SEMANTIFY: Unveiling Memes with Robust Interpretability beyond Input Attribution (Supplementary Material)

Dibyanayan Bandyopadhyay[1] , **Asmit Ganguly**[1] , **Baban Gain**[1] and **Asif Ekbal**[2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Patna, India
[2]School of AI and Data Science, Indian Institute of Technology, Jodhpur, India
{dibyanayan, asmit.ganguly, gainbaban, asif.ekbal}@gmail.com

A Proof

Theorem 1. *With very small step size γ, the condition $\boldsymbol{e}^T \cdot \nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > \boldsymbol{e}^T \cdot \nabla_{\boldsymbol{m}} f(\boldsymbol{m})\rho$, where $\rho > 1$, holds true.*

Proof. We prove that $h(\boldsymbol{m})$ is an increasing function of $\boldsymbol{m}$. We assume the following function $h(\boldsymbol{m})$ as,

$$h(\boldsymbol{m}) = f(\boldsymbol{m}^+) - f(\boldsymbol{m}) \tag{1}$$

where, $\boldsymbol{m}$ is the multimodal representation obtained from multimodal encoder, $\hat{y} = f(\boldsymbol{m})$ is the model predicted class. Now for $\boldsymbol{m}^+ > \boldsymbol{m}$

$$\begin{aligned} \delta h = h(\boldsymbol{m}^+) - h(\boldsymbol{m}) = \\ f(\boldsymbol{m}^{++}) - f(\boldsymbol{m}^+) - (f(\boldsymbol{m}^+) - f(\boldsymbol{m})) \end{aligned} \tag{2}$$

Notationally, we denote (assume), $\nu = \nabla_{\boldsymbol{m}} f(\boldsymbol{m})$ and, $\boldsymbol{m} \to \boldsymbol{m}^+$ further entails $\nu \to \nu^+$.

$$\delta h = \frac{\gamma^2}{\gamma^2} \delta h$$

As, $\boldsymbol{m}^+ = \boldsymbol{m} + \gamma \nabla_{\boldsymbol{m}} f(\boldsymbol{m})$ and we act at the regime where $\gamma \to 0$,

$$\begin{aligned} \delta h = \frac{\gamma^2}{\gamma} \lim_{\gamma \to 0} (\frac{f(\boldsymbol{m}^+ + \gamma \nu^+) - f(\boldsymbol{m}^+)}{\gamma} - \\ \frac{(f(\boldsymbol{m} + \gamma \nu) - f(\boldsymbol{m}))}{\gamma}) \end{aligned} \tag{3}$$

$$\delta h = \frac{\gamma^2}{\gamma} \lim_{\gamma \to 0} (f'(\boldsymbol{m}^+; \nu^+) - f'(\boldsymbol{m}; \nu))$$

As $\nu^+ \to \nu$, when $\gamma \to 0$

$$\delta h = \gamma^2 \lim_{\gamma \to 0} \frac{f'(\boldsymbol{m} + \gamma \nu; \nu) - f'(\boldsymbol{m}; \nu)}{\gamma}$$

By the definition of directional derivative of $f(\boldsymbol{m})$ in the direction of ν,

$$\delta h = \gamma^2 \lim_{\gamma \to 0} \frac{f'(\boldsymbol{m} + \gamma \nu).\nu - f'(\boldsymbol{m}).\nu}{\gamma}$$

$$\delta h = \gamma^2 \lim_{\gamma \to 0} \frac{\nu_i \partial_{\boldsymbol{m}_i} f(\boldsymbol{m} + \gamma \nu) - \nu_i \partial_{\boldsymbol{m}_i} f(\boldsymbol{m})}{\gamma}$$

$$\delta h = \gamma^2 \nu_i \partial_{m_i m_j} f(\boldsymbol{m}) \nu_j$$

$$\delta h = \gamma^2 \nu^T \boldsymbol{H}(f)(\boldsymbol{m}) \nu$$

As $f(\boldsymbol{m})$ is strongly convex, we know its Hessian $\boldsymbol{H}(f)(\boldsymbol{m})$ at any point is positive definite.

