I know what you MEME! Understanding and Detecting Harmful Memes with Multimodal Large Language Models

Yong Zhuang[†*], Keyan Guo[§*], Juan Wang[†‖], Yiheng Jing[†], Xiaoyang Xu[†],
Wenzhe Yi[†], Mengda Yang[†], Bo Zhao[†], Hongxin Hu[§]
[†]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University,
[§]University at Buffalo,
Email: [†]{yong.zhuang, jwang, yihengjing, xiaoyangx, wenzhey, mengday, zhaobo}@whu.edu.cn
[§]{keyanguo, hongxinh}@buffalo.edu

***Abstract*—Memes have become a double-edged sword on social media platforms. On one hand, they facilitate the rapid dissemination of information and enhance communication. On the other hand, memes pose a risk of spreading harmful content under the guise of humor and virality. This duality highlights the need to develop effective moderation tools capable of identifying harmful memes. Current detection methods, however, face significant challenges in identifying harmful memes due to their inherent complexity. This complexity arises from the diverse forms of expression, intricate compositions, sophisticated propaganda techniques, and varied cultural contexts in which memes are created and circulated. These factors make it difficult for existing algorithms to distinguish between harmless and harmful content accurately. To understand and address these challenges, we first conduct a comprehensive study on harmful memes from two novel perspectives: *visual arts* and *propaganda techniques*. It aims to assess existing tools for detecting harmful memes and understand the complexities inherent in them. Our findings demonstrate that meme compositions and propaganda techniques can significantly diminish the effectiveness of current harmful meme detection methods. Inspired by our observations and understanding of harmful memes, we propose a novel framework called HMGUARD for effective detection of harmful memes. HMGUARD utilizes adaptive prompting and chain-of-thought (CoT) reasoning in multimodal large language models (MLLMs). HMGUARD has demonstrated remarkable performance on the public harmful meme dataset, achieving an accuracy of 0.92. Compared to the baseline, HMGUARD represents a substantial improvement, with accuracy exceeding the baselines by 15% to 79.17%. Additionally, HMGUARD outperforms existing detection tools, achieving an impressive accuracy of 0.88 in real-world scenarios.**
Disclaimer. This paper contains harmful content, which has the potential to be offensive and may disturb readers.

I. INTRODUCTION

Memes have become a widely used and captivating medium on social media, often employed to disseminate ideas, cultures, trends, and events [1], [2]. They are now a prominent form of online expression, typically combining images and text to deliver messages in a concise, engaging, and impactful way. Memes on the Internet possess unique characteristics, such as susceptibility to parody, incorporation of intertextuality, viral propagation, and evolution over time [3], [4]. Online users utilize the characteristics of memes for humor or ridicule. However, there is a more concerning side: the same characteristics can be exploited by malicious individuals to create and spread memes containing explicit or implicit harmful content on social media, often evading detection [5]. For example, Fig. 1 illustrates a harmful meme that combines multiple sub-images to demonstrate how the term "China Virus" assimilates other expressions of the virus, such as "COVID19", "Coronavirus", and "SARS-COV". This process of assimilation contributes to the stigmatization of a particular nationality. Harmful memes pose significant threats to society by causing discomfort, stigmatization, or even harm to individuals [6]–[8]. Furthermore, such memes have the potential to negatively impact public online experiences, contribute to cyber radicalization [9], and even incite real-world crimes [10].

Fig. 1: A harmful meme example.

* Equal contribution.
‖ Corresponding author.

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.24415
www.ndss-symposium.org

Consequently, there is an urgent need for methods that can effectively detect harmful memes.

Meta, previously known as Facebook, recently introduced the "Hateful Memes Detection Challenge", highlighting that the proficiency of deep learning-based artificial intelligence to identify hateful memes still falls significantly short of human-level discernment [6]. This competition has ignited substantial research interest in the field of hateful and harmful meme detection [11]–[14]. Despite these efforts, existing detection tools have struggled to achieve satisfactory performance, largely due to the inherent complexities of harmful memes that remain poorly understood. Several works have speculated and hypothesized about the failures of detection and the challenges brought about by harmful memes, such as the lack of advanced reasoning ability of existing tools [14], [15], and the neglect of important features in image caption extraction methods [6], [16]. Nevertheless, a significant gap persists in systematic analysis, as current research predominantly focuses on addressing the limitations of detection models rather than delving into understanding the underlying complexities specific to harmful memes that make them so challenging to detect.

Encouragingly, recent research on memes has highlighted the importance of visual arts [17] and propaganda techniques [18]–[20] in understanding memes, shedding new light on harmful meme detection. Specifically, in the realm of visual arts, composition—defined as the organization of visual elements such as panel count and image scale—holds significant importance. The challenge in detecting harmful memes arises because subtle changes in composition can shift a meme's perceived meaning and emotional impact, often allowing harmful intent to be obscured within seemingly harmless visuals. Similarly, propaganda techniques, which employ strategic rhetorical and psychological tactics to influence opinions or behaviors towards specific objectives, introduce additional detection challenges for detection. These techniques can mask manipulative content within persuasive rhetoric, complicating the identification of harmful intentions. Addressing these complexities requires integrating visual arts and propaganda techniques into meme detection frameworks, emphasizing reasoning-based methodologies.

In this work, we present the first systematic investigation into the challenges in detecting harmful memes, with a focus on their inherent complexity. Our findings, evidenced by a low true-positive rate, reveal that existing detection tools are inadequately equipped to tackle these challenges effectively. We conduct a thorough analysis of the factors contributing to memes' inherent complexity, focusing on their composition and the use of propaganda techniques. Our analysis indicates that complex meme compositions, such as stitching images (*i.e.,* combining multiple images into one meme to deliver a complete message), significantly undermine the effectiveness of existing harmful meme detection tools. Additionally, the employment of propaganda techniques in memes further complicates detection efforts by embedding harmful content within sophisticated rhetorical and psychological triggers. These findings underscore the need for sophisticated detection methods that can understand and address both compositional complexities and the subtleties of propaganda techniques.

Building on our new understanding and integrating existing knowledge about harmful memes, we design and develop a novel framework, HMGUARD [1], specifically tailored to detect harmful memes effectively. HMGUARD is the first harmful meme detection framework that utilizes adaptive prompting [21] and a chain-of-thought (CoT) reasoning strategy [22] with multimodal large language Models (MLLMs). It properly leverages MLLMs' capability to integrate multimodal semantics with sophisticated reasoning abilities, effectively addressing the complexities of harmful memes. Meanwhile, we designed a CoT reasoning strategy named HMCOT, which decomposes the process of harmful meme detection into seven steps, targeting different aspects of the challenges posed by harmful memes. As a result, our framework achieves a state-of-the-art (SOTA) performance, with an accuracy score of 0.92 in harmful meme detection.

The key contributions of this paper are as follows:

- **New understanding of harmful memes from novel perspectives.** This study presents a novel understanding of the challenges posed by harmful memes. Our findings reveal the multifaceted challenges related to complex meme compositions, such as stitching images. In addition, propaganda techniques embedding harmful content within sophisticated rhetorical and psychological triggers also pose intractable challenges to harmful meme detection. These insights shed light on new prerequisites for enhancing inspection tools and underpin the development of innovative frameworks.
- **New framework for harmful meme detection.** We design and develop a new harmful meme detection framework called HMGUARD. HMGUARD is a novel harmful meme detection framework that utilizes adaptive prompting and CoT reasoning in MLLM to realize zero-shot adaption and multimodal complex reasoning, effectively alleviating the challenges brought by harmful memes.
- **Extensive evaluation of HMGUARD.** The evaluation results show that our system achieves the most advanced accuracy rate of 0.92 and F1-score of 0.91 in detecting harmful memes, and all the evaluation indexes exceed the highest level of the baselines. For the hateful meme dataset, our system's detection accuracy is 24.64% higher than the state-of-the-art (SOTA) benchmarks, and 41.67% higher on the F1-score. Furthermore, the experimental results indicate that on two public datasets, the prompt strategy proposed in this paper significantly enhances performance by 15.28% to 96% compared to the MLLM-based method with a generalized prompt. In real-world scenarios, our framework has proven effective in detecting harmful memes prevalent on social media platforms and achieved an accuracy of 0.88 and an F1-score of 0.86.

II. BACKGROUND AND RELATED WORK

A. Harmful Memes

In recent years, the digital landscape has witnessed the emergence of a new and rapidly proliferating form of harmful

[1]Our framework is available at https://github.com/koi-yong/HMGuard.

content: harmful memes [7]. Unlike traditional harmful content, harmful memes often employ humor and satire, making it more challenging to discern their intent and mitigate their impact [15]. The potency of these memes lies in their ability to encapsulate complex and often insidious messages in a format that is easily digestible, highly shareable, and capable of bypassing conventional content moderation systems due to their nuanced and context-dependent nature [4].

Harmful memes are rapidly facilitating the dissemination of hate speech [23], misinformation, and extremist ideologies [5]. Their capacity to cloak harmful content in layers of irony and cultural references makes them particularly attractive to younger demographics, creating significant challenges in limiting their reach and impact [5]. The detrimental effects of these memes often extend beyond the digital sphere, inciting real-world actions, contributing to individual radicalization, and deepening social divisions [10].

B. Harmful Meme Detection

Harmful memes pose significant challenges for automated detection due to their complex interpretation, rapid evolution, and the integration of visual and textual content. The intricacies of cultural references and humor make it difficult to understand their intent without context, while their swift evolution surpasses the capabilities of detection tools [5]. Moreover, the seamless fusion of visual and textual elements necessitates a comprehensive analysis to effectively detect and mitigate their potential harm [24], [25]

Pramanick et al. [7] formally defined the harmful meme concept and demonstrated its dependence on contextual factors. The complex nature of memes, which often rely on multiple modalities, makes it challenging to yield good performance only using unimodal detection methods like BERT [26] or VGG19 [27]. From the initial single pipeline feature analysis evolved to using traditional pre-trained encoders to derive the image and text representations and focusing on designing new methods to fuse multimodal data [28]. To better fuse the relationship between modes, the detection methods based on prompt learning use the template to extract the key information in memes [11], [12], to predict the mask decision target better.

