CAuSE: Decoding Multimodal Classifiers using Faithful Natural Language Explanation

Dibyanayan Bandyopadhyay[1] **Soham Bhattacharjee**[1]
Mohammed Hasanuzzaman[2] **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna
{dibyanayan, sohambhattacharjeenghss}@gmail.com, asif@iitp.ac.in
[2]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast
m.hasanuzzaman@qub.ac.uk

Abstract

Multimodal classifiers function as opaque black box models. While several techniques exist to interpret their predictions, very few of them are as intuitive and accessible as natural language explanations (NLEs). To build trust, such explanations must faithfully capture the classifier's internal decision making behavior, a property known as *faithfulness*. In this paper, we propose *CAuSE* (Causal Abstraction under Simulated Explanations), a novel framework to generate faithful NLEs for any pretrained multimodal classifier. We demonstrate that *CAuSE* generalizes across datasets and models through extensive empirical evaluation. Theoretically, we show that *CAuSE*, trained via interchange intervention, forms a causal abstraction of the underlying classifier. We further validate this through a redesigned metric for measuring causal faithfulness in multimodal settings. *CAuSE* surpasses other methods on this metric, with qualitative analysis reinforcing its advantages. We also perform detailed error analysis to pinpoint the failure cases of *CAuSE*[1].

1 Introduction

Multimodal classifiers (e.g. VisualBERT (Li et al., 2020)) integrate information from multiple modalities, such as images, text, and audio, and classify input into a predefined set of classes. These models are vital in a range of applications. For instance, given radiology reports in free-text format and corresponding chest X-ray images, a multimodal classifier can be trained to predict whether a patient has COVID-19 (Baltrušaitis et al., 2019).

However, their widespread adoption depends on whether we can trust their predictions. This requires the development of interpretability techniques that explain *how* the classifier came to its prediction. Input attribution methods aim to identify features or concepts in the input that influence the classifier's decision. Although these techniques provide useful insights, they face two significant limitations: (i) *Lack of natural language explanations (NLEs)*: Their outputs are typically low-level and not conveyed in natural language, which hampers interpretability (Sundararajan et al., 2017); and (ii) *Causal faithfulness*: They often do not capture a true causal link between the input and the prediction of the model (Bandyopadhyay et al., 2024b; Chattopadhyay et al., 2019). Crucially, faithfulness is essential to trust the predictions of a model.

To address these limitations, we propose **CAuSE (Causal Abstraction under Simulated Explanations)**, a novel post-hoc framework for generating *causally faithful* NLEs for arbitrary *frozen* multimodal classifier. CAuSE specifically targets *discriminative* classifiers with an encoder-classification head architecture, which are found to have been deployed in today's production systems (Megahed et al., 2025; Ji et al., 2025). Unlike generative models, these discriminative classifiers lack native explanation generation capabilities, necessitating post-hoc interpretability frameworks like CAuSE. Figure 1 shows an example of CAuSE explaining a *frozen* multimodal offensiveness classifier decision for a meme input.

At the core of CAuSE is a pretrained language model ϕ, which is guided by the hidden states of the classifier to generate a natural language explanation. CAuSE uses a novel loss function, grounded in interchange intervention training of Geiger et al. (2021), that enforces causal faithfulness of the generated explanations (see Section 5 for results and analysis). We also propose a variant of a widely used causal faithfulness metric (Atanasova et al., 2023), termed **CCMR**

[1]Code is available at https://github.com/newcodevelop/CAuSE

Transactions of the Association for Computational Linguistics, vol. 14, pp. 829–851, 2026. https://doi.org/10.1162/TACL.a.686
Action Editor: Roi Reichart. Submission batch: 7/2025; Revision batch: 10/2025; Published 6/2026.

(Counterfactual Consistency via Multimodal Representation), tailored for evaluating faithfulness of NLEs in multimodal contexts. Under this metric, CAuSE demonstrates strong performance on benchmark datasets such as i) *e-SNLI-VE* (Do et al., 2021), which is a dataset of image premises and text hypotheses labeled with entailment, contradiction, or neutral labels; ii) *Hateful Memes (HM)* (Kiela et al., 2020) in which the task is to predict whether an input meme is offensive or not; and iii) *VQA-X* (Park et al., 2018), which is a visual question answering dataset coupled with gold explanations. All these datasets are publicly available for research.

We conduct extensive qualitative analyses to examine: i) where and how CAuSE succeeds in producing causally faithful NLEs (§5.5.1 and §5.5.2), ii) typical failure cases (§5.5.3), and iii) general trends observed in error analysis (§5.6).

Our contributions are *three-fold*: (1) a framework for generating faithful, post-hoc NLEs for multimodal classifiers; (2) a novel loss function that enforces causal faithfulness; and (3) an extensive empirical evaluation demonstrating the effectiveness of CAuSE in producing causally faithful NLEs.

1.1 Scope and Applicability

CAuSE serves dual purposes as both a deployable tool and a methodological framework with broader implications.

Immediate application. As explained in the Introduction, CAuSE targets discriminative multimodal classifiers, which, unlike generative models, lack native explanation generation capabilities. Simulating these classifiers with separate generative models often produces plausible but unfaithful explanations that fail to reflect actual model reasoning (Madsen et al., 2024; Turpin et al., 2023). CAuSE addresses this through faithful explanation generation, demonstrated across various classifiers studied later (§Section 5).

Broader implications. Beyond practical deployment, CAuSE demonstrates how causal abstraction principles that are typically validated on simpler networks can be scaled to complex transformer-based architectures with an application to faithful explanation generation that is empirically shown to be both task-agnostic and largely architecture-agnostic. This positions CAuSE as a blueprint for future explainability research requiring causal faithfulness.

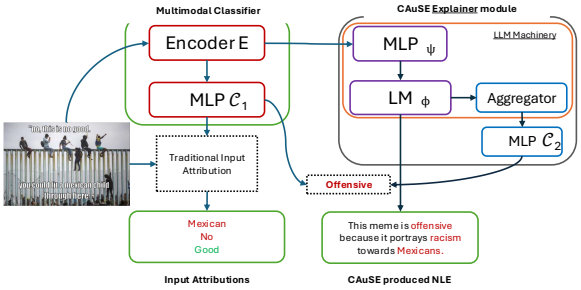

Figure 1: This figure shows the abstract schematic of CAuSE and how it explains a frozen *discriminative* multimodal classifier at inference. The internal components of CAuSE are i) an MLP ψ, ii) A language model (LM) ϕ, and iii) an aggregator followed by a classifier $\mathcal{C}_2$. *The input to the multimodal encoder is a meme which is composed of both text and image (separately not shown).*

2 Causal Abstraction and Interchange Intervention

2.1 Causal Abstraction

In Geiger et al. (2021), the authors introduced the concept of causal abstraction for neural models. They define a neural network, N_1, as a causal abstraction of a higher-level causal model, N_2, if the neural representations of N_1 exhibit the same causal outcomes as the corresponding high-level variables in N_2. This alignment is achieved through the Interchange Intervention Training (IIT) objective.

This type of alignment learned through IIT is referred to as *causal abstraction* in the literature. IIT ensures a systematic correspondence between interventions in neurons in N_1 and mapped variables in N_2. Unlike a traditional teacher-student training objective, which merely teaches the student to mimic the teacher's output, causal abstraction ensures that the student model internally mirrors the teacher's decision-making process. In particular, IIT guarantees that interventions on variables in N_2 yield analogous effects when applied to the associated neurons in N_1, thereby aligning the causal structure of the two models.

2.2 Interchange Intervention

To perform an interchange intervention, we begin by passing a "base" input b through the network N_1, randomly selecting a neuron $i_1 \in N_1$ and

freezing its activation at the value $i_1(b)$. Simultaneously, a second input, referred to as the "source" input s, is also passed through N_1, but with the activation of neuron i_1 held fixed at $i_1(b)$ instead of its native activation under s. The resulting output from this modified forward pass is denoted by $y_{i_1}^{INT}(N_1)$.

Assuming a one-to-one correspondence between each neuron i_1 in N_1 and a subset of variables $S(i_1)$ in a higher-level model N_2, we say that N_2 is a causal abstraction of N_1 if the following condition holds:

$$S(i_1) = \{i_2 \in N_2 : [y_{i_2}^{INT}(N_2) = y_{i_1}^{INT}(N_1)]\}$$

In other words, after intervention in i_1 and i_2, output of N_1 must match that of N_2, under the same interchange intervention.

Let $P_{y_{i_1}^{INT}(N_1)}$ and $P_{y_{i_2}^{INT}(N_2)}$ denote the probability distributions over the model outputs under interchange intervention for N_1 and N_2, respectively. Following Geiger et al. (2021), causal abstraction is encouraged by minimizing the loss of IIT, defined as the Kullback–Leibler (KL) divergence between these two distributions.

$$\mathcal{L}_{IIT} = D_{KL}(P_{y_{i_1}^{INT}(N_1)} \| P_{y_{i_2}^{INT}(N_2)}) \quad (1)$$

where $D_{KL}(P\|Q)$ denotes the KL divergence between distributions P and Q.

3 Methodology

3.1 Setup and Objective

The goal of CAuSE is to provide post-hoc natural language explanations for the predictions of a pretrained multimodal classifier M. Any arbitrary discriminative multimodal classifier M is composed of an encoder E and a multilayered perceptron (MLP) $\mathcal{C}_1$. *Once trained, M is kept frozen; our method does not alter or retrain it.*

Given an input pair $z = (t, v)$ of text and image, E computes a joint representation $c = E(t, v) \in \mathbb{R}^{m\times 1}$, which is then passed to $\mathcal{C}_1$ to obtain a logit vector $\hat{l} \in \mathbb{R}^L$ over L classes. The output prediction is $y_1 = \text{softmax}(\hat{l})$.

3.2 Explainer Module (CAuSE Framework)

CAuSE is a framework consisting of an explanation generation module, referred to as the **Explainer**, along with a set of loss functions used to train it. The Explainer is composed of four key components: (i) a multi-layer perceptron (MLP) ψ; (ii) a language model ϕ responsible for generating natural language explanations (NLEs); (iii) an aggregator module $\mathcal{A}$, which processes and transforms the token-level outputs of ϕ into a fixed-length feature representation; and (iv) a classifier $\mathcal{C}_2$, which is structurally identical to the original classifier $\mathcal{C}_1$ and is trained to replicate its predicted label y_1.

Training Objective for ϕ To generate faithful explanations grounded in the classifier M's internal reasoning, we condition the language model ϕ on its hidden representation $c = E(z)$. Since c may not align with ϕ's input embedding space, we project it using a simple MLP ψ, and use $\psi(c)$ as the embedding for the initial **BOS** (x_0) token. The model ϕ is then fine-tuned using the standard causal language modeling objective:

$$\mathcal{L}_\phi = -\sum_{i=1}^{T} \log P_\phi(x_i | x_{i-1}, \ldots, x_0) \quad (2)$$

We use GPT-2 (small) (Radford et al., 2019) as ϕ. We fine-tune it (rather than training from scratch) on plausible human-annotated explanations from e-SNLI-VE, Hateful Memes, and VQA-X datasets, as oracle explanations for the classifier's decisions are unavailable. Directly conditioning on raw input z leads to unfaithful outputs, as pretrained LMs cannot inherently incorporate internal model states of M. As shown in Table 5, naïve fine-tuning on pairs $(z, \text{explanation})$ results in poor faithfulness. Our approach overcomes this by explicitly grounding generation in $c = E(z)$.

