SGoT-R1: Social Graph of Thought Reasoning-Enhanced Multimodal Large Language Model for Harmful Meme Detection

Xiuxian Wang, Yuting Su, Wenhui Li*, Xiaowen Wang, Zhuojun Li, Anan Liu*

School of Electrical and Information Engineering, Tianjin University, 300072, China
wxx19971219@163.com, ytsu@tju.edu.cn, liwenhui@tju.edu.cn,
wxwen@tju.edu.cn, 2021202347@tju.edu.cn, anan0422@gmail.com

Abstract

Internet memes serve as widely distributed multimodal social content that conveys complex ideas through metaphorical expressions, often containing harmful implications that make accurate harmful meme detection an important problem. Reasoning knowledge extracted from large language models plays a crucial role in recent advances in harmful meme detection. However, these methods only perform reasoning analysis on memes from a single opinion, ignoring that memes are essentially products of group consensus, where their true meaning interpretation highly depends on the collision and aggregation process of diverse user viewpoints. To address this problem, we propose a Social Graph of Thought Reasoning Enhancement (SGoTRE) framework for harmful meme detection. The SGoTRE contains three key steps: First, through multi-agent simulation technology, we obtain diverse chains of thought that represent the parsing logic of users from different backgrounds toward memes, authentically restoring the diversity characteristics of group cognition. Second, we construct a Social Graph of Thought (SGoT) that effectively integrates multi-chain reasoning processes and structurally expresses the consensus and diversity of viewpoints among users. Finally, we utilize the SGoT for cognitive distillation, internalizing multi-opinion reasoning logic into a single multimodal large model (SGoT-R1) to achieve efficient and interpretable harmful meme detection. Experimental results show that SGoT-R1 significantly improves detection performance on mainstream datasets. Particularly on the most challenging FHM dataset, SGoT-R1 achieves an 8.9% improvement over state-of-the-art models.

Introduction

The rise of social media gives birth to a new type of multimodal entity—memes. As products of collective consensus, memes typically combine visual elements with concise text, conveying user emotions and cultural appeals through metaphorical expression(Shifman 2013). This type of content has strong sharing properties and can rapidly spread across various online platforms. While memes are often presented in humorous forms, some are strategically used to spread harmful content. These harmful memes can attack individuals or groups based on targeting identity information such as race, gender, and religion. Their widespread dissemination not only intensifies online public opinion conflicts but may also distort public perception, leading to the formation of misguided value orientations. Therefore, developing algorithms that can accurately detect memes containing harmful metaphors is crucial for eliminating potential social risks in online environments.

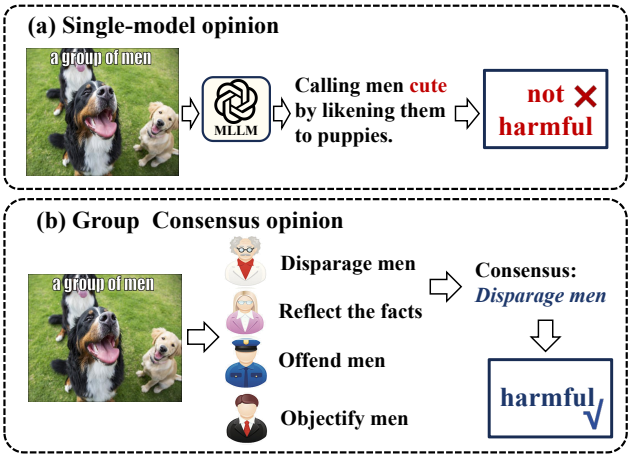

Figure 1: Comparison between the Single-Model opinion and the Group Consensus opinion.

With the diversification of detection methods and continuous improvement of benchmark datasets, the field of harmful meme detection makes significant progress(Kiela et al. 2020; Pramanick et al. 2021b,a). Early research mainly focuses on capturing interaction mechanisms of multimodal content (such as matching relationships between text and image content)(Pramanick et al. 2021b; Kiela et al. 2019; Cao et al. 2023b; Su et al. 2025), learning association patterns between features of different modalities. When these patterns match known harmful characteristics (such as strong associations between malicious text and provocative images), content is classified as harmful memes. However, the annotation process for harmful meme detection data re-

*Corresponding Author

quires annotators to have rich social experience and strong contextual understanding abilities, resulting in extremely high annotation costs and making it difficult to create large-scale, high-quality labeled datasets. Due to this limitation, models struggle to effectively learn the deep associations between images and text, which naturally leads to insufficient generalization ability.

With the rapid advancement of large language models (LLMs), recent studies have utilized their strong reasoning and extensive knowledge to aid in harmful content detection (Lin et al. 2023; Hee and Lee 2025). However, these methods often rely on a single model's opinion, which can introduce cognitive bias. Memes, as products of collective consensus (Shifman 2013), derive their metaphorical meanings from the shared understanding and cultural negotiation within communities. A single model cannot fully capture this consensus-forming process, leading to potential misjudgments when analyzing complex memes. For instance, Figure 1 shows a meme that disparages men by comparing them to dogs. In Figure 1(a), the single-model approach misclassifies the meme as "expressing cuteness" and fails to detect its harmful intent. In contrast, Figure 1(b) adopts a group consensus opinion, aggregating multiple user opinions. While individual views may differ, the majority agree the meme intends to satirize men. This demonstrates that group consensus-based methods are more accurate and robust in detecting complex metaphors and harmful content.