It is worth mentioning that ExplainHM is one of the SOTA existing works in the field of harmful meme detection. ExplainHM [29] leverages the debate capability of large language models to generate and debate explanations from different perspectives and then uses a smaller model to judge the harmfulness by synthesizing these debates with the multimodal content of memes.

Existing harmful meme detection tools only focused on marginal improvements in detection performance, failing to address the challenges inherent in the nature of harmful memes. Moreover, these tools lack the capability for *reasoning* [30], which is crucial for understanding and moderating the threats posed by harmful memes.

C. Multimodal Large Language Models and Chain-of-Thought Prompting

With the rapid development of the visual-language model (VLM) and large language model (LLM) in recent years, accumulating works improve text-image semantic fusion using these two advanced models. Visual Question Answering (VQA) tasks in VLM allow for the extraction of abundant feature information from memes in the form of queries. However, due to the lack of complex reasoning ability and rich background knowledge in the language model, VLM still has deviation in understanding harmful memes [13]. LLMs have rich background knowledge and more advanced complex reasoning abilities. However, due to insufficient information provided by the vision extractor, the model's understanding of image features and meanings may not be in place [14]. The integration of LLMs with visual capabilities has led to the emergence of multimodal large language models (MLLMs), such as GPT-4 [16]. MLLMs have demonstrated impressive performance on various vision-language tasks, including image-context reasoning, conceptual understanding, preference distillation, and embodied reasoning [31]. These advanced MLLMs stand out for their exceptional interpretability, adaptability, and augmented contextual insight, presenting numerous prospects for addressing complex and novel challenges in the intersection of vision and language. In our study, we leverage the capabilities of MLLMs to identify and analyze the complex patterns of harmful memes within explainability and interpretability.

Chain-of-thought (CoT) [22] prompting breaks questions into a sequence of reasoning steps before arriving at a final answer. LLMs can perform various reasoning tasks by using chain-of-thought prompting, which guides them to find answers through step-by-step demonstrations. Shao et al. [32] introduce synthetic prompting, a method that leverages a few handcrafted examples to prompt the model to generate more examples by itself and selects effective demonstrations to elicit better reasoning. Yoran et al. [33] introduce Multi-Chain Reasoning (MCR), an approach that prompts LLMs to meta-reason over multiple chains of thought rather than aggregate their answers. With the rapid development of multimodal technology and MLLMs, Zhang et al. [34] propose a multimodal-CoT that incorporates language (text) and vision (images) modalities into a two-stage framework that separates rationale generation and answer inference. In our study, we innovatively extend the application of CoT to the field of harmful meme detection, utilizing the reasoning capabilities of CoT to understand and detect harmful memes more effectively.

III. Threat Model

Harmful memes are a potent medium for the dissemination of misinformation, incitement of violence, and propagation of discrimination and hate speech. Due to their often humorous or satirical packaging, these memes can evade traditional scrutiny and moderation, making them particularly insidious tools for influencing public opinion and behavior. The rapid propagation capabilities of these memes through social networks exacerbate their potential impact, necessitating robust detection and mitigation strategies.

In the context of harmful memes, adversaries typically include online users who intentionally or unintentionally share or distribute them. Their motivations range from seeking to influence public discourse to sowing discord or unrest. The targets of harmful memes are generally the broader online community, which can include vulnerable or marginalized

groups who are disproportionately affected by the negative consequences of these memes.

Social media platforms like X and Meta offer content control and moderation mechanisms that empower users to report sensitive or inappropriate content [35], [36]. Such tools are crucial first lines of defense in mitigating the spread of harmful memes. However, the effectiveness of these mechanisms varies widely due to cultural, gender, and racial sensitivities, which can influence the perception and reporting of potentially harmful content. This variability often places a significant burden on platform administrators and automated detection tools, which must navigate these complex and nuanced landscapes to identify and mitigate harmful content effectively. The primary challenge in combating harmful memes lies in the timely and accurate detection of both explicit and implicit harmful content. Current automated tools can struggle with the subtleties and contextual nuances of memes, leading to under-detection or false positives. Moreover, the reliance on user reports can result in inconsistent moderation across different regions and communities.

Therefore, this work aims to bridge the gap between current automated detection methodologies and human knowledge. By employing advanced machine learning techniques combined with insights derived from human moderators, we design a novel system capable of understanding the semantic subtleties and context-specific nuances in harmful memes. This system will effectively detect harmful memes, thereby facilitating the creation of safer and more respectful online environments.

IV. Measurement and Observation

In this section, we present our studies focused on examining the effectiveness of existing methods for detecting harmful memes, as well as understanding the challenges involved in this detection process. These studies are critical in identifying the limitations of current technologies and paving the way for the development of more advanced and accurate approaches.

A. Data Preparation

In our study, we aim to understand the composition challenges of harmful memes. To this end, we utilize two well-established datasets in the field of harmful meme detection: HarMeme [7] and Meta Hateful Memes (FHM) [6].

HarMeme. The HarMeme dataset contains original memes that were actually shared on social media, and most of the content is related to COVID-19 [7]. The dataset categorizes these memes into three groups: "very harmful", "somewhat harmful", and "harmless". To ensure comparability with prior studies, we merge the "very harmful" and "somewhat harmful" categories into a single "harmful" category, following the evaluation settings of recent works [11]–[14], [24]. This adjustment transforms the task from a three-class classification to a binary classification problem.

Meta Hateful Memes (FHM). The FHM dataset was developed and disseminated by Meta during the Hateful Memes Challenge, which focuses on identifying hateful memes [6]. Compared to harmful memes, hateful memes target entities mainly based on personal attributes [24]. In this study, we choose this dataset as instances of content designed to inflict harm, assisting in the formulation of detection strategies for a wider spectrum of harmful content.

TABLE I: Overview of datasets.

Dataset	# Memes	# Harmful	# Harmless
HarMeme	289	110	179
FHM	711	422	289
Total	1000	532	468

In our study, we utilized the test sets from these two datasets to investigate the challenges of existing methods to detect harmful memes. To ensure the reliability of the data, we first cleaned the dataset by removing the repetitive memes. Then, we excluded the memes with text that was unrecognizable. In addition, by eliminating the redundancies and ambiguities, we ensured that each meme in our collection was distinct and contributed unique value to the dataset. Consequently, we compiled a refined dataset comprising 1,000 memes as depicted in TABLE I. This process not only streamlines the analysis but also enhances the accuracy of any insights derived from the data.

B. Failure of Existing Detection Methods

In order to understand the effectiveness of existing harmful meme detection systems, we measured the state-of-the-art detection methods (ExplainHM, discussed in § II-B) and two advanced MLLMs (LLaVa [37] and GPT-4 [16]) with the HarMeme dataset. Due to the universality and effectiveness of these detection methods and MLLMs, they can be considered representatives of existing technologies that can be used for harmful meme detection.

Specifically, ExplainHM uses a prompt in an LLM, "*Given the meme, with the Text: [T] embedded in the Image, and the following two meme rationales: (1) Harmless: [r^{hl}]; (2) Harmful: [r^{hf}], is this meme harmless or harmful?*", alongside the meme in question, as input. ExplainHM will analyze the aspects of both harmful and harmless, and summarize the decision through a response in the form of debate. For the chosen MLLMs, our methodology involves supplying a general prompt, "*Is this meme harmful or harmless?*". The MLLMs then generate a response that concludes their decision, classifying the meme as either "harmful" or "harmless".

In the experiments, we use the True Positive Rate (TPR) as our measurement metric, focusing on the model's effectiveness in detecting harmful content. TPR, also known as sensitivity or recall, is defined as the ratio of true positives to the total number of actual positives (*i.e.,* the sum of true positives and false negatives). Here, true positives represent harmful memes correctly identified by the model, while false negatives refer to harmful memes misclassified as harmless.

The low TPR values reported in TABLE II highlight that existing tools have significant room for improvement in detecting harmful memes. For instance, ExplainHM, an LLM-based method, achieves only 57.72% TPR. Similarly, the selected MLLMs perform poorly, with TPRs ranging from just 16.94% to a maximum of 52.42%. This underperformance

TABLE II: Existing methods in detecting harmful memes.

Detection Methods	TPR
ExplainHM	57.72 %
LLaVa	16.94 %
GPT-4	52.42 %

stems from a predominant focus on advancing application models rather than addressing the inherent challenges posed by harmful memes, which severely limits the effectiveness of existing detection methods.

C. Harmful Meme Detection Challenges

Building on our previous studies, we aim to explore the challenges posed by harmful memes and develop a novel method for their effective detection. Our work motivation focuses on the research of three challenges: *multimodal semantic fusion*, *meme composition*, and *meme propaganda technique*.

- *Multimodal semantic fusion* refers to the semantic fusion of memes containing text and image information. The nuanced interplay between text and image can convey subtle or overt harmful content. We discuss this in §IV-D that the fusion of text and visual semantics brings challenges to harmful meme detection.
- *Meme composition* refers to the organization of visual elements in a picture from the perspective of visual arts. Meme composition can subtly alter the perceived meaning and emotional impact, hiding harmful intent. We verify its effects on harmful meme detection in §IV-E.
- *Meme propaganda technique* can be defined as a form of communication that aims to influence the opinions or actions of people towards a specific goal [18]. Meme propaganda techniques introduce another layer of complexity by cloaking manipulative content with strategic communication, further complicating the identification of harmful content. we discuss the challenges it brings to harmful meme detection in §IV-F.

D. Multimodal Semantic Fusion

Early methods for harmful meme detection relied heavily on the capabilities of traditional pre-trained encoders to extract features from each modality independently [38]–[40], where separate channels for text and image data were analyzed without considering the integrative aspects of multimodal content. Recent advancements have introduced sophisticated models such as VisualBERT [41] and VL-T5 [42], which utilize multimodal pre-training and fusion techniques that can integrate textual and visual modalities from memes. However, such multimodal models are limited in their ability to understand the semantic interplay between different modalities. Specifically, memes often have the unique property where the text does not directly caption the image, and the relationships between the modalities can be ironic or contradictory. This makes understanding memes face the challenge of multimodal semantic complexity.

