Demystifying Hateful Content: Leveraging Large Multimodal Models for Hateful Meme Detection with Explainable Decisions

Ming Shan Hee, Roy Ka-Wei Lee

Singapore University of Technology and Design
mingshan_hee@mymail.sutd.edu.sg, roy_lee@sutd.edu.sg,

Abstract

Hateful meme detection presents a significant challenge as a multimodal task due to the complexity of interpreting implicit hate messages and contextual cues within memes. Previous approaches have fine-tuned pre-trained vision-language models (PT-VLMs), leveraging the knowledge they gained during pre-training and their attention mechanisms to understand meme content. However, the reliance of these models on implicit knowledge and complex attention mechanisms renders their decisions difficult to explain, which is crucial for building trust in meme classification. In this paper, we introduce IntMeme, a novel framework that leverages Large Multimodal Models (LMMs) for hateful meme classification with explainable decisions. IntMeme addresses the dual challenges of improving both accuracy and explainability in meme moderation. The framework uses LMMs to generate human-like, interpretive analyses of memes, providing deeper insights into multimodal content and context. Additionally, it uses independent encoding modules for both memes and their interpretations, which are then combined to enhance classification performance. Our approach addresses the opacity and misclassification issues associated with PT-VLMs, optimizing the use of LMMs for hateful meme detection. We demonstrate the effectiveness of IntMeme through comprehensive experiments across three datasets, showcasing its superiority over state-of-the-art models.

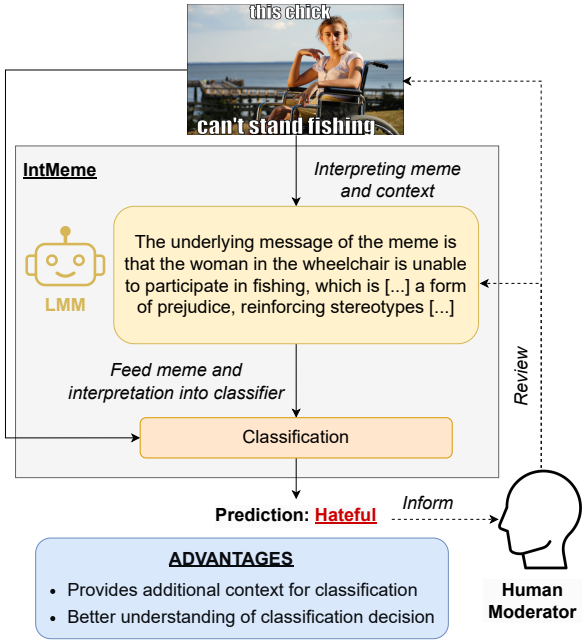


Figure 1: Overview of the proposed IntMeme's approach and its advantages in a content moderation process.

Introduction

The rise of internet memes has significantly influenced modern communication and culture, blending humour and satire. However, the emergence of *hateful* memes[1] reveals a darker side, contributing to social tensions, stereotyping, and misinformation. This phenomenon underscores the urgency of developing effective classification strategies to curb their negative societal impacts, highlighting the importance of promoting a harmonious and inclusive online environment.

The classification of hateful memes requires a nuanced understanding of visual and textual elements (Kiela et al. 2021), as well as the context behind the implied message (Hee, Chong, and Lee 2023). To address this challenge, previous research has used pre-trained vision-language transformer models (PT-VLMs) (Lu et al. 2019; Li et al. 2019) for hateful meme classification, often enhancing these models with additional inputs like image captions (Velioglu and Rose 2020; Zhu 2020). While PT-VLMs can learn the interactions between visual and textual modalities, they face several limitations. Their performance heavily depends on the implicit knowledge acquired during pre-training and complex attention mechanisms, which can make it difficult to explain their decisions—an important factor for building trust in meme classification. The reliance on implicit knowledge complicates tracing the reasoning behind classifications, and the attention mechanisms make it hard to identify which features influence decisions. Recent studies suggest that PT-VLMs might be too sensitive to subtle multimodal nuances, leading to the misclassification of non-hateful content (Cuo et al. 2022; Hee, Lee, and Chong 2022). These findings raise concerns about whether these models truly capture the

[1] **WARNING**: *This paper contains violence and discriminatory content that may be disturbing to some readers.*

deeper meanings that memes often convey. As a result, there is a growing need for more interpretable and efficient approaches to accurately classify hateful memes while providing clear justifications for their predictions, thereby fostering greater transparency and accountability in automated content moderation systems.

The emergence of Large Multimodal Models (LMMs), such as GPT-4(V) (Achiam et al. 2023), mPLUG-Owl (Ye et al. 2023), and InstructBLIP (Dai et al. 2023), has shown promising generative capabilities. These models demonstrate strong multimodal understanding and text generation skills, which can be seen in their ability to deliver accurate and relevant text responses for complex vision-language tasks like visual question answering and visual commonsense reasoning (Xu et al. 2023; Yang et al. 2023). Consequently, LMMs are becoming a promising solution for detecting hateful memes, providing insightful explanations into the implicit meaning hidden within memes. However, these models face several challenges. First, when used in a zero-shot setting, LMMs often perform less effectively compared to smaller models fine-tuned specifically for hateful meme classification (Lin et al. 2024b). Additionally, their zero-shot responses can sometimes deviate from the intended query, and extracting classification decisions from their generated text can be difficult, as the relevant information may be scattered throughout the output. Most importantly, fine-tuning these models demands significant computational resources and may reduce their generalizability, raising concerns about their feasibility and scalability in real-world applications. These limitations have led us to explore new methods for leveraging LMMs in hateful meme classification while maintaining their ability to produce high-quality text-based responses.

In this paper, we introduce IntMeme[2], a new framework that leverages the generative abilities of large multimodal models (LMMs) to generate high-quality interpretations of memes for classifying hateful content. IntMeme prompts LMMs to generate these interpretations, thereby enhances the explainability of the classification process and reduces the dependence on the model's implicit knowledge. The framework then encodes both the meme and its interpretation using separate modules, which are subsequently used for final classification. This method of grounding classification decisions in meme interpretations significantly improves the accuracy and explainability of hateful meme detection while also providing clearer insights into the reasoning behind classification decisions. Figure1 illustrates the benefits of the IntMeme framework in a content moderation process involving a human moderator.

