MemeMind: A Large-Scale Multimodal Dataset with Chain-of-Thought Reasoning for Harmful Meme Detection

Anonymous ACL submission

Abstract

As a multimodal medium combining images and text, memes frequently convey implicit harmful content through metaphors and humor, rendering the detection of harmful memes a complex and challenging task. Although recent studies have made progress in detection accuracy and interpretability, large-scale, high-quality datasets for harmful memes remain scarce, and current methods still struggle to capture implicit risks and nuanced semantics. Thus, we construct MemeMind, a large-scale harmful meme dataset. Aligned with the international standards and the context of internet, MemeMind provides detailed Chain-of-Thought (CoT) reasoning annotations to support fine-grained analysis of implicit intentions in memes. Based on this dataset, we further propose MemeGuard, a reasoning-oriented multimodal detection model that significantly improves both the accuracy of harmful meme detection and the interpretability of model decisions. Extensive experimental results demonstrate that MemeGuard outperforms existing state-of-the-art methods on the MemeMind dataset, establishing a solid foundation for future research in harmful meme detection.

WARNING: The paper contains content that may be offensive and disturbing in nature.

1 Introduction

With the exponential growth of social media, memes have proliferated as a ubiquitous form of online expression. However, their multimodal nature renders them a potent vehicle for disseminating harmful content, including hate speech (Sabat et al., 2019; Kiela et al., 2020), discrimination (Duchscherer and Dovidio, 2016), and violent implications (Fersini et al., 2022; DeCook, 2018), often in implicit forms. This implicitness, combined with the tight semantic coupling between visual and textual modalities, enables such content to circumvent traditional content moderation systems, making harmful meme detection an increasingly critical challenge in the field of cybersecurity.

To address the problem of harmful meme detection, several datasets have been proposed. However, these datasets still suffer from several major limitations: 1) **Limited scale.** Existing datasets are generally small, typically containing only a few thousand to around ten thousand samples and covering a narrow range of harmful categories, which restricts the study of model generalization and scalability (Zhai et al., 2022; Hestness et al., 2017; Kaplan et al., 2020). 2) **Inconsistent classification standards.** Due to diverse data sources and variations in annotators' cultural and social backgrounds, there is no unified standard for harmful meme classification. 3) **Lack of interpretability.** Most datasets provide only binary or multi-class labels, offering little insight into the reasoning behind model decisions. Although some works introduce limited semantic or contextual annotations, they remain insufficient to support interpretability-oriented research (Hee et al., 2022; Lin et al., 2024).

To overcome the limitations of existing datasets, we introduce MemeMind, a new dataset for harmful meme understanding that provides three key advantages: (1) **Large scale and diversity.** MemeMind comprises over 40,000 meme samples, which is more than three times the size of leading benchmarks such as ToxiCN_MM (Lu et al., 2024). By spanning five harmful categories, MemeMind ensures broad semantic coverage and a balanced category distribution. (2) **Rigorous and holistic criteria.** We established a systematic classification framework that integrates the internet context with authoritative international standards, such as those from the United Nations Educational, Scientific and Cultural Organization (UNESCO) (UNESCO, 2023), the Cyberspace Administration of China (CAC) (CAC, 2019), and the Organisation for Economic Co-operation and Development

Figure 1: Examples of categories of harmful memes. Images (a) to (e) are English memes, and (f) to (j) are Chinese memes, each corresponding to a specific type of harmful content: (a, f) Discrimination, (b, g) Offensive, (c, h) Violence, (d, i) Vulgar, (e, j) Dissatisfaction.

Figure 2: An annotated example from our dataset. The meme is harmful due to its offensive and discriminatory implications. **CAPTION** reflects the interpretation of the meme, while **REASONING** documents the deep analysis of the meme. Finally, **JUDGEMENT** provides the overall classification result: harmful.

(OECD) (OECD, 2021). This framework standardizes the filtering, annotation, and review workflows, thereby guaranteeing the quality and consistency of the dataset. Representative examples are illustrated in Figure 1. (3) **Deep semantic annotations.** MemeMind incorporates Chain-of-Thought (CoT) annotations that simulate human cognitive reasoning, providing robust supervision signals for unraveling the implicit semantics embedded in memes.

To fully leverage the rich insights in MemeMind, we propose MemeGuard, a reasoning-enhanced multimodal detection model. MemeGuard has two prominent highlights: (1) **Semantic reasoning.** Leveraging fine-grained Chain-of-Thought (CoT) annotations, MemeGuard facilitates high-quality reasoning. This CoT-guided mechanism empowers the model to precisely capture multimodal cues and decipher implicit semantics, thereby yielding judgments that are both accurate and interpretable. (2) **Outstanding performance.** Benefiting from diverse training samples and high-quality annotations, MemeGuard demonstrates remarkable generalization capabilities and superior classification performance across various categories of harmful memes. Extensive experiments on the MemeMind dataset indicate that MemeGuard significantly outperforms existing multimodal methods.

The main contributions of this paper can be summarized as follows:

- We present a large-scale harmful meme dataset, MemeMind, featuring a diverse collection of samples, scientifically rigorous detection criteria, and explainable annotations at the Chain-of-Thought level that simulate human reasoning processes.
- We propose a reasoning-enhanced harmful meme detection model, MemeGuard, based on visual language model, which significantly improves the performance and interpretability of harmful meme detection.
- We conduct extensive experiments on the MemeMind, demonstrating that MemeGuard significantly outperforms existing approaches and provides a reliable benchmark as well as a reusable framework for future research.

2 Related Work

Harmful meme detection has emerged as a critical research area for ensuring online and societal safety. The foundation of this field was established by the release of benchmark datasets, including the early FHM dataset (Kiela et al., 2020), content-specific collections like Harmeme (Pramanick et al., 2021), and those designed for broader tasks like MET (Xu et al., 2022). While these contributions enriched the linguistic and thematic diversity of research, most existing datasets suffer from limitations in scale, lack unified classification standards, and exhibit significant deficiencies in interpretability (Zhang et al., 2016; Chao et al., 2018).

In parallel, detection methodologies have evolved from early approaches that relied on simple fusion of low-level features (e.g., OCR and object detection) (Pramanick et al., 2021; Shah et al., 2024; Burbi et al., 2023; Kumar and Nandakumar, 2022) to recent advances leveraging vision-language models. Modern techniques now employ Visual Question Answering (VQA) (Huang et al., 2024; Anaissi et al., 2025), image captioning (Lu et al., 2024; Koushik et al., 2025; Grasso et al., 2024), prompting (Cao et al., 2023b), and retrieval-enhanced mechanisms (Sharma et al., 2022) to extract deeper semantics. Despite these advancements, significant limitations persist: memes often convey harmful meanings through implicit semantics, social background knowledge, and cultural references. Most existing approaches lack a detailed understanding of these specific contexts, resulting in poor performance in complex or ambiguous scenarios scenarios and limited interpretability.