In this section, we examine the challenges multimodal models face in achieving semantic fusion to understand harmful memes. For this analysis, we use the HatReD dataset [42], an extension of the FHM dataset that includes additional semantic annotations to enhance its utility for multimodal hateful meme detection. Human experts have carefully annotated the dataset to facilitate deeper semantic fusion and comprehension of hateful memes. We adopt BERTScore [43] as the evaluation metric and use the experimental results from Lin et al. [44] as baselines. The BERT-based scoring system evaluates semantic similarity by comparing a reference sentence x (the human interpretation of a meme) with a candidate sentence $\hat{x}$ (the model's interpretation). It computes three metrics: recall R_{BERT}, precision P_{BERT}, and the F1-score F_{BERT}.

R_{BERT} evaluates how well each token in x is represented in $\hat{x}$ by averaging the maximum cosine similarities between corresponding tokens:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{1}$$

P_{BERT} determines how effectively tokens in $\hat{x}$ capture the semantics of x, calculated similarly by averaging maximum cosine similarities:

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \tag{2}$$

F_{BERT} is the harmonic mean of R_{BERT} and P_{BERT}, integrating both metrics to assess the model's accuracy and completeness:

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{3}$$

TABLE III illustrates that traditional multimodal models like VisualBERT and VL-T5 exhibit significant gaps compared to human-level performance in interpreting the multimodal semantics of memes.

Remark 1: Multimodal Semantic Fusion Challenge

We observe that multimodal semantic fusion presents a challenge for understanding harmful memes due to the model's limited ability to capture the interactions between modalities.

TABLE III: The BERTScore of different tools for interpreting the meaning of memes.

Model	BERTScore		
	P_{BERT}	R_{BERT}	F_{BERT}
VisualBERT	0.5	0.45	0.47
VL-T5	0.47	0.41	0.45
LLaVA	0.77	0.80	0.79
GPT-4	**0.84**	**0.83**	**0.83**

However, MLLMs, such as LLaVA [37], and GPT-4 [16], have shown considerable improvements over traditional models. Notably, GPT-4's achieved increases of 68%, 84.44%, and 76.60% in P_{BERT}, R_{BERT}, and F_{BERT}, respectively, compared to the highest-performing traditional models.

These results highlight the ability of MLLMs to provide multimodal fusion interpretations of memes that are semantically closer to human-annotated interpretations.

Our work aims to develop a novel detection framework based on MLLMs that can achieve a comprehensive understanding of the contextual relationships and cultural implications between modal semantics during modality fusion, thereby addressing the challenges of multimodal semantic fusion in harmful meme detection.

E. Meme Composition

Measurement Framework. In our study, inspired by the measurement framework proposed by Ling et al. [17], we employed four meme compositions to analyze memes in the prepared dataset. As a result, four essential compositions were identified, as depicted in Fig. 2. The details for each composition are as follows:

- *Type of the images.* The types of images employed in memes can be categorized into three types. First, illustration images, which include drawings, paintings, or any form of printed artwork, are characterized by their artistic creation. Second, photo images are defined by their origin in camera photography. Third, screenshot images pertain specifically to visuals captured directly from a computer screen.
- *Scale.* In terms of scale, memes can be classified into three categories: Close-up, Medium shot, and Long shot. A close-up tightly frames a person or object, drawing attention to specific details. A medium shot provides a balanced view, giving equal emphasis to the subject and its background. In contrast, a long shot captures a wide scene where the subject becomes less distinguishable, shifting the focus to the overall environment rather than any single element.
- *Movement.* Movements depicted in memes can be classified into three distinct categories. First, Physical movement encompasses any form of motion captured within the image. Second, Emotional movement is conveyed through facial movement or body language that reveals underlying emotions. Lastly, Causal movement refers to a sequence of movements in which an action from one entity (the sender) causes a reaction or a set of actions from another entity (the recipient). For example, strong sunlight (the cause) leads people to squint or cover their eyes (the effect).
- *Number of panels.* According to the number of panels presented, a meme can be categorized as Single-panel memes, which are composed of only one image, and Stitching memes are memes composed of a series of images that are no less than two images.

We annotated the dataset using the methodology proposed by Ling et al. [17]. The details can be found in AppendixA.

Fig. 2: Examples of memes with different compositions.

Result and Observation. We evaluate the performance of harmful meme detectors by providing memes with varying compositions as inputs and observing the detection outcomes. TABLE IV presents the effectiveness of SOTA harmful meme detectors and MLLMs, using the True Positive Rate (TPR) as the evaluation metric.

Based on the number of panels, we observe that single-panel memes achieved a TPR exceeding 50%, whereas stitched images had an average TPR of only 35.18%. This significant difference underscores the increased complexity that stitched images pose for detection models, making such memes more challenging to interpret accurately. Here, a "significant" concern arises when the TPR falls below 50% for image types that account for more than 10% of the dataset, as reflected in TABLE IV.

For the types of images, the TPR values across different categories consistently remained above 50%, aligning with the overall detection results in TABLE II. This suggests that image types do not significantly affect the effectiveness of harmful meme detection.

In terms of scale, TPR levels for close-up and medium shots were consistently above the overall average, with close-up shots achieving a notably high TPR of 85.71%. In contrast, long shots performed poorly, with significant fluctuations in TPR. However, given the limited number of long-shot samples in the dataset (only 4%), we suspect this under-performance may be due to insufficient representation. Thus, we cannot conclusively determine whether scale, particularly long shots, significantly impacts harmful meme detection.

In terms of movement, physical and emotional movement

TABLE IV: TPR of different tools for detecting harmful memes across various meme composition categories.

Category	Subcategory	Proportion of Meme	True Positive Rate (TPR)			
			ExplainHM	LLaVa	GPT-4	Avg.
Number of Panels	Single	67%	67.03%	18.18%	61.54%	52.35%
	Stitching	33%	42.42%	16.48%	33.3%	35.18%
Type of the Images	Illustration	17%	71.43%	14.29%	66.67%	57.15%
	Photo	50%	63.04%	17.39%	58.7%	51.09%
	Screenshot	33%	73.91%	17.39%	65.22%	54.35%
Scale	Close-up shot	20%	75%	42.86%	85.71%	68.75%
	Medium shot	76%	63.64%	12.99%	59.74%	50%
	Long shot	4%	80%	0%	0%	40%
Movement	Physical movement	56%	61.36%	18.18%	63.64%	52.27%
	Emotional movement	39%	67.74%	16.13%	61.29%	52.42%
	Causal movement	5%	75%	0%	25%	37.5%

Fig. 3: A meme example with propaganda techniques.

in memes are readily detected by existing tools. However, the causal movement shows significant fluctuations and relatively low detection rates. This inconsistency is likely due to the small proportion of causal movement data in the dataset (only 5%) as well, making it challenging to draw definitive conclusions from further research in this category.

Remark 2: Meme Composition Challenge

We observe that meme composition challenges the interpretation and detection of harmful memes, particularly with stitched images, which complicate understanding visually.

Our work aims to design a novel harmful meme detection framework with MLLMs to analyze hiding harmful intent from a visual art perspective. Thus, we strive to alleviate the challenges brought by complex meme composition, especially for stitching images.

F. Meme Propaganda Technique

Meme propaganda techniques embed complex socio-political messages within seemingly innocuous media, always utilizing sophisticated rhetorical strategies and psychological triggers [45]. These techniques can be exploited by malicious attackers to make the expression of harmful content in memes more subtle and less detectable, thereby bypassing conventional content moderation systems and influencing viewers without immediate detection.

In this study, we investigate and deploy twenty-two propaganda techniques that are commonly used in expressing opinions and emotions on social media platforms from previous research work [18], including 'Name calling or labeling', 'Appeal to fear/ prejudices', 'Whataboutism', 'Misrepresentation of someone's position', 'Flag-waving', 'Causal oversimplification', 'Black-and-white fallacy or dictatorship', 'Reductio ad hitlerum', 'Smears', 'Loaded language', 'Doubt', 'Exaggeration/ Minimisation', 'Slogans', 'Appeal to authority', 'Thought-terminating cliche', 'Repetition', 'Obfuscation, Intentional vagueness, Confusion', 'Presenting irrelevant data', 'Bandwagon', 'Appeal to strong emotions', and 'Transfer'. The definitions and explanation details of each technique are presented in Appendix B.

Fig. 3 displays an example meme utilizing propaganda techniques. The meme manipulates viewer perceptions through a calculated use of propaganda techniques that trigger emotional reactions and distort facts. By employing an 'appeal to strong emotions' (i.e. Using images with strong positive/negative emotional implications to influence an audience), it bypasses rational thinking, tapping directly into fear and anger by visually linking modern Democrats with historically extreme groups. This not only stokes existing fears and prejudices but also subtly suggests that these extreme dangers are relevant today, warping the viewer's understanding of the political landscape. It also uses the propaganda techniques of 'name-calling' (i.e. Labeling the object of the propaganda campaign as something that the target audience fears, hates, finds undesirable, loves, or praises) and 'smears' (i.e. Damaging or calling into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups) further discredit Democrats by labeling them as extremists or racists, which undermines their credibility and makes it difficult for viewers to approach their policies or statements objectively. Furthermore, the technique of 'transfer'—evoking an emotional response by projecting positive or negative qual-

TABLE V: The TPR of different tools for detecting harmful memes with and without propaganda techniques.

Catetory	Proportion of Meme	True Positive Rate (TPR)			
		ExplainHM	LLaVa	GPT-4	Avg.
w/o propaganda techniques	57.3%	75%	17.31%	60%	50.77%
w/ propaganda techniques	42.7%	53.85%	15%	48.08%	38.98%

ities (praise or blame) of a person, entity, object, or value onto another to make the latter more acceptable or to discredit it—amplifies this effect by associating the negative qualities of these historical groups with modern Democrats, leading viewers to subconsciously perceive them as sharing similar reprehensible qualities, despite the lack of any factual basis. Through these methods, the meme aims to distort public opinion, exacerbating political polarization and fostering an environment of mistrust and misinformation.

To measure the challenge posed by propaganda techniques, we employ the datasets mentioned above with SOTA tools, focusing on assessing the impact of propaganda techniques on the performance of harmful meme detection. As presented in TABLE V, on average, compared to the meme without (w/o) propaganda technique, the meme with (w/) propaganda techniques in memes resulted in an 11.79% reduction in the TPR, suggesting that the use of meme propaganda techniques makes hate memes more difficult to detect.