Also by the definition of positive semi-definiteness, $\boldsymbol{a}^T.\boldsymbol{H}(f)(\boldsymbol{z}).\boldsymbol{a} > 0$, where $\boldsymbol{a}$ are non-zero real vectors. Considering $\boldsymbol{a}$ as $\boldsymbol{e}$ as we assumed $\boldsymbol{e}$ are non-zero vectors, we can write $\nu^T.\boldsymbol{H}(f)(\boldsymbol{m}).\nu > 0$. Also $\gamma > 0$. So by definition $\delta h > 0$. As h is an increasing function of $\boldsymbol{m}$ so, $\nabla_{\boldsymbol{m}} h(\boldsymbol{m}) > 0$, which further entails:

$$\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > \nabla_{\boldsymbol{m}} f(\boldsymbol{m}) \tag{4}$$

We have $\hat{y} = f(\boldsymbol{m})$ and $\boldsymbol{m}^+ = \boldsymbol{m} + \gamma \nabla_{\boldsymbol{m}} \hat{y}$ for $\gamma > 0$ and γ is a scalar, called step size. As $f(\boldsymbol{m})$ is strongly convex,

$$f(\boldsymbol{m}^+) \geq f(\boldsymbol{m}) + {\nabla_{\boldsymbol{m}}}^T \hat{y}(\boldsymbol{m}^+ - \boldsymbol{m}) + \epsilon \|\boldsymbol{m}^+ - \boldsymbol{m}\|^2 \tag{5}$$

where $\epsilon > 0$. So we get,

$$f(\boldsymbol{m}^+) \geq f(\boldsymbol{m}) + (\gamma + \epsilon \gamma^2) \|{\nabla_{\boldsymbol{m}}}^T \hat{y}\|^2 \tag{6}$$

As $\gamma + \epsilon \gamma^2 > 0$, so

$$f(\boldsymbol{m}^+) \geq f(\boldsymbol{m}) + K \|{\nabla_{\boldsymbol{m}}}^T \hat{y}\|^2 \tag{7}$$

where K is a positive real number.

By Cauchy-Schwartz inequality, assuming e^T and $\nabla_{\boldsymbol{m}} \hat{y}$ are not aligned,

$$e^T.(\nabla_{\boldsymbol{m}} \hat{y}) < \|e\| \|{\nabla_{\boldsymbol{m}}}^T \hat{y}\| \tag{8}$$

So, by Equation 7 and 8 and assuming e^T and $\nabla_{\boldsymbol{m}} \hat{y}$ are not orthogonal,

$$e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) + K e^T.\nabla_{\boldsymbol{m}} (\frac{e^T.\nabla_{\boldsymbol{m}} \hat{y}}{\|e\|})^2 \tag{9}$$

$$\begin{aligned} e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) + \\ 2K(e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m})) e^T.\nabla_{\boldsymbol{m}} (e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m})) \end{aligned} \tag{10}$$

$$e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m})(1 + 2Ke^T.\nabla_{\boldsymbol{m}}(e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}))) \quad (11)$$

Taking gradient w.r.t $\boldsymbol{m}$ and making sure the inequality still holds come from Equation 4.

Also, by the definition of Hessian Matrix $\boldsymbol{H}(f)(\boldsymbol{m})$ of $f(\boldsymbol{m})$, we can write,

$$e^T.\nabla_{\boldsymbol{m}}(e^T.\nabla_{\boldsymbol{m}} f(m)) = \boldsymbol{e}^T.\boldsymbol{H}(f)(\boldsymbol{m}).\boldsymbol{e}$$

(The above identity can be seen true by expansion of terms)

As $f(\boldsymbol{m})$ is assumed to be strongly convex, we know its Hessian $\boldsymbol{H}(f)(\boldsymbol{m})$ at any point is positive definite.

Also by the definition of positive semi-definiteness, $\boldsymbol{a}^T.\boldsymbol{H}(f)(\boldsymbol{z}).\boldsymbol{a} > 0$, where $\boldsymbol{a}$ are non-zero real vectors. Considering $\boldsymbol{a}$ as $\boldsymbol{e}$ as we assumed $\boldsymbol{e}$ are non-zero vectors, we can write $\boldsymbol{e}^T.\boldsymbol{H}(f)(\boldsymbol{m}).\boldsymbol{e} > 0$, which implies, $e^T.\nabla_{\boldsymbol{m}}(e^T.\nabla_{\boldsymbol{m}} f(m)) > 0$

From the above argument, and Equation 11,

$$e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m})\rho \quad (12)$$

where ρ is always greater than 1.