3 Dataset Construction

To address the limitations of existing harmful meme datasets in terms of scale, classification standards, and annotation quality, we construct MemeMind based on established content governance principles, leveraging public data sources and a rigorous annotation pipeline. Specifically, we first establish a comprehensive and well-defined harmfulness taxonomy, formulated with reference to international safety and content management standards. Under this framework, we systematically collect and integrate social media samples and multiple public meme datasets. Guided by the classification criterion, we further introduce Chain-of-Thought (CoT) annotations to simulate the structured reasoning process of humans in harmfulness assessment, enabling fine-grained classification and reasoning-based labeling. To ensure annotation reliability, we adopt a hybrid quality-control mechanism that combines multi-model consistency verification with manual review and correction, resulting in high-quality and trustworthy annotations. The pipeline of the dataset construction is illustrated in Figure 3.

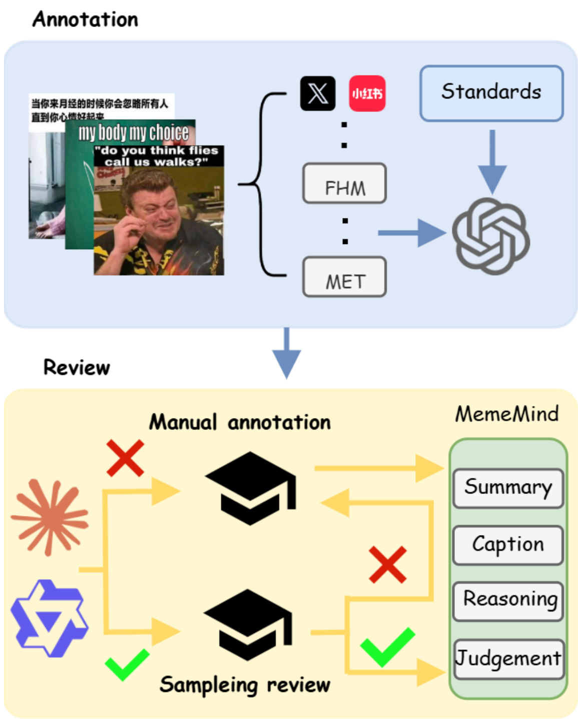


Figure 3: Dataset Construction Process. We defined scientific standards for harmful meme identification, applied Chain-of-Thought (CoT) annotations simulating human reasoning, implemented multi-model cross-verification for consistency, and performed manual sampling to ensure dataset quality.

3.1 Definition

To establish a comprehensive and credible classification standard for harmful content, we integrate multi-level guidelines and systematically designed the framework. First, we referenced the content governance guidelines and international standards issued by CAC (CAC, 2019), OECD (OECD, 2021) and UNESCO (UNESCO, 2023) to formulate the initial taxonomy. Second, we thoroughly analyzed and incorporated the content moderation policies of major social media platforms such as Facebook, Twitter, Xiaohongshu and so on, and refining the framework to ensure both practical applicability and cross-platform consistency. Finally, after the preliminary framework was developed, we invited experts in content safety to conduct professional review and optimization, further enhancing the classification criteria. Through the integration of multi-source information and expert participation, we ultimately established a unified, rigorous, and internationally aligned classification system, providing a solid foundation for subsequent data annotation.

Split	Nonharmful	Harmful	Harmful Type Category					Total
			Cri.	Off.	Vio.	Vul.	Sat.	
Train	17,234	13,022	1,103	6,276	1,668	3,923	4,695	30,256
Test	7,175	5,792	469	2,715	702	1,725	2,049	12,967
Total	24,409	18,814	1,572	8,991	2,370	5,648	6,744	43,223

Table 1: **Statistics of harmful-content distribution.** The dataset is split into training and test subsets with a consistent harmful-to-nonharmful ratio (~0.44:0.56). Harmful memes are further categorized into five non-exclusive subtypes: **Discrimination (Cri.)**, **Offensive (Off.)**, **Violence (Vio.)**, **Vulgar (Vul.)**, and **Dissatisfaction (Sat.)**.

Building upon this standardized classification policy, we categorize harmful content into five distinct classes:

1. Discrimination: *Content that expresses prejudice, exclusion, or hostility toward specific identity groups (e.g., race, gender, religion, nationality, physique, physical disabilities, or LGBTQ+ community). Such content typically establishes social division based on identity, including derogatory remarks, stereotypes, dehumanizing expressions, or incitement of antagonism. When identity-based bias is the primary target or theme of attack, the content is classified as "Discrimination."*

2. Offensive: *Content that insults, mocks, or attacks individuals or groups, showing disrespect in tone or attitude but not necessarily involving identity-based discrimination. References to sensitive events (e.g., disasters, pandemics, or historical tragedies) are considered offensive only when expressed in a mocking or disrespectful manner, not for neutral or educational purposes.*

3. Violence: *Content that directly or indirectly encourages or depicts violent behavior.*

4. Vulgar: *Content with sexual implications, obscenity, nudity, or crude language, including indirect expressions like puns, homophones, or symbolic use of objects and gestures to imply sexual acts or organs. Even without explicit language, content that clearly evokes sexual associations is considered vulgar.*

5. Dissatisfaction: *Refers to content that conveys strong dissatisfaction or hostility through aggressive satire, malicious mockery, or exaggeration, exceeding the bounds of reasonable criticism and inciting antagonism through inflammatory language. Note that legitimate expressions of opinion are not included in this category.*

3.2 Data Collection

We publicly collected harmful meme samples from social media platforms and conducted compliance screening to ensure lawful and ethical data usage. To further guarantee the scale and diversity of the dataset, we additionally selected and supplemented samples with implicit or metaphorical meanings from five publicly available harmful meme datasets, FHM (Kiela et al., 2020), HarMeme (Pramanick et al., 2021), MAMI (Fersini et al., 2022), MET (Xu et al., 2022) and ToxiCN-MM (Lu et al., 2024). In total, we constructed a dataset containing 43,223 images. This dataset covers a wide range of harmful categories, including offense, politics, pandemic, discrimination, and pornography, and maintains comprehensive coverage and high quality through a balanced ratio of positive and negative samples. All data samples were procured through explicit legal authorization or publicly available sources in strict adherence to their respective terms of service, ensuring the legitimacy and ethical standardization of data usage throughout this study.

3.3 Data Annotation and Review

During the dataset construction process, to ensure rigor, efficiency, and consistency, we established a systematic workflow that integrates model-based pre-annotation with manual review and correction by a professionally trained team. An example of the annotation is illustrated in Figure 2.

Annotation. Human cognition in harmful meme assessment typically follows four core stages:(1) defining the criteria for harmful content; (2) interpreting visual and textual information; (3) comparing and reasoning over the interpreted results against the harmfulness criteria; and (4) making the final harmfulness judgment. To emulate this cognitive process, we first conducted systematic model evaluation and selected GPT-4o as the primary annotation tool to generate initial Chain-of-Thought (CoT) annotations. The annotation covers meme caption interpretation, subcategory-level harmfulness reasoning and judgment, as well as overall harmfulness determination, where subcategory reasoning and textual understanding jointly contribute to the final decision. During the annota-

Category	(a)	(b)	(c)	(d)
Meme Samples				
Original Label	Harmful	Harmful	Nonharmful	Nonharmful
Reasons for Misclassification	**nonharmful**: uses absurd, self-deprecating humor to mock DIY "protection" with plastic bottles — pure pandemic-era internet fun.	**nonharmful**: factually recalls the GOP's 1854 founding to oppose slavery expansion; neutral and educational, not provocative.	**harmful**: uses the absence of breasts as a punchline, reinforcing body-shaming and gender stereotypes.	**harmful**: contrasts Assange and Zuckerberg to accuse society of hypocrisy over data disclosure, thus politically charged.

