel contribution in this field. Despite its simplicity, the causal effect of the modalities is neither investigated nor formally defined in the existing literature. Empirically, Experiment I unveils model biases overlooked by traditional methods like attention attribution scores (Hao et al., 2021; Hee et al., 2022) without consideration for modality.

For Experiment II, our exploration of few-shot LLM performance has provided an understanding of how LLM adapts to different levels of input information, shedding light on their capabilities and limitations in various scenarios. Applying meta-gradients has allowed us to assess the attribution of attention weights, adding granularity to the interpretability landscape. Evaluating the effectiveness of the causal task over the causal evaluation of LLMs is challenging since it is a new concept. Nonetheless, this could be a valuable benchmark for future model evaluations. Despite relying on a specific instruction prompt for the causal task, we could adapt the design for broader applications. For example, with a simple modification to the prompt, we could test LLMs with multi-class meme classification (Davidson et al., 2017).

6. Conclusion

We posit that hateful meme detection transcends mere classification, gravitating towards intersectional causal effect analysis. Our evaluations spanned various Transformer architectures in unique settings. To ensure our approach's universality, extending our evaluations to other hateful memes datasets (Gomez et al., 2020; Das et al., 2023) will be pivotal. In the quest for broader insights, exploring diverse challenges, such as the intersectionality in multimodal medical analyses (Azilinon et al., 2023), will be part of our future work. For scalability, utilizing more of the power of LLMs will be promising for confounder extraction and response evaluation.

7. Ethical Considerations

In this research, we aim to develop innovative analytical methods for identifying and mitigating the proliferation of hateful memes, a pressing concern given the complex interplay of text and imagery in propagating hate speech online. The nature of this endeavor necessitates rigorous ethical scrutiny, especially concerning the selection, utilization, and presentation of these memes within our scholarly work and its broader dissemination. Herein, we delineate the principal ethical considerations guiding our study.

The hateful memes we examine originate from a prior investigation (Kiela et al., 2020). We direct readers to the original study for insights into the ethical measures employed during the dataset's compilation. Our engagement with this dataset is underpinned by a firm commitment not to propagate or validate the adverse messages it encompasses.

Our sample selection approach is predicated on a causality framework detailed in the *Confounder Extraction* section, ensuring a comprehensive examination of hate speech manifestations across diverse community targets. This methodology underscores our commitment to a nuanced analysis that refrains from generalizations or biases.

A pivotal aspect of our ethical strategy is to reconcile the imperative of methodological transparency with the necessity to limit harm. Consequently, we exhibit restraint in our presentation of hateful memes. Specifically, Figure 5 is the sole instance within our publication where an actual hateful meme is depicted. We have exercised meticulous care to ensure that neither the accompanying text description nor the figure caption disseminates any form of hate speech.

This strategy is emblematic of our broader ethical stance, emphasizing the conscientious handling of sensitive content. Our research is animated by a profound dedication to combating hate speech in all its forms, reflecting an unwavering commitment to ethical research practices that respect the dignity of all individuals and communities. Through this work, we aspire not only to advance the field of hate speech detection but also to contribute meaningfully to creating more inclusive and respectful digital spaces.

8. Limitations

This study's primary limitation concerns the unverified generalizability of its findings. Hateful memes represent an evolving area of concern that necessitates extensive, openly accessible datasets for comprehensive analysis and validation. Our research endeavors to tackle this challenge, yet the broader scope for future exploration is highlighted by the potential applications of our findings, as detailed in the *Discussion* and the *Conclusion* sections.

A further constraint is the linguistic homogeneity of the dataset employed, with the Hateful Memes Challenge dataset comprising exclusively English-language textual content. This presents a critical limitation in the context of the global escalation of extremism, where hate speech proliferates across linguistic boundaries. The detection of multilingual hate speech thus emerges as a crucial area for future research, necessitating methodologies capable of navigating language-specific nuances and cultural contexts.

Additionally, the field of hate speech detection faces resource limitations, notably in the size and diversity of available datasets. Hateful speech datasets are generally small, restricting the depth and breadth of training data for machine learning models. We believe future studies could utilize LLMs as dataset curators.