In turn, this would entail that if $e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) \rightarrow 0$, then $e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > 0$, which signifies the **Alignment-Optimization Correlation**. □

B Intuition on Alignment-Optimization Correlation

The intuition and motivation behind the development of the model is to get some introspection on the causal relation of the decision of the model (that is the interpretability of the decision of the model) to potential keywords. Although the retrieved keywords (interchangeably called tokens) cannot compare with LLM-generated explanations, we do not use any LLM, as it may introduce hallucination and LLMs are notorious for generating fluent yet unfaithful text (This fact is already cited in the main paper). This area is to be further researched upon as stated in the limitation section. In the next paragraph, we describe in detail the intuition behind the idea of Alignment-Optimization Correlation.

Sampled tokens should be conductive toward the interpretability objective. More concretely, moving the multimodal representation $\boldsymbol{m}$ (which is the main driving force behind the classification performance of the LLM) towards $\boldsymbol{e}$ (sampled token embedding) should increase the model's confidence on the predicted class $\hat{y} = f(\boldsymbol{m})$. Let us denote, $\boldsymbol{m}^+ = \boldsymbol{m} + \gamma\nabla_m f(m)$. So if $\boldsymbol{m}$ moves in the direction of the gradient of $f(\boldsymbol{m})$ wrt $\boldsymbol{m}$, $f(\boldsymbol{m})$ would be maximized at the fastest rate. Consequently, if $\boldsymbol{e}$ and $\nabla_{\boldsymbol{m}} f(m)$ becomes aligned then we can say moving $\boldsymbol{m}$ in the direction of $\boldsymbol{e}$ would mean that the confidence in the predicted class increases and that the sampled tokens are interpretable.

There is a caveat to the whole proposition. Making $\boldsymbol{e}$ and $\nabla_{\boldsymbol{m}} f(m)$ fully aligned would strip off any extra information that $\boldsymbol{e}$ may capture other than alignment with the gradient. So, to strike a balance between information carrying capacity of $\boldsymbol{e}$ and alignment with the gradient, we would intuitively sample $\boldsymbol{e}$ such that $\boldsymbol{e}$ and the gradient are near orthogonal. Mathematically, $e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) \rightarrow 0$.

Making them near orthogonal implies moving $\boldsymbol{m}$ in the direction of $\boldsymbol{e}$ would not be very conducive towards the interpretability. So we devise a delicate tradeoff. We show that even if $e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) \rightarrow 0$, $e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > 0$.

This condition implies that even though $\boldsymbol{e}$ and the gradient are near orthogonal at the current time-step (which intuitively should be the case considering the $\boldsymbol{e}$ should carry as diverse information as possible), the successive gradients of $f(\boldsymbol{m})$ wrt $\boldsymbol{m}$ and $\boldsymbol{e}$ are aligned. So from later steps, $\boldsymbol{e}$ becomes aligned to the optimization objective. More simply, sampling tokens from the orthogonal direction to $\nabla_m f(m)$ would preserve diverse information of those tokens and successive gradient ascends of $\boldsymbol{m}$ towards $\boldsymbol{e}$ would make the the value of the predicted class to increase, subsequently showing that tokens sampled which are diverse ($e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}) \rightarrow 0$) are also interpretable ($e^T.\nabla_{\boldsymbol{m}} f(\boldsymbol{m}^+) > 0$).

So this criterion of sampling tokens from the orthogonal space of $\nabla_{\boldsymbol{m}} f(\boldsymbol{m})$ strikes as a delicate balance between interpretability objective from the optimization perspective and sampling diverse tokens. This is what we call Alignment-Optimization Correlation. This step facilitates sampling a diverse yet interpretable set of keywords, the utility of which is verified through various experimental setups shown in the main paper.