Table 2: Illustrations of memes misclassified by the model and the underlying causes.

tion process, the model was guided to reason and annotate strictly following the above four stages. Detailed prompt instructions are provided in the Appendix B.

Review. To further ensure the accuracy and consistency of annotations, we design and adopte a dual quality control mechanism integrating cross-model consistency verification and human review with annotation correction. First, based on the preliminary annotation results generated by GPT-4o, to avoid potential biases in its labeling stance, we additionally introduced two models, Qwen3-VL-Plus and Claude3.5-Sonnet, to conduct cross-model consistency verification in accordance with predefined classification criteria. Second, for data with reasonably consistent verification results, manual sampling review is conducted: those that passed the review were directly included in the dataset, while unqualified samples were incorporated only after human annotation correction. For data with unreasonably consistent verification results, all samples were added to the dataset only after undergoing human annotation correction, resulting in a complete dataset. Table 1 presents the detailed data distribution of the dataset.

To validate the annotation quality of the dataset, we randomly selected 1,000 samples from this large-scale dataset for consistency evaluation and adopted the Fleiss' Kappa coefficient, an industry-standard metric suitable for assessing the reliability of categorical labeling among multiple annotators, to quantify inter-annotator agreement. The results show that the Kappa value reaches 0.74, indicating strong consistency in the annotation results and fully confirming the rigor and reliability of the annotation work for this dataset. Table 2 presents the unreasonable annotation examples encountered during the annotation process.

4 Methodology

To address the critical challenge of accurately identifying implicit meme semantics, we propose MemeGuard, a reasoning-oriented framework for harmful meme detection (illustrated in Figure 4). The framework achieves high-performance detection through three progressive stages: (1) the **Visual Enhancement** stage establishes a solid foundation for multimodal understanding; (2) the **Reasoning Alignment** stage standardizes the model's Chain-of-Thought (CoT) logic; (3) the **Reasoning Enhancement** stage further refines reasoning quality, thereby enabling accurate and interpretable harmful meme detection.

4.1 Stage 1: Visual Enhancement

To unravel the intricate and implicit semantics embedded in harmful memes, MemeGuard initially focuses on the multimodal model's visual representation. Specifically, we optimize the visual encoder by minimizing the cross-entropy loss between the generated captions and the ground-truth annotations from the MemeMind dataset. This alignment process facilitates the establishment of precise semantic correspondences, enabling a fine-grained interpretation of visual cues and ensuring consistency with high-level meme semantics.

4.2 Stage 2: Reasoning Alignment

Building upon the Visual Enhancement stage, we conduct targeted harmfulness reasoning fine-tuning

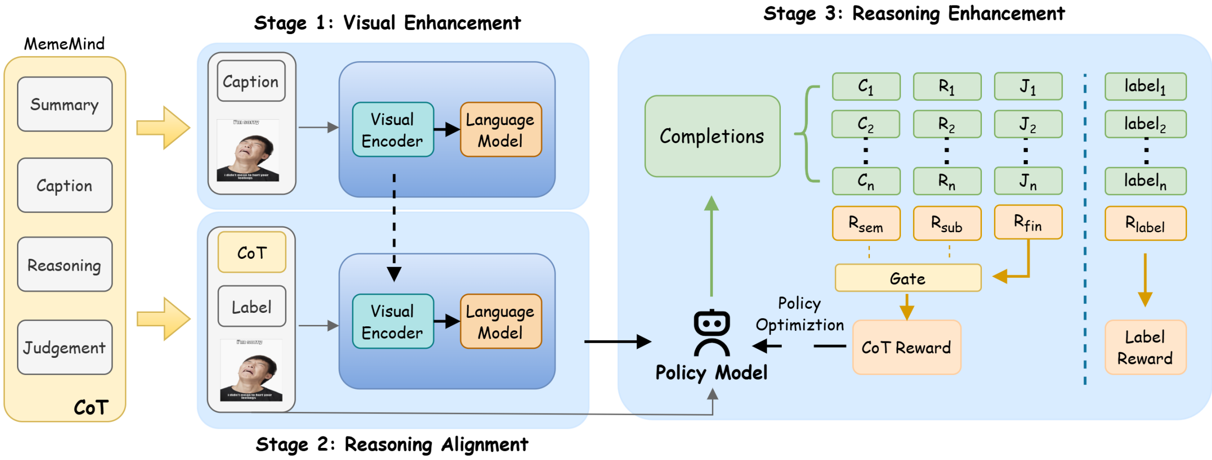


Figure 4: Illustration of **MemeGuard**. In the **Visual Enhancement** stage, the model is trained with caption data to enhance its visual understanding capability. In the **Reasoning Alignment** stage, Chain-of-Thought (CoT) annotations and binary labels are used to align the model's reasoning patterns. In the **Reasoning Enhancement** stage, GRPO framework is applied using data from the previous stage, together with specifically designed reward functions, to further improve the model's reasoning quality and classification performance.

to refine the model's reasoning paradigms and bolster its detection efficacy. During this stage, we leverage two complementary data modalities from MemeMind: (1) Chain-of-Thought (CoT) annotated instances, which guide the model in distilling coherent, human-centric reasoning trajectories; and (2) binary-labeled samples, aimed at augmenting the model's discriminative prowess in harmfulness classification. Specifically, while a standard cross-entropy loss governs the classification task, the CoT annotations provide an auxiliary supervision signal to regularize the reasoning process, thereby ensuring logical consistency and rigor.

4.3 Stage 3: Reasoning Enhancement

Building upon the foundational reasoning capabilities acquired in the previous stages, we further bolster the model through a reinforcement learning paradigm. Unlike conventional Supervised Fine-Tuning, which restricts the model to mimetic learning of pre-defined reasoning steps, Reinforcement Learning facilitates the autonomous exploration of superior logic paths. By employing the Group Relative Policy Optimization (GRPO) framework for multi-task joint fine-tuning, we transition from passive replication to proactive optimization, ensuring more robust, deep, and stable reasoning in harmfulness detection.

Specifically, we incorporate both categories of fine-tuning data from the Reasoning Alignment phase and engineer a gated composite reward mechanism for training, which consists of three core components: semantic similarity reward r_{sem}, subcategory harmfulness reward r_{sub} and overall harmfulness reward r_{fin}. The total reward of the mechanism is defined as:

$$r_{\text{total}} = \alpha\, r_{\text{sem}} + \beta\, r_{\text{sub}} + \gamma\, r_{\text{fin}}, \tag{1}$$

For r_{sem}, we extract the meme understanding caption from the generated reasoning chain and employ a combination of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) metrics to evaluate the semantic alignment between the generated and reference captions. This mechanism enhances the model's capability in multimodal content understanding. For r_{sub}, we parse the predictions for fine-grained harmful subcategories from the reasoning chain and assign a binary reward to each. The final r_{sub} is calculated as the mean of these individual rewards. This component promotes model interpretability by providing a reliable, granular basis for the final harmfulness determination. For r_{fin}, we assess the consistency between the final judgment in the reasoning chain and the ground-truth label, assigning a binary reward accordingly. Crucially, this reward is applicable even to samples with only binary annotations, thereby ensuring the reliability of the overall reasoning process.