In summary, while this study contributes valuable insights into detecting and mitigating hateful memes, it also underscores the need for further research. Addressing the limitations related to dataset generalizability, linguistic diversity, and the scarcity of training data are pivotal steps toward developing more effective and universally applicable solutions for combating online hate speech. Exploring innovative methods, such as LLM-based dataset curation, represents a promising direction for overcoming these challenges.

From a theoretical standpoint, the groundwork of our in-context learning analysis relies upon the principles of simplified linear attention (Irie et al., 2022; Dai et al., 2023). However, this foundation's direct applicability to conventional Transformer models invites scrutiny. Consequently, a more nuanced interpretation (Ren and Liu, 2023) may be imperative for advancing our understanding in future investigations.

9. Acknowledgments

CyberAgent Inc.[3] partially supported this work, enabling our conference participation. We are profoundly grateful to the anonymous reviewers, whose constructive feedback significantly enriched our discussions and contributed to refining our research. Yosuke Miyanishi thanks J.X., H.W., and T.Y. for their insightful comments that significantly improved our study.

10. References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Op-

[3] https://www.cyberagent.co.jp/

tuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, Anchorage AK USA. ACM.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations*, Toulon, France.

Mikhael Azilinon, Julia Makhalova, Wafaa Zaaraoui, Samuel Medina Villalon, Patrick Viout, Tangi Roussel, Mohamed M. El Mendili, Ben Ridley, Jean-Philippe Ranjeva, Fabrice Bartolomei, Viktor Jirsa, and Maxime Guye. 2023. Combining sodium MRI , proton MR spectroscopic imaging, and intracerebral EEG in epilepsy. *Human Brain Mapping*, 44(2):825–840.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From Query Tools to Causal Architects: Harnessing Large Language Models for Advanced Causal Discovery from Data. *arXiv preprint*.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. 2016. Causally Interpreting Intersectionality Theory. *Philosophy of Science*, 83(1):60–81.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rui Cao, Ziqing Fan, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2021. Disentangling Hate in Online Memes. *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147.

Yidong Chai, Yonghang Zhou, Weifeng Li, and Yuanchun Jiang. 2021. An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1.

Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the Bud: Detection, Diffusion and Mitigation of Hate Speech on Social Media. *ACM SIGWEB Newsletter*, 2022(Winter):1931–1745.

Dushyant Singh Chauhan. 2020. All-in-One: A Deep Attentive Multi-task Learning Framework for Humour, Sarcasm, Offensive, Motivation, and Sentiment on Memes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 281–290, Suzhou, China. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via Language Model In-context Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12375, pages 104–120. Springer International Publishing, Cham.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X. Wang, and Eric Schulz. 2023. Meta-in-context learning in large language models. *arXiv preprint*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers. In *Findings of the Association for Computational Linguistics*, pages 4005–4019. Association for Computational Linguistics.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting Hate Speech in Multi-modal Memes. *arXiv preprint*.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1014–1023.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515, Montreal, Canada.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Tanvi Deshpande and Nitya Mani. 2021. An Interpretable Approach to Hateful Meme Detection. In *ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 723–727.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Manas Gaur, Keyur Faldu, and Amit Sheth. 2021. Semantics of the Black-Box: Can Knowledge Graphs Help Make Deep Learning Systems More Interpretable and Explainable? *IEEE Internet Computing*, 25(1):51–59.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Snowmass Village, CO, USA. IEEE.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. In *The 35th AAAI Conference on Artificial Intelligence*. AAAI Press.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On Explaining Multimodal Hateful Meme Detection Models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655, Virtual Event, Lyon France. ACM.

Andreas Holzinger. 2021. Explainable AI and Multi-Modal Causability in Medicine. *i-com*, 19(3):171–179.

Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion*, 71:28–37.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A Multimodal Dataset for Detecting Hateful Memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*. Association for Computational Linguistics.