The significant impact of propaganda techniques in harmful meme detection underscores the need for advanced analytical approaches that transcend basic content filters and surface interpretation [18]. It requires a deeper semantic analysis capable of identifying nuances such as sarcasm, satire, or double entendre, which are prevalent in memes.

Remark 3: Meme Propaganda Technique Challenge

We observe that the meme propaganda technique poses challenges for detecting harmful content, as it makes the expression more subtle and less detectable.

Our research aims to design a novel detection framework based on MLLMs, which uncovers the harmful content hidden within complex rhetorical strategies, thereby alleviating the challenges posed by meme propaganda techniques.

V. HMGuard Design

A. Design Intuition

Detecting harmful memes is a complex contextual understanding and decision-making task that necessitates intricate inference processes. As highlighted in our previous study (see Section IV), we observed that various factors, including meme fusion, meme composition, and the use of meme propaganda techniques, pose significant challenges to the detection of harmful content. To explore the potential of MLLM to complete reasoning-based multimodal detection tasks, we introduce the CoT prompt, which breaks the task into multiple intermediate steps and deduces the final decision through the intermediate outputs. According to the intermediate reasoning prompts, the model attains a deeper and more thorough understanding of the problem to make a more accurate final decision. These intermediate prompts need to be carefully designed based on important elements such as the definition and characteristics of harmful memes, allowing the model to determine whether a meme is harmful in a clear, step-by-step process in which each step considers the output of the intermediate step to generate the output. In this way, not only is the reasoning ability of LLM well applied to MLLM in implementing reasoning-based decisions to identify hate memes, but it also has outstanding scalability.

B. Overview of HMGuard

The overview of our framework, HMGuard, is shown in Fig. 4. Based on our previous research on harmful memes, we developed HMCoT Prompts, a novel reasoning-based chain-of-thought prompting strategy designed to address the complex challenges posed by harmful memes, such as multimodal semantic fusion, meme composition, and meme propaganda technique. This strategy aims to facilitate reasoning-based final decisions through prompts that consider different factors, thereby enabling precise detection of harmful memes. For the MLLM, we need to perform meme domain alignment and task-specific adaptation before running the HMCoT Prompts to tune the model for understanding memes and the task of harmful meme detection. In the next stage, we use the MLLM to execute HMCoT Prompts, and the responses generated by the MLLM are analyzed to extract answers for each of the HMCoT Prompts. In the final stage, we utilize all the intermediate responses to determine whether the input meme contains harmful content.

C. HMCoT Prompting

Addressing the multifaceted challenges brought by the complexity of harmful memes requires our prompt to be adept at leveraging MLLMs to effectively make detection decisions. According to our investigation and discussion in Section IV, we have clarified the definition of harmful memes, the inadequacy of existing detection methods, and important relevant factors, such as compositions and propaganda techniques. Therefore, the detection of harmful memes is a complicated process that needs to be analyzed and evaluated for each important relevant factor. HMGuard realizes this detection process by leveraging MLLMs with delicately designed CoT prompts. This method systematically addresses the complex task of detecting online hate by treating each of the factors as a distinct subproblem. In the end, the results from these sub-questions and intention verification are integrated to formulate a comprehensive decision regarding the harmfulness of the memes.

1) Crafting HMCoT *Prompts.:* HMCoT divides harmful meme detection into seven major steps as depicted in Fig. 5: (1) Meme Domain Alignment and Task-specific Adaption; (2) Surface Meaning Identification; (3) Fusion Meaning Identification; (4) Composition Meaning Identification; (5) Propaganda Meaning Identification; (6) Intention Verification; (7) Final Decision.

Meme Domain Alignment and Task-specific Adaption. In the context of harmful meme detection, initiating the process

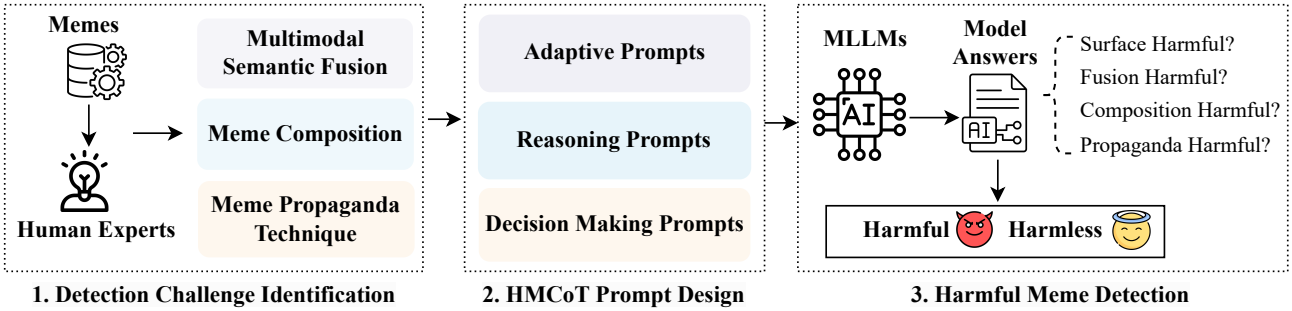

Fig. 4: Overview of HMGUARD.

with meme domain alignment and task-specific adaptation is crucial because it equips the model with the necessary cultural and contextual understanding to accurately interpret the nuanced interplay of visuals and text that memes embody [46], [47]. These ensure that the model is sensitive to the varied interpretations across different cultures, while the adaptation focuses the model capabilities on the specific challenge of identifying subtle and overt harmful content, thus enhancing the precision and reliability of the detection process.

In tools based on LLMs, domain alignment and task-specific adaptation are usually represented by specific prompting strategies, aimed at guiding the model's understanding and response. For meme domain alignment, we use the prompt *"This is a meme, using text and images for humor or satire, shaped by culture and contexts"* to guide the model in analyzing the interplay of images and text within social and cultural environments and contexts.

For task-specific adaptation, we adopt the prompt *"You are a content moderation specialist. Your task is to pinpoint any instances of hate speech, explicit violence, discrimination, or any other type of content that may be considered harmful"*. This prompt directs the model to concentrate on the specialized task of detecting harmful content, thereby enhancing the performance of the identification process.

Surface Meaning Identification. As the initial sub-problem within the HMCoT framework, our objective is to diminish the redundancy in the chains of thought that follow. Consequently, in cases where the imagery and text within a meme overtly contain harmful content, we return results in advance. According to this, we devise a guiding question Q1: *"What are the words and images contained in the meme, and do the semantics of the text and image directly convey harmful content?"*.

Fusion Meaning Identification. Given the complex relationship between the semantics of text and imagery in memes, it is necessary to further consider the semantics expressed after multimodal fusion. According to previous studies, we have found that some memes' text and images do not have obvious harmful content, and may even seem unrelated, but when combined, they can become harmful. Therefore, in this module, we employ a guiding question in Q2:*"What are the explicit or implicit relationships between text and image, and does the relationship between the text and the image potentially reveal harmful content?"*.

Composition Meaning Identification. From § IV-E, we understand the challenges of detecting harmful memes involving stitching images. In this module, we first pose the question Q3a: *"Is the meme a stitching image?"* to determine the number of panels. If the answer is yes, we proceed to question Q3b; if no, we move on to the next module. Q3b is formulated upon recognizing that the meme exists in the form of stitching images, guiding the model to further analyze the relationship between each image, such as narrative sequence, chronological order, etc. We use the guiding question Q3b: *"Consider the relationship of images, and understand whether the stitching images are trying to express explicit or implicit harmful content."* to further examine the meaning presented at the meme composition level, such as the comparison of before and after images, cause-and-effect relationships, etc., thereby preventing the model's understanding of memes from being trapped in local features.

Propaganda Meaning Identification. In this module, we propose the following question prompts to identify the propaganda meaning of the meme, Q4: *"Are the following propaganda techniques used to explicitly or implicitly express harmful content?"*, and we provide with choices mentioned in §IV-F.

Intention Verification. Through the analysis of the subquestions above, the model has a more profound and comprehensive understanding of memes. We hope to verify the intention of the meme in this module. We use question Q5: *"Does the meme intend to have any targeted derogatory, humiliating, insulting, satirical, or disparaging meaning?"*

Final Decision. Given the context provided by the answers to all the previous subquestions, we define the final subquestion Final Decision: *"Combining the analysis from the previous questions, please make the final decision on whether this meme is harmful or harmless. You need to make sure that your answers are consistent with the questions above."* as a comprehensive decision task. In order to alleviate the problem of faithfulness hallucination [14], namely, inconsistency between the content generated by the model and the context.

2) Leveraging MLLMs for Processing HMCoT *Prompts:* Our system leverages an MLLM to operationalize our prompting strategy in the following way. Given image input X_{meme} and supplemented by the HMCoT prompts, characterized as the prompt text input X_{prompt}, the output is computed as,

$$\hat{y} = argmax\ p(y|X_{meme}, X_{prompt}), \tag{4}$$

Using our HMCoT prompts, we decompose the primary problem of harmful meme detection into a series of sub-

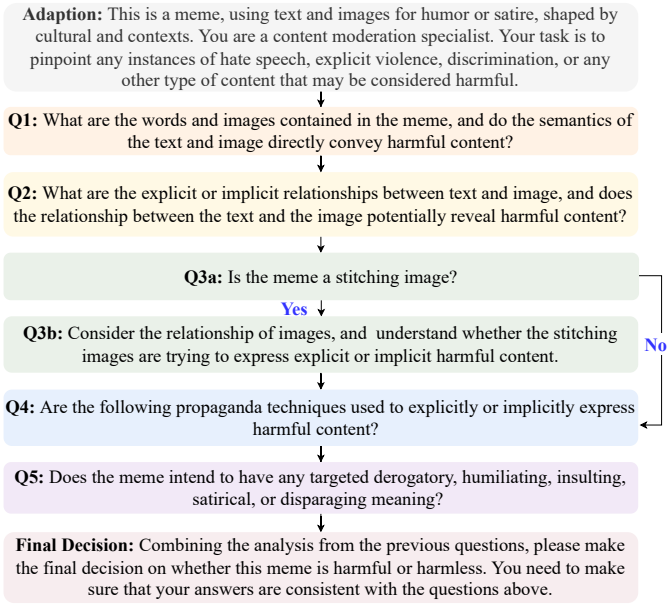

Fig. 5: HMCoT prompting strategy.

problems. This enables a structured and sequential approach to decision-making, where the final output $\hat{y}$ is a culmination of insights derived from intermediate states. The process can be represented as $\hat{y} \leftarrow p_n \leftarrow p_{n-1} \leftarrow ... \leftarrow p_1$, where $p_1, ..., p_{n-1}, p_n \in P$ illustrate a systematic progression through various stages of reasoning. To be specific, the steps are as follows and follow a dimensional order.