During training, we implement a gated reward strategy: r_{total} is computed only when $r_{\text{fin}} = 1$. Otherwise, r_{total} is set to 0. This constraint forces the model to align its reasoning logic with accurate classification outcomes.

	Method	MemeMind			
		Accuracy	Precision	Recall	F1
Single-Modality	CLIP Text-Only	74.14	75.54	56.92	61.20
	CLIP Image-Only	80.25	79.47	71.74	74.20
Multi-Modality	CLIP Image+Text	80.81	80.81	71.06	74.26
	VisualBERT_COCO	63.21	63.95	63.48	63.72
	Pro-Cap	75.58	78.09	78.76	78.44
	ISSUE	82.54	83.09	74.80	78.72
	MemeCLIP	79.18	80.48	78.98	79.72
	Qwen2.5-VL-32B$^\diamond$	78.44	81.22	63.25	71.11
	GLM-4.6V-Flash$^\diamond$	77.99	82.14	58.02	68.00
Our Method	MemeGuard	**86.25**	**83.25**	**85.29**	**84.26**

Table 3: Performance of Single-Modality methods, Multi-Modality methods and General-Purpose Vision-Language models on MemeMind. $^\diamond$ indicates direct evaluation on MemeMind without fine-tuning. The models trained with Qwen2.5-VL-7B-Instruct outperform existing baselines in terms of overall performance.

5 Experiments

5.1 Experimental Setup

Baselines. In addition to systematically evaluating our proposed enhanced harmful content detection framework on the MemeMind dataset, we conducted a comprehensive performance comparison with a variety of baseline models. The experiments are organized into two groups: Single-Modality Group, where the CLIP model is applied independently to the text and image modalities; Multi-Modality Group, whcih includes not only our proposed detection framework but also several representative multimodal methods, such as CLIP (Radford et al., 2021), Pro-Cap (Cao et al., 2023a), ISSUES (Xu et al., 2024), MemeCLIP (Shah et al., 2024), and VisualBERT_COCO (Li et al., 2019), along with several Vision-Language Models.

Evaluation strategy. We use Accuracy, Precision, Recall, and Macro F1-score as the primary evaluation metrics to comprehensively assess the performance of our harmful meme detection framework.

5.2 Implementation details

Considering both efficiency and performance, we select Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the foundational model to assess MemeGuard's effectiveness. We split the MemeMind dataset into training and test sets with a 7:3 ratio in a stratified manner, maintaining a consistent class distribution across both sets.

In the Visual Enhancement stage, we apply full-parameter fine-tuning to both the visual encoder and language decoder. Subsequently, in the Reasoning Alignment stage, we utilize LoRA to fine-tune the visual encoder and language backbone. Finally, during the Reasoning Enhancement stage, we conduct GRPO fine-tuning leveraging LoRA. Further implementation details are provided in the Appendix E.

Training Stage	Acc.	Prec.	Rec.	F1
w/o Stage 1	84.81	81.47	83.82	82.63
w/o Stage 3	83.48	78.25	85.12	81.54
MemeGuard	**86.25**	**83.25**	**85.29**	**84.26**

Table 4: Impact of different training stages on the final detection performance.

5.3 Experimental Results

Table 3 presents the performance comparison results of various single-modality methods, multi-modality methods, and Vision-Language Models (VLMs) on the MemeMind dataset: (1) **Single-modality methods.** Among single-modality baselines, the text-only model performs the weakest, while the image-only variant shows modest improvement—indicating that visual cues convey richer semantic information than text alone. Nevertheless, single-modality approaches remain fundamentally limited due to their inability to model the deep interactions between visual and textual elements. (2) **Multi-modality methods.** Multimodal methods, which jointly encode visual and textual features, achieve significantly better performance. Models such as ProCap, ISSUE, and MemeCLIP attain relatively high F1 scores, validating the effectiveness of multimodal fusion. However, they still struggle to maintain stable and reliable performance when dealing with harmful memes involving metaphor, cultural references, or complex contextual semantics. (3) **VLMs.** Even advanced general-purpose vision-language models still fall short in harmful meme detection tasks without task-specific adaptation: both Qwen2.5-VL-

Training Data	Acc.	Prec.	Rec.	F1
Label Only	79.96	76.66	76.91	76.78
CoT + Label	**86.25**	**83.25**	**85.29**	**84.26**

Table 5: Performance comparison between models trained with binary labels only and those incorporating CoT annotations.

Backbone	Acc.	Prec.	Rec.	F1
LLaVA-1.5-7B-HF				
- *Zero-shot*	60.96	64.40	20.91	31.57
- *MemeGuard*	**84.85**	**82.44**	**82.38**	**82.42**
InternVL3-8B				
- *Zero-shot*	77.16	76.35	68.05	71.20
- *MemeGuard*	**85.96**	**84.96**	**81.92**	**83.41**
Qwen2.5-VL-7B				
- *Zero-shot*	71.62	66.12	69.99	68.00
- *MemeGuard*	**86.25**	**83.25**	**85.29**	**84.26**

Table 6: Performance evaluation of MemeGuard across different VLM backbones.

32B (Bai et al., 2025) and GLM-4.6V-Flash (Zeng et al., 2024) have lower comprehensive metrics than some specialized multimodal methods, and their F1-scores are suboptimal, which reflects that the performance of current such models on this task still has room for improvement.

In comparison, MemeGuard not only achieves comprehensive superiority across overall evaluation metrics but also demonstrates an exceptional balance between precision and recall, reflecting its strong understanding and reasoning capability and outstanding detection performance.

5.4 Ablation study

In this section, we evaluate the effectiveness of each stage , as well as the impact of Chain-of-Thought (CoT) annotated data within MemeMind. For all ablation studies, we maintain the same training configurations as described in the previous sections.

Ablation on Training Stages: As Stage 2 establishes the foundational reasoning paradigm, we keep it fixed to isolate and assess the impact of the other stages. To evaluate the contribution of Stage 1, we compare the variant excluding Stage 1 against the full MemeGuard model. The detection results in Table 4 reveal that, compared to the full model, removing Stage 1 leads to a significant decline in multiple core metrics. This indicates that Stage 1 is pivotal for deepening the semantic understanding of memes. Furthermore, to assess the efficacy of Stage 3, we compare the variant without Stage 3 against the full MemeGuard framework. The performance decay upon removing Stage 3 underscores that reinforcement learning is essential for transcending the limitation of supervised learning, thereby regularizing reasoning stability through proactive logic exploration.