We denote the full explanation generation pipeline ($F = \mathcal{A} \circ \phi \circ \psi$), as the *LLM machinery*. Thus, *Explainer* is composed of both *LLM Machinery* and $\mathcal{C}_2$ ($\mathcal{C}_2 \circ F$).

Aggregator $\mathcal{A}$ The logit output of ϕ is a tensor of shape $(1, T, V)$, where V is the vocabulary size and T is the sequence length. We first sum over the time axis to obtain a vector $x \in \mathbb{R}^V$. This is then transformed via a feed-forward network into a vector of size $\mathbb{R}^{m\times 1}$ to match the input dimension expected by $\mathcal{C}_2$.

Classifier $\mathcal{C}_2$ $\mathcal{C}_2$ is a replica of $\mathcal{C}_1$, receiving the previously aggregated explanation representation from $\mathcal{A}$ as input. It is trained to (i) match $\mathcal{C}_1$'s prediction y_1 via a teacher-student loss, and (ii) ensure causal alignment through the IIT loss (see Section 3.3).

During training, the entire Explainer is optimized, with each component serving a distinct purpose (detailed in the next subsection). At inference, only the fine-tuned ϕ (along with ψ) is used.

3.3 Causal Alignment and Behavioral Simulation

The central goal of CAuSE is to ensure that the explanations produced by the Explainer are faithful to the behavior of the pretrained multimodal classifier $M = \mathcal{C}_1 \circ E$. This is accomplished in two steps: first, we make the Explainer simulate the decision of M; second, we ensure that the Explainer serves as a causal abstraction of M. Below, we define the core concepts and describe how CAuSE achieves these goals.

Simulation of Classifier Behavior We first formalize the notion of simulation.

Definition (Simulation). Two neural networks A and B are said to simulate each other if, when given the same input z, they produce the same output: $A(z) \approx B(z)$.

In our case, the encoder E produces a multimodal representation $c = E(z)$ from an input $z = (t, v)$. This representation is passed directly to $\mathcal{C}_1$ to obtain a prediction $y_1 = \mathcal{C}_1(c)$. The same representation c is also passed to the Explainer module. To ensure that the Explainer mimics $\mathcal{C}_1$ on the same input, we train it using a teacher-student loss:

$$\mathcal{L}_{TS} = D_{KL}(P_{\mathcal{C}_1}(y_1) \| P_{\mathcal{C}_2}(y_2)) \tag{3}$$

By minimizing $\mathcal{L}_{TS}$, we ensure that $y_2 = \mathcal{C}_2(F(c)) \approx \mathcal{C}_1(c) = y_1$. Since both the systems receive the same input $c = E(z)$, and produce the same output, the Explainer (i.e., $\mathcal{C}_2 \circ F$) simulates the classifier $\mathcal{C}_1$ under the definition above.

Causal Abstraction via Identical Networks While simulation ensures output-level agreement, it does not guarantee that the Explainer behaves like M under counterfactual interventions. For this, we seek a stronger condition: that the Explainer forms a *causal abstraction* of M. To establish this, we first introduce the notion of causally identical neural networks.

Definition (Causally Identical Networks). Given two neural networks A and B of the same architecture, we say they are identical if: i) they share the same weights: $\mathcal{W}_A = \mathcal{W}_B$, and ii) they are trained under the IIT objective that enforces alignment of their outputs under counterfactual substitutions (Equation 1). Under the definition posed in Section 2.2, they become causal abstraction of each other.

To make $\mathcal{C}_2$ identical to $\mathcal{C}_1$ using the above definition, we use two objectives: i) a weight-regularization loss to encourage weight similarity (using their frobenius norm):

$$\mathcal{R}_{\text{match}} = \|\mathcal{W}_{\mathcal{C}_1} - \mathcal{W}_{\mathcal{C}_2}\|_F \tag{4}$$

and ii) An IIT loss that encourages counterfactual agreement:

$$\mathcal{L}_{IIT} = D_{KL}(P^{\text{INT}}_{\mathcal{C}_1}(y_1) \| P^{\text{INT}}_{\mathcal{C}_2}(y_2)) \tag{5}$$

When both $\mathcal{R}_{\text{match}}$ and $\mathcal{L}_{IIT}$ are minimized, $\mathcal{C}_2$ becomes *causally identical* to $\mathcal{C}_1$, by virtue of the above definition.

From Classifier Alignment to Explainer Abstraction. We now connect the alignment between $\mathcal{C}_1$ and $\mathcal{C}_2$ to the abstraction property of the Explainer.

Theorem. If $\mathcal{C}_2$ is identical to $\mathcal{C}_1$ (as per the definition above), then the Explainer, comprising the LLM machinery and $\mathcal{C}_2$, becomes a causal abstraction of the pre-trained classifier $M = \mathcal{C}_1 \circ E$.

Proof Sketch. In Lemma 1, we prove that under $\mathcal{L}_{IIT}$, $F(E(z)) = E(z)$.

We consider three types of causal interventions: (i) **Input-level interventions**: Between E and F, we assume a mapping $\alpha(z) = E(z)$. Given that $\mathcal{C}_2(F(E(z))) = \mathcal{C}_1(E(z))$ (by Lemma 1), the outputs from the model M and the explainer remain consistent under such interventions. (ii) **Intermediate representation interventions**: Since F acts as the identity function on $E(z)$ (by Lemma 1), interventions on $E(z)$ and $F(E(z))$ lead to equivalent outputs. (iii) **Interventions at the neuron level**: When interventions are applied within $\mathcal{C}_1$ and $\mathcal{C}_2$, the outputs remain aligned due to the shared architecture, training based on the IIT objective and the weight regularization objective. Taken together, these cases demonstrate that the explainer and the model M are causally equivalent under aligned interventions.

Proof. See Appendix §A for specific proof assumptions and formal details. □

This establishes that not only does the Explainer match the classifier's outputs on observed exam-

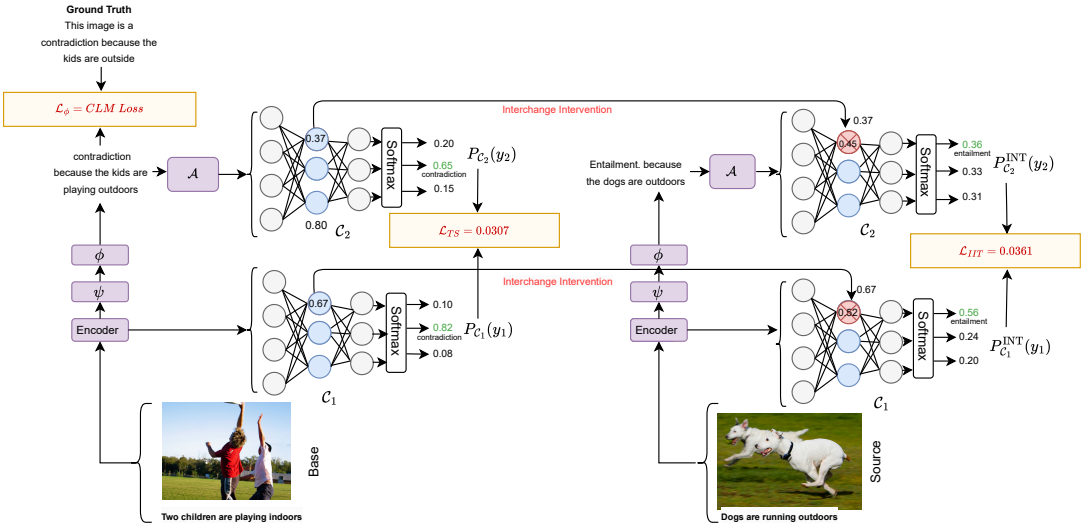

Figure 2: This *toy* diagram shows the training process of the CAuSE framework, specifically the calculations of $\mathcal{L}_{TS}$ and $\mathcal{L}_{IIT}$ losses on a sample from e-SNLI-VE dataset. The input vectors to $\mathcal{C}_1$ and $\mathcal{C}_2$ are arbitrary but depicts a real scenario.

ples (via simulation posed by $\mathcal{L}_{TS}$), but also that it faithfully reflects the causal structure of the M.

Summary. CAuSE theoretically ensures both behavioral and causal alignment between the Explainer and the classifier M: i) $\mathcal{L}_{TS}$ ensures the Explainer simulates M's decision by matching predictions on the same encoder representation. ii) $\mathcal{L}_{IIT}$ and $\mathcal{R}_{\text{match}}$ make $\mathcal{C}_2$ causally identical to $\mathcal{C}_1$. This makes the overall Explainer a causal abstraction of the multimodal classifier $M = \mathcal{C}_1 \circ E$ (because of the result of the theorem posed above).

Overall Loss Function The final loss is:

$$\mathcal{L}_{\text{CAuSE}} = \mathcal{L}_\phi + \mathcal{L}_{TS} + \mathcal{L}_{IIT} + \mathcal{R}_{\text{match}} \quad (6)$$

Through a hypothetical example, the calculation of several components of the loss function is depicted in Figure 2. A walkthrough of the Figure is described in detail in Section 3.4 for reference.

Overall Training and Inference Protocol. i) We assume the base classifier $M = (\mathcal{C}_1 \circ E)$ is already trained until convergence. Then we freeze its weights. *Training the classifier M is not a part of CAuSE training framework.* ii) We separately train the Explainer $(\psi, \phi, \mathcal{A}, \mathcal{C}_2)$ using $\mathcal{L}_{\text{CAuSE}}$, keeping M frozen. iii) At inference time, we generate NLE from ϕ for an input z.

Clarification on Explainer Role. It is important to clarify that while the Explainer is trained to simulate the predictions of $\mathcal{C}_1$, it is designed as a *causal abstraction* of the entire classifier M. In our framework, we do not assume any inherent dependence between *A being a causal abstraction of B* and *A simulating B*. Causal abstraction and simulation are treated as independent properties.

Nonetheless, we observe that combining simulation with causal abstraction yields better empirical results than simulation alone. Specifically, the mimicking faithfulness performance of the Explainer, quantified using both F1 and CCMR, improves in majority of the cases when the causal abstraction losses $\mathcal{L}_{\text{IIT}}$ and $\mathcal{R}_{\text{match}}$ are added to the simulation loss $\mathcal{L}_{\text{TS}}$ (also referred to as $\mathcal{L}_{\text{CAuSE}}$), as opposed to using $\mathcal{L}_{\text{TS}}$ alone (§Table 2).

Broad Neural Coverage under $\mathcal{L}_{IIT}$. At each step of minimizing $\mathcal{L}_{IIT}$, we sample 20% of neurons uniformly randomly from $\mathcal{C}_1$ and $\mathcal{C}_2$ simultaneously and perform interchange intervention on them. Sampling a subset of neurons rather than performing interchange intervention on all of them is the norm in literature (Geiger et al., 2021) as that would prevent the network from being degenerate-where the network always outputs values for base input. Sampling a fraction of neurons at each backpropagation step probabilistically ensures broad coverage of network's internal representation. With a very high probability, after a finite number of backpropagation steps, all neurons in $\mathcal{C}_1$ and $\mathcal{C}_2$ would be sampled at least once

and interchange intervention would be done on all of them. Details of such a bound can be found in the Appendix Section C.

3.4 Walk-through Example

Figure 2 illustrates the training process of CAuSE on an e-SNLI-VE sample, involving three loss components: $\mathcal{L}_\phi$, $\mathcal{L}_{TS}$, and $\mathcal{L}_{IIT}$. Note that the following values are illustrative and derived from a simplified toy version of the CAuSE components, but the procedure remains identical in the full framework.