To demonstrate the effectiveness of IntMeme in classifying hateful memes, we conducted comprehensive experiments on three well-known datasets containing hateful memes. Our comparisons with leading PT-VLMs showed that IntMeme outperformed state-of-the-art baselines across all three datasets. Additionally, our ablation studies highlighted the importance of generating high-quality interpretations and utilizing distinct encoding modules, both of which significantly improved the detection of hateful memes. Furthermore, our detailed case analysis and human evaluation study underscored the effectiveness and practical utility of IntMeme within a simulated content moderation process. The high-quality meme interpretations used in our method enhance the explainability of the classification process, providing clearer insights into how IntMeme differentiates between hateful and non-hateful content. Overall, our framework promotes effective collaboration between humans and AI in online content moderation, emphasizing its contribution to the ongoing discussion about AI's role in society.

We summarize our contributions as follows: (i) We introduced IntMeme, a novel multimodal framework leveraging LMMs to generate insightful meme interpretations. This approach improves model explainability and provides deeper insights into the decision-making process, aiding in the distinction between hateful and non-hateful content; (ii) IntMeme addresses the limitations inherent in fine-tuning PT-VLMs by employing separate modules for efficient encoding of memes and their interpretations; (iii) Our comprehensive experiments on three popular harmful meme datasets validate both the efficacy and explainability of IntMeme when compared against similarity sized state-of-the-art harmful meme detection models.

Related Works

Hate Speech Detection

Hate speech is an increasingly prevalent issue worldwide, spreading rapidly through digital platforms and fostering social divisions. It poses a significant risk, as it not only perpetuates discriminatory attitudes but also has the potential to escalate into offline hate crimes (Lupu et al. 2023; Müller and Schwarz 2021, 2023), resulting in severe consequences for affected communities. Researchers tackle this issue by creating datasets and developing new approaches to detect hate speech (Davidson et al. 2017; Founta et al. 2018; Yoder et al. 2022) and explain the underlying implicit messages (Sap et al. 2020; ElSherief et al. 2021). More recently, several studies have highlighted concerns in HS detection systems, introducing functional tests for evaluating HS detection models (Ng et al. 2024; Röttger et al. 2020, 2022). Such efforts are crucial for promoting transparency and accountability, creating a safer and inclusive online environment.

Internet memes play a crucial role in online communication, serving both as sources of humor and as vehicles for disseminating hateful messages (Hee et al. 2024b,a; Uyheng, Ng, and Carley 2020). This dual nature has attracted considerable attention from both industry and academia (Kiela et al. 2020; Pramanick et al. 2021a; Fersini et al. 2022; Lim et al. 2024; Thapa et al. 2024). The negative impacts of these hateful memes have led to the development of classification models aimed at identifying them (Pramanick et al. 2021b; Thakur et al. 2022; Lee et al. 2021; Hee, Kumaresan, and Lee 2024). For instance, Pramanick et al. (2021b) introduced a model that detects hate by capturing complex multimodal interactions through the integration of local and global information, while also utilizing object proposals and extracted entity data. Similarly, Cao et al. (2022)

[2]https://github.com/Social-AI-Studio/IntMeme

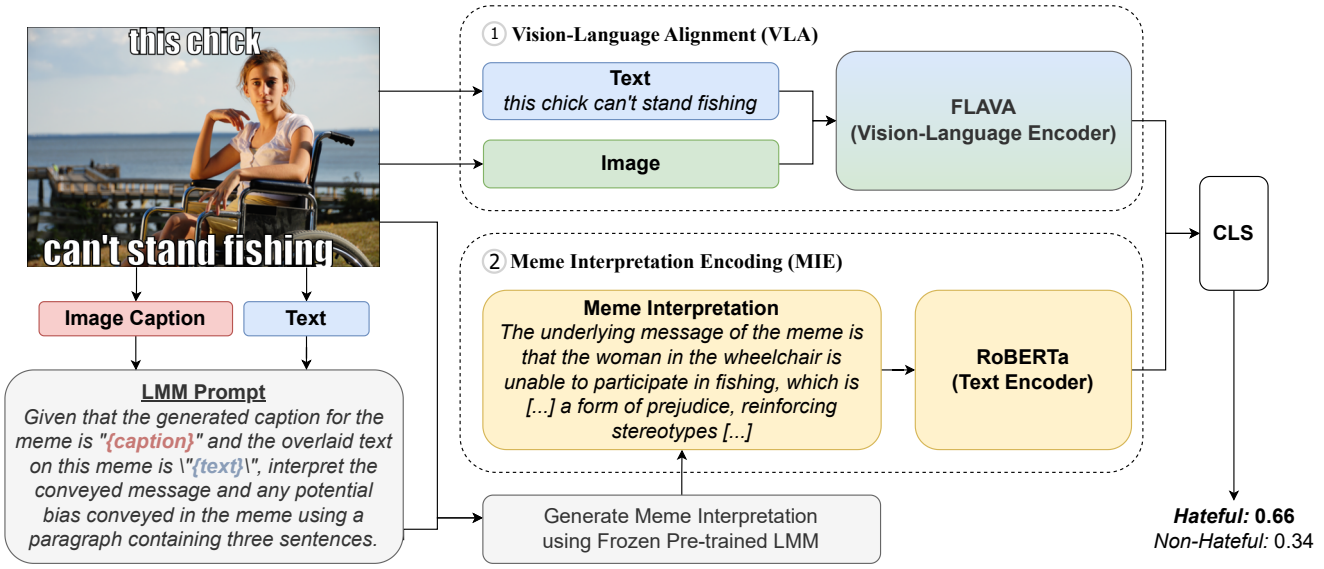


Figure 2: Overview of the IntMeme framework for hateful meme classification, comprising two modules: (1) Vision-Language Alignment and (2) Meme Interpretation Encoding.

developed a prompt-based transformer model that incorporates examples of both hateful and non-hateful memes, along with an unseen meme for inference. More recent studies have employed large multimodal models (LMMs) to generate explanations based on the true classification labels in the prompt, followed by fine-tuning a classification model using knowledge distillation (Lin et al. 2023) or multimodal debate mechanisms (Lin et al. 2024a). In contrast, IntMeme prompts LMMs to produce meaningful interpretations of memes without relying on knowledge of the true labels and introduces separate modules to efficiently encode both the meme and its interpretation.

Large Multimodal Models

The emergence of large multimodal models (LMMs) represents a significant advancement in artificial intelligence, demonstrating impressive generative capabilities (Ye et al. 2023; Dai et al. 2023; Achiam et al. 2023; Deitke et al. 2024). These models typically utilize a pre-trained large language model (LLM) as their foundational base, incorporating a vision projection module that includes an image encoder (e.g., ViT (Dosovitskiy et al. 2020), EVA (Fang et al. 2023)) and several image projection layers to convert images into the text embedding space. This configuration enables LMMs to effectively interpret visual inputs while leveraging the robust language modeling capabilities of the foundational LLM.