Step 1. Meme Domain Alignment and Task-Specific Adaption. We first conduct the MLLM to achieve meme domain alignment and task-specific adaption, so that it can understand the input meme. H_D is the attention feature generated from the interaction between the adaptive prompt and the meme's image and text features, representing the model's understanding of the meme in the adapted domain. The operator symbolizes the act of conditioning, wherein the model assimilates and processes the meme X_{meme} in conjunction with adaption prompt D.

Step 2. Surface Meaning Identification. Next, we prompt the MLLM to check whether the text and images in the meme contain any obvious harmful content.

$$A_1 = argmax\ p(a|H_D, Q_1). \tag{5}$$

Step 3. Fusion Meaning Identification. Then, we prompt the MLLM to analyze the complex interaction in the text and images contained in the meme and check for any implicit harmful content.

$$A_2 = argmax\ p(b|H_D, Q_2). \tag{6}$$

Step 4. Composition Meaning Identification. Next, we prompt the MLLM to determine if the meme involves stitching images. If so, we will execute Q3b, analyzing the relationships between the images and checking for any implicit harmful content. If not, we will proceed directly to the next step.

$$\begin{aligned} A_{3a} &= argmax\ p(c|H_D, Q_{3a}), \\ A_{3b} &= argmax\ p(d|H_D, Q_{3b}). \end{aligned} \tag{7}$$

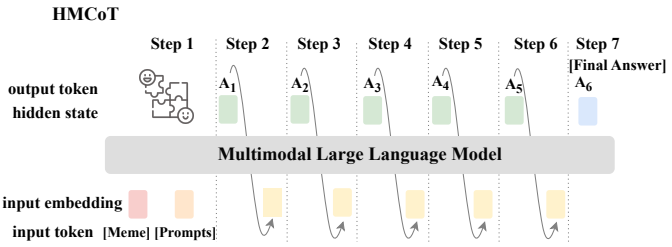

Fig. 6: The flowchart of MLLM processing HMCoT.

Step 5. Propaganda Meaning Identification. Next, we prompt the MLLM to determine if the meme uses propaganda techniques to reflect sensitive attributes.

$$A_{4a} = argmax\ p(e|H_D, Q_4). \tag{8}$$

Step 6. Intention Verification. Next, we prompt the MLLM to verify if the meme has a harmful intent.

$$A_5 = argmax\ p(g|H_D, Q_5). \tag{9}$$

Step 7. Final Decision. The final decision is made by prompting the MLLM to output a conclusion based on the input sentence and the previous output.

$$A_6 = argmax\ p(h|H_D, A_1, A_2, A_{3a}, A_{3b}, A_4, A_5, Q_6), \tag{10}$$

$$\hat{y} = \begin{cases} harmful, & \text{if } A_6 = \text{harmful}, \\ harmless, & otherwise. \end{cases} \tag{11}$$

Fig. 6 illustrates the workflow of the MLLM processing HMCoT, which is designed to detect harmful content in memes through a series of thoughtfully structured steps. The process starts by adapting the model to the specific nuances of meme content, ensuring accurate context recognition. As the flow progresses, it systematically breaks down the task into subquestions. Each step builds on the previous one, gradually forming a comprehensive understanding of the meme's potential harmfulness. In the last step, all these insights are integrated to make a well-informed decision about the presence of harmful content. This structured approach ensures a thorough and precise detection process, minimizing the risk of overlooking potentially harmful content.

VI. Implementation and Evaluation

In this part, we first discuss the implementation of our approach, followed by experiments to evaluate the effectiveness of our approach to detect harmful memes. Furthermore, we conduct an ablation study of the proposed HMCoT to verify the contribution of each prompt to harmful memes detection. Next, we analyze the cases that our approach fails to classify. Finally, we run HMGuard on the data collected on the social media platform. Our evaluation goals are as follows:

- Examining the effectiveness of HMGuard by comparing it with existing benchmarks (§VI-C)
- Examining the effectiveness of the adaptive prompts of HMGuard. (§VI-D)
- Examining the effectiveness of the reasoning-based prompts identification process of HMGuard. (§VI-E)

TABLE VI: Comparison of the detection performance with existing benchmarks.

Detector	FHM				HarMeme			
	Accuracy	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
MOMENTA	0.61	0.66	0.51	0.57	0.77	0.69	0.45	0.55
HateDetectron	0.69	0.54	0.73	0.58	0.8	0.77	0.62	0.68
MR.HARM	0.58	0.34	0.65	0.45	0.8	0.56	0.82	0.66
ExplainHM	0.48	0.35	0.55	0.48	0.73	0.25	0.62	0.71
GPT-4	0.61	0.55	0.64	0.6	0.74	0.72	0.5	0.69
HMGUARD	**0.86**	**0.88**	**0.83**	**0.85**	**0.92**	**0.83**	**0.98**	**0.91**

Note: Underline represents the best results in baselines; **Bolding** represents the best results among all approaches.

- Examining the effectiveness of HMGUARD on "In-the-Wild" Samples. (§VI-F)

A. Implementation Details

In this section, we present the implementation details of HMGUARD. We utilize the GPT-4 model (gpt-4-vision-preview) as our preferred MLLM for the large-scale execution and evaluation of HMCOT prompts. We select this model due to its status as one of the most advanced MLLMs available. It has demonstrated great contextual understanding and reasoning capabilities, making it highly suitable for our application. All of our experiments are conducted using four NVIDIA A100 40GiB GPUs. We utilize the publicly available harmful memes datasets as described in § IV-A, supplemented by an additional 300 popular memes collected from Pinterest. This expanded dataset serves as a test case representing "in-the-wild" memes, providing a real-world scenario to evaluate the performance of our models.

B. Baselines

We evaluate HMGUARD with 5 baselines to conduct a comparative evaluation: (1) MOMENTA [24], a multimodal harmful meme detection system, deploying both local and global multimodal fusion mechanisms for harmful meme detection. (2) HateDetectron [48] is the winning system of Meta's Hateful Meme Challenge, which employs a method that involves expanding the training set through the discovery of similar datasets on the web to refine a pre-trained vison-language model (VisualBERT). (3) MR.HARM [14] is a large language model-based harmful meme detection system, making use of the textual content understanding ability of LLM by inputting the explanation of the meme and the embedded text. (4) ExplainHM [29] is a system that utilizes the argumentative capabilities of LLMs to produce and evaluate explanations from diverse viewpoints, then uses a smaller model to judge the harmfulness by synthesizing these debates with the multimodal content of memes. (5) GPT-4 [16], an advanced MLLM with advanced reasoning capabilities.

C. Comparisons with Existing Benchmarks

In this experiment, we assess the effectiveness of HMGUARD compared with existing benchmarks for harmful meme detection. MOMENTA is developed by the publisher of the Harmeme dataset, while HateDetectron is developed by the winning team in the Meta Hateful Meme Challenge. MR.HARM and ExplainHM represent the most advanced LLM-based detection methods, the former utilizing the reasoning power of LLMs, the latter utilizing the debating power of LLMs. In addition, we introduce GPT-4 as representatives of MLLMs with the generalized prompt "Please classify the meme as harmful or harmless."

TABLE VI shows our proposed method, called HMGUARD, for benchmark performance on the HarMeme and FHM datasets. First, we delve into the comparative results on the FHM dataset, a repository of hateful memes. Within the existing tools, HateDetectron achieved the highest accuracy of 0.69, yet its F1-score was a modest 0.58, indicating significant room for improvement in the successful hateful meme detection. HMGUARD, achieved a SOTA accuracy of 0.86 and an F1-score of 0.85. Compared to the baseline of GPT4 with a generalized prompt, HMGUARD register improvements ranging from 29.69% to 60% across various metrics.

For the HarMeme dataset, a repository of harmful memes, MR.HARM achieved the highest accuracy of 0.8 and the highest recall of 0.82 within existing tools, yet the F1-score was only 0.66. Building upon this, HMGUARD reached an impressive accuracy of 0.92, a recall of 0.98, and an F1-score of 0.91.

GPT-4 with the generalized prompt shows comparable performance in harmful meme detection compared to other baselines, indicating that GPT-4 has a good interpretative ability for multimodal semantics. Still, with the deployment of the HMCOT prompting strategy, HMGUARD significantly outperforms it by 15.28% to 96% across various metrics.

TABLE VII: Comparing TPR with baselines from different meme categories.