Step 1: Causal Language Modeling Loss ($\mathcal{L}_\phi$) The base input z is passed through the frozen encoder of M to obtain the multimodal representation $E(z)$. This is then projected via ψ and provided to the language model ϕ, which generates a natural language explanation. The explanation generation step is supervised using a standard causal language modeling loss as shown in Equation 2.

Step 2: Teacher–Student Loss ($\mathcal{L}_{TS}$) The explanation from ϕ is passed through the aggregator $\mathcal{A}$ and input to classifier $\mathcal{C}_2$. Let the output distributions from $\mathcal{C}_1$ and $\mathcal{C}_2$ be:

$$P_{\mathcal{C}_1}(y_1) = [0.10,\ 0.82,\ 0.08]$$

$$P_{\mathcal{C}_2}(y_2) = [0.20,\ 0.65,\ 0.15]$$

Using KL divergence, the teacher–student loss is:

$$\mathcal{L}_{TS} = D_{\mathrm{KL}}(P_{\mathcal{C}_1}(y_1) \,\|\, P_{\mathcal{C}_2}(y_2)) = 0.0307$$

Step 3: Interchange Intervention Loss ($\mathcal{L}_{IIT}$) The source input is encoded similarly. During this pass, neuron activations from the base pass (e.g., 0.67 from $\mathcal{C}_1$ and 0.37 from $\mathcal{C}_2$) are used to interchange (overwrite) corresponding activations for the source input. This is called interchange intervention. The resulting predictions after intervention are:

$$P^{\mathrm{INT}}_{\mathcal{C}_1}(y_1) = [0.56,\ 0.24,\ 0.20]$$

$$P^{\mathrm{INT}}_{\mathcal{C}_2}(y_2) = [0.36,\ 0.33,\ 0.31]$$

The IIT loss is:

$$\mathcal{L}_{IIT} = D_{\mathrm{KL}}(P^{\mathrm{INT}}_{\mathcal{C}_1}(y_1) \,\|\, P^{\mathrm{INT}}_{\mathcal{C}_2}(y_2)) = 0.0361$$

Training and Inference Only the explainer modules (ψ, ϕ, A, $\mathcal{C}_2$) are updated. The base classifier $M = \text{Encoder} + \mathcal{C}_1$ remains frozen. During inference, we use only ψ and ϕ to generate faithful explanations conditioned on $E(z)$.

3.5 Generalizability and Retraining in CAuSE

CAuSE is dataset-agnostic and requires no significant architectural changes across tasks. Its core components, ϕ, Aggregator $\mathcal{A}$, and $\mathcal{C}_2$; are both *necessary* and *sufficient*: necessary, because each component is essential to ensure causal abstraction between the classifier M and the Explainer; and sufficient, because no structural modifications are needed when switching to a new dataset or classifier.

Minor adjustments include: (i) setting the output dimension of $\mathcal{C}_2$ to match the number of labels (e.g. 2 for Hateful Memes; 3 for e-SNLI-VE; 3,129 for VQA-X), and (ii) adapting ψ to the hidden size of $c = E(x)$. Retraining is required for each (dataset D, classifier M) pair, as CAuSE is designed to explain a specific frozen model M trained on D.

This retraining is lightweight, only $\sim$450M parameters are trained and inference uses only $\sim$270M parameters, much smaller than large VLMs like PaLiGemma (Beyer et al., 2024) or LLaVA (Liu et al., 2024), which still underperform CAuSE (Table 5).

Overall, CAuSE is broadly applicable as long as hidden states $E(z)$ from M and ground-truth explanations for D are available. Adaptation to different tasks requires minimal effort with little architectural changes.

4 Counterfactual Consistency via Multimodal Representation (CCMR)

Existing faithfulness metrics, such as CCT (Siegel et al., 2024) and Explanation Mention (Atanasova et al., 2023) are designed for unimodal text models and rely on discrete text perturbations to test whether changes in input lead to corresponding changes in explanations. However, extending this to multimodal settings is nontrivial: there is no clear notion of discrete perturbation for image inputs, and fused text-image representations make it hard to isolate the effect of individual modalities. A text change may not be reflected in the explanation even if the model's decision process is faithful, due to how multimodal fusion influences model behaviour. To address this, we propose CCMR, a metric based on continuous perturbations in multimodal representation space, enabling more reliable faithfulness evaluation.

4.1 CCMR Score Calculation

Counterfactual Faithfulness (Atanasova et al., 2023) : Consider an input text $Z = \{w_0, w_1, \ldots, w_i, \ldots, w_{n-1}, w_n\}$, where w_i is the i-th word or token. A counterfactual input Z^C is created by replacing w_i with w_i^c:

$$Z^C = \{w_0, w_1, \ldots, w_i^c, \ldots, w_{n-1}, w_n\}.$$

For a model, which generates natural language explanations (NLEs), let X and X^C be the NLEs for Z and Z^C, respectively. The NLE X is considered faithful if $w_i^c \in X^C$, meaning the explanation reflects the discrete input change.

This test has two limitations for CAuSE: i) it does not support continuous input changes, and ii) X^C may not be a proper counterfactual, as it might not change the model's (M) prediction class. We address these issues by constructing counterfactual representations as follows:

Let $z \in \mathcal{T}$ be a vector representation of a data point from the test set. The corresponding counterfactual input z' for M satisfies:

$$z' = \arg\min_{z' \in \mathcal{T}} d(z, z') \text{ s.t. } \mathcal{C}_1(E(z)) \neq \mathcal{C}_1(E(z'))$$

where d is Euclidean distance metric, and $\mathcal{C}_1(z)$ denotes the output class of M given input z. The counterfactual z' can be expressed as: $z' = z + \mu$, where $\mu = z' - z$ is the perturbation. Given $z' = z + \mu$ as input to M, let us assume its output class is y_1^{CF} .

Inspired by Atanasova et al. (2023), we define an explanation from ϕ as faithful if it reflects the change in M's output under a counterfactual input. Let $E(z)$ be the input to the Explainer F when M receives input z. We construct a potential counterfactual for the Explainer as $E(z) + \nu$, and ensure it is valid for M by solving an optimization on ν such that $\mathcal{C}_1(E(z) + \nu) = \mathcal{C}_1(z + \mu)$.

Since $\mathcal{C}_2 \circ F$ is a causal abstraction of M, a counterfactual for M should induce a corresponding counterfactual in the explainer. We input $E(z) + \nu$ to F and obtain the predicted class from ϕ as y_2^{CF} (in inference, both prediction and explanations are obtained from ϕ and not from $\mathcal{C}_2$). The Counterfactual Consistency via Multimodal Representation (CCMR) score is computed as the F1 score between y_2^{CF} and y_1^{CF}, across test examples.

Importantly, we do not feed z', a known counterfactual for M from the test set, to the explainer. Doing so would let even a naïve mimic of M achieve high CCMR. Instead, we optimize for ν to construct a non-test representation of a counterfactual, ensuring the CCMR score reflects true causal consistency rather than mere imitation.

Algorithm 1 shows the calculation of the CCMR score and Table 2 shows the "faithfulness" performance of components of the the proposed CAuSE framework.

Algorithm 1: CCMR Score for the pre-trained classifier M and the explainer

Input: Data-point $\mathbf{z} \in \mathcal{T}$
Function `CounterFactual(z)`:
  $\mathbf{z}' \leftarrow \arg\min_{\mathbf{z}' \in \mathcal{T}} d(\mathbf{z}, \mathbf{z}')$ s.t. $\mathcal{C}_1(E(\mathbf{z})) \neq \mathcal{C}_1(E(\mathbf{z}'))$;
  $\mu \leftarrow \mathbf{z}' - \mathbf{z}$; `// Compute the perturbation`
  Optimize ν **such that** $\mathcal{C}_1(E(\mathbf{z}) + \nu) = \mathcal{C}_1(E(\mathbf{z} + \mu))$
  return $E(\mathbf{z}) + \nu$, $\mathbf{z}'$
Procedure *Calculate CCMR Score*:
  ZList $\leftarrow \emptyset$;
  XList $\leftarrow \emptyset$;
  while $\mathcal{T} \neq \emptyset$ **do**
    Sample $\mathbf{p} \in \mathcal{T}$; `// Draw a new data point`
    $\mathbf{q}', \mathbf{p}' \leftarrow$ `CounterFactual`$(\mathbf{p})$;
    ZList $\leftarrow$ ZList $\cup \{\phi(F(\mathbf{q}'))\}$; `// Append` y_2^{CF} `to the list`
    XList $\leftarrow$ XList $\cup \{\mathcal{C}_1(E(\mathbf{p}'))\}$; `// Append` y_1^{CF} `to the list`
    $\mathcal{T} \leftarrow \mathcal{T} - \{\mathbf{p}\}$;
  return F_1(XList, ZList) ; `// Return F1 score between XList and ZList`

Composite CCMR. Varying ν may push $E(z) + \nu$ to become an out-of-distribution (OOD) sample for the Explainer. This is especially true if the Explainer is not a causal abstraction of M. In such cases, ϕ may fail to generate coherent predictions containing any of the output classes. Table 2 reports the percentage of feasible generations and their corresponding CCMR (across faithful generations). The composite CCMR (shown as c-CCMR) is defined as the harmonic mean of these two and reflects the true causal faithfulness of CAuSE and its counterparts.

5 Results and Analysis

5.1 Dataset and Experimentation

Datasets. The e-SNLI-VE and VQA-X datasets provide human-annotated explanations for each image–text pair, which we use as ground-truth instances for training CAuSE. In contrast, the Facebook Hateful Meme (HM) dataset lacks such annotations. To address this, we generate synthetic

explanations using the GPT-4o[2] language model for the offensive class, followed by manual verification and post-processing. These curated explanations serve as ground-truth for training CAuSE. The details of this data creation process are provided in Appendix D.

However, not all instances are used for generating explanations. The availability of training explanations is filtered based on the accuracy of the underlying multimodal classifier M. For example, if M achieves 70% accuracy on a batch of 1,000 samples, only the 700 correctly predicted examples, along with their corresponding explanations, are retained for training CAuSE on ground-truth prediction and explanations. For 300 misclassified examples, CAuSE is trained only to mimic M's prediction. This filtering ensures that CAuSE learns to explain M's actual predictions, not the dataset labels. This filtering procedure is described in Algorithm 2.

Algorithm 2: Filtering Dataset for CAuSE

Input: Test dataset $\mathcal{D} = \{(x_i, y_i, e_i)\}_{i=1}^{N}$, classifier M
Output: Filtered dataset D' used to train CAuSE
Procedure *FilterCorrectPredictions*:
  D' $\leftarrow \emptyset$;
  foreach $(x_i, y_i, e_i) \in \mathcal{D}$ **do**
    $\hat{y}_i \leftarrow M(x_i)$; `// Predict using classifier`
    if $\hat{y}_i = y_i$ **then**
      D' $\leftarrow$ D' $\cup \{(x_i, \hat{y}_i, e_i)\}$; `// Keep correctly predicted samples`
    else
      D' $\leftarrow$ D' $\cup \{(x_i, \hat{y}_i)\}$
  return D' ; `// Return filtered dataset`

Experimentation. The experiments were performed on a Kaggle kernel with PyTorch version 2.1.2 and a single P-100 GPU, with a random seed of 42 maintained for all runs. VQA-X required a single A-100 GPU. Additionally, visual language model (VLMs) baselines were implemented using PEFT[3] and LoRA (Hu et al., 2022).