Kazuki Irie, Robert Csordas, and Schmidhuber Jürgen. 2022. The Dual Form of Neural Networks Revisited: Connecting Test Time Predictions to Training Patterns via Spotlights of Attention. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, Baltimore, Maryland, USA.

Gargi Joshi, Rahee Walambe, and Ketan Kotecha. 2021. A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access*, 9:59800–59821.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and

Large Language Models: Opening a New Frontier for Causality. *arXiv preprint*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Thirty-Fourth Annual Conference on Neural Information Processing Systems*, Red Hook, NY, USA.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The Hateful Memes Challenge: Competition Report. *Proceedings of Machine Learning Research*.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking Intersectional Biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *The 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *16th European Conference On Computer Vision*.

Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–24.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A Multimodal Framework for the Detection of Hateful Memes. *arXiv preprint*.

Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. 2022. Causal Reasoning Meets Visual Representation Learning: A Prospective Study. *Machine Intelligence Research*, 19(6):485–511.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: Multimodal Learning with Severely Missing Modality. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*.

R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9:100169.

Diego Marcos, Sylvain Lobry, and Devis Tuia. 2019. Semantically Interpretable Activation Maps: What-where-how explanations within CNNs. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4207–4215.

Niklas Muennighoff. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arXiv preprint*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, Nashville, TN, USA. IEEE.

Judea Pearl. 2001. CAUSALITY: MODELS, REASONING, AND INFERENCE. *Cambridge University Press*, page 11.

Ruifeng Ren and Yong Liu. 2023. In-context Learning with Transformer Is Really Equivalent to a Contrastive Learning Pattern.

Donald B. Rubin. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3).

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. 2019. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. *AI for Social Good Workshop at NeurIPS 2019 (short paper)*.

Pedro Sanchez, Jeremy P. Voisey, Tian Xia, Hannah I. Watson, Alison Q. O'Neil, and Sotirios A.

Tsaftaris. 2022. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice. IEEE.

Kinshuk Sengupta and Praveen Ranjan Srivastava. 2021. Causal effect of racial bias in data and machine learning algorithms towards user persuasiveness &amp; discriminatory decision making: An Empirical Study. *arXiv preprint*.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the Victims Targeted by Harmful Memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Munacinga Simatele and Martin Kabange. 2022. Financial Inclusion and Intersectionality: A Case of Business Funding in the South African Informal Sector. *Journal of Risk and Financial Management*, 15(9):380.

Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, Seoul Republic of Korea. ACM.

Varun Sundaram, R.Srinivas Pavan, Shashank Kandaala, and R. Ravinder Reddy. 2022. Distinguishing Hate Speech from Sarcasm. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1–5, Goa, India. IEEE.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328. PMLR.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Thirty-First Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. 2022. A Review of Causality for Learning Algorithms in Medical Image Analysis. *Machine Learning for Biomedical Imaging*, 1(November 2022 issue):1–17.

Johannes von Oswald, Eyvind Niklasson, Randazzo, Ettore, Sacramento, Jo\~{a}o, Mordvintsev, Alexander, Zhmoginov, Andrey, and Vladymyrov, Max. 2023. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 1464, page 24, Honolulu, HI, USA. JMLR.org.

Matej Vuković and Stefan Thalmann. 2022. Causal Discovery in Manufacturing: A Structured Literature Review. *Journal of Manufacturing and Materials Processing*, 6(1):10.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Multi-Modal Learning with Missing Modality via Shared-Specific Feature Modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15878–15887, Vancouver, BC, Canada. IEEE.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal intersectionality for fair ranking. In *2nd Symposium on Foundations of Responsible Computing*. FORC.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.

A. Appendix

A.1. Further Exploration for Local Explainability

Sample analysis for Oscar (Fig. 6) shows a similar trend to UNITER (Fig. 5). Interestingly, VisualBERT does not attend to the key components (woman or cargo) in the image, supporting its bias towards textual information.