Based on the above analysis, we attempt to propose a corresponding method from the perspective of group consensus. The core idea of this method can be divided into three progressive steps: (1) Diverse user opinion acquisition: The foundation of group consensus lies in the exchange and integration of diverse user viewpoints. Thus, we first comprehensively collect users' varied interpretations of memes, providing a data basis for subsequent consensus modeling. (2) User opinion consensus modeling: In forming group consensus, user opinions display both commonalities and significant differences. Simply aggregating these views often overlooks important distinctions and fails to reflect how consensus evolves. Therefore, we use a graph structure to connect shared elements among opinions while preserving existing differences, fully reconstructing the consensus formation process. (3) Social cognitive reasoning distillation: Repeatedly invoking large models for group reasoning is time- and resource-intensive. By leveraging knowledge distillation, we embed group reasoning logic into a single model, which greatly reduces computational costs while maintaining analytical performance.

To achieve these goals, we propose the Social Graph of Thought Reasoning Enhancement (SGoTRE) framework for harmful meme detection. Through this framework, we trained a Social Graph of Thought Reasoning-enhanced multimodal large language model (SGoT-R1). Our SGoTRE enhances the powerful metaphorical reasoning capabilities of multimodal large language models (MLLMs) by introducing the Social Graph of Thought (SGoT). Specifically, we first utilize multi-agent systems based on MLLMs to simulate users' thinking processes when analyzing memes according to different user profiles, generating diverse user chains of thought. Subsequently, we fuse and structurally process these diverse user chains of thought to construct SGoT that reflects group consensus and viewpoint diversity. Finally, we propose a novel multimodal large model training paradigm: taking memes as input, SGoT as the reasoning process, and ground-truth category labels as output. The SGoT-R1 trained through this paradigm can restore the cognitive processes of multi-opinion users, achieving interpretable detection of harmful memes.

Our key contributions are summarized as follows:

- We propose a novel SGoTRE framework for harmful meme detection that systematically addresses the limitations of existing single-opinion methods in cognitive bias by integrating the diversity and consensus of user opinions, effectively capturing collective cognitive patterns in meme comprehension.
- We introduce SGoT-R1, a multimodal large language model that distills multi-agent cognitive reasoning processes into a single unified architecture through SGoT. This design maintains reasoning opinion diversity while eliminating expensive online multi-agent computation, achieving a balance between performance and efficiency.
- Comprehensive experiments on three benchmark datasets validate the advanced performance of SGoT-R1. On the challenging FHM dataset, our method achieves an 8.9% accuracy improvement over state-of-the-art approaches while providing interpretable reasoning explanations.

Related Work

Research on multimodal information interaction relationships primarily follows two technical paradigms: one approach adopts traditional dual-stream architectures, performing classification and detection based on visual and textual feature fusion (Kiela et al. 2020, 2019; Pramanick et al. 2021b; Cao et al. 2023b; Su et al. 2025); the other focuses on domain adaptation of pre-trained multimodal models to improve task-specific performance (Lippe et al. 2020; Pramanick et al. 2021a; Yang et al. 2022; Cao et al. 2023c; Muennighoff 2020; Sharma et al. 2022). Although these methods achieve significant progress in performance metrics, their decision-making processes still suffer from insufficient interpretability. To address this issue, (Hee, Lee, and Chong 2022) proposes a reverse reasoning framework to reveal the formation mechanism of image-text associations in hate content recognition. Meanwhile, attempts to solve problems through enhanced data resources also face major obstacles. Although recent studies attempt to construct datasets with metaphorical content annotations, the high cost of manual annotation severely constrains dataset scale, thereby limiting model generalization capability.

Recent research explores leveraging the rich knowledge reserves and powerful reasoning capabilities of pre-trained vision-language models or large language models to generate auxiliary knowledge for enhancing hate content detection and improving model generalization (Cao et al. 2023a; Lin et al. 2023, 2024; Hee and Lee 2025). However, these methods still have key limitations: they often treat large

models as a single expert opinion, and the generated reasoning knowledge exhibits obvious singularity. However, memes are essentially products of group consensus, and their interpretation relies on the collision and aggregation of different user viewpoints. Single-opinion reasoning fails to match the collective nature of memes, potentially leading to judgment bias.

Method

Problem Statement

We define the harmful meme detection dataset as a collection containing multiple memes, where each meme $M = \{I, T\}$ consists of an image I and a text sequence T. This study transforms the harmful meme detection task into a natural language generation problem: the model takes the meme text T and image I as input and generates a label output $Y \in \{\text{harmful}, \text{not harmful}\}$, thereby explicitly determining whether the meme is harmful.