Dataset	Train Split	Test Split
e-SNLI-VE	9000	1000
Hateful Memes	6997	1000
VQA-X	5997	960

Table 1: Train-test splits for the datasets.

[2] https://chatgpt.com/
[3] https://github.com/huggingface/peft

5.2 Automatic Evaluation

CAuSE is evaluated across two verticals: i) Faithfulness, and ii) Plausibility.

Faithfulness measures the alignment between predictions of the Explainer variants (e.g. CAuSE or ϕ or $\phi + TS$) [4] and M for a specific input. It is quantified by F1 score between CAuSE predictions and predictions from M. Under the counterfactual regime, the normal F1 score is replaced by CCMR score. The process of calculating CCMR score is already illustrated in Section 4.1.

Plausibility measures the semantic similarity of the generated NLEs from the Explainers to the ground-truth (GT) explanations. Note that we do not have access to *oracle* explanations of the classifier decisions. GT explanations work as a proxy for unavailable *oracle* explanations. BLEU (Papineni et al., 2002) and BERTScore (Zhang* et al., 2020) are used to measure plausibility. n-gram BLEU score is shown as BLEU-n.

Faithfulness-plausibility tradeoff. GT explanations may not reflect how the classifier M actually arrives at its decisions (i.e. *oracle* explanations). As a result, an Explainer trained to be a causal abstraction of the M (via $\mathcal{L}_{CAuSE}$) can achieve high predictive alignment with M, as measured by metrics like F1 or CCMR. However, its generated explanations may score lower on BLEU or BERTScore when compared to the human-provided GT explanations, since the classifier's decision-making process might differ from human reasoning.

This leads to a *faithfulness–plausibility tradeoff*, high faithfulness (to the classifier) may come at the cost of plausibility and vice versa. Ideally, if we had access to the classifier's actual reasoning process as an oracle, we could directly measure alignment. In its absence, F1/CCMR captures faithfulness, while BLEU/BERTScore reflects plausibility. This tradeoff is evident from Tables 2 and 3. ϕ achieves higher plausibility (cf. Table 3) at the cost of lower faithfulness (cf. Table 2) compared to CAuSE, perfectly illustrating the tension between these two metrics.

An Illustration of tradeoff. We illustrate the faithfulness-plausibility tradeoff on three datasets through specific examples, taking Qwen-VL (Bai et al., 2023) as M. Note that we replace traditional language modeling (LM) head of Qwen-VL with

[4] CAuSE, ϕ, and ϕ+TS refer to Explainers trained with the loss functions $\mathcal{L}_{CAuSE}$, $\mathcal{L}_{\phi}$, and $\mathcal{L}_{\phi} + \mathcal{L}_{TS}$, respectively.

an MLP layer ($\mathcal{C}_1$). In these examples, ϕ generated explanations are plausible and align with GT explanation, but for predictions, they are unfaithful, as they do not match the output of the underlying base model M. Conversely, for the same examples, CAuSE produces *informative* explanations that are faithful to the base model's predictions, yet less aligned with GT explanations (less plausible). These contrasting behaviours clearly demonstrate the inherent tradeoff between faithfulness and plausibility. Figure 3 illustrates these.

(i) VQA-X: In a VQA-X example, the model (M) answers "yes" to "Is a storm going on?". While ϕ's explanation ("the answer is in the sky above the clouds") is highly plausible due to its similarity to the human explanation, it is unfaithful as it omits M's prediction. In contrast, CAuSE's explanation ("yes because it is black and white") is faithful by explicitly retaining the model's output but less plausible. Importantly, CAuSE is more diagnostic, as it exposes a potential reliance of M on a superficial shortcut (grayscale appearance) rather than semantic cues (e.g., storm clouds).

(ii) e-SNLI-VE: For the hypothesis "The men are fishermen", which M predicts as an entailment, ϕ's explanation suggests a neutral outcome and thus does not reflect the model's label, indicating low faithfulness. Conversely, CAuSE's explanation perfectly mirrors the model's prediction. Although ϕ's explanation is unfaithful, it is more plausible than CAuSE's because it has greater similarity to the human explanation, from which the CAuSE explanation is completely off. This clearly shows the faithfulness-plausibility tradeoff.

(iii) Hateful Meme: The gold label is not offensive, while the classifier M predicts offensive. This mismatch truncates the gold explanation to match the predicted label (refer to Section 5.1, Algorithm 2), inflating the plausibility of ϕ's explanation ("not offensive because not offensive") that is more similar to the gold explanation than CAuSE's explanation. CAuSE instead matches the classifier output ("offensive because it promotes harmful stereotypes"), thus faithful. Despite lower similarity to the gold explanation, CAuSE is more informative as it reveals the purported reasoning behind misclassification of M.

5.3 Baselines

VLM baselines. Post-hoc interpretability techniques typically focus on identifying implicit (e.g., Integrated Gradients (Sundararajan et al., 2017)) or explicit concepts (e.g., Semantify (Bandyopadhyay et al., 2024a)) rather than generating coherent natural language explanations (NLEs). Visual language models (VLMs) can leverage zero-shot/few-shot prompting to generate plausible NLEs yet they are not guaranteed to be faithful (Turpin et al., 2023; Madsen et al., 2024).

CAuSE differs from prompting and fine-tuning VLMs by initializing generation from the multimodal encoder's hidden state, directly grounding NLEs in the classifier's decision process and yielding greater faithfulness, an ability absent in naïve prompting or fine-tuning.

Nevertheless, we evaluate prompt based and fine-tuned VLMs as baselines for comparison with CAuSE. These models perform significantly worse across all metrics (see Table 5).

Performance of Baselines.

(a) Naive Prompting: We evaluated zero-shot and few-shot ($k = 0/2$) prompting using PaLiGemma and LLaVA, where VLMs were prompted with in-context examples showing classifier M's decisions. However, the generated explanations showed poor faithfulness on Hateful Memes and e-SNLI-VE. Although the models showed an improved faithfulness for VQA-X, the plausibility scores were very poor.

(b) Fine-Tuning: Fine-tuning VLMs to imitate the classifier M led to substantial improvements over prompting. These models are trained to reproduce the results of M given the same inputs, effectively learning to approximate its decision function.

5.4 Ablation studies

What is the role of loss functions other than $\mathcal{L}_\phi$?

In Table 2, CAuSE achieves the highest F1 score across almost all experiments, except for a few exceptions (four out of twelve cases in total).

In terms of causal abstraction, as measured by composite CCMR, CAuSE also obtains the highest scores in most cases, with a few exceptions of VQA-X and Hateful meme datasets (two out of twelve cases in total). However, for plausibility (cf. Table 3), CAuSE ranks second in most combinations highlighting the faithfulness-plausibility trade-off we have mentioned in section 5.2.

IIT ensures causal abstraction between M and the explainer. Consequently, we observe CAuSE obtains a slightly higher composite CCMR score

M	Ablations	Hateful Meme				e-SNLI-VE				VQA-X			
		F1	CCMR	% gen.	c- CCMR	F1	CCMR	% gen.	c-CCMR	F1	CCMR	% gen.	c-CCMR
CLIP+MFB	ϕ	99.00	**83.00**	10.20	18.16	93.69	**79.89**	17.72	42.90	**84.16**	44.79	**86.97**	59.13
	$\phi + TS$	98.80	67.86	5.71	10.53	81.08	60.19	31.47	41.33	76.88	44.92	85.10	58.80
	$CAuSE$	**99.30**	55.40	**50.10**	**52.65**	**95.30**	67.36	**87.68**	**76.19**	74.53	**46.27**	86.67	**60.37**
VisualBERT	ϕ	99.50	38.15	39.58	38.85	98.80	69.20	51.65	59.15	87.50	6.17	70.93	11.35
	$\phi + TS$	**100.00**	40.05	41.35	40.69	94.89	64.08	**74.48**	**68.89**	79.80	5.27	80.12	9.89
	$CAuSE$	99.40	**47.94**	**55.63**	**51.50**	**98.91**	**69.40**	59.09	63.83	**87.83**	**26.84**	**80.19**	**40.22**
FLAVA	ϕ	**99.40**	36.36	2.37	4.45	97.40	83.77	50.71	63.18	**97.40**	33.33	55.00	41.51
	$\phi + TS$	96.90	34.97	**64.00**	**45.23**	97.40	82.92	**57.11**	67.64	83.02	19.52	56.56	29.02
	$CAuSE$	99.30	**49.68**	16.76	25.06	**97.70**	**87.20**	56.40	**68.50**	80.44	**35.22**	**62.08**	**44.94**
Qwen-VL	ϕ	68.60	38.49	99.03	55.43	93.29	65.10	97.07	77.93	59.79	10.44	**74.79**	18.32
	$\phi + TS$	61.30	48.79	99.87	65.56	85.59	63.29	77.33	69.61	26.23	4.42	70.84	8.32
	$CAuSE$	**70.60**	**49.28**	**100.00**	**66.02**	**94.09**	**68.00**	**99.65**	**80.84**	**61.15**	**18.40**	65.10	**28.69**

Table 2: *Faithfulness* Results for various classifiers (M), using CLIP+MFB, VisualBERT, FLAVA, and Qwen-VL. The table reports the effect of ablating individual loss components from Equation 6. All scores are shown in %.

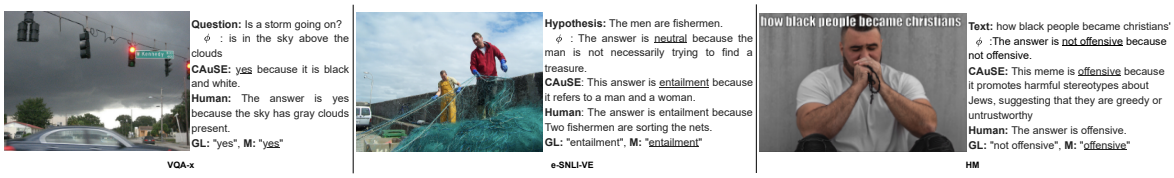

Figure 3: Examples illustrating the faithfulness–plausibility trade-off across the three datasets. In each case, ϕ produces a more plausible explanation for an incorrect prediction (i.e. unfaithful), while CAuSE generates a less plausible but more faithful (reflecting a correct prediction) explanation. Both ϕ and CAuSE seek to emulate the classifier's output and provide justifications consistent with the predicted answer. **GL:** Ground-truth label, **M:** Prediction from M.

and F1 score compared to its counterparts (ϕ, and $\phi + TS$) in most cases (cf. Table 2). This result provides empirical evidence that our proposed loss, $\mathcal{L}_{IIT}$, improves causal faithfulness.

Is the faithfulness of CAuSE classifier agnostic? Yes it is *largely classifier agnostic* since it maintains faithfulness for all the evaluated models in most settings. We have used two representative models from two fusion mechanisms widely used in literature, CLIP+MFB for late-fusion and VisualBERT for early-fusion. Also, two modern classifiers are used: FLAVA (Singh et al., 2022) and Qwen-VL (Bai et al., 2023)[5]. As noted earlier, we use Qwen-VL by replacing its language modeling (LM) head layer with a multi-layered perceptron (MLP) for classification. For all of these models, CAuSE obtains a better CCMR score compared to the counterparts (ϕ or $\phi + TS$) in the majority of cases. Refer to Table 2 for reference.