C Detailed System Design

As we understood from our proposed framework, our system is comprised of two modules: i) Multimodal encoder followed by a ii) Language Model (LM). The LM works as the classifier. In the main paper Multimodal encoding is adequately explained but how the LM is getting its input and acting as the classifier warrants further scrutinized and mathematical outlook. The LM gets both the meme text (T) of length n and meme caption (C) of length m as input along with the multimodal representation $\boldsymbol{m}$. To achieve this, we first concatenate T and C into one text stream (S). This text stream S is passed through the LM embedding layer which converts it into a $\mathbb{R}^{k \times n}$ dimensional matrix (let us call it $E = [t_1, t_2, ...t_n, c_1, c_2, ..., c_m]$), where $k = n + m$ is the combined length of meme text T and caption C, and $n = 1024$, is the GPT-2 [Radford *et al.*, 2019] token embedding vector dimension. This E constitutes the vector representation of the input text. For further understanding of the visual context of the meme, we concatenate E with $\boldsymbol{m}$ (multimodal representation obtained from the previous multimodal encoder block), giving the multimodal context of the input meme. The final dimension of the concatenated input vector I to LM is $\mathbb{R}^{(k+1) \times n}$, which is denoted as $I = [t_1, t_2, ...t_n, c_1, c_2, ..., c_m, \boldsymbol{m}]$.

The LM is trained by the CLM objective which in turn maximizes the following maximum likelihood objective:

$$\mathcal{L}_{CLM} = -\sum_{i=1}^{k+1} \log p(x_i | x_{l<i})$$

,

where x_i are the textual tokens corresponding to the input I and the corresponding token pertaining to the multimodal representation $\boldsymbol{m}$ is denoted by x_{k+1}. We force the LM to

decode '-' as x_{k+1}, which is considered as a *special token* corresponding to the multimodal representation $\boldsymbol{m}$.

D Comparison to Multimodal LLM (LLaVA-1.5 13B)

We compare our method with LLaVA-1.5-13B [Liu *et al.*, 2023] in terms of *plausibility* and *interpretability* of the retrieved keywords. To achieve the task, we manually annotate 100 test samples with ground truth keywords. *Plausibility* refers to the semantic match of retrieved keywords to that of the ground truth keywords and we measure *interpretability* using the LAS score.

Metrics	LLaVA	Ours
Cosine Sim	0.56	**0.64**
LAS	**0.15**	0.09

Table 1: Plausibility (measured by cosine similarity) and Interpretability (measured by LAS) for LLaVA extracted keywords compared to extracted keywords from our framework.

In Table 1, we compare our proposed method to LLaVA-1.5. Our method slightly outperforms LLaVA-1.5 in *plausibility*, and for *interpretability*, LLaVA-1.5 performs slightly better than our model. However, the average scores for LAS and plausibility between LLaVA and our framework remain similar, indicating that our framework captures semantic nuances and explainability capabilities comparable to an existing MLLM, despite having significantly fewer parameters (30-X less).

To compare fairly against MLLM (specifically LLaVA-1.5), we resort to the following steps:

We input a meme to LLaVA and ask it to generate whether it is offensive by giving it the prompt: "*Is this meme offensive*?" The output in natural language is followed by another question: *"Give four one-word keywords that summarize your explanation.*" which outputs four keywords just like our model.

Subsequently, we compare the LAS score of the LLaVA-produced output to that of our proposed model. A higher LAS score obtained by our model reflects higher faithfulness and thus higher model interpretability. We do not choose LLaVA-1.5 to output natural language text as it will be more fluent than a set of keywords and thus would be unfair to compare against our method. Secondly, our main objective is to retrieve a set of **faithful keywords** that can exemplify the model decision well *rather than generating explainable and fluent natural language text*.

Additionally, we pick 100 memes and manually annotate each one of them with a set of four keywords. We measure cosine similarity between the bag-of-word of set A (retrieved from our model) and ground truth set B. We call the average cosine similarity for this case for 100 memes as sim_o. Similarly, we measure the average cosine similarity between set C (retrieved from our LLaVA-1.5) and ground truth set B, which we call sim_l. $sim_o > sim_l$ shows that our method fairs better in terms of plausibility.

So, in summary when considering the richness of extracted keywords, our method fairs better for both faithfulness (i.e. interpretability) and plausibility (i.e. explainability) than state-of-the-art MLLM models like LLaVA.