Overview

In this section, we introduce a novel SGoTRE framework for harmful meme detection. Based on the SGoTRE, we train a multimodal large language model SGoT-R1 with multi-opinion user cognitive reasoning capabilities. The overall structure of our SGoTRE is illustrated in Figure 2.

Diverse User Opinion Acquisition

To replicate the diversity of user opinions in real-world scenarios, this module simulates the cognitive and reasoning patterns of different groups through a multi-agent system. The specific process is as follows:

User Pool Given that users' occupations significantly influence their sociocultural backgrounds, value orientations, and decision-making logic (Berthet 2022), we adopt occupation as the primary criterion for user categorization. Specifically, we directly utilize the real Twitter user data (including user profiles and tweet histories) already collected by the social simulator HiSim (Mou, Wei, and Huang 2024). Subsequently, we apply the method proposed in (Gao et al. 2024) to identify and classify users' occupations, ultimately categorizing them into ten occupational groups. To ensure the representativeness and diversity of the simulated population, we further screen high-influence users within each occupational group—measuring influence by the number of likes and retweets their posts receive. For each occupation, we select the top-h most influential representative users, constructing a simulated user set $U = \{u_1, u_2, \ldots, u_{N_u}\}$ consisting of N_u users in total. The opinions of these representative users are considered to reflect the typical stances and cognitive tendencies of their respective occupational groups.

Agent-based Human-like Simulation To enable agents to accurately simulate the target users, we inject corresponding user profile information into each agent. The injected information includes: (1) user profile (such as occupational background, age, and interests), and (2) comment habits (such as tone, commonly used expressions, and preferences regarding sensitive topics). Based on these profiles, we design a prompt P_1 to guide the agent in generating user-consistent reasoning processes. The agent is required to output a structured chain of thought (CoT) comprising four steps. The prompt P_1 is as follows:

> Please simulate the user based on {user profile} and {commenting habits}, and generate an interpretation chain for the meme according to the following reasoning steps:
> 1. Content Interpretation: Extract the surface meaning of both image and text elements;
> 2. Image-Text Relationship Analysis: Assess the relationship between the image and text;
> 3. Stance Identification: Specify the attitude towards individuals or groups involved in the meme;
> 4. Metaphor Analysis: Based on the above, uncover the deeper metaphors implied by the image-text combination.

Following these steps, each agent simulates the logical reasoning of a specific user when interpreting a given meme $M_i = \{I_i, T_i\}$, where I_i is visual information of the i-th meme and T_i is textual information of the i-th meme. This process yields a user chain of thought $C_{i,j} = \{c^1_{i,j}, c^2_{i,j}, c^3_{i,j}, c^4_{i,j}\}$, where $C_{i,j}$ denotes the chain of thought produced by the j-th user for the i-th meme, $c^1_{i,j}$ denotes the 1-th step of $C_{i,j}$. The above steps can be formalized as follows.

$$P_1 = [t_1, prof_j, habit_j] \tag{1}$$

$$C_{i,j} = \text{MLLM}(M_i, P_1), \tag{2}$$

where t_1 is the instruction of P_1, $prof_j$ is the j-th user's profile, $habit_j$ is the j-th user's comment habit.

User Opinion Consensus Modeling

After obtaining N_u chains of thought generated by agents for each meme, this module constructs a Social Graph of Thought map that embodies both consensus and diversity through feature encoding, clustering analysis, and linkage association. The specific process is as follows:

Feature Encoding For each four-step reasoning chain generated by an agent—including image-text interpretation, image-text joint analysis, stance judgment, and metaphor analysis—we treat the reasoning result of each step as an independent node $c^k_{i,j}$. We use the GME embedding model to encode all nodes, converting the textual reasoning content into low-dimensional dense vector representations $\mathbf{v}^k_{i,j}$. This process provides a numerical basis for subsequent clustering and correlation analysis.

$$\mathbf{v}^k_{i,j} = f_{\text{GME}}\left(c^k_{i,j}\right) \quad k = 1, 2, 3, 4, \tag{3}$$

where $\mathbf{v}^k_{i,j}$ is the feature vector representation of the $c^k_{i,j}$.

Clustering Analysis Based on the feature encoding above, we perform clustering at each reasoning step for every meme to identify consensus viewpoints.

For each meme and each reasoning step, we collect all feature vectors generated by N agents:

$$\mathcal{V}^k_i = \left\{\mathbf{v}^k_{i,j} \mid j = 1, 2, \ldots, N_u\right\} \tag{4}$$