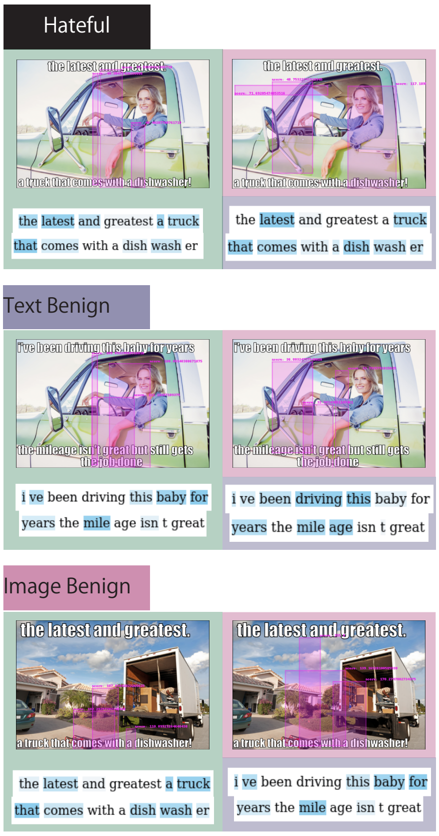
Hateful
the latest and greatest
a truck that comes with a dishwasher!
the latest and greatest a truck
that comes with a dish wash er
Text Benign
i've been driving this baby for years
the mileage isn't great but still gets the job done
i ve been driving this baby for
years the mile age isn t great
Image Benign
the latest and greatest.
a truck that comes with a dishwasher!
the latest and greatest a truck
that comes with a dish wash er
i ve been driving this baby for
years the mile age isn t great

Figure 6: Sampled derived from Oscar.

A.2. Breakdown of Attention Attribution Score

The attention attribution score $attr$ (Eq. 3) is the product of the attention weight matrix and the integral of the gradient. To see the separate impact, we replaced the $attr$ term of the $MIDAS$ equation (Eq. 9) with the attention $MIDAS_{att}$ or the gradient $MIDAS_{grad}$ for comparison. In general, $MIDAS_{att}$ (Fig. 8-10) shows a more similar trend to the original $MIDAS$ than $MIDAS_{grad}$ (Fig. 11-13). This result implies that the attention weights decide the model's strategy, while the gradient adjusts the impact of the individual component.

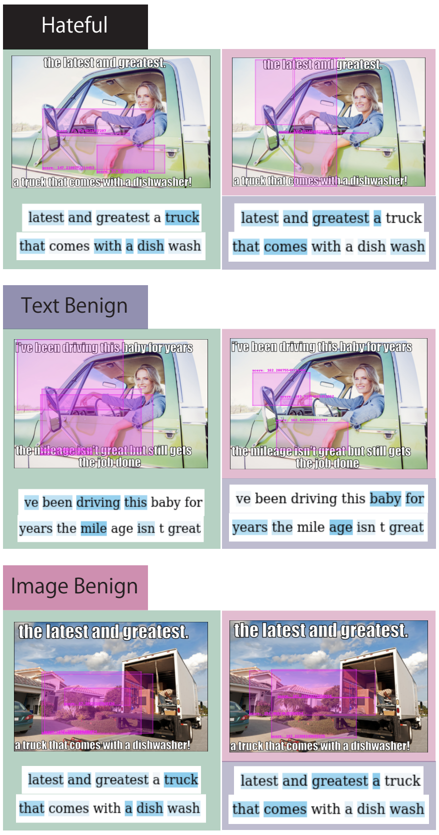
Hateful
the latest and greatest
a truck that comes with a dishwasher!
latest and greatest a truck
that comes with a dish wash
Text Benign
i've been driving this baby for years
the mileage isn't great but still gets the job done
ve been driving this baby for
years the mile age isn t great
Image Benign
the latest and greatest.
a truck that comes with a dishwasher!
latest and greatest a truck
that comes with a dish wash

Figure 7: Samples derived from VisualBERT.