Some extracted keywords are shown in Table 2. The relevant memes are shown in Figure 1.

E Zero-shot generalization

We assess the generalization capability of our models on HarMeme [Pramanick *et al.*, 2021] dataset in a zero-shot manner. This dataset contains out-of-distribution (COVID-19-related) memes when compared to our training dataset, which constitutes memes from various domains (e.g. racism, politics, etc). Our method achieves a zero-shot F1 score of 64.91% for the offensive meme classification objective, which is better than random performance. The retrieved keywords are mostly explainable and can be adequately used to interpret the model.

Meme ID	Ours
2067	disorder, clinic, effect, clinically
1516	socialist, criminals, economic
5425	hero, villain, horror, alien
5582	refugee, america, forced, presidential
5595	revealed, autistic, legitimate, diagnostic
5528	genocide, situation, acting, considered
5649	comedy, triggered, crisis, liberals
0581	responsible, widespread, delusional, genetic

Table 3: Example memes from HarMeme dataset. We use our model in a zero-shot manner to extract the salient keywords using our proposed framework.

Some example data points are shown in Table 3 for which we show the interpretable keywords extracted from our proposed framework. The corresponding memes can be seen in Figure 2. We observe that for a particular meme, the corresponding keywords adequately describe the meme.

F Modality Importance

Meme classification is a multimodal task and it is essential to quantify the importance of each modality for the corresponding explanation generated from our model. We qualitatively analyze four memes (corresponding memes are shown in Figure 3) where our framework is compared to text-only and image-only baselines. These baselines are formed by only passing either textual or visual representation from the CLIP [Radford *et al.*, 2021] to the downstream *average-pooling* layer. From the qualitative analysis (Table 4) the following things can be inferred: 1. Multimodal (Ours) representation achieves the best quality of retrieved keywords as expected. Though it is closely rivaled by text-only representation, for the last two memes (Id: 91768 and 13875), where our method fails to classify the offensiveness class, the text-only baselines perform pretty poorly too. Even the extracted keywords do not make much sense.

2. Image-only baselines perform pretty poorly. The performance is even poorer than text-only baselines. This is probably due to how the CLIP-based filtering stage is designed,

IDs	GT	LLaVA-1.5-13B	Ours
01924	oppression, racism, anti-feminism, suicide-bombing	Offensive, Violent, Comparison, Women's rights	racism, oppression, slavery, feminists
96328	racism, anti-white, derogatory, hospital	racism, discrimination, harmful, offensive	whites, bigot, refugee, racism
39862	anti-protestanism, inability, mocking, sarcasm	Offensive, disrespectful, insensitive, inappropriate	pedestrians, ter, pro, logic

Table 2: Sample keywords retrieved from LLaVA and our method for sample memes.

Figure 1: Sample memes corresponding to the meme ids in Table 2. Memes are sorted from left to right according to their IDs in Table.

Meme	Ours	Text-only	Image-only
01276	*philanthrop words encourage happiness*	*philanthrop words encourage happiness*	*philanthrop words encourage happiness*
98724	*jews holocaust nazis hitler*	*nazis adolf holocaust jews*	*Adolf jew holocaust nazis*
91768	*jew holocaust hitler*	*die ger dictator propaganda*	*jew holocaust hitler german*
13875	*cats cat lunch eat*	*cats cat lunch sat*	*none*

Table 4: Sample Qualitative analysis of our proposed method's output w.r.t unimodal baselines.

where the image part of the meme is compared to the text part of the meme. This downstream process might lead to non-sensical output when visual representation is used as the unimodal counterpart of textual representation.

3. We tabulate the classification performance of these baselines in Table 5. As per our intuition, multimodal learning beats both unimodal (both textual and visual) representation-based baselines.

Metrics	Text	Vision	Multimodal
Acc	70.29	66.35	75.64
F1	67.77	64.38	73.46

Table 5: Unimodal and Multimodal performance of our proposed method for the task of offensiveness detection. Note we report final classification result as obtained from the LLM, not from the intermediate classifier layer.

G Scale of Relatedness and Exhaustiveness

Relatedness is defined as how much a set of generated keywords is relevant to the content of the input meme, i.e. if all generated keywords are relevant to the meme, the score is five, and if none of them are related, the score is one. The scores between two and four are given when some of the keywords are relevant or partially relevant. In-between scores are subjectively rated by the evaluators depending on their understanding.