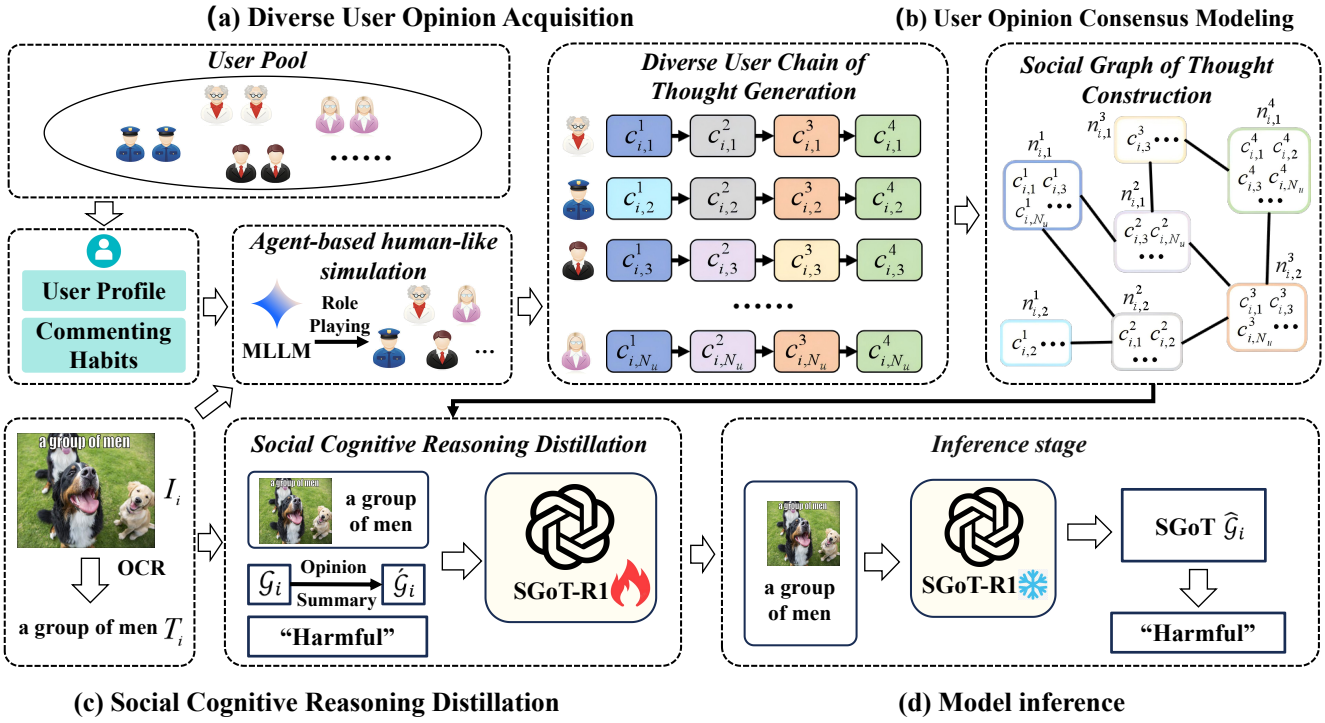

Figure 2: An overview of our SGoTRE. Based on the SGoTRE framework, we train a large multimodal language model called SGoT-R1. For memes composed of images and text sequences, SGoT-R1 first generates a SGoT to analyze the meme as part of its reasoning process, and then outputs a label indicating whether the meme is harmful.

We then apply hierarchical clustering (Dasgupta and Long 2005) to $\mathcal{V}_i^k$:

$$\mathcal{N}_i^k = \text{Cluster}\left(\mathcal{V}_i^k\right) \tag{5}$$

$$\mathcal{N}_i^k = \{n_{i,g}^k \mid g = 1, 2, ..., G_i^k\} \tag{6}$$

where $\mathcal{N}_i^k$ denotes the set of consensus nodes identified for i-th meme at reasoning step k, $n_{i,g}^k$ denotes the g-th consensus node of $\mathcal{N}_i^k$, G_i^k is the number of consensus nodes in the k-th step of the CoT for i-th meme.

The optimal number of clusters is determined by maximizing the silhouette coefficient (Rousseeuw 1987), which measures the cohesion and separation of the clusters. Each cluster represents a group of agents who share similar reasoning at the corresponding step. Thus, for each meme and each reasoning step, we obtain multiple consensus nodes, which collectively reflect the diversity and consensus of user opinions.

Social Graph of Thought Construction Following the logical structure of the four-step reasoning chain ("content interpretation → image-text relationship analysis → stance identification → metaphor analysis"), we establish cross-step connections between consensus nodes as follows:

For i-th meme M_i, at each reasoning step k ($k = 1, 2, 3, 4$), let the set of consensus nodes be $\mathcal{N}_i^k = \{n_{i,g}^k\}$.

We add a edge from consensus node $n_{i,g}^k$ to $n_{i,h}^{k+1}$ if there exists at least one agent j whose reasoning chain passes through both clusters, i.e.,

$$\exists\, j \in \{1, 2, \ldots, N_u\}, \quad \mathbf{v}_{i,j}^k \in n_{i,g}^k, \quad \mathbf{v}_{i,j}^{k+1} \in n_{i,h}^{k+1} \tag{7}$$

Under this condition, we define the set of cross-step edges for M_i as:

$$\mathcal{E}_i = \left\{ \left(n_{i,g}^k,\, n_{i,h}^{k+1}\right) \,\middle|\, \exists\, j,\ \mathbf{v}_{i,j}^k \in n_{i,g}^k,\ \mathbf{v}_{i,j}^{k+1} \in n_{i,h}^{k+1},\ k = 1, 2, 3 \right\} \tag{8}$$

By preserving the multi-consensus-node structure at each step, we ultimately construct a SGoT for each meme:

$$\mathcal{G}_i = \left(\bigcup_{k=1}^{4} \mathcal{N}_i^k,\ \mathcal{E}_i \right) \tag{9}$$

This graph encapsulates both intra-layer clustering relations and inter-layer logical links: the diversity of opinions is captured by the multiple consensus nodes in each layer, while the cross-layer edges represent the logical continuity of group reasoning chains.