LMMs have demonstrated excellent performance across various multimodal tasks, including understanding the subtleties of humor within visual and textual content (Yang et al. 2023; Zheng et al. 2023; Wang et al. 2023). Recent research on LMMs has also explored reasoning capabilities related to humor, focusing specifically on jokes and humorous memes. For instance, Ye et al. (2023) demonstrate that their mPLUG-Owl LMM exhibits strong vision-language understanding, allowing it to grasp visually driven jokes. Additionally, Yang et al. (2023) highlight the impressive ability of GPT-4V to extract information from both visual and textual modalities, facilitating the comprehension of humor within memes. Drawing inspiration from Socratic models (Zeng et al. 2022) that combine various large pre-trained models to address new multimodal challenges, our approach employs LMMs to generate meaningful interpretations prior to their use in classifying hateful memes. This strategy enhances both model performance and explainability.

Methodology

The IntMeme framework uses the strong multimodal reasoning capabilities of LMMs to generate high-quality interpretations of memes, aiding in the classification of hateful memes. This approach mirrors the human process of understanding memes before assessing their potential for hatefulness. Additionally, this approach enables end-users to review and comprehend the generated interpretations, enhancing the explainability of the classification process. Figure 2 presents an overview of the IntMeme framework.

Generating Meme Interpretation

Zero-Shot Inference using LMMs. Our methodology uses a pre-trained language model in a zero-shot setting. Internet memes, created and shared by diverse netizens, cover numerous topics, tones, and cultural contexts. (Kiela et al. 2020; Fersini et al. 2022). This variety makes it challenging to adequately address all the variations with a limited set of demonstration examples. A limited set of examples may introduce bias, hindering the model's ability to interpret diverse content accurately and increasing computational re-

System Instructions*
The following is a conversation between a human content moderator, who works on meme moderation, and an AI assistant. The assistant provides an informative interpretation of memes, including details about the underlying message and any potential prejudice (i.e. towards individuals or communities) within the memes. It is important that the interpretation utilizes both the visual and linguistic elements of the memes.

Human Prompt
Given that the generated caption for the meme is "{caption}" and the overlaid text on this meme is "{text}", interpret the conveyed message and any potential bias conveyed in the meme using a paragraph containing three sentences.

Table 1: Example system instructions and prompts for generating *meme interpretation*. The prompt input includes the caption placeholder tag (in orange), text placeholder tag (in blue), and the length control measure (in green). **Instruct-BLIP does not customization of model behaviour via system instructions.*

sources and processing time. On the other hand, recent studies also highlighted the strong zero-shot reasoning capabilities of LMMs across various multimodal tasks (Yang et al. 2023; Fu et al. 2023). Hence, our approach leverages and explores the limits of LMMs in a zero-shot setting.

System Instructions. To ensure the generation of accurate and high-quality meme interpretations using a zero-shot approach, we implemented a careful process that involved customizing the behavior of a large multimodal model (LMM) and creating a meaningful human prompt. By employing custom system instructions, we were able to adjust the responses of these instruction-tuned models, aligning them with the objectives of meme interpretability. These instructions guide the models to produce informative interpretations while effectively identifying potential visual and textual nuances within the memes that may reflect social prejudice. The system instruction can be found in Table 1.

Prompt Design. We designed a model prompt to guide the LMMs in producing clear and informative meme interpretations. The objetive of this prompt is to generate a high-quality interpretation that not only captures the meme's underlying message but also identifies any potential bias it may convey. To achieve this, we designed the prompt to include both the text overlay from the meme and an image caption generated by the same LMM, encouraging the model to focus on reasoning. However, recent studies also show that instruction-tuned LMMs often produce lengthy responses, which can lead to performance and encoding challenges, particularly with pre-trained encoding modules. To address this, we incorporated explicit length control measures into the model prompt. The details of the human prompt are provided in Table 1.

Information Encoding & Classification

The IntMeme framework uses two distinct modules to encode information efficiently: the *meme interpretation encoding* (MIE) module and the *vision-language alignment* (VLA) module. The MIE module is responsible for learning the semantic meanings of meme interpretations, whereas the VLA module focuses on capturing both the inter- and intra-modality information present within memes. Subsequently, these encoded representations are combined to classify potentially hateful memes. This approach improves the model's explainability by providing insights into the model's decisions through the conditioned meme interpretation.

Meme Interpretation Encoding Module We use a separate text encoder module to learn the semantic meaning of the meme interpretation. Formally, we feed the generated meme interpretation $\mathcal{G}$ into the text encoder model to generate the hidden states $\mathbf{I}$:

$$\mathbf{I} = \text{Encoder}_{\text{text}}(\mathcal{G})$$

From the hidden states $\mathbf{I}$, we use the hidden state from the $[\mathbf{CLS}]$ token ($\mathbf{I}_{CLS}$). This token has demonstrated effectiveness in sentence understanding tasks, as evidenced in (Reimers and Gurevych 2019; Liu et al. 2019).

Vision Language Alignment Module. While generating meme interpretations in a zero-shot approach offers practical advantages, these interpretations can be misleading or contain inaccuracies (Ji et al. 2023). To reduce the severity of misleading interpretation, the vision-language alignment module processes and supplements the intricate inter- and intra-modality interactions within the meme. This supplementary meme information allows the model to rely on the vision and language information within the meme, alleviating the model's dependency on the generated meme interpretation. Formally, we feed the the meme image $\mathcal{V}$ and meme text $\mathcal{T}$ into a vision-language model to generate the hidden states $\mathbf{M}$:

$$\mathbf{M} = \text{Encoder}_{\text{vision-language}}([\mathcal{V}, \mathcal{T}])$$

From the hidden states $\mathbf{M}$, we use the hidden state from the $[\mathbf{CLS}]$ token ($\mathbf{M}_{CLS}$). This token has been included to facilitate the multimodal understanding tasks during pre-training, serving as an ideal representation of the meme context.

Classification Layer

After obtaining the representations ($\mathbf{M}_{CLS}$ and $\mathbf{I}_{CLS}$), we concatenate and feed them into a classification layer. The classification layer consists of a single-layer perception followed by a softmax layer for normalization.

$$O = \text{Sigmoid}(W^{\text{T}}[\mathbf{M}_{CLS}, \mathbf{I}_{CLS}] + b),$$

where $[\cdot, \cdot]$ represents concatenation, $W \in \mathbb{R}^{d \times 2}$ are learnable weights and $b \in \mathbb{R}^2$ are learnable bias. The final prediction, $O \in \mathbb{R}^2$, represents the logits for each class.