Exhaustiveness is defined as the amount of coverage of the theme of a meme through keywords. If the meme can be completely explained with the generated keywords, the score should be five, and if the generated keywords are unable to convey any meaningful information about the meme and are insufficient to explain it, the score is one.

H Proposed Diversity Metrics

We define two metrics for measuring the diversity of the generated samples for a particular meme (referred to as Intra-sample diversity) and for the whole test set (referred to as Inter-sample diversity). Intra-sample diversity metric is used to measure how diverse a set of retrieved tokens is. As an example, let us suppose we have the following four keywords retrieved for a particular meme: {*Hitler, Adolf, Jews, WW2*}. Although the set of keywords correctly reflects the meme is related to WW2 and Nazi Germany, it is less diverse than another set: {*Nazi, Antisemitism, Holocaust, Hitler*} where the collected words are more exhaustive and diverse. High diversity would mean that the generated keywords are not related to each other and thus may not refer to a specific topic. So high diversity is not desirable. Similarly, too low diversity would also be undesirable because it entails the repetitive nature of the retrieved keywords.

Intra-sample diversity. Mathematically we first calculate the word vectors ($\boldsymbol{v}_i \in S$) inside a sample S for the word w_i by GLoVe [Pennington *et al.*, 2014]. The intra-sample diversity score (i_1) is then defined as:

$$i_1 = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{v}_i - \mu(\boldsymbol{v}_i)\|_2 \tag{13}$$

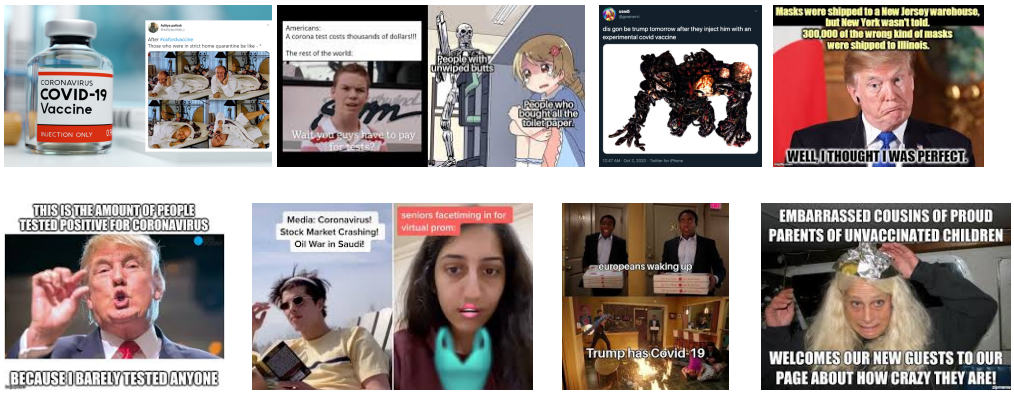

Figure 2: Sample memes corresponding to the meme ids in Table 3. Memes are sorted from left to right, top to bottom according to their IDs.

Figure 3: Memes corresponding to modality importance analysis

, where $\mu(\boldsymbol{v}_i)$ is the mean of the word vectors defined as $\mu(\boldsymbol{v}_i) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i$ and N is the number of samples (typically $N = 4$) retrieved for one particular meme.

Inter-sample diversity. This is a dataset-wide metric. We measure how similar or dissimilar (on average) two samples (a sample refers to N retrieved keywords specific to an input meme) are. This is defined similarly to Intra-sample diversity as:

$$i_2 = \frac{1}{M}\sum_{i=1}^{N} \|\mu(\boldsymbol{v}_i) - \psi(\mu(\boldsymbol{v}_i))\|_2 \tag{14}$$

, where $\psi(\boldsymbol{v}_i) = \frac{1}{M}\sum_{i=1}^{N} \boldsymbol{v}_i$ and M is the number of samples in the dataset.