Are explanations from various models useful? A critical question is whether CAuSE's explanations remain useful despite lower plausibility. To investigate this, two annotators independently evaluated 150 randomly sampled explanations from CAuSE and ϕ across three datasets using a 5-point Likert scale (1=poor, 2=fair, 3=good, 4=very good, 5=excellent) on two dimensions:

i) *Relatedness:* It assesses whether an explanation provides logically coherent justification for the model M's prediction, independent of human-like reasoning (or, GT explanation). For instance, CAuSE's VQA-X explanation "yes because it is black and white" earned high relatedness by meaningfully exposing the model's shortcut reasoning, despite diverging from human explanations referencing dark clouds (cf. left-most image of Figure 3).

ii) *Fluency:* It measures the grammatical quality of the generated explanation.

Annotators accessed inputs and predictions of M but were explicitly instructed not to consult ground-truth explanations to avoid bias towards human-like reasoning. Inter-annotator agreement was substantial for both metrics: Cohen's $\kappa = 0.68$ for relatedness and $\kappa = 0.76$ for fluency.

Table 4 shows CAuSE maintains comparable

[5] https://huggingface.co/Qwen/Qwen2-VL-2B

Dataset	M	Ablations	B-1	B-2	B-3	B-4	BERTScore
Hateful Meme	CLIP+MFB	ϕ	0.70	0.65	0.62	**0.59**	0.98
		$\phi + TS$	0.67	0.61	0.57	0.53	0.97
		$CAuSE$	**0.70**	**0.65**	**0.62**	0.58	**0.98**
	VisualBERT	ϕ	**0.71**	**0.68**	**0.65**	**0.63**	**0.97**
		$\phi + TS$	0.71	0.68	0.65	0.63	0.97
		$CAuSE$	0.67	0.62	0.59	0.56	0.96
	FLAVA	ϕ	**0.62**	**0.54**	**0.46**	0.38	0.93
		$\phi + TS$	0.53	0.46	0.40	0.34	0.91
		$CAuSE$	0.61	0.52	0.44	0.36	**0.93**
	Qwen-VL	ϕ	0.58	0.52	0.48	0.45	0.95
		$\phi + TS$	**0.68**	**0.62**	**0.57**	**0.54**	0.95
		$CAuSE$	0.55	0.49	0.44	0.40	**0.95**
e-SNLI-VE	CLIP+MFB	ϕ	**0.39**	**0.33**	**0.28**	**0.24**	**0.91**
		$\phi + TS$	0.34	0.28	0.24	0.20	0.90
		$CAuSE$	0.38	0.32	0.27	0.24	0.90
	VisualBERT	ϕ	**0.43**	**0.36**	**0.31**	**0.27**	**0.91**
		$\phi + TS$	0.38	0.32	0.27	0.23	0.90
		$CAuSE$	0.39	0.32	0.27	0.23	0.90
	FLAVA	ϕ	**0.43**	**0.36**	**0.31**	**0.27**	**0.91**
		$\phi + TS$	0.42	0.35	0.30	0.26	0.91
		$CAuSE$	0.40	0.32	0.26	0.22	0.90
	Qwen-VL	ϕ	**0.42**	**0.34**	**0.27**	**0.22**	**0.91**
		$\phi + TS$	0.39	0.31	0.24	0.20	0.91
		$CAuSE$	0.38	0.29	0.23	0.18	0.91
VQA-X	CLIP+MFB	ϕ	**0.43**	**0.36**	**0.31**	**0.25**	**0.92**
		$\phi + TS$	0.17	0.14	0.11	0.09	0.88
		$CAuSE$	0.24	0.19	0.15	0.12	0.88
	VisualBERT	ϕ	**0.57**	**0.51**	**0.46**	**0.41**	**0.94**
		$\phi + TS$	0.33	0.27	0.22	0.18	0.90
		$CAuSE$	0.33	0.27	0.23	0.19	0.90
	FLAVA	ϕ	**0.55**	**0.52**	**0.49**	**0.45**	**0.95**
		$\phi + TS$	0.48	0.40	0.35	0.30	0.90
		$CAuSE$	0.13	0.11	0.09	0.07	0.83
	Qwen-VL	ϕ	0.31	**0.24**	**0.17**	**0.10**	**0.89**
		$\phi + TS$	0.20	0.12	0.08	0.05	0.85
		$CAuSE$	**0.23**	0.17	0.11	0.07	0.88

Table 3: *Plausibility* results for various classifiers (M) showing the role of ablating various loss components in Equation 6.

utility to ϕ: slightly higher relatedness (2.49 vs. 2.29 for ϕ) but slightly lower fluency (2.79 vs. 3.09 for ϕ). The minimal score difference across both metrics indicates that CAuSE's enhanced faithfulness (as measured through CCMR and F1) incurs a negligible cost in perceived usefulness, validating that CAuSE explanations, which purportedly expose the model's true reasoning, retain value for understanding discriminative classifier decisions.

5.5 Qualitative studies

5.5.1 MLLM-as-a-Judge Evaluation

To better understand the relative strengths of CAuSE and its ablated variants, we adopt the *MLLM-as-a-Judge* framework (Chen et al., 2024), which leverages a multimodal large language model (MLLM) to assess which of several candidate outputs best aligns with a reference model's reasoning, which in this case, is the frozen multimodal classifier M (VisualBERT).

We perform a tournament-style evaluation

Dataset	Model (M =Qwen-VL)	Relatedness	Fluency
Hateful Meme	ϕ	2.10	**3.13**
	CAuSE	**2.23**	2.90
e-SNLI-VE	ϕ	**2.87**	**3.09**
	CAuSE	2.70	2.91
VQA-X	ϕ	1.90	**3.05**
	CAuSE	**2.53**	2.55

Table 4: Human evaluation scores for relatedness and fluency across datasets. The model M is Qwen-VL.

Dataset	Baselines	F1	B-1	B-2	B-3	B-4	BERTScore
Hateful Meme	LLaVA ($k=0$)	58.44	0.090	0.010	0.010	0.010	0.889
	LLaVA ($k=2$)	46.55	0.120	0.020	0.010	0.010	0.864
	PaLiGemma ($k=0$)	45.51	0.110	0.020	0.010	0.010	0.856
	PaLiGemma ($k=2$)	47.89	0.170	0.040	0.020	0.020	0.878
	PaLiGemma (FT)	72.33	**0.410**	0.270	0.150	0.090	0.891
	LLaVA (FT)	**72.38**	0.400	**0.270**	**0.170**	**0.130**	**0.894**
e-SNLI-VE	LLaVA ($k=0$)	33.12	0.220	0.070	0.030	0.020	0.876
	LLaVA ($k=2$)	35.77	0.220	0.070	0.030	0.010	0.869
	PaLiGemma ($k=0$)	38.78	0.270	0.090	0.030	0.010	0.851
	PaLiGemma ($k=2$)	31.21	0.180	0.050	0.020	0.010	0.861
	PaLiGemma (FT)	**64.90**	0.190	0.040	0.010	0.010	**0.866**
	LLaVA (FT)	64.29	**0.220**	**0.080**	**0.030**	**0.020**	0.859
VQA-X	LLaVA ($k=0$)	88.12	0.045	0.022	0.011	0.002	0.642
	LLaVA ($k=2$)	91.11	0.033	0.014	0.007	0.004	0.653
	PaLiGemma ($k=0$)	35.05	0.004	0.001	0.001	0.000	0.653
	PaLiGemma ($k=2$)	35.74	0.004	0.001	0.001	0.000	0.651
	PaLiGemma (FT)	94.08	**0.112**	**0.034**	**0.026**	**0.024**	**0.712**
	LLaVA (FT)	**94.11**	0.070	0.022	0.011	0.008	0.679

Table 5: VLM-based baselines. FT denotes fine-tuned model. $k = 0/2$ shows $0/2$ shots prompting results.

where the MLLM ranks the natural language explanations generated by: (i) *CAuSE*: CAuSE framework with full loss $\mathcal{L}_{CAuSE}$, (ii) $\phi + TS$: CAuSE framework trained with $\mathcal{L}_{\phi} + \mathcal{L}_{TS}$ (i.e., without $\mathcal{L}_{IIT}$), and (iii) ϕ: CAuSE framework trained with $\mathcal{L}_{\phi}$ only.

For each test example, we first compute the classifier's LIME (Ribeiro et al., 2016) attributions, highlighting the most influential text tokens and image regions used by M to make its prediction. The MLLM then receives the input, the predicted classifier label, and the corresponding LIME attributions, along with the three candidate explanations. It is asked to rank the candidates based on how well they reflect the classifier's prediction and saliency map.

Ranking-based scores are then assigned as follows:

(i) If the ranking is $CAuSE > \phi{+}TS > \phi$, the assigned scores are 2, 1, and 0, respectively. (ii) If $CAuSE = \phi + TS > \phi$, the scores are 1.5, 1.5, and 0. (iii) If $CAuSE > \phi + TS = \phi$, the scores are 2, 0.5, and 0.5, and (iv) If all explanations are judged equally good, each method receives a score of 1.

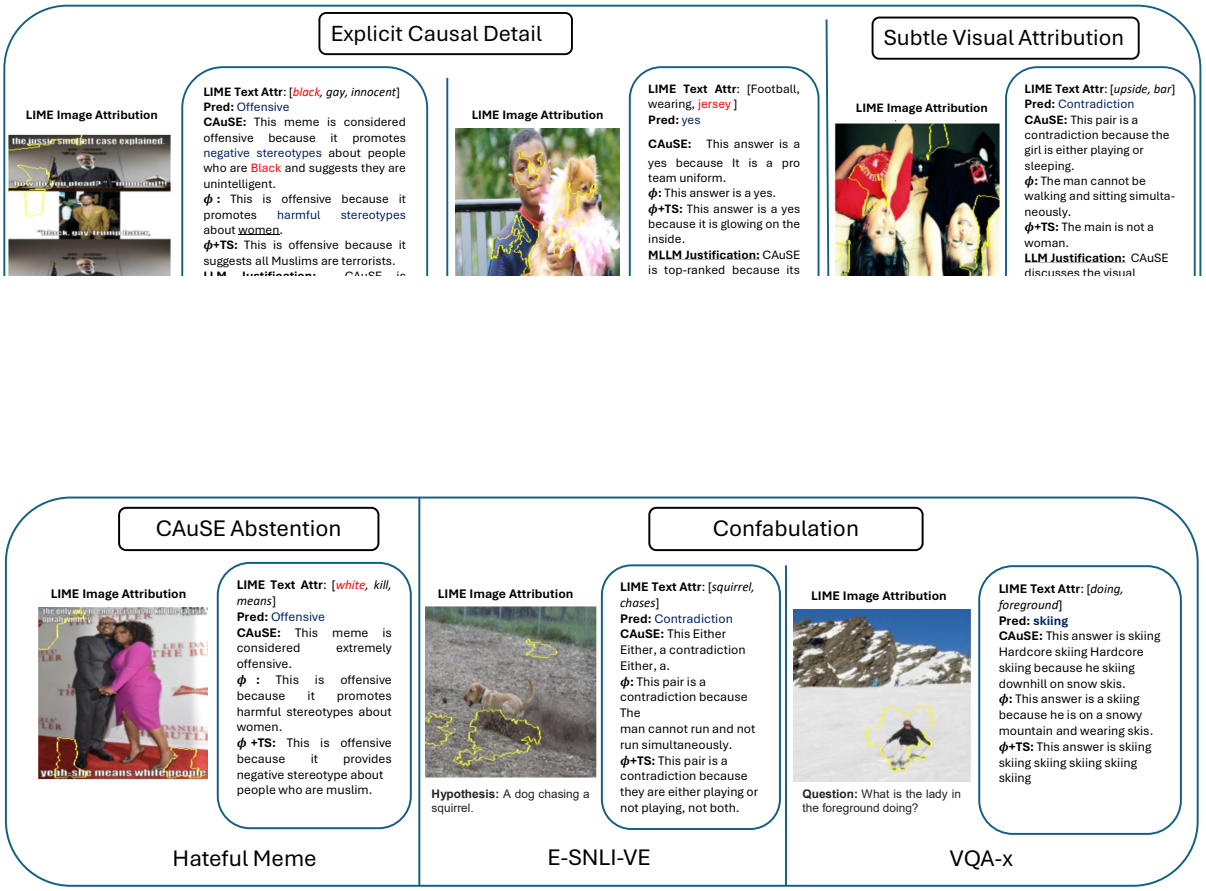

Figure 5: Representative dataset examples where CAuSE underperforms ablations in *fluent* and *coherent* generation.