Category of Harmful Meme		**HMGUARD**	**Improvement**
Number of Panels	Single	97.44%	30.41%
	Stitching images	96.88%	54.46%
Type of the Images	Illustration	99.99%	28.56%
	Photo	97.44%	34.40%
	Screenshot	95%	21.09%
Scale	Close-up shot	99.99%	14.28%
	Medium shot	96.92%	33.28%
	Long shot	99.99%	19.99%
Movement	Physical movement	94.59%	30.95%
	Emotional movement	99.99%	32.25%
	Causal movement	99.99%	24.99%
w/o propaganda techniques		99.99%	24.99%
w/ propaganda techniques		97.78%	43.93%

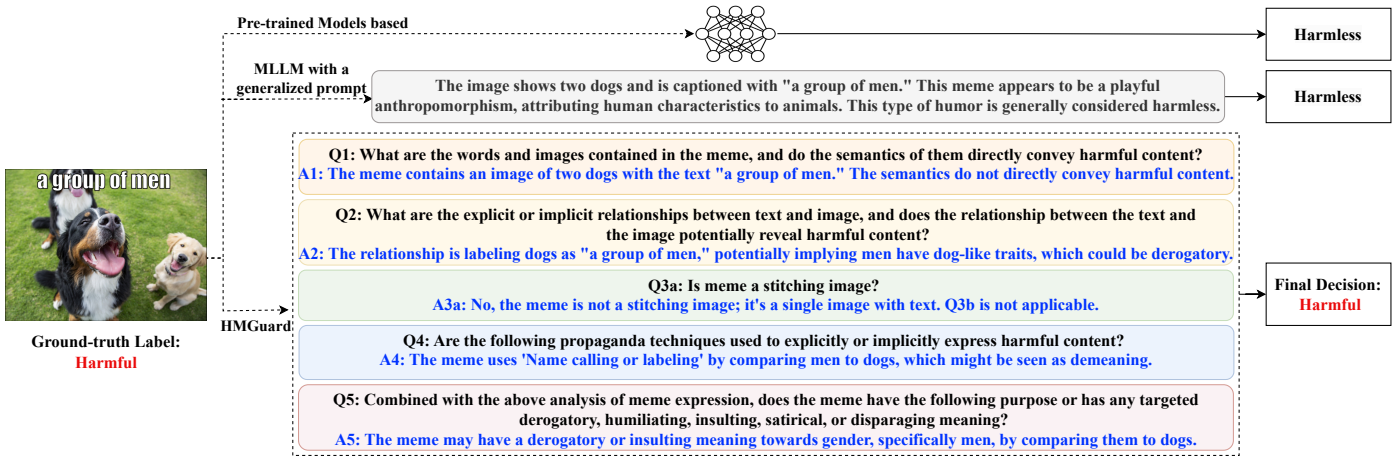

Fig. 7: A case study of HMGUARD comparing with a pre-trained based model and an MLLM with the generalized prompt.

Our approach significantly enhances its performance by exploiting the reasoning capability of MLLMs. To more comprehensively demonstrate the effectiveness of HMGUARD, we have detailed a representative case in Fig. 7. This case study includes the failure process of methods based on a pre-trained model and MLLM with a generalized prompt, contrasting them with the successful strategy employed by HMGUARD, which leverages a series of sub-questions to infer the true label.

Furthermore, we want to investigate if HMGUARD is capable of understanding and addressing the challenges of meme compositions and meme propaganda techniques that previous methods failed, as depicted in § IV-E and § IV-F. TABLE VII shows that HMGUARD exhibits substantial enhancements across all factors relative to the baseline.

Notably, HMGUARD effectively address the most significant challenges identified in Section IV, particularly for factors ”stitching images” and ”with propaganda techniques”, where it achieves remarkable improvements of 54.46% and 43.93%, respectively. Moreover, even in cases where the baseline performance is already satisfactory, such as illustration and close-up shot, HMGUARD still demonstrates significant enhancements in detection rates. This highlights HMGUARD's robustness and effectiveness in detecting harmful memes across various types and features, overcoming challenges that other tools struggle to handle.

D. Effectiveness of Adaptive Prompts

In this experiment, we conduct an ablation study to evaluate the effectiveness of adaptive prompts for HMCOT. Specifically, we test HMGUARD with and without (w/o) the adaptive prompts in HMCOT on the HarMeme dataset, which corresponds to the first step outlined in § V-C. Our adaptive prompts consist of two modules: meme domain alignment and task-specific adaptative. The first module employs prompts to guide the MLLM for context adjustment to align with the meme scenario, while the second module uses prompts to direct the MLLM to clearly understand and adapt to the downstream task of harmful meme detection. Specifically, we conduct an ablation study on the adaptive prompts as a whole, followed by separate ablation studies for each module. As shown in TABLE VIII, the introduction of adaptive prompts lead to an 8.24% increase in accuracy and a 54.24% improvement in F1-score. The inclusion of meme domain alignment results in a 22.97% increase in the F1-score, while the inclusion of task-specific adaption results in a 65.45% increase in the F1-score. Given the imbalance in the dataset, where harmless memes are more prevalent than harmful ones, the significant improvement in the F1-score demonstrates the importance of the two modules in our adaptive prompts. This suggests that adaptive prompts enhance the overall efficacy of HMGUARD in detecting harmful memes by steering the MLLM towards context adjustment and focusing on knowledge relevant to harmful meme detection through task-specific adaptation, thus facilitating a narrowed target search space, and the interaction between the two amplifies the overall effect.

TABLE VIII: Ablation study of adaptive prompts

Prompt	Accuracy	F1-score
HMCOT w/o Adaptive Prompts	0.85	0.59
HMCOT w/o Meme Domain Alignment	0.86	0.74
HMCOT w/o Task-specific Adaption	0.87	0.55
HMCOT Prompts	**0.92**	**0.91**

TABLE IX: Ablation study of our reasoning prompts

Prompt	Accuracy	F1-score
Only adaptive Prompts	0.77	0.57
HMCOT w/o M1	0.78	0.68
HMCOT w/o M2	0.78	0.61
HMCOT w/o M3	0.81	0.65
HMCOT w/o M4	0.76	0.62
HMCOT w/o M5	0.83	0.75
HMCOT Prompts	**0.92**	**0.91**

E. Effectiveness of reasoning-based Prompts

In this experiment, we evaluate the effectiveness of the reasoning prompts of HMCOT on the HarMeme dataset.

Specifically, we verify the validity of each module of HMCoT introduced in §V-C. In the TABLE IX, *M1* refers to *Surface Meaning Identification*. *M2* refers to *Fusion Meaning Identification*, *M3* refers to *Composition Meaning Identification*, *M4* refers to *Propaganda Meaning Identification* and *M5* refers to *Intention Verification*. These results indicate that reasoning prompts have led to a 19.48% increase in accuracy and a 59.65% enhancement in F1-score, suggesting that decomposing complex problems into sub-problems for reasoning within MLLMs is highly beneficial for harmful meme detection. Furthermore, each module of HMCoT contributes to the improvement of harmful meme detection performance, with M4 providing the largest contribution to accuracy, improving it by 21.05%, and M2 offering the most significant boost to the F1-score, with an increase of 49.18%.

F. Running HMGuard on "In the Wild" Samples

To further demonstrate the effectiveness and robustness of our proposed framework in real-world scenarios, we collected memes "in the wild" from social media platforms and conducted high-quality manual annotations. Subsequently, we utilize this collected dataset to evaluate the performance of HMGuard and compare it with SOTA methods.

Collection of memes from Pinterest. Pinterest is a social media site that groups images into collections based on similar themes. The search function returns images based on user-defined descriptions and tags. Therefore, we collect memes from Pinterest using keyword search terms to determine whether the returned images are likely harmful or harmless. In total, we obtained 512 memes published on Pinterest and removed repetitive or blurred memes. There are 300 memes that remain and are used to evaluate the expansibility of HMGuard, and we call this data set HMW.

Annotation of the memes. We appoint two authors of this work to participate in the data annotation because they have sufficient background knowledge and a well-deserved academic ethic. To establish inter-rater reliability, we developed a common code book for labeling memes in the collection as harmful or harmless memes, and the codebook is provided in Appendix C. We randomly selected 150 samples and required two authors to annotate them independently using a code book. The 150 samples are divided into two parts: 100 samples of the first two parts are used for discussion to reach a consistent label, and the last 50 samples are used to test the final coding consistency. After two rounds of discussion, the two authors reach 100% agreement on the coding. In total, the dataset includes 102 harmful memes and 198 harmless memes.

Experiment settings. The experiment settings are fundamentally the same as mentioned in Section VI. We use *gpt-4-vision-preview* as the base model for experiments. For hyperparameters, we set the temperature as the default value of 1. We use two evaluation measures: Accuracy and macro-F1, which are better when higher values are used. Since the test set is imbalanced, measures by macro-F1 are more relevant.

TABLE X: The comparison result on "in-the-wild" samples.

Detector	Accuracy	Percision	Recall	F1-score
HateDetectron	0.7	0.53	0.52	0.51
MR.HARM	0.73	0.57	0.52	0.5
HMGuard	**0.88**	**0.83**	**0.89**	**0.86**

Experiment results. As shown in TABLE X, HMGuard also achieves high performance on the HMW dataset, with an accuracy of 0.88 and an F1-score of 0.86. This shows that HMGuard also has significant advantages in the detection of harmful memes in the wild.

VII. Discussion

Limitations. The datasets deployed in our study only contain memes with embedded text in English. Expanding our research framework to different languages will provide a more comprehensive understanding of harmful memes in various linguistic regions and improve detection capabilities. Additionally, our "in the wild" evaluation is conducted using a limited number of memes collected from Pinterest. To further validate the effectiveness of HMGuard in real-world scenarios, we plan to expand our data collection to include more social media platforms in future research.

Robustness of HMGuard. In our analysis, we observed that adversarial attacks targeting natural language processing (NLP) components of harmful memes present significant challenges for existing detection tools. Recognizing their real-world relevance, we evaluated the robustness of HMGuard against such adversarial attacks. Given the limited availability of adversarially modified harmful memes in public datasets, we improved our dataset by introducing perturbations. Specifically, we selected 15 harmful memes with sensitive words embedded from the FHM dataset and applied four types of NLP-based adversarial manipulations: letter addition, letter deletion, letter swapping, and space insertion [49], and examples are shown in Appendix D1. This augmentation process produced 60 adversarial examples, which were later used to test HMGuard. Our framework demonstrated strong robustness, achieving a remarkable detection rate of 95% on the augmented dataset. More detailed evaluation results are presented in Appendix D2.

Integration with other MLLMs. Our framework, HMGuard, with its adaptable and transferable architecture, can be deployed on different MLLMs. We conduct an additional experiment leveraging an open-source MLLM, LLaVA, to deploy HMGuard and evaluate it with the HMW dataset. Using the LLaVA-v1.6-34b version, the accuracy for harmful meme detection achieves 0.78. These results are not as impressive as the performance of HMGuard shown in TABLE X due to insufficient reasoning capabilities resulting from model structure and size. Nevertheless, these results not only surpass the capabilities of the SOTA current detection tools but also significantly outperform the LLaVA with a generalized prompt in the real-world scenario. We believe that with updates to LLMs, such as LLaMA3 [50], the reasoning capabilities of MLLMs will be elevated to the next level.