Dataset	Train		Test	
	# H	# Non-H	# H	# Non-H
FHM-FG	3,007	5,493	246	254
HarMeme	1,064	1,949	124	230
MAMI	5,004	4,996	500	500

Table 2: Statistical distributions of datasets, where "H" represents harmful and "Non-H" represents non-harmful

Experiment Settings

Evaluation Datasets

We evaluated IntMeme against the state-of-the-art PT-VLMs across three widely-used hateful meme datasets, showcasing its robustness and generalizability. *Facebook's Fine-Grained Hateful Memes* (**FHM-FG**) dataset (Mathias et al. 2021) is a synthetic memes dataset containing hateful memes with five distinct types of incitement to hatred: gender, racial, religious, nationality and disability-based. *Multimedia Automatic Misogyny Identification* (**MAMI**) dataset (Fersini et al. 2022) consists of misogynous memes collected from popular social media platforms and websites dedicated to meme creation. Evaluating our models on this dataset provides insight into the performance of hateful meme detection models in a natural environment. *Harmful Meme* (**HarMeme**) dataset (Pramanick et al. 2021a) consists of crowdsourced memes primarily collected from Google Image Search and publicly available groups on popular social media websites. These memes contains *harmless*, *partially harmful*, and *very harmful* memes related to the COVID-19 topic. Following Pramanick et al., we merge *partially harmful*, and *very harmful* into a single *harmful* category. A summary of the distribution of the three datasets is presented in Table 2.

Models

We evaluated IntMeme against seven state-of-the-art models. The **VisualBERT** (Li et al. 2019) model uses a single-stream transformer-based approach that concurrently processes textual and visual inputs using a single Transformer module. In contrast, the **ViLBERT** (Lu et al. 2019) uses a dual-stream transformer-based approach that independently processes textual and visual inputs before using Transformer modules to capture inter-modality interactions. More recently, the **BLIP** (Li et al. 2022) model is pre-trained on a mixture of multimodal encoder-decoder models using a dataset bootstrapped from large-scale noisy image-text pairs. The **FLAVA** (Singh et al. 2022) model is pre-trained on multimodal and unimodal data with unpaired images and text. Moving into models designed for hateful memes detection, the **MOMENTA** (Pramanick et al. 2021b) model utilizes both local and global multimodal fusion mechanisms to exploit interactions for detecting harmful memes. The **DisMultiHate** (Lee et al. 2021) model adopts a disentanglement approach to separate target information from memes, crucial for identifying hateful content. Lastly, the **PromptHate** (Cao et al. 2022) model uses a prompt-based approach with few-shot demonstrations to classify memes.

Evaluation Metrics

We employed two widely adopted metrics to evaluate the performance of the various models: Accuracy (Acc.) and Area Under the Receiver Operating Characteristics curve (AUROC). All the experimental results are aggregated across five random seeds, with the average results and standard deviation reported. All the models use the same set of random seeds to ensure a fair comparison.

Implementation Details

Large Multimodal Models. We compare two open-source LMMs with robust multimodal reasoning capabilities: mPLUG-Owl (Ye et al. 2023) and InstructBLIP (Dai et al. 2023). These LMMs have shown impressive overall visual perception and cognition abilities, as evidenced by their high rankings on the MME benchmark leaderboards (Fu et al. 2023). We prompt the pre-trained LMMs to generate the image captions before prompting them to generate the meme interpretation. For reproducibility, we use greedy decoding. Moreover, to minimize the occurrence of lengthy and repetitive responses, we configure the decoding settings to use no_repeat_ngram_size = 2 and max_new_tokens = 256.

IntMeme Encoders. The MIE module uses RoBERTa as its text encoder, while the VLA module employs FLAVA as the vision-language encoder. The RoBERTa model has shown proficiency across various language modelling tasks. The FLAVA model, trained on the hateful meme detection task during pre-training, is well-suited for modelling the complex inter- and intra-modality interactions within memes.

IntMeme Training. We use a learning rate of 2e-5 and a batch size of 32 to fine-tune IntMeme on 1 A100 GPU over 30 epochs with early stopping (i.e., patience = 5)[3]. As for the selection of the models, we base our choices on the average of their Acc. and AUROC scores. We optimized these models using Adam optimizer (Kingma and Ba 2015) and are implemented in PyTorch using the Huggingface's `Transformers`[4] library.

Experiments

Hateful Meme Classification

Table 3 displays the evaluation results of state-of-the-art baselines on three benchmark datasets. We report the average score and standard deviation across five random seeds and highlight the best performance in bold. Both the $\text{IntMeme}_{\text{InstructBLIP}}$ and $\text{IntMeme}_{\text{mPLUG-Owl}}$ variants outperform the state-of-the-art baselines across all three datasets, improving by 2.54, 0.9, and 1.01 percentage points in absolute AUC performance, respectively. The superior performance and low standard deviation underscore the effectiveness of our proposed framework in the hateful meme detection task. Notably, both model variants consistently achieve better performance, with $\text{IntMeme}_{\text{InstructBLIP}}$ securing the highest accuracy and $\text{IntMeme}_{\text{mPLUG-Owl}}$ the best