Final tournament scores are obtained by aggregating these per-sample scores across the entire test set. As shown in Table 7, CAuSE with the full loss consistently outperforms its ablations across all three datasets.

5.5.2 Qualitative Examples from the MLLM Tournament

To make it clear *why* CAuSE wins, Figure 4 shows three representative examples (*with* M as VisualBERT) where CAuSE was judged best by our blind MLLM critic. Each row gives: a LIME image heatmap attribution, the top important words from the text as per LIME, the classifier's prediction, three candidate explanations, and the MLLM's free-text justification. We can clearly observe that CAuSE can capture most of the nuances of the classifier attribution, denoted as *Explicit causal detail*, and also point out the visual cues while generating the explanation, denoted as *Subtle visual attribution*. For instance, CAuSE identifies that the girl is playing in the third image, which aligns with the LIME image attribution, highlighting the girl's hands in a playful mode.

5.5.3 Failure Modes of CAuSE

In many scenarios, when subtle reasoning is needed which is not obvious from both textual and visual inputs, CAuSE fails to produce a coherent explanation. Figure 5 shows three such cases.

Illustrative Examples. In the first example, due to the implicit nature of offense (racism), CAuSE ablations fail to accurately reason why the meme is offensive (implicit racism), although correctly highlighting it was closely related to "stereotype". CAuSE does not explain anything at all except that the meme is offensive.

In the second example, the hypothesis states: "Dog chasing a squirrel." The image does not clearly indicate whether the dog is actually chasing a squirrel, making the ground-truth label inherently ambiguous. Although the classifier predicts contradiction, it does so with low confidence (probability of 0.52). As a causal abstraction of

Image	CAuSE	ϕ only	ϕ+TS	Ground Truth (GT)	M	$\hat{y}$
 A juggler is performing outdoors.	entailment because several people are standing in front of a building.	entailment because A man with a mustache is a type of man.	entailment because A man is a man.	A juggler is juggling clubs at an outdoor plaza.	VisualBERT	Entailment
	entailment because a man and running along a track is performing a stunt while jumping over a hurdle.	entailment because A man is a man.	entailment because The man is a man and the three people are the three people.	A juggler is juggling clubs at an outdoor plaza.	CLIP+MFB	Entailment
	offensive because it promotes negative stereotypes about people who are Muslim.	offensive because it promotes negative stereotypes about people who are Muslim.	offensive because it promotes a harmful stereotype about Muslims.	It promotes anti-Semitism and hatred towards Jewish people.	VisualBERT	Offensive
	This meme is offensive.	This meme is offensive.	This meme is offensive because it promotes a number of harmful stereotypes and prejudices.	It promotes anti-Semitism and hatred towards Jewish people.	CLIP+MFB	Offensive
 What kind of room is this?	This answer bathroom is a bathroom because A toilet sits along the back wall.	This answer is a bathroom because There is a toilet and toilet paper.	This answer is a bathroom bath	This answer is a bathroom because I see a toilet and a tub.	VisualBERT	Bathroom
	The answer is bathroom because A toilet sits next to a sink.	The answer is bathroom because the toilet and sink are both in the room.	The answer is bathroom shower	This answer is a bathroom because I see a toilet and a tub.	CLIP+MFB	Bathroom

Table 6: Error Analysis: These cases demonstrate errors that are prevalent in baselines as well as CAuSE. For all such cases, the classifier predicted class and the class predicted by the explainers are the same but the explanation does not *completely* reflect the underlying scenario.

Dataset	CAuSE	ϕ only	ϕ+TS
HM (VisualBERT)	**1.08**	0.98	0.93
e-SNLI-VE (VisualBERT)	**1.15**	0.96	0.67
VQA-X (VisualBERT)	**1.01**	0.87	0.85

Table 7: MLLM Tournament scores for the Explainers on HM and eSNLI-VE, and VQA-X datasets.

the classifier, CAuSE reflects this uncertainty in its explanation by using terms like "Either". However, it ultimately fails to generate a coherent explanation. In contrast, the other two variants, although mostly incorrect (especially ϕ), manage to produce a more plausible explanation. The third example demonstrates CAuSE's poor fluency. Although its reasoning is correct and plausible, the explanation contains grammatical mistakes and is a clear case of confabulation. In contrast, ϕ generates more natural and fluent text.

5.6 Generic Error Analysis

In Table 6, we selected three examples, one from each dataset, to illustrate two broad categories of cases where both baselines and CAuSE fail in generating logically valid explanations.

Lack of fine-grained representation. In the first two examples from e-SNLI-VE dataset, the hypothesis states: "A juggler is performing outdoors," and the premise is entailed, as supported by the ground-truth explanation: "A juggler is juggling clubs at an outdoor plaza." While CAuSE and its counterparts correctly match the classifier's prediction (both for VisualBERT and CLIP+MFB), the generated explanations fall short; CAuSE omits key details (e.g., the word "juggling"), and others produce incorrect outputs. These errors likely stem from limited object-level details in the initial representation c used by CAuSE. The last two examples from the VQA-X dataset exhibit a similar issue. Although CAuSE and comparable methods produce predictions consistent with the ground truth and generate largely accurate explanations, they fail to attribute the label to the presence of the bathtub, which is only faintly visible in the input image.

Implicit semantic category. In the third and fourth examples (from Hateful Meme dataset), although CAuSE correctly predicts the output class as offensive, it does so for the wrong reasons. This is true for other counterparts as well. Due to the explicit content in the meme, all the models, including CAuSE diagnose that the meme is offensive based on general stereotypes, but they could not point out "antisemitism", as neither the image nor the text explicitly conveys the historical context of the Holocaust, where six million Jews were killed. Without this implicit context, no model can pinpoint antisemitism present in this meme.

6 Related Work

Interpretability. Interpretability is crucial for building trust in AI systems within human society. Techniques like LIME, SHAP and RISE (Ribeiro et al., 2016; Lundberg and Lee, 2017; Petsiuk et al., 2018) explain classifier predictions by providing feature-level explanations for local interpretability. Although model-agnostic, these methods lack global interpretability, which is addressed by GALE (van der Linden et al., 2019), where local explanations are aggregated into a global model understanding. Approaches like SmoothGrad (Smilkov et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) utilize input gradients for model explanation, while CAM (Zhou et al., 2016) highlights critical pixels for decision making in visual classification. Counterfactual generations (Chang et al., 2019; Mothilal et al., 2020; Goyal et al., 2019) also offer insights into the inner working of the model by revealing decision boundaries. However, most of these methods often overlook implicit features behind model decisions and lack natural language explanations. To address these limitations, we propose a novel framework for classifier explanation which generates both *faithful* and *plausible* (human-like) natural language outputs.

Causal Interpretability. Causal interpretability refers to the ability to explain a model's decisions by identifying the cause-effect relationships between input features and the model's output. Feder et al. (2022) demonstrated how incorporating causal reasoning in NLP tasks can improve model predictions and enhance interpretability by going beyond simple correlations between input features and outputs. Further works by Geiger et al. (2021); Vig et al. (2020); Meng et al. (2022) have focused on causal abstraction and causal mediation analysis, helping to create causally faithful models and identify direct and indirect causal factors behind certain model behaviors. In addition to generating counterfactuals, testing models on counterfactual inputs is another critical aspect of understanding model behavior. Since creating exact counterfactuals is challenging (Abraham et al., 2022; Calderon et al., 2022), recent research has focused on approximations (Geiger et al., 2021) or counterfactual representations (Feder et al., 2021; Elazar et al., 2021; Ravfogel et al., 2021). Our proposed counterfactual metric CCMR is inspired by these counterfactual representations. Moreover, most of the existing works focus on single modality (Feder et al., 2021; Goyal et al., 2020). In contrast, the natural language explanation provided by our framework is model and task-agnostic, and capable of handling multimodal inputs.

7 Conclusion and Future Work

In this paper, we presented *CAuSE* (Causal Abstraction under Simulated Explanation), a novel framework for generating causally faithful natural language explanations for discriminative multimodal classifiers. By integrating *Interchange Intervention Training* (IIT) with a Language Model (LM) based module, CAuSE addresses the limitations of existing interpretability methods, ensuring explanations are directly tied to the classifier's reasoning. Our proposed CCMR score highlights CAuSE's improved faithfulness performance on datasets like e-SNLI-VE, Hateful Memes, and VQA-X.

While CAuSE demonstrates robust task-agnostic performance, future work will focus on enhancing fine-grained object-level representations and extending the framework to temporal data, such as video and audio. Additionally, we aim to explore how *self-supervised* learning and deeper integration of implicit cultural knowledge can further improve the framework's scalability and contextual understanding in real-world applications. Currently, CAuSE is tailored for discriminative classifiers and adapting it for generative models is an area for future research.

Acknowledgements

Dibyanayan Bandyopadhyay acknowledges support from the Prime Minister's Research Fellowship (PMRF). The authors gratefully acknowledge the partial support from the project titled "COILD-Centre of Indian Language Data", sponsored by Meity, Govt. of India. The authors also thank the anonymous reviewers and the Action Editors for their constructive and valuable feedback, which helped improve the quality of this work.

References

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

vances in Neural Information Processing Systems, volume 35, pages 17582–17596. Curran Associates, Inc.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*, abs/2308.12966. Version 3.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Dibyanayan Bandyopadhyay, Asmit Ganguly, Baban Gain, and Asif Ekbal. 2024a. 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.

Dibyanayan Bandyopadhyay, Mohammed Hasanuzzaman, and Asif Ekbal. 2024b. Seeing through VisualBERT: A causal adventure on memetic landscapes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10715–10731, Miami, Florida, USA. Association for Computational Linguistics.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint*, abs/2407.07726. Version 2.