Ethical consideration. We collect and annotate data from social platforms to test the effectiveness of HMGuard on "in the wild" samples. This task is performed by the two authors. Regarding harmful content, we made it clear before collecting and annotating that harmful content would be present in the data. In our paper, we have taken measures to minimize the

presence of harmful content. Regarding user privacy, we ensured that only meme data was collected from social platforms, without involving any user account information.

Deployment. Our framework can potentially be deployed to alleviate other digital content safety issues, such as unsafe images and deepfake detection. Integrating visual arts analytical frameworks into reasoning-based decision systems can effectively extend their applicability to identify and mitigate the risk of harmful content manifested through visual media in other domains. In addition, our research confirms the effectiveness of the CoT strategy in the field of harmful meme detection. This approach also holds potential for application in various other cyberspace security domains, such as online hate moderation [51], [52], unsafe image detection [53], and vulnerability discovery [54]. These cases demonstrate that the interaction between CoT and large-scale models effectively addresses complex real-world problems requiring intricate reasoning. We encourage technicians and researchers in the field of cyber security to further explore and understand the principles of cyber security technologies, combining CoT with advanced large models to mitigate more challenging issues.

Future work. To advance the capabilities of our framework, we plan to extend its functionality to support multilingual meme detection, addressing the growing demand for robust detection across diverse languages. This enhancement is particularly vital given the global nature of memes, where humor, cultural references, and context differ significantly between languages. A multilingual extension would significantly broaden the applicability and effectiveness of HMGUARD. Moreover, we are looking forward to exploring a broader range of harmful content detection capabilities, including the detection of AI-generated images and artwork. The rapid evolution of content generation technologies necessitates detection systems like HMGUARD to adapt dynamically to these emerging complexities, ensuring their continued relevance and efficacy. Additionally, we aim to investigate advanced reasoning and knowledge acquisition techniques to enhance the framework's decision-making capabilities in intricate scenarios. Promising approaches such as Graph-of-Thought [55] and Retrieval-Augmented Generation [56] present opportunities to augment HMGUARD's cognitive depth. By leveraging these methods, the system can achieve a more nuanced understanding of meme context and subtlety, ultimately improving detection accuracy and classification robustness.

VIII. CONCLUSION

In this work, we have conducted a comprehensive study to uncover the limitations of existing harmful meme detection tools and the challenges inherent in detecting harmful memes. This investigation highlights the pressing need for a robust and innovative detection framework. Our findings, for the first time, reveal the detrimental impact of meme compositions and propaganda techniques on detection accuracy, shedding light on previously overlooked complexities. Leveraging these critical insights, we have introduced the novel HMGUARD framework, specifically designed to address these challenges in harmful meme detection. Evaluation results demonstrate that HMGUARD effectively interprets and detects harmful memes, outperforming existing methods and addressing a critical gap in real-world applications. This work represents a significant step forward in moderating harmful content and fostering safer online environments. In the future, we aim to extend HMGUARD to support multilingual meme detection and enhance its ability to detect AI-generated content. These developments will address the evolving challenges posed by harmful media. Additionally, we plan to integrate advanced approaches to improve detection accuracy and adaptability.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive comments. This work is supported in part by the National Natural Science Foundation of China under Grants No. 61872430, 61402342, and the Key R&D Project of Hubei Province, China under Grants No.2023BAB165, No.2022BAA041.

REFERENCES

[1] "Meme," in *Merriam-Webster's Collegiate Dictionary*, 2003.

[2] Wikipedia contributors, "Meme — Wikipedia, the free encyclopedia," 2024, [Online; accessed 24-January-2024]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Meme&oldid=1196758187

[3] ——, "Internet meme — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Internet_meme&oldid=1195029354, 2024, [Online; accessed 24-January-2024].

[4] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, and S. Zannettou, "On the evolution of (hateful) memes by means of multimodal contrastive learning," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 293–310.

[5] S. Sharma, F. Alam, M. S. Akhtar, D. Dimitrov, G. Da San Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty *et al.*, "Detecting and understanding harmful memes: A survey," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 5597–5606.

[6] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[7] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2783–2796.

[8] N. Vishwamitra, K. Guo, S. Liao, J. Mu, Z. Ma, L. Cheng, Z. Zhao, and H. Hu, "Understanding and Analyzing COVID-19-related Online Hate Propagation Through Hateful Memes Shared on Twitter," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '23. New York, NY, USA: Association for Computing Machinery, 2024, p. 103–107. [Online]. Available: https://doi.org/10.1145/3625007.3630111

[9] M. Bilewicz and W. Soral, "Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization," *Political Psychology*, vol. 41, pp. 3–33, 2020.

[10] R. Hatzipanagos, "How online hate turns into real-life violence," https://www.washingtonpost.com/nation/2018/11/30/, 2018.

[11] J. Ji, W. Ren, and U. Naseem, "Identifying creative harmful memes via prompt based approach," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 3868–3872.

[12] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, "Prompting for multimodal hateful meme classification," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 321–332.

[13] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5244–5252.

[14] H. Lin, Z. Luo, J. Ma, and L. Chen, "Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[15] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, "Prompting for multimodal hateful meme classification," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 321–332.

[16] OpenAI, "Gpt-4v(ision) system card," 2023. [Online]. Available: https://openai.com/research/gpt-4v-system-card

[17] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Dissecting the meme magic: Understanding indicators of virality in image memes," *Proceedings of the ACM on human-computer interaction*, vol. 5, no. CSCW1, pp. 1–24, 2021.

[18] D. Dimitrov, B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino *et al.*, "Detecting propaganda techniques in memes," in *ACL-IJCNLP 2021-59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 2021, pp. 6603–6617.

[19] V. S. Greene, ""deplorable" satire: Alt-right memes, white genocide tweets, and redpilling normies," *Studies in American Humor*, vol. 5, no. 1, pp. 31–69, 2019.

[20] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, and P. Nakov, "A survey on computational propaganda detection," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4826–4832.

[21] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 623–18 633.

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[23] P. C. d. Q. Hermida and E. M. d. Santos, "Detecting hate speech in memes: a review," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 833–12 851, 2023.

[24] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Momenta: A multimodal framework for detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4439–4455.

[25] K. Guo, W. Zhao, M. Jaden, V. Vishwamitra, Z. Zhao, and H. Hu, "Understanding the Generalizability of Hateful Memes Detection Models Against COVID-19-related Hateful Memes," *International Conference on Machine Learning and Applications*. [Online]. Available: https://par.nsf.gov/biblio/10399964

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," in *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, 2022, pp. 171–183.

[29] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, "Towards explainable harmful meme detection through multimodal debate between large language models," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 2359–2370.

[30] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the Underlying Meaning of Multimodal Hateful Memes," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 5995–6003, aI for Good. [Online]. Available: https://doi.org/10.24963/ijcai.2023/665

[31] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal large language models: A survey," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 2247–2256.

[32] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Synthetic prompting: Generating chain-of-thought demonstrations for large language models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 30 706–30 775. [Online]. Available: https://proceedings.mlr.press/v202/shao23a.html

[33] O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant, "Answering questions by meta-reasoning over multiple chains of thought," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 5942–5966. [Online]. Available: https://aclanthology.org/2023.emnlp-main.364

[34] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[35] Facebook, "How to limit sensitive content that you see on instagram," 2024, [Online; accessed 9-April-2024]. [Online]. Available: https://www.facebook.com/help/251027992727268?cms_id=251027992727268

[36] P. Paudel, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Lambretta: learning to rank for twitter soft moderation," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 311–326.

[37] Liu, Haotian and Li, Chunyuan and Wu, Qingyang and Lee, Qingyang, "Llava: Large language and vision assistant," 2023. [Online]. Available: https://llava-vl.github.io/

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

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for multi-modal tracking," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 3492–3500.

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

[42] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 5995–6003.

[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[44] M. Lin, C. Dai, and T. Guo, "A method for generating explanations of offensive memes based on multimodal large language models," *Journal of Computer Research and Development*, vol. 61, no. 5, pp. 1206–1217, 2024.

[45] F. M. G. De Leon and R. Ballesteros-Lintao, "The rise of meme culture: Internet political memes as tools for analyzing Philippine propaganda," *Journal of Critical Studies in Language and Literature*, vol. 2, no. 4, pp. 1–13, 2021.

[46] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," *Domain adaptation in computer vision applications*, pp. 1–35, 2017.

[47] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, "Sniffer: Multimodal large language model for explainable out-of-context misinformation detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 052–13 062.

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

[49] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible nlp attacks," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1987–2004.

[50] Meta, "Build the future of ai with meta llama 3," 2024, [Online; accessed 23-April-2024]. [Online]. Available: https://llama.meta.com/llama3/

[51] N. Vishwamitra, K. Guo, F. T. Romit, I. Ondracek, L. Cheng, Z. Zhao, and H. Hu, "Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 181–181.

[52] K. Guo, A. Hu, J. Mu, Z. Shi, Z. Zhao, N. Vishwamitra, and H. Hu, "An Investigation of Large Language Models for Real-World Hate Speech Detection," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 1568–1573.

[53] K. Guo, A. Utkarsh, W. Ding, I. Ondracek, Z. Zhao, G. Freeman, N. Vishwamitra, and H. Hu, "Moderating illicit online image promotion for unsafe user-generated content games using large vision-language models," in *USENIX Security Symposium (USENIX Security)*, 2024.

[54] Y. Nong, M. Aldeen, L. Cheng, H. Hu, F. Chen, and H. Cai, "Chain-of-thought prompting of large language models for discovering and fixing software vulnerabilities," *arXiv preprint arXiv:2402.17230*, 2024.

[55] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk *et al.*, "Graph of thoughts: Solving elaborate problems with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 682–17 690.

[56] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 754–17 762.

[57] Wikipedia contributors, "Cohen's kappa — Wikipedia, the free encyclopedia," 2024, [Online; accessed 8-April-2024]. [Online]. Available: https://en.wikipedia.org/wiki/Cohen%27s_kappa

APPENDIX

A. Data Annotation for meme composition category

1) Data Annotation: To accurately identify the composition of each meme, we utilize GPT-4 [16] with a few-shot learning cue strategy to enhance its understanding and annotation capabilities.