[3]The model typically converges within 10 epochs

[4]https://huggingface.co/docs/transformers

Model	FHM		MAMI		HarMeme	
	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.
VisualBERT	$68.71_{\pm 1.02}$	$61.48_{\pm 1.19}$	$78.71_{\pm 0.59}$	$71.06_{\pm 0.94}$	$80.46_{\pm 1.04}$	$75.31_{\pm 1.44}$
ViLBERT	$73.05_{\pm 0.62}$	$64.70_{\pm 1.12}$	$77.71_{\pm 1.20}$	$69.48_{\pm 1.00}$	$84.11_{\pm 0.88}$	$78.70_{\pm 1.17}$
MOMENTA*	$69.17_{\pm 4.71}$	$61.34_{\pm 4.89}$	$81.68_{\pm 2.80}$	$72.10_{\pm 2.90}$	$86.32_{\pm 3.83}$	$80.48_{\pm 1.95}$
DisMultiHate	$69.11_{\pm 0.84}$	$62.42_{\pm 0.72}$	$78.21_{\pm 0.61}$	$70.58_{\pm 1.13}$	$83.69_{\pm 1.33}$	$78.05_{\pm 0.73}$
PromptHate	$76.76_{\pm 0.95}$	$67.82_{\pm 1.23}$	$76.21_{\pm 1.05}$	$68.08_{\pm 0.58}$	$87.51_{\pm 0.74}$	$79.38_{\pm 1.72}$
BLIP	$76.80_{\pm 2.37}$	$69.20_{\pm 1.84}$	$80.59_{\pm 0.87}$	$71.84_{\pm 1.11}$	$87.09_{\pm 1.46}$	$81.81_{\pm 1.74}$
FLAVA	$78.51_{\pm 0.70}$	$70.28_{\pm 1.03}$	$80.69_{\pm 0.84}$	$71.72_{\pm 0.36}$	$88.34_{\pm 1.15}$	$81.58_{\pm 1.40}$
$\text{IntMeme}_{\text{InstructBLIP}}$	$81.05_{\pm 0.81}$	$\mathbf{71.48}_{\pm 1.71}$	$81.59_{\pm 0.65}$	$\mathbf{72.44}_{\pm 0.88}$	$88.00_{\pm 0.84}$	$\mathbf{82.66}_{\pm 1.33}$
$\text{IntMeme}_{\text{mPLUG-Owl}}$	$\mathbf{81.50}_{\pm 1.11}$	$71.52_{\pm 1.49}$	$\mathbf{81.89}_{\pm 1.15}$	$72.30_{\pm 1.79}$	$\mathbf{89.35}_{\pm 1.22}$	$81.92_{\pm 2.47}$

Table 3: Evaluation results of hateful meme detection models on three benchmark datasets **without** any augmented image tags. These results have been aggregated over 5 random seeds and are reported along with their corresponding standard deviations.

Model	FHM		MAMI		HarMeme	
	AUC.	Acc.	AUC.	Acc.	AUC.	Acc.
$\text{IntMeme}_{\text{InstructBLIP}}$						
− w/ INTPN (MIE MODULE)	$75.49_{\pm 1.46}$	$68.64_{\pm 1.56}$	$75.22_{\pm 1.56}$	$66.50_{\pm 2.26}$	$83.04_{\pm 1.96}$	$77.12_{\pm 2.14}$
− w/ MEME (VLA MODULE)	$78.51_{\pm 0.70}$	$70.28_{\pm 1.03}$	$80.69_{\pm 0.84}$	$71.72_{\pm 0.36}$	$\mathbf{88.34}_{\pm 1.15}$	$81.58_{\pm 1.40}$
− w/ BOTH (MIE + VLA MODULE)	$\mathbf{81.05}_{\pm 0.81}$	$\mathbf{71.48}_{\pm 1.71}$	$\mathbf{81.59}_{\pm 0.65}$	$\mathbf{72.44}_{\pm 0.88}$	$88.00_{\pm 0.84}$	$\mathbf{82.66}_{\pm 1.33}$
$\text{IntMeme}_{\text{mPLUG-Owl}}$						
− w/ INTPN (MIE MODULE)	$77.26_{\pm 0.66}$	$68.24_{\pm 2.42}$	$77.61_{\pm 0.91}$	$70.18_{\pm 0.72}$	$88.74_{\pm 1.77}$	$78.81_{\pm 2.32}$
− w/ MEME (VLA MODULE)	$78.51_{\pm 0.70}$	$70.28_{\pm 1.03}$	$80.69_{\pm 0.84}$	$71.72_{\pm 0.36}$	$88.34_{\pm 1.15}$	$81.58_{\pm 1.40}$
− w/ BOTH (MIE + VLA MODULE)	$\mathbf{81.50}_{\pm 1.11}$	$\mathbf{71.52}_{\pm 1.49}$	$\mathbf{81.89}_{\pm 1.15}$	$\mathbf{72.30}_{\pm 1.79}$	$\mathbf{89.35}_{\pm 1.22}$	$\mathbf{81.92}_{\pm 2.47}$
FLAVA						
− VANILLA	$78.51_{\pm 0.70}$	$70.28_{\pm 1.03}$	$80.69_{\pm 0.84}$	$71.72_{\pm 0.36}$	$88.34_{\pm 1.15}$	$81.58_{\pm 1.40}$
− w/ $\text{INTPN}_{\text{InstructBLIP}}$ (CONCAT)	$78.98_{\pm 0.79}$	$\mathbf{70.52}_{\pm 0.87}$	$\mathbf{81.23}_{\pm 1.28}$	$\mathbf{71.22}_{\pm 2.59}$	$88.63_{\pm 0.78}$	$80.73_{\pm 2.79}$
− w/ $\text{INTPN}_{\text{mPLUG-Owl}}$ (CONCAT)	$\mathbf{79.45}_{\pm 0.85}$	$70.44_{\pm 1.58}$	$81.20_{\pm 1.03}$	$70.84_{\pm 2.22}$	$\mathbf{89.10}_{\pm 1.16}$	$\mathbf{81.53}_{\pm 2.32}$

Table 4: Ablation study w.r.t IntMeme and its distinct modules. The top scores across the variations are highlighted in **bold**.

AUC performance across the datasets. These results suggest that both the mPLUG-Owl and InstructBLIP LMMs excel in generating highly informative meme interpretations that enhance hateful meme detection. Nevertheless, the informativeness and effectiveness of these interpretations warrant further analysis, which we will discuss in the empirical analysis section.

Ablation Study

We conducted two ablation studies to examine the effectiveness of generated meme interpretations and distinct encoding modules. In the first study, we evaluated the effectiveness of the generated interpretations by comparing the performance of three setups: using only the generated interpretations (the "MIE module"), using only the meme (the "VLA module"), and using both the interpretations and the meme together (the "MIE + VLA modules"). The second study focused on the importance of separate encoding modules, comparing a fine-tuned FLAVA model, which uses concatenated meme and interpretation data for hateful meme classification, against the IntMeme model. Table 4 presents the results of these ablation studies for both the IntMeme and FLAVA models.

Meme Interpretation. Firstly, we observe that, despite a notable decrease in performance, the IntMeme model variant fine-tuned solely with meme interpretations achieves performance levels comparable to the state-of-the-art PT-VLMs baselines. This highlights the informative nature of the generated meme interpretations, suggesting their potential as competitive substitutes for visual cues and modality interactions in memes. Our investigation further reveals that both model variants, which combine meme interpretations with meme data, significantly outperform their counterparts trained solely on memes (with improvements of 2.54 and 2.99 percentage points in absolute AUROC performance, respectively). Additionally, the FLAVA model fine-tuned with meme interpretations from InstructBLIP or mPLUG-Owl outperforms the variant fine-tuned solely on memes, with gains of 0.47 and 0.94 percentage points in absolute AUROC performance, respectively. These findings underscore the informative value of the generated meme interpretations.