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.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2019. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. 2019. Neural network attributions: A causal perspective. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 981–990. PMLR.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6562–6595. PMLR.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10563–10571.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-SNLI-VE: Corrected visual-textual entailment with natural language explanations. *arXiv preprint*, abs/2004.03744. Version 3.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

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.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. 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 1479–1491, Online. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2020. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint*, abs/1907.07165. Volume 2.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xiaofeng Ji, Faming Gong, Nuanlai Wang, Yanpu Zhao, Yuhui Ma, and Zhuang Shi. 2025. Pixel-level semantic parsing in complex industrial scenarios using large vision-language models. *Information Fusion*, 116:102794.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1224–1234.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *arXiv preprint*, abs/1907.03039. Version 1.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Fadel M. Megahed, Ying-Ju Chen, Bianca Maria Colosimo, Marco Luigi Giuseppe Grasso, L. Allison Jones-Farmer, Sven Knoth, Hongyue Sun, and Inez Zwetsloot. 2025. Adapting OpenAI's CLIP model for few-shot image inspection in manufacturing quality control: An expository case study with multiple application examples. *arXiv preprint*, abs/2501.12596. Version 2.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized input sampling for explanation of black-box models. *arXiv preprint*, abs/1806.07421. Version 3.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–546, Bangkok, Thailand. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. *arXiv preprint*, abs/1706.03825. Version 1.

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 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Los Alamitos, CA, USA. IEEE Computer Society.

A Proofs

Lemma 1. *(Simulation) Considering the weights of $\mathcal{C}_2$ and $\mathcal{C}_1$ remain the same throughout the training process, using IIT between them ensures $F(c) = (\mathcal{A} \circ \phi \circ \psi)(c) = E(t, v)$, where $c = E(t, v)$. Informally, the LLM machinery ($F = (\mathcal{A} \circ \phi \circ \psi)$) simulates the behavior of E.*[6]

Proof. This proof establishes the principle in a simplified and idealized setting with a two-neuron linear representation, demonstrating that the IIT training objective drives F to behave as an identity function. Also, note $b = (t, v)$

We first assume that the weights of $\mathcal{C}_1$ and $\mathcal{C}_2$ are w and w', respectively, and they remain fixed throughout training. Using $\mathcal{R}_{match}$, $w = w'$. For simplicity, we assume both $\mathcal{C}_1$ and $\mathcal{C}_2$ are simple two-layer fully connected feedforward neural networks with two dimensional linear logits.

Define the two logits for C_1 under the interchange intervention as $o_1 := \sum_{i=1}^{2} w_{1i} E(b)_i$ and $o_2 := \sum_{i=1}^{2} w_{2i} E(s)_i$, and for C_2 as $o_1^F := \sum_{i=1}^{2} w'_{1i} F(E(b))_i$ and $o_2^F := \sum_{i=1}^{2} w'_{2i} F(E(s))_i$. By IIT objective and condiering, $w = w'$, on softmax output logits, we have

$$\frac{e^{o_1}}{e^{o_1} + e^{o_2}} = \frac{e^{o_1^F}}{e^{o_1^F} + e^{o_2^F}}. \tag{7}$$

Here, b and s are the base and source inputs, respectively, for E. Correspondingly, $E(b)$ and $E(s)$ are the base and source inputs for F.

Assume, for contradiction, that $E(b) \neq F(E(b))$. Then there exists some j such that $E(b)_j \neq F(E(b))_j$. Let, $E[b] = [p, q]$ and $F[E[b]] = [\rho_1 p, \rho_2 q]$. Here we implicitly assumed that the explainer mapping acts componentwise on encoder outputs, i.e. for any $z \in \mathbb{R}^2$ we write $F(z) = \phi(z) \odot z$, where $\phi(z) = (\rho_1(z), \rho_2(z))$ consists of continuous scalar-valued functions. Further, we restricted attention to encoder outputs in the domain of interest U (i.e., $E(x) \in U$), which is a non-empty open subset of encoder representations ($\mathcal{S} = \{E(x) : x \in D_E\} \subset \mathbb{R}^2$), where $\rho_1(z)$ and $\rho_2(z)$ are constant (locally on U), so $F([p, q]) = [\rho_1 p, \rho_2 q]$ for fixed scalars ρ_1, ρ_2 on U.

Let (s_i, b_j) pairs be sampled from the Cartesian product $D_E \times D_E$. Assuming IIT equalities to hold over all $(s_i, b_j) \in D_E \times D_E$ exactly (KL=0), and restricting our attention to diagonal pairs (b, b), $E(s) = E(b)$.

So, considering $E(s) = E(b)$, and denoting, $w_{11} = \beta, \quad w_{12} = \gamma, \quad w_{21} = \delta, \quad w_{22} = \epsilon$, we can write Equation 7 as:

$$\begin{aligned} &\beta p(1-\rho_1) + \gamma q(1-\rho_2) \\ &= \underbrace{\log\left(\exp(\beta p + \gamma q) + \exp(\delta p + \epsilon q)\right)}_{\mathcal{Q}} \\ &\quad - \underbrace{\log\left(\exp(\beta\rho_1 p + \gamma\rho_2 q) + \exp(\delta\rho_1 p + \epsilon\rho_2 q)\right)}_{\mathcal{Q}^*} \end{aligned} \tag{8}$$

Similarly, considering intervention on the second set of neurons:

$$\delta p(1-\rho_1) + \epsilon q(1-\rho_2) = \mathcal{Q} - \mathcal{Q}^* \tag{9}$$

Equations 8 and 9 both hold true if any one of the following conditions is satisfied:

i) $\beta = \delta$ and $\gamma = \epsilon$

ii) $E(b)$ is a zero vector

iii) $\rho_1 = \rho_2 = 1$

The first condition is impossible (degenerate), because even though weights of $\mathcal{C}_1$ and $\mathcal{C}_2$ are the same, the weights of $\mathcal{C}_1$ are fixed, and pairwise equality of weights for $\mathcal{C}_2$ is not enforced during training. The second condition is false because $E(b) \in U$ is fixed for any b and is not a zero vector. Assuming U is non-degenerate so that p and q vary independently on U and using that U is open, the above identities can hold for all $(p, q) \in U$ only if $1 - \rho_1 = 1 - \rho_2 = 0$, i.e. $\rho_1 = \rho_2 = 1$ on U, implying

$$E(b) = F(E(b))$$

□

Theorem 1. *(Causal Abstraction) If $\mathcal{C}_1$ and $\mathcal{C}_2$ become identical (their weights are equal and they are a causal abstraction of each other), the explainer becomes a causal abstraction of M and* vice versa.

Proof. We treat E and F as black boxes, assuming E is injective on the domain of interest $\mathcal{X}_0 \in U$ to ensure a well-defined left-inverse β, which can be seen as a standard identifiability assumption for

[6] $A \circ B$ denotes the composition operation between A and B.

Notations	Meaning
E	Encoder of the multimodal classifier (VB, CLIP+MFB)
$\mathcal{C}_1$	The feed forward layer of the multimodal classifier
M	The multimodal classifier, CAuSE seeks to explain ($C_1 \circ E$)
ψ	The linear layer or MLP that projects the encoder representations into the LM
ϕ	The LM (GPT-2) finetuned in our framework.
$\mathcal{A}$	The aggregator that takes the LM output and aggregates it into vectors.
$\mathcal{C}_2$	The feed forward layer of CAuSE, which is identical to $\mathcal{C}_1$
F (LLM machinery)	This is the combination of the MLP, LM and the Aggregator ($A \circ \phi \circ \psi$)
Explainer	All the components taken together ($C_2 \circ F$)
CAuSE	The Explainer along with its losses and the training paradigm.

Table 8: The revised notations that are used throughout the paper

causal abstraction. Since $\mathcal{C}_1$ and $\mathcal{C}_2$ share the same architecture, we permit neuron-level interventions on them while treating other components as observable only via inputs, outputs, and intermediate states. To verify if the explainer forms a valid causal abstraction of M, we analyze three cases:

We assume there exists a correspondence between inputs of E, called x and inputs of F, called z. For intervention on x, $do(x = x')$, we intervene z to z' as $do(z = z')$, but we cannot arbitrarily choose z', as a plausible causal abstraction between M and explainer with respect to the input would require for each high-level intervention $do(x = x')$, there is a corresponding low-level intervention $do(z = \alpha(x'))$ whose result agrees with the high-level model. Following the standard definition of causal abstraction, we only align interventions on z of the form $z = \alpha(x)$. Arbitrary z-values lie outside the image of α and thus have no corresponding high-level intervention, so they are excluded from the abstraction guarantee. E is a good choice for α, because it is already a mapping between x to z and if we choose $\alpha(x) = E(x)$, then $y_2 = \mathcal{C}_2(F(\alpha(x))) = \mathcal{C}_2(F(E(x))) = \mathcal{C}_2(E(x)) = \mathcal{C}_1(E(x)) = y_1$.

If we intervene x to x', we will also intervene z to $\alpha(x')$. Also, $y_1^{do(x=x')} := \mathcal{C}_1(E(x'))$ Under this assumption: $y_2^{do(z=z')} := \mathcal{C}_2(F(z')) = \mathcal{C}_2(F(\overline{(do(z' = \alpha(x')))})) = \mathcal{C}_2(E(x')) = \mathcal{C}_1(E(x')) = y_1^{do(x=x')}$

We restrict our analysis to z' in the image of E, i.e. $z' = E(x')$ for some x'. Then there exists a partial inverse $\beta : E(\mathcal{X}_0) \to X$, $\beta(E(x')) = x'$, or $\beta(z') = x'$ for all $z' \in Im(E)$.

So if we intervene on z to make it z', we assume that intervened z' is also an image of some intervened x', and $x' = \beta(z')$. Denote, $y_2^{do(z=z')} := \mathcal{C}_2(F(z'))$. Under these assumptions, $y_1^{do(x=x')} := \mathcal{C}_1(E(x')) = \mathcal{C}_1(E(\overline{do(x' = \beta(z'))})) = \mathcal{C}_1(E(\beta(z'))) = \mathcal{C}_1(F(\overline{E(\beta(z'))})) = \mathcal{C}_1(F(z')) = \mathcal{C}_2(F(z')) = \mathcal{C}_2(F(z')) = y_2^{do(z=z')}$.

This is easier to do for intermediate representations. Let us denote the intermediate representation of F as $k_F = F(z)$, where $z = E(x)$. We transform $F(E(x))$ to $F(E(x))'$ as $do(k_F = k'_F)$. Before this transformation, as $F(E(x)) = E(x)$, we know the intermediate representation for E (i.e. k_E) would be equal to $k_F = F(E(x))$ too. So, $k_E = k_F$. As we have $k_E = k_F$, hence all interventions on k_E and k_F are interchangeable.

So, $k'_E := k'_F$, if $do(k_F = k'_F)$, and $k'_F := k'_E$, if $do(k_E = k'_E)$. So, under $do(k_F = k'_F)$, $y_2^{do(k_F=k'_F)} := \mathcal{C}_2(k'_F) = \mathcal{C}_1(k'_F) = \mathcal{C}_1(k'_E) = y_1^{do(k_E=k'_E)}$. Similarly, $y_1^{do(k_E=k'_E)} := \mathcal{C}_1(k'_E) = \mathcal{C}_2(k'_E) = \mathcal{C}_2(k'_F) = y_2^{do(k_F=k'_F)}$.

If a neuronal intervention is performed inside $\mathcal{C}_1$ or $\mathcal{C}_2$, the outputs will still match, because $\mathcal{C}_1$ and $\mathcal{C}_2$ are trained to be causally equivalent via the IIT loss and are structurally identical, ensured by R_{match}. Note that any input intervention at the level of $\mathcal{C}_1$ and $\mathcal{C}_2$ falls under the previous case for intervention on intermediate representation, and is already covered.

□

B VLM self-explanation inconsistency vs CAuSE

VLM can be a candidate for generating the explanation of an already trained classifier but as empirically shown in Section 5.3, VLMs suffer from lower F1 score even after finetuning, due to its non access to the representation from M. Through an example below (Figure 6), we also try to explain the VLM failure example under a counterfactual setting while our method CAuSE succeeds.