Social Cognitive Reasoning Distillation

To address the high computational cost of multi-agent online reasoning, we distill the multi-opinion consensus reasoning logic embedded in the SGoT into a single MLLM. This allows the model to restore the cognitive processes of users with diverse opinions and achieve interpretable detection of harmful memes. The specific process is as follows:

Opinion Summary In a single consensus node, the repeated occurrence of highly similar user opinions leads to information redundancy and increases the computational costs of the model during both training and inference phases. To address this issue, we input the diverse user opinions within each consensus node into a LLM, converting the collection of multiple opinions within the node into a single opinion summary. This process can be formulated as follows:

$$\acute{n}_{i,g}^{k} = LLM([P_2, n_{i,g}^{k}]) \tag{10}$$

$$\acute{\mathcal{N}}_{i}^{k} = \{\acute{n}_{i,g}^{k} \mid g = 1, 2, ..., G_i^k\} \tag{11}$$

$$\acute{\mathcal{G}}_i = \left(\bigcup_{k=1}^{4} \acute{\mathcal{N}}_i^k,\ \mathcal{E}_i \right) \tag{12}$$

Where P_2 is the prompt of opinion summary, and detailed information can be found in the supplementary material.

Based on the annotated meme dataset, training samples are constructed in the form of triplets $(M_i, \acute{\mathcal{G}}_i, Y_i)$ for each sample, where M_i denotes the input, $\acute{\mathcal{G}}_i$ represents the reasoning process, and Y_i corresponds to the output.

Model Distillation Based on the constructed triplet dataset $(M_i, \acute{\mathcal{G}}_i, Y_i)$, this paper proposes a knowledge distillation method. The goal is to distill the collective reasoning ability of multi-agent systems into a single MLLM. The process is as follows:

First, we treat the SGoT $\acute{\mathcal{G}}_i$, generated by multiple agents, as the teacher signal. This graph serves as supervision for the model's reasoning process. The student model $\mathcal{F}_{stu}$ is a single MLLM, which takes the meme's visual and textual information M_i as input. The model is required to sequentially output a predicted SGoT $\hat{\mathcal{G}}_i$ (as the intermediate reasoning result), and then produce the final harmfulness label $\hat{Y}_i$. During training, the student model needs to reconstruct the group reasoning path and accurately classify the harmfulness label. To improve parameter efficiency and reduce computational cost, we apply LoRA for fine-tuning. LoRA inserts trainable low-rank matrices into selected layers of the model. By only updating a small number of additional parameters, the model adapts efficiently to specific tasks. The input-output flow of the distillation process can be formalized as the following equation.

$$\left(\hat{\mathcal{G}}_i,\ \hat{Y}_i\right) = \mathcal{F}_{stu}(M_i) \tag{13}$$

where $\hat{\mathcal{G}}_i$ is the intermediate reasoning result of $\mathcal{F}_{stu}$. $\hat{Y}_i$ is the final detection result of $\mathcal{F}_{stu}$.

Training Objective and Optimization

This work uses a joint cross-entropy loss for end-to-end optimization. The loss function constrains the model's output on two objectives:

Social Graph of Thought Generation Loss: The teacher signal's SGoT is serialized into a text sequence $\acute{\mathcal{G}}_i = (w_1, w_2, \ldots, w_T)$, and the student model generates the SGoT is serialized into a text sequence $\hat{\mathcal{G}}_i = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_T)$. At each time step t, the student model outputs a probability distribution $\hat{p}_t^{(v)}$ over the vocabulary. The graph cross-entropy loss is defined as:

$$\mathcal{L}_{\text{graph}}(\acute{\mathcal{G}}_i, \hat{\mathcal{G}}_i) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v=1}^{V} \mathbf{1}(w_t = v) \cdot \log \hat{p}_t^{(v)} \tag{14}$$

where V is the vocabulary size.

Harmful Label Classification Loss: For binary harmfulness classification, the ground-truth label is denoted as $Y_i \in \{0, 1\}$, and the student model predicts the probability $\hat{Y}_i \in [0, 1]$ of being harmful. The binary cross-entropy loss is defined as:

$$\mathcal{L}_{\text{cls}}(Y_i, \hat{Y}_i) = -\left(Y_i \cdot \log(\hat{Y}_i) + (1 - Y_i) \cdot \log(1 - \hat{Y}_i)\right) \tag{15}$$

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{graph}}(\acute{\mathcal{G}}_i, \hat{\mathcal{G}}_i) + \mathcal{L}_{\text{cls}}(Y_i, \hat{Y}_i) \tag{16}$$

Datasets	Train		Test	
	Harmful	Not-harmful	Harmful	Not-harmful
FHM	3050	5450	500	500
Harm-C	1064	1949	124	230
Harm-P	1486	1534	173	182

Table 1: Statistical distributions of datasets used for evaluation.