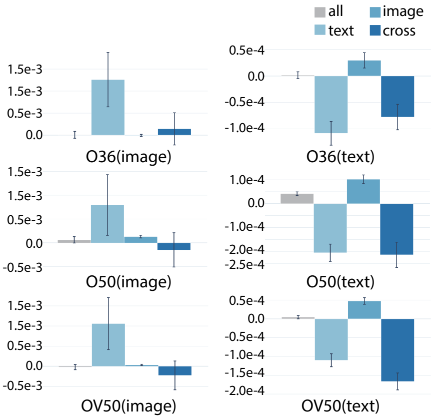

Figure 8: Oscar $MIDAS_{att}$.

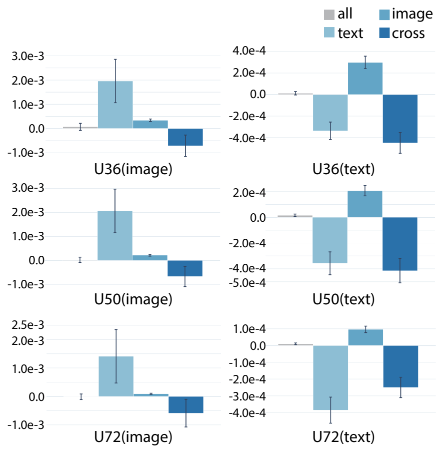

Figure 9: UNITER $MIDAS_{att}$.

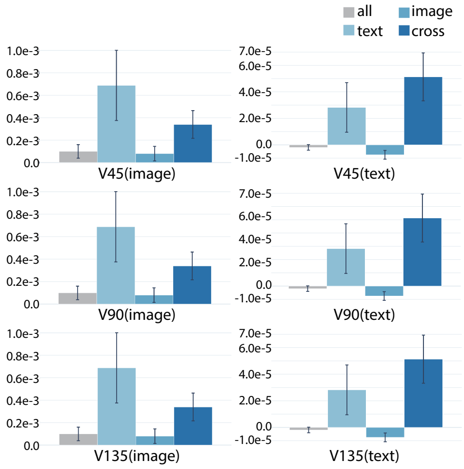

Figure 10: VisusalBERT $MIDAS_{att}$.

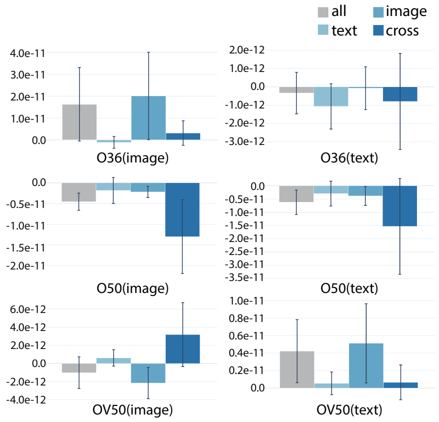

Figure 11: Oscar $MIDAS_{grad}$.

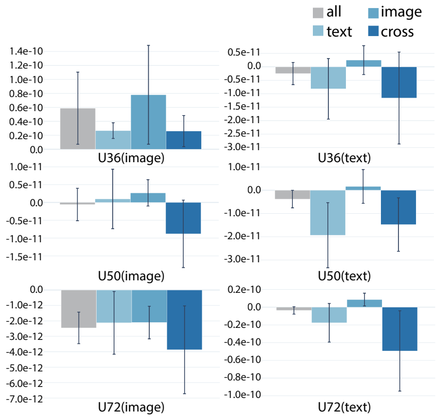

Figure 12: UNITER $MIDAS_{grad}$.

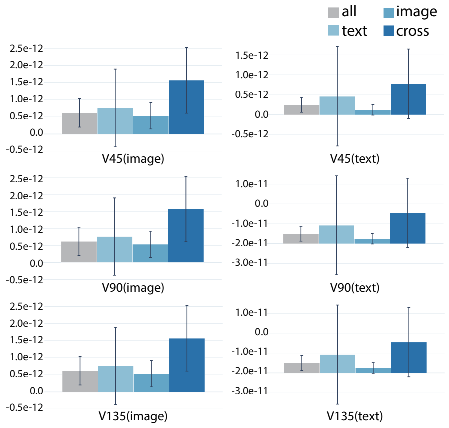

Figure 13: VisualBERT $MIDAS_{grad}$.