This figure demonstrates that our model provides more faithful post-hoc explanations under counterfactual scenarios. Unlike LLMs, which often fail to justify their original explanations when the input is minimally changed, our model updates its explanation to reflect the classifier's altered decision

Part ① illustrates that the LLM is unfaithful to its own explanation. It claims that modifying a specific aspect of the image would change its classification from offensive to non-offensive. However, even after altering the input as suggested, the model's label remains unchanged. This demonstrates that standard VLMs often fail to remain consistent with their own explanations. In contrast, our proposed framework CAuSE exhibits robust behavior in the same scenario. When presented with a counterfactual instance ③—i.e., a minimally edited input that causes the M's output to change (see M's output in ② and ③)—finetuned ϕ (trained with $\mathcal{L}_{CAuSE}$) updates its explanation accordingly.

C Bound for Full Neuron Coverage

We aim to find the number of training repetitions, N, required to sample and intervene on all n neurons in the network with a high probability, $1-\delta$. During each repetition, we uniformly sample a fraction $p_s = 0.2$ of the neurons at random.

Derivation of the Bound

The probability that a specific neuron is *not* sampled in a single round is $(1-p_s)$. Since the sampling rounds are independent, the probability that a particular neuron i is not sampled in any of the N repetitions is $P(E_i) = (1-p_s)^N$.

To ensure all neurons are covered, we must bound the probability that at least one neuron is missed. Using the union bound (Boole's inequality), we can express this probability as:

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i) = n(1-p_s)^N \quad (10)$$

We require this failure probability to be no greater than a small threshold δ, leading to the inequality $n(1-p_s)^N \leq \delta$. Solving for N by taking the natural logarithm of both sides yields:

$$(1-p_s)^N \leq \frac{\delta}{n} \quad (11)$$

$$N\ln(1-p_s) \leq \ln(\delta) - \ln(n) \quad (12)$$

$$N \geq \frac{\ln(n) - \ln(\delta)}{-\ln(1-p_s)} \quad (13)$$

Since $\ln(1-p_s)$ is negative for $0 < p_s < 1$, the inequality is reversed during the final step of the derivation.

Application to Our Network

For the classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, the total number of neurons is $n = 557,440$. Setting a desired success probability of 99.999% (i.e., $\delta = 10^{-5}$), and using $p_s = 0.2$, we can apply Equation 13 to calculate the required number of repetitions:

$$N \geq \frac{\ln(557440) - \ln(10^{-5})}{-\ln(0.8)} \approx 110.9$$

Thus, we need at least $N = 111$ repetitions. With a batch size of 16, this corresponds to processing $111 \times 16 = 1776$ training samples. As the training sets for both of our datasets contain more than this number of samples, we can be confident that all neurons are sampled and intervened upon within a single training epoch.

D Hateful Meme Classification Explanation Dataset Preparation

The dataset was created in a similar approach to the methods used in works such as (Feng et al., 2021; Schick and Schütze, 2021). Specifically, GPT-4o LLM was used to generate explanations for each meme in the Facebook Hateful meme dataset(Kiela et al., 2020).

Evaluation scores			Relevance Distribution (%)			Kappa
Model	Relevance	Fluency	NR	PR	HR	Relevance
GPT-4o	2.34	0.94	18	33	49	0.44

Table 9: Human evaluation results for LLM generated explanation. Here, NR: Not relevant, PR: Partially relevant, HR: Highly relevant.

There were 3 main steps, (i) **zero-shot template-based prompt design**, (ii) **inferencing by the LLM**, and (iii) **manual test set creation using post-processing.** The template-based prompt design incorporates a caption generated by the state-of-the-art captioning model OFA(Wang et al., 2022), along with visual entities retrieved from the Google Cloud Vision API[7]. An example

[7] https://cloud.google.com/vision

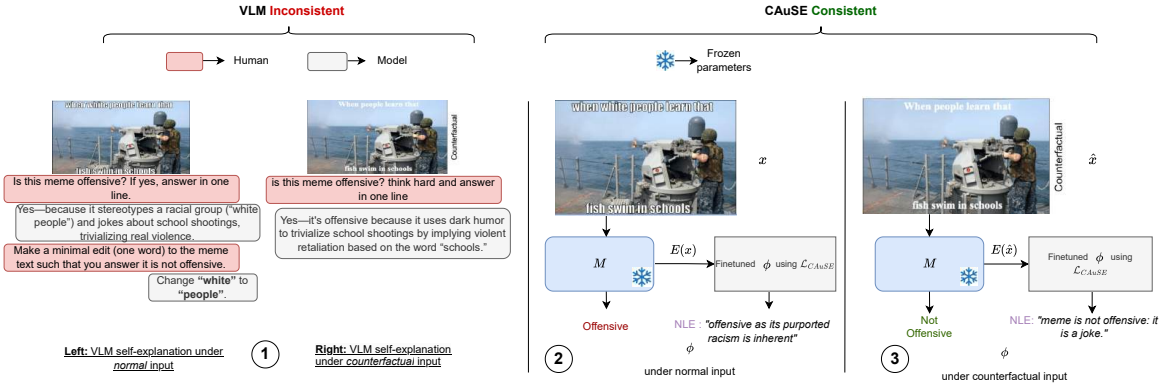

Figure 6: ① *Left:* Given a meme, a VLM (GPT-5.2 (https://chatgpt.com/)) predicts a label and provides a corresponding explanation. ① *Right:* Under a counterfactual input, the explanation remains unchanged despite the altered input, indicating model unreliability. ② and ③ illustrate how the NLE adapts to the counterfactual input along with the prediction. The NLE generated by CAuSE is therefore faithful to the multimodal classifier's prediction, as defined by Atanasova et al. (2023). The example input, the counterfactual generation process and the VLM inference pattern in ① are adopted from Bandyopadhyay et al. (2024b)

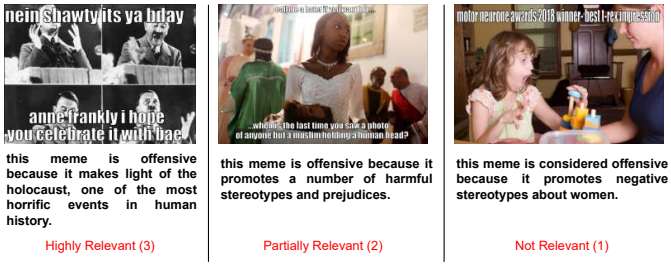

Figure 7: Example of relevance score used for evaluation of generated explanation from LLM.

prompt-response pair is shown in Figure 8.

In step two, standard LLM greedy inferencing is employed to generate an explanation (*LLM response*) for a given prompt. To ensure the quality of the LLM-generated explanations, three evaluators were hired to check the quality of the explanations for a collection of 200 sampled memes. The evaluators are chosen from a pool of Masters' CS students in the age range of 22-27, who have previous similar annotation experience. They were asked to rate the **relevance** of the generated explanation on a scale of *1 (Not relevant), 2 (Partially relevant) or 3 (Highly relevant)*, based on how well it reflected the reason for offensiveness (cf. Figure 7), and the **fluency** of the explanation on a *continuous scale from 0 to 1*, based on how grammatically correct it was (Desai et al., 2022; Kayser et al., 2021).

The complete evaluation statistics are reported in Table 9, which shows that the quality of the generated data is quite good. The Cohen kappa (Cohen, 1960) score obtained is 0.44, indicating fair agreement among the raters.

Finally, from the offensive class examples in the test set, **every LM explanation was post-processed** *if* the LLM annotation does not properly reflect the ground truth reason behind a meme's purported offensiveness. To achieve this, the three evaluators (the same as the above) were specifically instructed to i) trim a longer explanation into a single sentence, ii) remove redundant information from the LLM responses, and iii) rewrite generated explanations that obtain low relevance scores to make them more relevant to the input meme.

Categorization of post-processing. Out of 596 offensive examples, in the test set, 297 examples had to be manually post-edited (giving the percentage of post-edited samples to 49.8%) upon manual inspection, which means all of the not relevant (18%) and almost all of the partially relevant (33%) samples were manually post-edited.

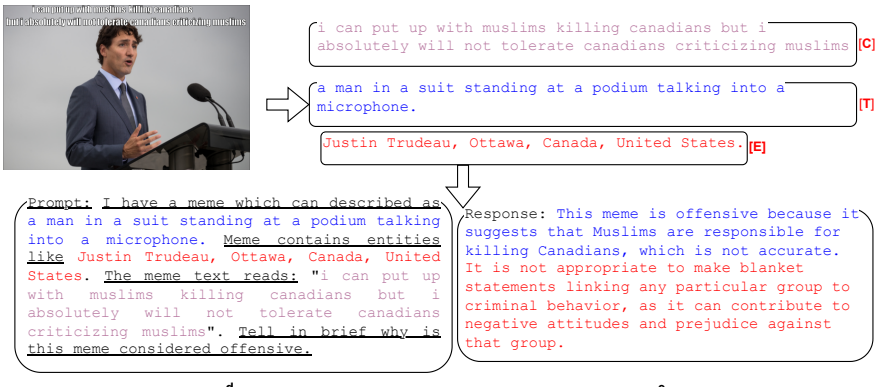

Figure 8: Dataset annotation process via template-based prompt design for LLMs. [T] refers to the caption, [C] refers to the meme text, and [E] is the set of extracted entities. All of them are combined to

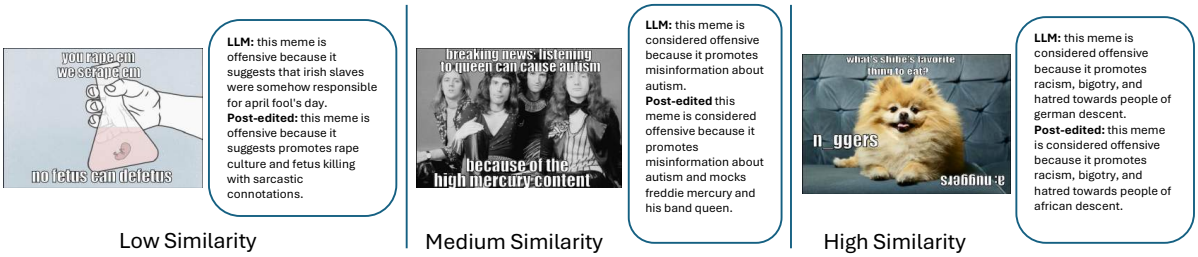

Figure 9: Example post-edits by similarity categorization between before-after edit samples.

We categorize samples into three groups based on BERTScore similarity between the original and post-edited text: low (below the 25th percentile, < 0.925), medium (25th-75th percentile, 0.925–0.973), and high (above the 75th percentile, > 0.973). Notably, the high 25th percentile cutoff of 0.925 indicates that even the low-similarity group required only minor post-editing. Figure 9 shows examples from each group.

Annotation portal. We make use of an annotation portal shown in Figure 10 for post-processing and generating post-processed gold explanation.

The 'Gold explanation' field in the annotation portal refers to the LLM-generated explanation. In Figure 10, the LLM-generated explanation does not properly reflect the exact reason behind the meme's offensiveness. This is where the 'Edited Reference' field comes into play. This field is kept to make changes to the LLM-generated 'Gold Explanation', such that this properly reflects the offensiveness behind a meme. We finally use the edited sentence using the gold-standard annotation as test set used for evaluating all our models.

Figure 10: Depiction of the annotation portal.