Experiments

Datasets and Metrics

We use three publicly available meme datasets for evaluation: (1) FHM (Kiela et al. 2020), (2) Harm-C (Pramanick et al. 2021a), and (3) Harm-P (Pramanick et al. 2021b). Harm-C and Harm-P contain memes related to COVID-19 and U.S. politics, respectively. FHM is released by Facebook as part of a challenge for crowdsourced multimodal harmful meme detection to address the problem of harmful speech, as shown in Table **??**. In FHM, each meme is labeled as either harmful or harmless, while Harm-C and Harm-P are originally annotated with three categories: very harmful, somewhat harmful, and not harmful. For a fair comparison, following the settings of recent related work (Hee and Lee 2025), we merge the very harmful and somewhat harmful categories into a single harmful class. To ensure fair and comprehensive evaluation, we adopt two widely utilized metrics in harmful meme detection: Accuracy (Acc) and Macro-F1 score (Su et al. 2025; Lin et al. 2023).

Experimental settings

To generate diverse user chains of thought for each meme, we employ the multimodal large language model Gemini-1.5 (Team et al. 2024) developed by Google, specifically the version Gemini-1.5-flash. This version achieves performance comparable to state-of-the-art MLLMs while significantly reducing economic and computational costs. To ensure the reproducibility of results, we set the temperature parameter to 0.01 and the maximum generation length to 1024.

Type	Model	Year	FHM		Harm-C		Harm-P	
			Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Unimodal	Text BERT	NeurIPS 2019	57.1	41.5	70.2	66.3	80.1	78.4
	Image-Region	CVPR 2016	52.3	34.2	68.7	63.0	73.1	72.8
Multimodal	MMBT-Region	NeurIPS 2019	65.1	61.9	73.5	67.1	82.5	80.2
	MOMENTA	ACL 2021	61.3	57.5	83.8	82.8	89.8	88.3
	MaskPrompt	EMNLP 2022	72.9	65.2	84.5	81.5	88.2	87.1
	HateSieve	NAACL 2025	73.5	71.6	83.6	83.1	88.8	88.5
Knowledge-enhanced	Pro-Cap	ACM MM 2023	74.9	71.7	85.0	83.2	89.3	87.9
	MR.HARM	EMNLP 2023	75.4	75.1	86.2	85.4	89.6	89.6
	ExplainHM	WWW 2024	75.6	75.4	87.0	86.4	90.7	90.7
	IntMeme	AAAI 2025	71.5	71.2	81.9	80.4	87.3	86.1
Collective Consensus	SGoT-R1 (Ours)	2025	**84.5**	**84.3**	**89.0**	**88.7**	**91.3**	**91.3**

Table 2: harmful meme detection performance comparisons on FHM, Harm-C and Harm-P dataset

Model	Time	FHM		Harm-C		Harm-P	
		ACC	F1	ACC	F1	ACC	F1
GPT 4o							
w/ zero-shot	3.7s	68.8	68.3	67.5	60.3	63.1	64.5
w/ SGoT	15.3s	72.1	71.8	73.2	70.3	80.6	80.3
Gemini 1.5 Flash							
w/ zero-shot	3.0s	60.2	58.9	66.1	64.2	61.6	55.5
w/ SGoT	14.9s	67.8	62.5	70.4	68.9	78.2	78.4
SGoT-R1-2B	0.67s	79.4	79.1	87.3	86.8	89.0	88.7
SGoT-R1-7B	1.62s	**84.5**	**84.3**	**89.0**	**88.7**	**91.3**	**91.3**

Table 3: Evaluation the effectiveness of SGoT in harmful meme detection

Model	FHM		Harm-C		Harm-P	
	ACC	F1	ACC	F1	ACC	F1
Qwen2VL-2B						
w/ zero-shot	54.2	53.9	56.7	51.5	53.9	21.8
w/ SFT	76.2	76.2	82.5	81.1	80.3	79.7
w/ SGoTRE	79.4	79.1	87.3	86.8	89.0	88.7
Qwen2VL-7B						
w/ zero-shot	63.8	63.2	64.1	63.0	55.5	22.9
w/ SFT	78.6	78.1	85.9	83.2	85.9	86.3
w/ SGoTRE	**84.5**	**84.3**	**89.0**	**88.7**	**91.3**	**91.3**

Table 4: Evaluation the effectiveness of our SGoTRE in harmful meme detection. We report the zero-shot and SFT performance of the baseline model and compare it with SGoT-R1

Our SGoT-R1 trains on a computing platform equipped with an NVIDIA A100 GPU (80GB), an AMD EPYC 9754 CPU with 128 cores, 256GB of RAM, and 1TB of disk storage. The training runs for 20 epochs with a batch size of 4. We set the learning rate to 5e-5, the LoRA rank to 256, alpha to 512, and dropout to 0.05. We use the AdamW optimizer (Loshchilov and Hutter 2017).