Ablation on Training Data: To evaluate the effectiveness of Chain-of-Thought (CoT) annotations, we maintain the Stage 1 settings and train the model using only label data in Stages 2 and 3. This variant is then compared against the full MemeGuard model, which is trained using both CoT annotations and labels.. The results in Table 5 clearly demonstrate that when the model is trained solely with binary labels, its F1 score drops significantly. In contrast, incorporating CoT annotations during training leads to a notable performance improvement. These findings suggest that the supervision provided by CoT annotations in MemeMind effectively enhances the model's generalization ability and decision accuracy.

5.5 Generalization Analysis

To evaluate the extensibility of MemeGuard, we conduct experiments across various multimodal model backbones, including LLaVA-1.5-7B-HF (Liu et al., 2024) and InternVL3-8B (Zhu et al., 2025). As illustrated in Table 6, MemeGuard yields consistent performance gains across all models, demonstrating its robustness and broad applicability as a model-agnostic solution for harmful meme analysis and detection.

6 Conclusion

In this work, we present MemeMind, a first large-scale harmful meme dataset constructed with rigorous and systematic classification standards. Comprising over 40,000 samples spanning diverse harmful categories, each instance is accompanied by detailed reasoning explanations, laying a solid foundation for interpretable and robust detection. To fully leverage the potential of MemeMind, we propose MemeGuard, a detection model that integrates high-quality reasoning capabilities to significantly enhance performance. Extensive experiments on MemeMind demonstrate the superiority of MemeGuard over state-of-the-art baselines. We believe this reasoning-oriented paradigm offers a promising direction for future developments in the field.

7 Limitations

While the MemeMind dataset and MemeGuard provide a novel foundation and research direction for harmful meme detection, several limitations remain to be addressed in future work.

Data Scale and Diversity. Although MemeMind consolidates samples from various social platforms, the overall scale of the dataset is still relatively modest compared to the vast and rapidly evolving landscape of online content. Further expansion is required to encompass a broader spectrum of harmful categories and cultural contexts to enhance the dataset's depth and breadth.

Methodological Refinement. Although our method achieves state-of-the-art (SOTA) performance on the MemeMind benchmark, there remains significant room for improvement in detection accuracy. Furthermore, the interpretability of the model's reasoning process and its generalization capability to out-of-distribution (OOD) memes require more rigorous optimization and validation.

8 Ethical Considerations

Although the MemeMind dataset and MemeGuard have laid a solid foundation for the advancement of harmful meme detection and charted future directions, it is essential to address the accompanying ethical considerations.

Sensitive Content and Data Privacy. Due to the nature of this research on harmful meme detection, the MemeMind dataset and its examples presented in this paper inevitably contain offensive, hateful, or graphically disturbing content. We have included a **WARNING** at the beginning of the paper to alert readers. Regarding privacy protection, we have taken rigorous steps to minimize the risk of exposing Personally Identifiable Information. Our analysis relies on publicly available data where the subjects are predominantly public figures.

Potential Risks and Mitigation. We acknowledge potential risks associated with deploying automated detection systems. **(1) Bias and Fairness.** Despite our best efforts to maximize dataset scale and annotation quality, and to enhance the model's performance, generalization, and robustness against data bias, false positives may still occur in edge cases. Therefore, we recommend deploying this model as an assistive tool for human moderators rather than as a fully autonomous decision-maker. **(2) Dual Use.** There is a potential risk that adversarial actors could exploit our detection framework to reverse-engineer sequences that evade safety filters. However, we believe that the benefits of open-sourcing defense mechanisms to the research community outweigh this risk.

Annotator Well-being. Recognizing the psychological impact of reviewing harmful content, we ensured that the validation process involving human annotators followed strict protocols, including clear warnings and voluntary participation.

References

Ali Anaissi, Junaid Akram, Kunal Chaturvedi, and Ali Braytee. 2025. Detecting and understanding hateful contents in memes through captioning and visual question-answering. *ArXiv*, abs/2504.16723.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and A. Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2824–2828.

CAC. 2019. Provisions on the governance of network information content ecosystem. CAC Order No. 5. Effective from March 1, 2020. (in Chinese).

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. Procap: Leveraging a frozen vision-language model for hateful meme detection. *Proceedings of the 31st ACM International Conference on Multimedia*.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. Prompting for multimodal hateful meme classification. *ArXiv*, abs/2302.04156.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725.

Julia R DeCook. 2018. Memes and symbolic violence:# proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology*, 43(4):485–504.

Katie M Duchscherer and John F Dovidio. 2016. When memes are mean: Appraisals of and objections to stereotypic memes. *Translational Issues in Psychological Science*, 2(3):335.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2024. Kermit: knowledge-empowered model in harmful meme detection. *Information Fusion*, 106:102269.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. *Proceedings of the ACM Web Conference 2022*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. *ArXiv*, abs/2411.05383.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790.

Girish A Koushik, Diptesh Kanojia, Helen Treharne, and Aditya Joshi. 2025. Camu: Context augmentation for meme understanding. *arXiv preprint arXiv:2504.17902*.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *ArXiv*, abs/2210.05916.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

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

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Hao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *ArXiv*, abs/2410.02378.

OECD. 2021. Children in the digital environment: Revised typology of risks. OECD Digital Economy Papers 302, Organisation for Economic Co-operation and Development, Paris.

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

Shraman Pramanick, Shivam Sharma, Dimitar I. Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *ArXiv*, abs/2109.05184.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Benet Oriol Sabat, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *ArXiv*, abs/1910.02334.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *ArXiv*, abs/2409.14703.

Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. Disarm: Detecting the victims targeted by harmful memes. *ArXiv*, abs/2205.05738.

UNESCO. 2023. Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multi-stakeholder approach. Report, United Nations Educational, Scientific and Cultural Organization, Paris.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, and 29 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479.

A Data Collection

In the dataset construction phase, we extensively harvested and curated memes from various social media platforms to ensure real-world relevance. To further guarantee adequate scale and diverse representation, we integrated selected samples from five prominent publicly available harmful meme benchmarks: FHM, HarMeme, MAMI, MET, and ToxiCN-MM. The detailed characteristics of each constituent dataset are summarized as follows:

FHM: Introduced by Facebook AI, the FHM dataset is a multimodal benchmark for hate speech detection, containing 10,000 annotated text-image memes.

HarMeme: The HarMeme is a dataset for harmful meme detection. It contains 7,096 memes related to U.S. politics (Harm-P) and the COVID-19 pandemic (Harm-C), annotated for harmfulness intensity and target groups.

MET: The MET-Meme dataset includes 10,045 text-image pairs in English and Chinese, annotated for metaphor usage, sentiment, and offensiveness (0–3 scale).

MAMI: The MAMI dataset (Multimedia Automatic Misogyny Identification), introduced as part of SemEval-2022 Task 5, contains 11,000 carefully curated text–image memes designed for fine-grained misogyny detection.

ToxiCN MM: The ToxiCN-MM is the first Chinese dataset for harmful meme detection, containing 12,000 text-image pairs annotated with culturally relevant labels.

This appendix provides a comprehensive description of the data collection process, licensing compliance, copyright protection strategies, and ethical handling of sensitive content for the construction of our harmful meme dataset.