Initially, we selected a subset of 100 samples from the dataset as a benchmark to compare the effectiveness of GPT-4 annotations with manual annotations. In our few-shot learning prompt, we provide GPT-4 with 12 carefully selected examples that illustrate the various compositions in memes that we aim to identify. At the same time, this subset was independently annotated by two expert researchers using a detailed measurement framework to ensure the validity of our comparison and to achieve 100% agreement after discussion.

The consistency and accuracy of GPT-4 annotations relative to human annotators were quantitatively assessed using Cohen's kappa coefficient [57]. Notably, we achieved a 92% level of agreement, indicating *almost perfect* agreement between GPT-4's automatic annotation and the manual work of the experts. This high degree of agreement indicates that GPT-4, with the help of the few-shot learning facilitation strategy, is proficient in identifying the composition of memes.

Then, encouraged by the convincing results, we extend the use of GPT-4 along with specialized few-shot learning prompts to the task of annotating the remaining unlabeled memes in the data collection. This comprehensive annotation process resulted in the entire dataset consisting of 289 memes, carefully annotated. This fully annotated dataset is ready for subsequent experimental studies to provide a basis for in-depth analysis and understanding of memes.

Please read the instruction and answer questions according to the following sentences:
Q1: What the number of panels of the meme?
☐ Single panel: memes that are composed of only one image
☐ Multiple panels: memes that are composed of a series of images
Q2: : What the type of the images of the meme?
☐ Photo: a picture taken by a camera
☐ Screenshot: an image of a screenshot taken from a computer screen
☐ Illustration: a drawing, painting, or printed work of art
Q3: Which kind of scale the meme is?
☐ Close up: a shot that tightly frames a person or object
☐ Medium shot: a shot that shows equality between subjects and background
☐ Long shot: a shot where the subject is no longer identifiable and the focus is on the larges scene rather than on one subject
Q4: Which kind of movement is included in the meme?
☐ Physical movement
☐ Emotional movement
☐ Causal movement

Fig. 8: Codebook of meme composition category annotation.

B. Definitions and Examples of Propaganda Techniques

We list all the propaganda techniques used in memes that have been discussed in the Dimitrov et al. study [18] as follows:

1. Loaded language: Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

2. Name-calling or labeling: Labeling the object of the propaganda campaign as something that the target audience fears, hates, finds undesirable, loves, or praises.

3. Doubt: Questioning the credibility of someone or something.

4. Exaggeration / Minimisation: Either representing something in an excessive manner: making things larger, better, worse (e.g., the best of the best, quality guaranteed) or making something seem less important or smaller than it really is (e.g., saying that an insult was actually just a joke).

5. Appeal to fear/ prejudices: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases, the support is built based on preconceived judgments.

6. Slogans: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

7. Whataboutism: A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

8. Flag-waving: Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or an idea.

9. Misrepresentation of someone's position (Straw man): Substituting an opponent's proposition with a similar one, which is then refuted in place of the original proposition.

10. Causal oversimplification: Assuming a single cause or reason when there are actually multiple causes for an issue. This includes transferring blame to one person or group of people without investigating the complexities of the issue.

11. Appeal to authority: Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. We also include here the special case where the reference is not an authority or an expert, which is referred to as Testimonial in the literature.

12. Thought-terminating cliche: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract the attention away from other lines of thought.

13. Black-and-white fallacy or dictatorship: Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).

14. Reductio ad hitlerum: Persuading an audience to disapprove of an action or an idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.

15. Repetition: Repeating the same message over and over again, so that the audience will eventually accept it.

16. Obfuscation, Intentional vagueness, Confusion: Using words that are deliberately not clear, so that the audience may have their interpretations. For example, when an unclear phrase with multiple possible meanings is used within an argument and, therefore, it does not support the conclusion.

17. Presenting irrelevant data (Red Herring): Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

18. Bandwagon: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action."

19. Smears: A smear is an effort to damage or call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.

20. Glittering generalities (Virtue): These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or an issue.

21. Appeal to (strong) emotions: Using images with strong positive/negative emotional implications to influence an audience.

22. Transfer: Also known as association, this is a technique that evokes an emotional response by projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another one in order to make the latter more acceptable or to discredit it.

Please read the instruction and answer questions according to the following sentences:

Q1:What are the words and images contained in the meme, and do the semantics of the text and image directly convey harmful content?
Words:
Images:
☐ Yes ☐ No

Q2: What are the explicit or implicit relationships between text and image, and does the relationship between the text and the image potentially reveal harmful content?
Relationship:
☐ Yes ☐ No

Q3a: Is meme a stitching image?
☐ Yes ☐ No

Q3b: If the answer to Q3a is 'Yes', consider the relationship between the images, and combine the text to understand whether the stitching images are trying to express explicit or implicit harmful content.
☐ Yes ☐ No

Q4: Are the following propaganda techniques used to explicitly or implicitly express harmful content?
'Loaded language', 'Name calling or labeling', 'Doubt', 'Exaggeration/ Minimisation', 'Appeal to fear / prejudices', 'Slogans', 'Flag-waving', 'Misrepresentation of someone's position', 'Causal oversimplification', 'Appeal to authority', 'Thought-terminating cliche', 'Black-and-white fallacy or dictatorship', 'Reductio ad Hitler', 'Repetition', 'Obfuscation', 'Intentional vagueness', 'Confusion', 'Presenting irrelevant data', 'Bandwagon', 'Smears', 'Appeal to strong emotions', 'Transfer'
☐ Yes ☐ No

Q5: Does the meme intend to have ant targeted derogatory, humiliating, insulting, satirical, or disparaging meaning?
☐ Yes ☐ No

Final Decision:
Combined with the above analysis of meme expression, does the meme have the following purpose or has any targeted derogatory, humiliating, insulting, satirical, or disparaging meaning?
☐ Harmful ☐ Harmless

Fig. 9: Codebook of harmful meme annotation.

TABLE XI: The TPR of different tools for detecting harmful memes with and without adversarial attack.

Catetory	ExplainHM	LLaVa	GPT-4	HMGuard
w/o adversarial attack	60%	26.67%	46.67%	100%
w/ adversarial attack	51.67%	23.33%	38.33%	95%

C. Code Book for "in the wild" Memes Annotation

We assigned two authors of this study to participate in data annotation, leveraging their extensive background knowledge and strong academic integrity. To ensure inter-rater reliability, we developed a common codebook for classifying memes in our collection as either harmful or harmless. The codebook is illustrated in Fig. 9.

D. NLP-based adversarial attacks against HMGUARD

1) Examples of NLP-based adversarial manipulations: We applied four types of NLP-based adversarial manipulations, as illustrated in Fig. 10: letter addition, where extra letters are inserted into words; letter deletion, removing some letters; letter swapping, which rearranges letters within words; and space insertion, adding spaces within or between words. These techniques subtly alter text to test the resilience of models while keeping the text understandable to humans.

2) Evaluation results: From TABLE XI, it can be observed that existing tools exhibit failures in handling adversarial

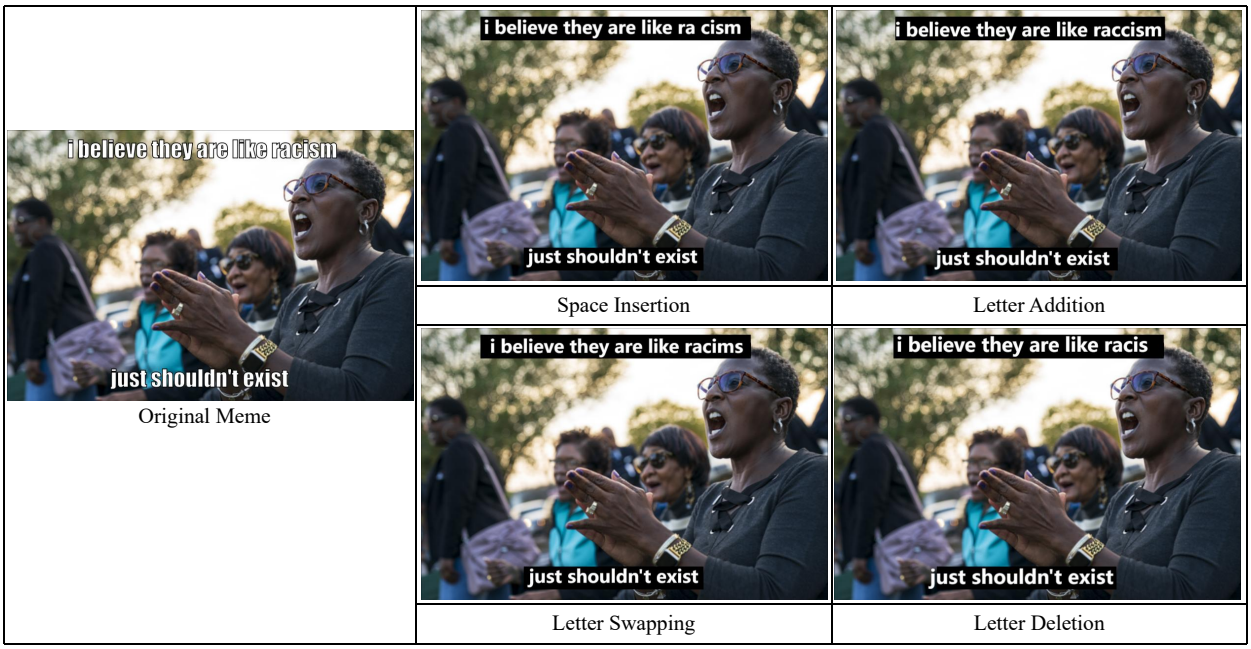

Fig. 10: Examples of four NLP-based adversarial manipulations.

examples, with the highest TPR reaching only 51.67%. Compared to the w/o adversarial attack scenario, all existing tools experience a noticeable drop in performance, indicating that NLP-based adversarial attacks in memes negatively impact harmful meme detection. In contrast, HMGUARD demonstrates strong robustness, achieving a TPR of 95% even in the presence of adversarial perturbations. Future work will further investigate the impact of adversarial examples on detection performance and explore effective mitigation strategies.