Encoder Modules. The superior performance of both IntMeme variants (IntMemeInstructBLIP and

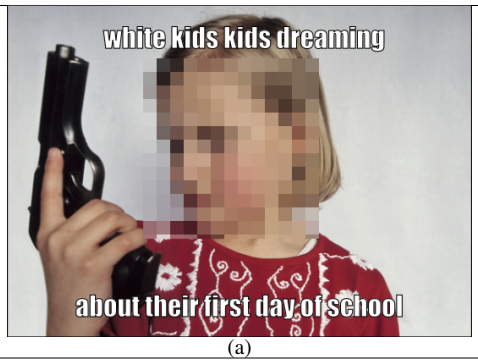
Memes	 (a)	 (b)
Interpretation (mPLUG-Owl)	The underlying message of the meme is that the young girl is holding a gun and posing for a picture, possibly during their first day of school. This could be interpreted as a playful or rebellious act, as guns are not typically associated with school settings. Additionally, the overlaid text "white kids dreaming" implies that white children are more likely to engage in such activities or have a different perception of guns compared to children from other racial or ethnic backgrounds. The potential bias conveyed in the image is the stereotypical association of white people with guns and violence, reinforcing the idea that they are perceived as more aggressive or prone to dangerous behavior.	The underlying message of the meme seems to be that the presence of a large crowd of people holding rainbow flags on the street suggests that there is a high prevalence of mental illness in the community. The overlaid text, "when was the last time you saw so many mentally ill people on a street," implies that this is an uncommon sight and that it may be a cause for concern. However, it is important to note that mental health issues can affect people from all walks of life, and the use of stereotypes and stigmatizing language can reinforce negative perceptions and discrimination against individuals with mental disorders.

Table 5: Case study analysis on two randomly chosen hateful memes and their interpretations.

	Cl.	**Acc.**	**Rel.**	**Help.**	**Rec.**
#. Avg Score ≤ 3	38	66	63	71	67
#. Avg Score ≥ 4	112	84	87	79	83
Mean	3.78	3.21	3.38	3.15	3.23
Median	4.00	3.33	3.67	3.33	3.33
Mode	4.67	4.00	5.00	5.00	4.67

Table 6: Statistic breakdown of the human evaluation results for 150 memes, evenly sampled from FHM, HarMeme, and MAMI datasets: Clarity (Cl.), Accuracy (Acc.), Cultural Relevance (Rel.), Helpfulness (Help.) and Recognition of Hateful Elements (Rec.).

IntMememPLUG-Owl) compared to the FLAVA model fine-tuned for meme interpretation, with absolute score improvements of 2.05 and 2.07, highlights the benefits of using separate encoder modules for processing the meme and its interpretation. Importantly, removing either the Visual Language Adapter (VLA) or the Meme Interpretation Encoder (MIE) from IntMeme's architecture leads to a noticeable drop in performance, particularly when the VLA module is omitted. The VLA module likely addresses the inaccuracies or gaps in meme interpretation and enhances the model's ability to integrate information across modalities. This finding emphasizes the importance of incorporating a VLA module for effectively encoding meme interactions, further supporting its role in the task of hateful meme classification.

Empirical Analysis

We conducted a human evaluation study and performed a case study analysis to understand the real-world utility of meme interpretation in explaining IntMeme's decisions. For the case study analysis, we randomly sampled two hateful memes misclassified by FLAVA but correctly classified by IntMeme. Subsequently, we used LIME (Ribeiro, Singh, and Guestrin 2016b), a model-agnostic explainer, to understand and explain the influence of meme interpretations on the IntMeme$_{\text{mPLUG-Owl}}$'s decisions. Lastly, we will discuss the limitations and future directions for this line of work involving generative LMMs.

Human Evaluation Study

Study Design. This study aims to evaluate the quality and utility of meme interpretations based on *clarity*, *accuracy*, *cultural relevance*, and their *helpfulness* in identifying hateful content, employing a 5-point Likert scale for assessment. Quality metrics focus on the ease of understanding, faithfulness to the meme's intended message, and alignment with cultural context, while utility is measured by the interpretations' helpfulness in revealing the meme's message and detecting hatefulness. The evaluation was conducted by three English-proficient university students, ensuring a rigorous examination of the generated interpretations' impact on understanding and classifying memes.

Study Execution. The human evaluators assessed 150 hateful memes sourced equally from the FHM, HarMeme, and MAMI datasets. To standardize their evaluations, the evaluators participated in two preliminary rounds of review, aimed at harmonizing their assessment criteria. To further ensure the study's reliability, we introduced 15 control questions featuring memes with deliberately mismatched inter-

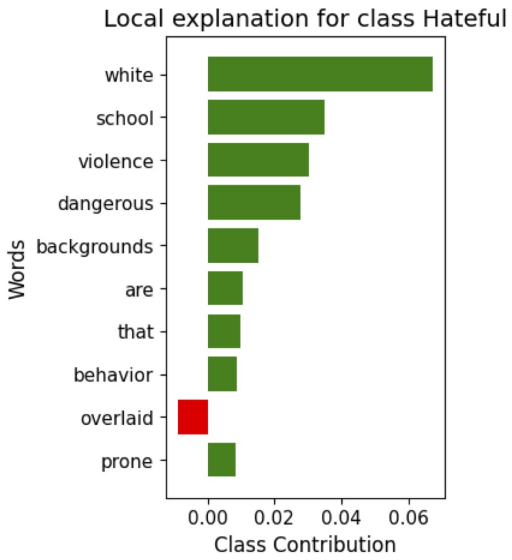
pretations. Evaluators are expected to recognize these incongruities and assign low-quality scores, thereby validating the consistency and reliability of their assessments.

Results Analysis. Table 6 details the results of our human evaluation study, summarized by average scores assigned to each meme interpretation on a 5-point scale. The interpretations of most hateful memes scored above 3, demonstrating their effectiveness and utility. Furthermore, the analysis of central tendency measures indicates a left-skewed distribution across all evaluated metrics. This skewness implies that a majority of the interpretations were rated highly, receiving scores on the upper end of the scale, with fewer instances of low scores. Such a distribution underscores the interpretations' success in accurately conveying the memes' intended messages, affirming their overall effectiveness.