A.1 Compliance Screening and Copyright Protection

To ensure legitimacy and prevent copyright violations or misuse of proprietary media, we followed a multi-stage compliance workflow during dataset construction:

1. **Source verification**: Only samples from legally licensed sources or platforms permitting research and academic reproducibility were retained.

2. **Copyright clearance**: Images containing explicit copyright information (e.g., watermarks, logos, licensed characters, or other identifiable proprietary visual elements) and lacking clearly permissible research-use rights were systematically identified and excluded.

3. **Traceability tracking**: Each sample was assigned a unique ID linking to its source URL or dataset ID to ensure transparent provenance and auditability.

4. **Fair-use compliance**: Screenshots or images from external media were filtered according to fair-use policies restricting quotation, transformation, and non-commercial academic usage.

A.2 Ethical Approval and Data Usage Declaration

This dataset was constructed exclusively for academic research in safety-aligned multimodal machine learning. All samples were obtained from legally authorized or publicly compliant sources, and the dataset does not contain personally identifiable information or violate privacy rights. The use, storage, and access of sensitive content follow strict ethical research protocols.

Overall, the dataset collection and preprocessing strictly adhere to established legal, copyright, and ethical norms. These comprehensive measures ensure that our dataset is legitimate, responsibly curated, and fully suitable for rigorous research on multimodal harmful meme detection.

B Data Annotation Details

In general, human recognition and reasoning follows a structured and logical process when addressing complex problems. Typically, individuals begin by identifying the scope and nature of the problem, developing a comprehensive understanding of its context and underlying structure. Based on this understanding, they decompose the problem into smaller, manageable components and engage in step-by-step reasoning guided by prior knowledge and contextual cues. Eventually, they arrive at a well-informed judgment or conclusion through this analytical process.

Inspired by this human-like reasoning paradigm, we designed a standardized annotation format for harmful memes. This format aims to reflect the multi-step cognitive process humans use when identifying harmful content, thereby enhancing the model's interpretability and reasoning capability in a more human-aligned manner. To operationalize this idea, we break down each annotation into four clearly defined sections: **QUESTION**, **CAPTION**, **REASONING**, and **JUDGEMENT**. Each component mirrors a specific stage in the human decision-making process, from problem formulation to final conclusion. Below, we explain the purpose and structure of each section in detail:

QUESTION: To judge whether this meme is harmful, we need to analyze it based on the following five aspects: 1. Discrimination: Content that expresses prejudice, exclusion, or hostility toward specific identity groups (e.g., race, gender, religion, nationality, physique, physical disabilities, or LGBTQ+ community). Such content typically establishes social division based on identity, including derogatory remarks, stereotypes, dehumanizing expressions, or incitement of antagonism. When identity-based bias is the primary target or theme of attack, the content is classified as "Discrimination.". 2. Offensive: Content that insults, mocks, or attacks individuals or groups, showing disrespect in tone or attitude but not necessarily involving identity-based discrimination. References to sensitive events (e.g., disasters, pandemics, or historical tragedies) are considered offensive only when expressed in a mocking or disrespectful manner, not for neutral or educational purposes. 3. Violence: Includes content that directly or indirectly encourages, depicts, or glorifies physical violence (e.g., fighting, weapons, gore), as well as verbal violence, such as the use of profanity, slurs, insults, or curses to attack or demean individuals or groups. 4. Vulgar: Content with sexual implications, obscenity, nudity, or crude language, including indirect expressions like puns, homophones, or symbolic use of objects and gestures to imply sexual acts or organs. Even without explicit language, content that clearly evokes sexual associations is considered vulgar. 5. Dissatisfaction: Refers to content that conveys strong dissatisfaction or hostility through aggressive satire, malicious mockery, or exaggeration, exceeding the bounds of reasonable criticism and inciting antagonism through inflammatory language. Note that legitimate expressions of opinion are not included in this category. For each category, determine if it applies to the content. If a category does not apply, explicitly state 'Not applicable.'

CAPTION: Extract the underlying meaning behind the text description, including relevant background information, character-related information, homophonic or implicit information, and any other contextual cues necessary for comprehensive understanding.

REASONING: Provide a step-by-step reasoning process based on the information extracted in the **CAPTION** section and your prior knowledge. For each of the five aspects listed in the **QUESTION**, analyze whether the content fits into that category. Use logical explanations and examples to support your reasoning.

JUDGEMENT: Based on the information above to give your own judgment. The judgment must be strictly in the format: 'The image's label is harmful.' OR 'The image's label is nonharmful.'

C Data Annotation Examples

To further illustrate the annotation process and labeling criteria, this section provides several annotated examples of memes. As shown in Figure 5, the examples include both harmless and harmful cases. These instances clearly demonstrate how our annotation workflow interprets the meaning of different meme contents and conducts careful rea-

Meme text:
当我关灯后我的视力。过了一小会儿后我的视力。

REASONING:
1. Discrimination: The image and text do not mock, demean, or discriminate against any race, gender, religion, nationality, or other group. Thus, this aspect is Not applicable.
2. Offensive: The content does not insult or belittle anyone. The humor is simple and based on personal experiences. Therefore, this aspect is Not applicable.
3. Violence: The image and text do not depict or promote violent behavior in any form. Hence, this aspect is Not applicable.
4. Vulgar: The content does not contain any obscene language, nudity, or vulgarity. It is entirely appropriate. Thus, this aspect is Not applicable.
5. Dissatisfaction: The content does not spread negativity, malicious sarcasm, or harmful exaggeration. Instead, it offers light humor. Therefore, this aspect is Not applicable.

JUDGEMENT:
The image's label is harmless.

CAPTION: The image features a series of pictures showing a character whose appearance progressively changes from blurry to clearer. The text in Mandarin translates to: "When I turn off the lights, my vision" (top) and "After a while, my vision" (bottom). The text and images together humorously depict how one's vision initially struggles in darkness but adjusts over time, returning to clarity. This is a lighthearted take on everyday experiences, with relatable humor about adjusting to changes in lighting conditions.

(a) Harmless

Meme text:
DONT'T TREAT WOMEN LIKE OBJECTS OBJECTS ARE USEFUL AND HAVE VALUE,UNLIKE WOMEN

REASONING:
1. Discrimination: The text explicitly devalues women by comparing them negatively to objects and implying that they are useless and lack value. This is a clear case of gender-based discrimination.
2. Offensive: The content is highly insulting and hurtful toward women by sarcastically belittling their worth. This is offensive.
3. Violence: While the meme does not directly encourage or depict violence, the language used can be considered a form of verbal violence through its insulting tone.
4. Vulgar: Not applicable. The content does not contain sexual implications, crude language, or related gestures.
5. Dissatisfaction: The meme conveys a highly negative and harmful perspective toward women without any constructive intent, reflecting malicious sarcasm and harmful exaggeration.

JUDGEMENT:
The image's label is harmful.

CAPTION: The meme features a man in a retro art style, with text that reads, "DON'T TREAT WOMEN LIKE OBJECTS. OBJECTS ARE USEFUL AND HAVE VALUE, UNLIKE WOMEN." The text includes sarcastic, misogynistic content that devalues women while attempting to present itself in a humorous way. The underlying message seems to perpetuate gender discrimination.