I Automatic Evaluation metrics

Let us assume the final explanation set $X^j = \{x_i^j\}_{i=1}^n$, which contains $n = 4$ final concepts for the j-th meme in the test set. Denoting $\hat{x}_i^j$ as their text representation (obtained from passing each concept through GLoVe word vectors), the BoW representation $\hat{X}^j$ of the set X^j would be $\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^j$, which is essentially a mean of the concept word representations.

Simulation of the original model. We train a simulator (an SVM model) on i) $\hat{X}^j$, ii) concatenation of $\hat{X}^j$ and m^j, denoted by $[\hat{X}^j; m^j]$, and iii) m^j, to predict the original model prediction $\hat{y}^j$. The jth superscript reflects the jth meme. Intuitively, the SVM simulates the original model based on the provided information (either one of cases (i), (ii), or (iii)). In Table 1 of the main paper, *F1 w/ exp* denotes the simulator performance in case (i) when only the attribute-ranked concepts X were used as simulator input. Similarly, *F1 w/ inp* denotes case (iii), where the multimodal representation (m^j) is used for model input. Lastly, *F1 w/ both* denotes case (ii).

Leakage Adjusted Simulatability. Simulatability of a simulator can be defined as the fraction of examples the simulator can correctly predict the original model's prediction when incorporating for *explanation and input* vs when incorporating for *input only*. Concretely based on the previous notations, simulatability (S) is defined as $\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\hat{y}^j|[\hat{X}^j; m^j]] - \mathbb{1}[\hat{y}^j|m^j]$, where N denotes input memes in the test set. The leakage-adjusted simulatability is measured by grouping N into two groups, i.e. leakage (N_1) and non-leakage (N_2) groups. In the leakage group, we opt for the test examples for which the explanation alone can predict the model outcome, such that $\mathbb{1}[\hat{y}^j|\hat{X}^j] = 1$. Similarly, the non-leakage group contains examples where the explanation is not sufficient alone for predicting the model outcome, such that, $\mathbb{1}[\hat{y}^j|\hat{X}^j] = 0$. The LAS score is the average of the simulatability scores (S_1 and S_2) across these two groups, where:

$$S_1 = \frac{1}{N_1}\sum_{j=1}^{N_1} \mathbb{1}[\hat{y}^j|[\hat{X}^j; m^j]] - \mathbb{1}[\hat{y}^j|m^j] \tag{15}$$

and

$$S_2 = \frac{1}{N_2} \sum_{j=1}^{N_2} \mathbb{1}[\hat{y}^j | [\hat{X}^j; m^j]] - \mathbb{1}[\hat{y}^j | m^j] \quad (16)$$

The LAS [Hase *et al.*, 2020] score is then defined as $LAS = \frac{1}{2}(S_1 + S_2)$

Essentially, Leakage-adjusted simulatability (LAS) measures how much the retrieved keywords contribute to predicting a model's original output while considering any potential direct leakage of the output class through these keywords. A higher LAS value indicates that the retrieved keywords are more interpretable when combined with the input, even when they alone are ineffective in predicting the output class.

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

In essence, a lower Sufficiency score indicates that the retrieved keywords contribute more effectively to predicting the model's class, making them more interpretable. Conversely, a higher Comprehensiveness score suggests that using retrieved keywords along with the meme input enhances the model's confidence in predicting the class compared to using the meme input alone. Therefore, lower Sufficiency and higher Comprehensiveness scores indicate better interpretability of the framework.

J Limitations

1. Task Agnosticism and Generalizability: While our research paper presents a task-agnostic method, it is crucial to acknowledge that the extent of its task-agnostic nature remains to be fully explored. The current study focuses on a specific task, and the generalizability of the proposed method to a broader range of tasks is an open question. Future research should investigate the applicability and effectiveness of the proposed approach across diverse tasks to establish its true task-agnostic capabilities.

2. Simple Keywords and Natural Language Output: The paper employs simple keywords for interpreting model decisions. While these keywords provide a foundational understanding, more in-depth exploration is necessary to refine the language and generate *precise natural language sentences* that effectively convey the nuances of the proposed model and the input meme used. Subsequent work should involve a detailed analysis and improvement of the language used, ensuring that the conveyed message is accurate, clear, and well-structured in the form of a grammatically correct sentence.

References

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

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

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Pramanick *et al.*, 2021] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Radford *et al.*, 2019] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.