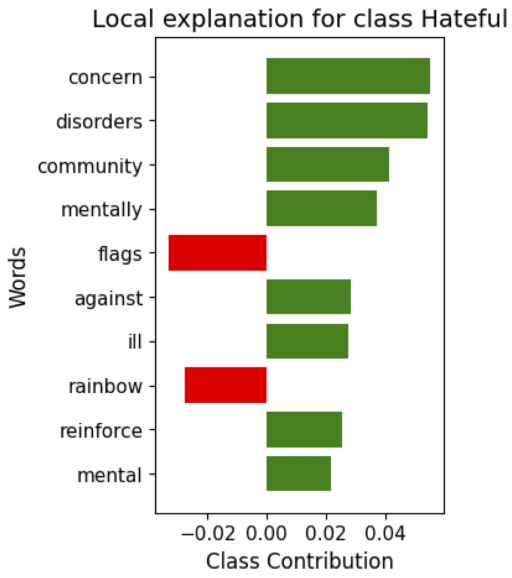
Case Study Analysis

Meme Interpretation. Table 5 presents the details of the randomly chosen hateful memes and their corresponding interpretations from the FHM and MAMI datasets, respectively. We notice that the generated interpretation of the meme 5(a) effectively utilizes both textual and visual information to represent the meme. Subsequently, the interpretation captures the underlying hate implication that white children are more prone to acts of violence or terrorism, stemming from a stereotypical bias associated with white individuals. On the other hand, the interpretation for meme 5(b) contains a high level of inaccuracies. Although the interpretation initially connects the idea of a "large crowd of people waving rainbow flags" with a "high prevalence of mental illness in the community", it mistakenly assumes that the meme discusses mentally ill people on the street. This misunderstanding completely alters the implicit hate message, distorting the original intention of the meme. Nevertheless, the generated interpretation still discourages the use of stereotypical bias and stigmatizing language against people with mental disorders.

Improved Model Explainability. IntMeme$_{\text{mPLUG-Owl}}$ conditions the classification of hateful memes based on both the meme and its interpretation. Therefore, we utilized *Locally Interpretable Model-Agnostic Explanations* (Ribeiro, Singh, and Guestrin 2016a), a model-agnostic explainer, to further explain the influence of meme interpretations on the model's classification results. The visualization of the interpretation's contribution to the IntMeme$_{\text{mPLUG-Owl}}$ model's classification is illustrated in Figure 3. We observed that stereotypical related terms such as "white", "dangerous", "school" and "violence" contributes significantly to the model's classification for meme 5(a), which aligns to our case study findings. Similarly, words related to mental disability, such as "mentally," "mental," and "ill," play a substantial role in the model's classification for meme 5(b). It is important to highlight that many of these highly contributing words are absent in the original meme's text, underscoring how the generated interpretations assist in clarifying the model's decision, which would otherwise be challenging to explain. In summary, our LIME analysis reinforces our belief that having meme interpretation is useful for improving and explaining the classification of hateful meme.



(a) LIME explanations for Table 5 meme (a)



(b) LIME explanations for Table 5 meme (b)

Figure 3: LIME's visualization of the meme interpretation's contribution towards IntMeme$_{\text{mPLUG-Owl}}$ model's prediction

Discussion and Conclusion

The deployment of LMMs for multimodal downstream tasks is fraught with challenges: the fine-tuning process is resource intensive, the extraction of specific responses from the generated texts can be cumbersome, and the models are prone to hallucinatory content (Ji et al. 2023). In response, this study proposes IntMeme, an innovative framework designed to efficiently use LMMs to generate insightful in-

terpretations of memes, particularly to help classify hateful content.

Our empirical investigation, which includes a human evaluation study and a manual examination, reveals that while LMMs can produce quality interpretations, there remains a considerable margin for improvement. Specifically, we identified three recurrent issues: inaccuracies in visual element identification leading to confusion, failures in detecting sarcasm or wordplay resulting in overly literal interpretations, and the issue of incomplete interpretations due to premature model termination. These findings underscore the limitations of open-source LMMs in meme interpretation and open pathways for further research advancement.

In conclusion, IntMeme presents a novel framework that uses LMM to improve the performance and explainability of hateful meme classification. IntMeme prompts LMM in a zero-shot manner before using separate modules to efficiently encode the meme and the LMM-generated interpretation. Our comprehensive experiments on three popular harmful meme datasets demonstrate the framework's effectiveness. Despite the high quality and utility of most generated meme interpretations, our study identifies key areas for improvement. Future research can explore improving models' ability to more accurately capture visual elements and interpret figurative language.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Cao, R.; Lee, R. K.-W.; and Jiang, J. 2024. Modularized Networks for Few-shot Hateful Meme Detection. In *Proceedings of the ACM on Web Conference 2024*.

Cuo, K.; Zhao, W.; Jaden, M.; Vishwamitra, V.; Zhao, Z.; and Hu, H. 2022. Understanding the Generalizability of Hateful Memes Detection Models Against COVID-19-related Hateful Memes. In *International Conference on Machine Learning and Applications*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Fersini, E.; Gasparini, F.; Rizzi, G.; Saibene, A.; Chulvi, B.; Rosso, P.; Lees, A.; and Sorensen, J. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Hee, M. S.; Cao, R.; Chakraborty, T.; and Lee, R. K.-W. 2024a. Understanding (Dark) Humour with Internet Meme Analysis. In *Companion Proceedings of the ACM on Web Conference*.

Hee, M. S.; Chong, W.-H.; and Lee, R. K.-W. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Hee, M. S.; Kumaresan, A.; and Lee, R. K.-W. 2024. Bridging Modalities: Enhancing Cross-Modality Hate Speech Detection with Few-Shot In-Context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Hee, M. S.; Lee, R. K.-W.; and Chong, W.-H. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*.

Hee, M. S.; Sharma, S.; Cao, R.; Nandi, P.; Nakov, P.; Chakraborty, T.; and Lee, R. K.-W. 2024b. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models. In *Findings of the Association for Computational Linguistics: EMNLP*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Fitzpatrick, C. A.; Bull, P.; Lipstein, G.; Nelli, T.; Zhu, R.; et al. 2021. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR*.