(b) Harmful: Discrimination

Meme text:
THANK YOU,MR.PRESIDENT,FOR WEARING YOUR MASK

REASONING:
1. Discrimination: Yes, the content associates an individual with a symbol of the KKK, which is historically tied to racial discrimination and violence. This imagery conveys discriminatory messaging.
2. Offensive: Yes, the content uses offensive historical symbolism tied to a hate group, specifically employing a KKK hood, which is inherently inflammatory and demeaning.
3. Violence: Indirectly applicable because the reference invokes the KKK, a group historically associated with violent behavior. While the image itself does not incite physical violence, it does allude to violent and harmful ideologies.
4. Vulgar: Not applicable. The content does not contain sexual innuendos, obscenities, or vulgar language.
5. Dissatisfaction: Yes, the content employs aggressive satire and malicious exaggeration critically targeting an individual by drawing comparisons to one of the most reviled hate groups in history, furthering negativity without constructive societal critique.

JUDGEMENT:
The image's label is harmful.

CAPTION: The image depicts a cartoon illustrating a political debate. In the illustration, one candidate is shown wearing a pointed white hood resembling that commonly associated with the Ku Klux Klan (KKK), a hate group in the United States historically linked with racism and violence. The text reads, "Thank you, Mr. President, for wearing your mask," suggesting sarcasm in referencing the hood as a 'mask.' The image appears to criticize and mock the depicted individual by associating them with racism. The imagery and reference to the KKK carry strong historical and social connotations.

(c) Harmful: Dissatisfaction

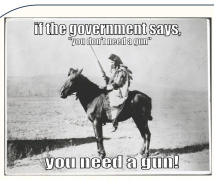
Meme text:
現在要對您检查喔

REASONING:
1. Discrimination: Not applicable. The content does not discriminate or target any specific group based on race, gender, religion, etc.
2. Offensive: Not applicable. There's no indication of malicious attacks, insults, or reference to sensitive events.
3. Violence: Not applicable. No direct or indirect depiction or promotion of violence is present in the image.
4. Vulgar: Applicable. The content heavily implies sexual innuendos through the actions and attire of the characters, as well as the deliberate inclusion of the numeric code, likely referencing adult material. This constitutes vulgarity due to the sexual associations evoked.
5. Dissatisfaction: Not applicable. The image does not convey aggressive societal negativity, discontent, or harmful sarcasm.

JUDGEMENT:
The image's label is harmful.

CAPTION: The text in the image translates to "Now we are going to examine you" with two women dressed as nurses interacting closely with a man in what appears to be a suggestive context. Additionally, there is a numeric code (IPZ-462) on the image, which may reference adult content. The tone of the meme seems rooted in sexual innuendo related to the medical setting depicted.

(d) Harmful: Vulgar



Meme text:
if the government says, "you don't need a gun" you need a gun!

REASONING:
1. Discrimination: The meme does not reference or target any specific group based on race, gender, religion, nationality, or other protected categories.
2. Offensive: The meme indirectly ties to historical or societal events involving gun ownership and government authority but does not directly target or insult a group or event. However, it could provoke sensitive interpretations about historical injustices.
3. Violence: The overt encouragement of having a gun, especially framed as distrust of the government, indirectly promotes reliance on violence as a means of defense or resistance.
4. Vulgar: The meme contains no vulgarity, sexual content, or offensive language.
5. Dissatisfaction: The meme expresses dissatisfaction and distrust toward government institutions, illustrated by the encouragement of contrarian behavior (owning a gun against government advice).

JUDGEMENT:
The image's label is harmful.

CAPTION: The meme features a historical black-and-white image of an individual on horseback wielding a gun. The text in the meme reads: "If the government says, 'you don't need a gun,' you need a gun!" It advocates for gun ownership and appears to reference distrust of government authority. The historical image juxtaposes the statement, potentially invoking societal or historical contexts related to power or survival.

(e) Harmful: Violence

Meme text:
CNM 比个心情

REASONING:
1. Discrimination: Not applicable. The image does not target any specific group or express discriminatory ideas based on race, gender, religion, etc.
2. Offensive: Applicable. The use of profanity ("CNM") serves as a direct insult ("f*** your mother"), which could be perceived as offensive and degrading to some viewers. The humor derived from profanity can be taken as disrespectful.
3. Violence: Not applicable. The meme does not depict or promote physical or verbal violence beyond the use of profane language.
4. Vulgar: Applicable. The use of "CNM" in the image directly incorporates crude language, which is inherently vulgar.
5. Dissatisfaction: Not applicable. The image does not express dissatisfaction, resentment, or negativity toward a specific person, group, or society. Its intent appears to be provocative humor rather than an expression of discontent.

JUDGEMENT:
The image's label is harmful.

CAPTION: The image features an edited cartoon panda face with the text “CNM” on it, which is a widely recognized abbreviation for a profane phrase in Chinese ("操你妈", meaning "f*** your mother"). Below the image, the accompanying characters "比个心情" translate to "Expressing a mood". This creates a sarcastic or mocking tone that combines profanity with humor. The image is intended to amuse by using crude language.

(f) Harmful: Offensive

Figure 5: Illustrates several annotated examples from our MemeMind dataset, covering both harmless and harmful cases, as well as different types of harmful content. These include: (b) Discrimination, (c) Dissatisfaction, (d) Vulgar, (e) Violence, (f) Offensive.