Comparison with state-of-the-art methods

To evaluate the effectiveness of our method, we conducted comprehensive comparative experiments against three categories of state-of-the-art baseline models across three benchmark datasets: (1) Unimodal harmful meme detection models: As shown in the first group of Table 2, these methods perform detection based solely on either image or text modality. The compared models include: Text BERT (Devlin et al. 2019) and Image-Region (He et al. 2016). (2) Multimodal harmful meme detection models: The second group in Table 2 presents models that fuse textual and visual information, demonstrating significant performance improvements over unimodal approaches. The compared models include: MMBT-Region (Kiela et al. 2019), MOMENTA (Pramanick et al. 2021b), MaskPrompt (Cao et al. 2023b) and HateSieve (Su et al. 2025). (3) Knowledge-enhanced harmful meme detection models: The third group in Table 2 includes methods that incorporate external knowledge through pre-trained vision-language models (VLMs) or multimodal large language models (MLLMs). The compared models include: Pro-Cap (Cao et al. 2023a), MR.HARM (Lin et al. 2023), ExplainHM (Lin et al. 2024) and IntMeme (Hee and Lee 2025).

Experimental results demonstrate that our proposed SGoT-R1 model achieves optimal performance across all benchmark datasets. Specifically, SGoT-R1 attains accuracy improvements of 8.9%, 2.0%, and 0.6% over the best baseline models on the FHM, Harm-C, and Harm-P datasets, respectively. Notably, the magnitude of performance improvement exhibits a positive correlation with dataset complexity and scale, with the most significant enhancement observed on the FHM dataset, which features the broadest thematic coverage, largest scale, and most complex implicit hate mechanisms. This performance gain validates the advantages of SGoT-R1—effectively capturing collective cognitive patterns in meme understanding by integrating both diversity and consensus from user opinions.