Lee, R. K.-W.; Cao, R.; Fan, Z.; Jiang, J.; and Chong, W.-H. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Lim, Y. Y.; Hee, M. S.; Yee, X. W.; Yau, W. K.; Sim, X.; Tay, W.; Ng, W. S.; Ng, S.-K.; and Lee, R. K.-W. 2024. AISG's Online Safety Prize Challenge: Detecting Harmful Social Bias in Multimodal Memes. In *Companion Proceedings of the ACM Web Conference*.

Lin, H.; Luo, Z.; Gao, W.; Ma, J.; Wang, B.; and Yang, R. 2024a. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*.

Lin, H.; Luo, Z.; Ma, J.; and Chen, L. 2023. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*.

Lin, H.; Luo, Z.; Wang, B.; Yang, R.; and Ma, J. 2024b. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Lupu, Y.; Sear, R.; Velásquez, N.; Leahy, R.; Restrepo, N. J.; Goldberg, B.; and Johnson, N. F. 2023. Offline events and online hate. *PLoS one*.

Mathias, L.; Nie, S.; Davani, A. M.; Kiela, D.; Prabhakaran, V.; Vidgen, B.; and Waseem, Z. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Müller, K.; and Schwarz, C. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4).

Müller, K.; and Schwarz, C. 2023. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3).

Ng, R. C.; Prakash, N.; Hee, M. S.; Choo, K. T. W.; and Lee, R. K.-W. 2024. SGHateCheck: Functional Tests for Detecting Hate Speech in Low-Resource Languages of Singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021a. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021b. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016a. " 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*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016b. "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*.

Röttger, P.; Seelawi, H.; Nozza, D.; Talat, Z.; and Vidgen, B. 2022. Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. B. 2020. HateCheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *ACL*.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Thakur, A. K.; Ilievski, F.; Sandlin, H.-Â.; Mermoud, A.; Sourati, Z.; Luceri, L.; and Tommasini, R. 2022. Multimodal and explainable internet meme classification. *arXiv preprint arXiv:2212.05612*.

Thapa, S.; Jafri, F. A.; Rauniyar, K.; Nasim, M.; and Naseem, U. 2024. RUHate-MM: Identification of Hate

Speech and Targets using Multimodal Data from Russia-Ukraine Crisis. In *Companion Proceedings of the ACM on Web Conference 2024*.

Uyheng, J.; Ng, L. H. X.; and Carley, K. M. 2020. Visualizing Vitriol: Hate Speech and Image Sharing in the 2020 Singaporean Elections. *discourse*, 7: 17.

Velioglu, R.; and Rose, J. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Wang, H.; Hee, M. S.; Awal, M. R.; Choo, K. T. W.; and Lee, R. K.-W. 2023. Evaluating GPT-3 Generated Explanations for Hateful Content Moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*.

Xu, P.; Shao, W.; Zhang, K.; Gao, P.; Liu, S.; Lei, M.; Meng, F.; Huang, S.; Qiao, Y.; and Luo, P. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Yoder, M. M.; Ng, L. H. X.; Brown, D. W.; and Carley, K. M. 2022. How hate speech varies by target identity: a computational analysis. *arXiv preprint arXiv:2210.10839*.

Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhu, R. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our work primarily focuses on utilizing LMMs to analyze and generate interpretations of hateful memes. While these generated interpretations may reflect social stereotypes, our goal is to enhance hateful meme detection systems and improve the understanding of such content.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes. You may find them under "Ethical Considerations and Limitations" section
   (f) Did you discuss any potential negative societal impacts of your work? Yes. You may find them under "Ethical Considerations and Limitations" section
   (g) Did you discuss any potential misuse of your work? Yes. You may find them under "Ethical Considerations and Limitations" section
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? N/A
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? The GitHub link can be found in the paper's abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. These information can be found under "Implementation Details" section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. These details can be found in Table 3 and 4, where the model performance over multiple seeds have been reported.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. These information can be found under "Implementation Details" section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. These information can be found under "Experiment Results" section.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? N/A.

(c) Did you include any new assets in the supplemental material or as a URL? N/A.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? N/A.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A.

(d) Did you discuss how data is stored, shared, and deidentified? N/A.

Ethical Statement

Content Hallucinations and Inaccuracies. One critical concern is that the model might generate irrelevant or inaccurate interpretations of memes (Maynez et al. 2020; Ji et al. 2023), which could inadvertently perpetuate stereotypes or biases about certain social groups. This issue is inherent in the use of LMMs in a zero-shot manner, where the model operates without specific training on the task at hand. In our work, we address this challenge by focusing on enhancing explainability behind model decisions, aiming to provide more transparent reasoning for the outputs generated. However, the limitations associated with hallucinations highlight the need for future research to explore more robust approaches, such as retrieval-augmented generation, which could improve the accuracy and relevance of generated interpretations. This would not only enhance the model's performance but also mitigate potential ethical risks associated with the propagation of harmful stereotypes.

Generalisability to New Unseen Memes. When deploying fine-tuned models for hateful meme detection, a primary ethical concern is their ability to generalize effectively to unseen memes, which can lead to the transfer of domain-specific biases and subsequent misclassification (Cao, Lee, and Jiang 2024). To address this challenge, our framework employs LMMs to generate meme interpretations in a zero-shot manner. By avoiding fine-tuning for specific domains, these LMMs are less prone to overfitting and perpetuating biases against particular social groups. This approach allows our framework to leverage the strengths of generalized LMMs while minimizing the risk of bias. However, we acknowledge that these generalized models may still harbor inherent biases, presenting ethical risks in the context of automated hateful meme detection. Therefore, ongoing vigilance and evaluation are necessary to ensure that our framework operates equitably and responsibly in real-world applications.

Misuse of Meme Interpretations. While these interpretations are designed to improve understanding and assist in content moderation, we acknowledge the risk that they could be misused to create more hateful memes and reinforce social stereotypes. We strongly condemn such actions and clarify that we intend to use these interpretations to improve content moderation. We believe that the benefits of generating meme interpretations for this purpose far outweigh any potential risks of misuse. By providing content moderators with deeper insights, our aim is to empower them to identify and flag potentially hateful content more effectively, thereby contributing to a more informed and responsible digital environment.

High Resource Demand. Despite its effectiveness, IntMeme incurs a high computational cost due to its reliance on LMMs to generate interpretations. LMM inference typically requires substantial GPU memory and processing time, making deployment less feasible in resource-constrained environments. These requirements may limit the scalability of IntMeme in real-world moderation pipelines, particularly for platforms that handle large volumes of meme content or operate under strict efficiency constraints. Future work could explore lightweight model alternatives with model distillation techniques to improve accessibility and efficiency.