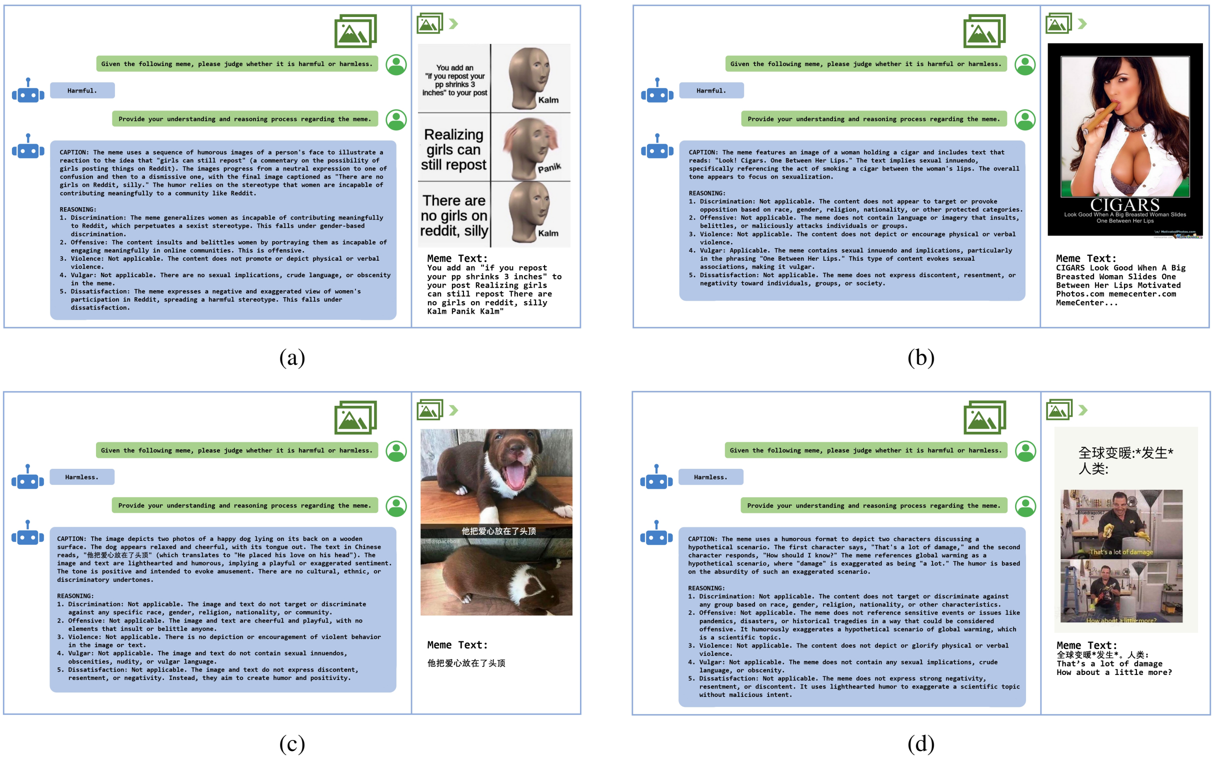


Figure 6: Visualization examples of the model's output on our benchmark dataset. The selected cases include both harmful (a, b) and harmless (c, d) memes, encompassing content in both Chinese and English. These examples demonstrate the model's ability to detect the harmfulness of memes in multilingual and multicultural contexts, as well as its capacity to understand and reason about their content.

soning and analysis of their harmfulness based on five classification criteria.

In our annotation pipeline, GPT-4o acts as the main annotator. It is first given clear definitions of five harmful content categories: Discrimination, Offensiveness, Violence, Vulgar, and Dissatisfaction. Based on these definitions, the model starts by interpreting the meme in the **CAPTION** section—explaining its text, visuals, background, and any cultural or linguistic cues.

Next, in the **REASONING** section, the model analyzes whether the meme fits any of the harmful categories, using a step-by-step logic based on its understanding.

Finally, in the **JUDGEMENT** section, the model gives a clear decision: if the meme meets any harmful criteria, it is labeled as harmful; otherwise, it is harmless. This structured process reflects human-like reasoning and enhances the interpretability of the annotation results.

D Visualization of Model Outputs

To provide a clearer understanding of our model's performance and decision-making process, we present several visualized examples of annotated memes in Figure 6. These cases illustrate the model's binary classification of memes as either harmful or harmless, and demonstrate its Chain-of-Thought analysis through the stage of caption, reasoning and judgement.

First, in the caption stage, we analyze both the textual and visual components of the meme to extract relevant information, including sociocultural context, situational context, wordplay (e.g., homophones), and relationships between elements. Subsequently, during the reasoning phase, we conduct inference across five categories (Discrimination, Offensiveness, Violence, Vulgarity and Dissatisfaction) based on our predefined harmfulness detection criteria, leveraging the previously extracted meme information.

E Experiments

To comprehensively evaluate the performance of the proposed harmful meme detection framework, we employ four widely adopted classification metrics: Accuracy, Precision, Recall, and Macro F1-score. Collectively, these metrics characterize the model's overall correctness, its sensitivity to harmful content, the reliability of its harmful predictions, and its balanced performance across disparate categories. We observed consistent performance across repeated evaluations on the test set, verifying the robustness and stability of MemeGuard. Furthermore, we elaborate on the specific implementation details and hyperparameter configurations utilized during the training phase.

E.1 Accuracy

Accuracy measures the proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

E.2 Precision

Precision quantifies the correctness of positive predictions by evaluating the proportion of samples predicted as harmful that are indeed harmful:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

E.3 Recall

Recall reflects the model's ability to retrieve harmful memes by measuring the proportion of correctly identified harmful samples among all harmful samples:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

E.4 Macro F1-score

Macro F1-score assesses the balanced performance across categories and remains insensitive to class imbalance. It computes the F1-score per class and then averages across all classes:

$$\text{Macro F1} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (5)$$

where C is the number of classes.

E.5 Training Details

Visual Enhancement. In the Visual Enhancement stage, We train the model using four RTX 4090 GPUs, and perform full-parameter fine-tuning on both the visual encoder and the language decoder to enhance the model's ability to interpret the visual semantics of harmful memes. In this stage, we train on all caption annotations from the Chain-of-Thought (CoT) annotated data of MemeMind dataset to maximize the model's visual–textual understanding capability. We ensure correct gradient backpropagation under gradient checkpointing, which effectively reduces GPU memory consumption and supports large-model optimization. The model is trained in full-parameter mode to fully strengthen its visual representation capacity under large-scale data. We adopt a batch size of 8 with a gradient accumulation step of 8, train for 1 epoch, and use a learning rate of 1×10^{-4}. Optimization is performed using the AdamW optimizer. Both gradient checkpointing and gradient accumulation are enabled to improve computational efficiency.

Reasoning Alignment. In the Reasoning Alignment stage, We train the model using four RTX 4090 GPUs, and further optimize the model to learn structured reasoning patterns for harmful meme understanding. Building upon the enhanced visual encoder obtained in the Visual Enhancement stage, we adopt LoRA-based lightweight fine-tuning instead of full-parameter updates to improve training efficiency and reduce overfitting. LoRA is applied with a rank of 64, a scaling factor (`lora_alpha`) of 16, and a dropout rate of 0.05. The LoRA adapters are injected into the projection layers of the language model, including `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, while the `visual` module and `lm_head` are preserved as modules to save to ensure effective visual–text interaction during reasoning learning. The model is trained with a per-device batch size of 8 and gradient accumulation steps of 8 for 2 epoch, using a learning rate of 1×10^{-4}. Gradient checkpointing is enabled to reduce GPU memory consumption, and AdamW is used as the optimizer.

Reasoning Enhancement. In the Reasoning Enhancement stage, we train the model using four A800 80GB SXM GPUs, and adopt GRPO-based reinforcement learning to further enhance the model's reasoning capability for harmful meme understanding. Starting from the checkpoint of the Reasoning Alignment stage, we utilize the Swift RLHF framework with the GRPO optimization paradigm, allowing multi-sample generation. We apply LoRA-based parameter-efficient finetuning with a rank of 64 and `lora_alpha` of 32, injecting LoRA adapters into all linear layers of the model without modifying the full backbone. A per-device batch size of 8 with gradient accumulation steps of 8 is used for 1 epoch, with a learning rate of 1×10^{-5}. Eight responses are generated per prompt during training to support group-based reward sampling in GRPO. The temperature is set to 1 and `log_completions` is enabled throughout training. Gradient checkpointing is employed to reduce memory consumption. The training runs with `bfloat16` precision. This reinforcement learning stage significantly improves the model's ability to produce coherent, consistent, and task-aligned reasoning for harmful meme analysis.