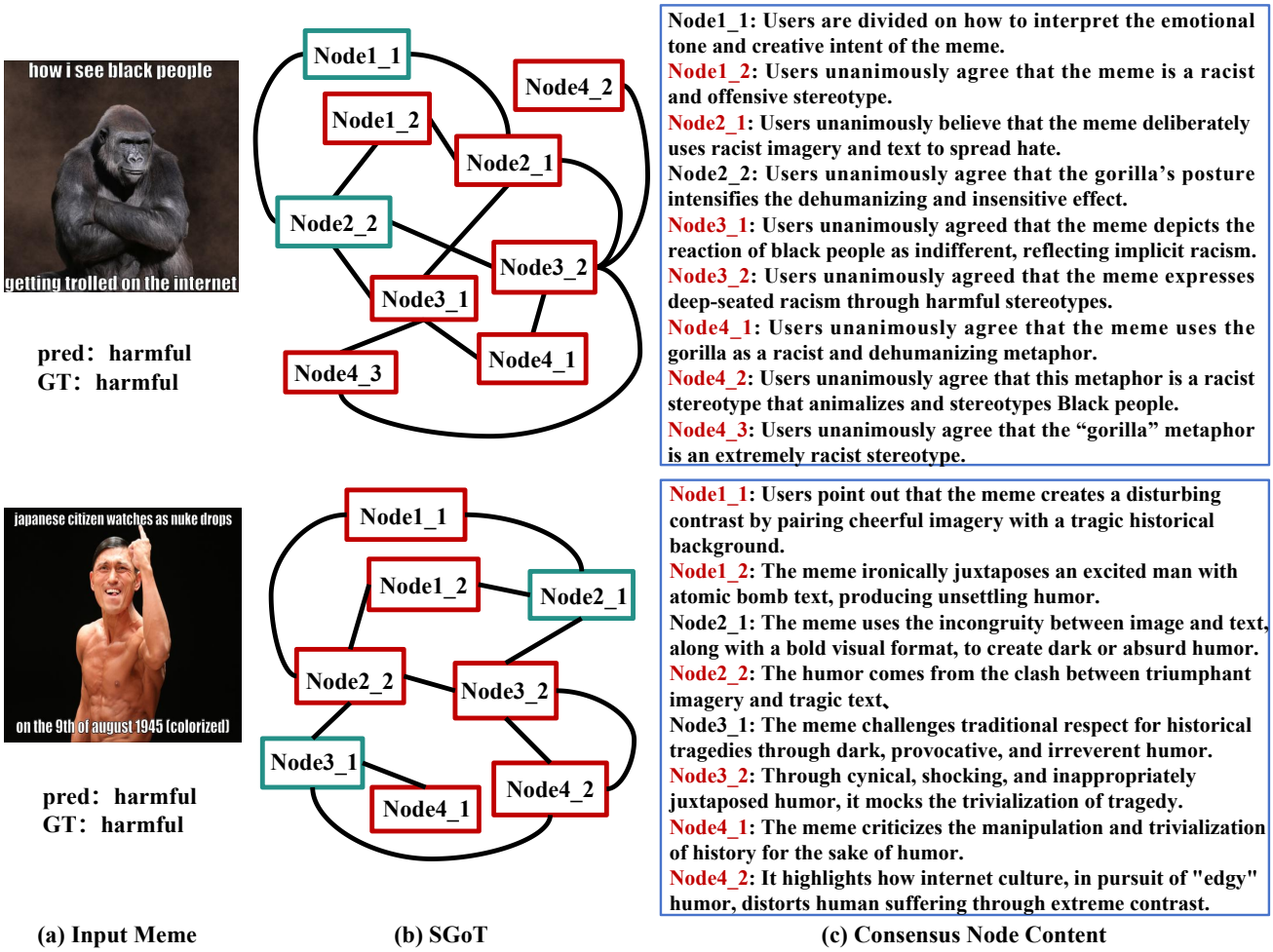

Figure 3: Examples of correctly predicted harmful memes and their corresponding SGoT are provided below. To clearly present core information from the verbose original content of each consensus node, we employ LLM to extract key viewpoints. In the diagram, consensus nodes aligned with harmful prediction results are highlighted in red

Ablation Study

Zero-Shot To evaluate whether SGoT improves harmful meme detection through structured collective cognition, we conducted zero-shot inference experiments. In these experiments, SGoT is used as a supplementary input to enhance baseline model performance, and results are compared with current mainstream baselines. Two representative multimodal large language models are selected as baselines: (1) GPT-4o (Hurst et al. 2024), a state-of-the-art multimodal model; and (2) Gemini 1.5 Flash, which balances performance and cost. As shown in Table 3, integrating SGoT leads to significant performance gains for both GPT-4o and Gemini 1.5 Flash, confirming the effectiveness of SGoT. Additionally, compared to directly generating SGoT with the baseline model, SGoT-R1 greatly reduces inference time.

Supervised Fine-Tuning To further validate the effectiveness of the SGoTRE method for harmful meme detection, we conducted supervised fine-tuning experiments. The SGoT serves as an intermediate reasoning step between the input and output, and we fine-tuned the multimodal models accordingly. The results are compared with both zero-shot and supervised fine-tuning baselines, using the 2B and 7B versions of Qwen2-VL (Wang et al. 2024) as baselines. As shown in Table 4, experimental results show that models using the SGoT approach achieve the best performance across all metrics, fully demonstrating the feasibility and effectiveness of distilling collective cognition reasoning into a unified framework.

Qualitative Analysis

As shown in Figure 3, we select two typical cases of harmful memes for qualitative analysis: one is a meme that mocks Black people using the image of a chimpanzee, and the other offends victims of the Japanese nuclear bombings by distorting human suffering. The experiment displays the SGoT generated by our SGoT-R1 and the consensus node content of SGoT, clearly presenting the reasoning process. The

results show that, in both cases, the majority of consensus nodes point to the presence of harmful metaphors in the memes, with only a few nodes (such as Node1_1 and Node2_2 in the first case) expressing ambiguous or non-harmful views. These results validate the effectiveness of the proposed model in predicting harmful memes and also demonstrate that, from the opinion of social consensus, it is possible to more clearly grasp the general public's judgment tendencies toward such harmful content.

Conclusion

In this paper, we propose a Social Graph of Thought Reasoning Enhancement (SGoTRE) framework to address the limitations of single-opinion reasoning in harmful meme detection. Our SGoTRE enhances multimodal large language models' metaphorical reasoning capabilities by introducing the Social Graph of Thought (SGoT). We first utilize multi-agent systems to simulate diverse user thinking processes and generate varied chains of thought, then construct SGoT to integrate these chains and model group consensus and viewpoint diversity, and finally distill the collective reasoning logic into a single model SGoT-R1 through a novel training paradigm. Extensive results demonstrate the superiority of our approach that achieves a new state-of-the-art on mainstream datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62425307, U21B2024, 62202327) and the Tianjin Municipal Natural Science Foundation of China (24JCQNJC01410).

References

Berthet, V. 2022. The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology*, 12: 802439.

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

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2023b. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2023c. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Dasgupta, S.; and Long, P. M. 2005. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4): 555–569.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.

Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; and Li, Y. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1): 1–24.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hee, M. S.; and Lee, R. K.-W. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 774–785.

Hee, M. S.; Lee, R. K.-W.; and Chong, W.-H. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, 3651–3655.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

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

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

Lin, H.; Luo, Z.; Ma, J.; and Chen, L. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mou, X.; Wei, Z.; and Huang, X. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*.

Muennighoff, N. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2022. DISARM: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738*.

Shifman, L. 2013. *Memes in digital culture*. MIT press.

Su, X.; Li, Y.; Inkpen, D.; and Japkowicz, N. 2025. A Context-Aware Contrastive Learning Framework for Hateful Meme Detection and Segmentation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 5201–5215.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yang, C.; Zhu, F.; Liu, G.; Han, J.; and Hu, S. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4505–4514.