

MemeLogos: Knowledge-Driven Retrieval-Augmented Reasoning for Harmful Meme Detection

Hexiang Gu, Saihui Hou, *Member, IEEE*, Changlan Liu, Min Ren, Huijia Wu, Hongxun Gu, and Zhaofeng He, *Member, IEEE*

Abstract—With the rapid proliferation of social media, harmful memes have become a growing threat to online ecosystems. Their harmfulness is often implicitly conveyed through complex interactions among visual, textual, and contextual cues, making accurate detection challenging. Recent studies have introduced Retrieval-Augmented Generation (RAG) and external knowledge into harmful meme detection. However, existing methods mainly retrieve similar instances or auxiliary information, limiting their ability to capture reusable harmfulness knowledge and underlying reasoning mechanisms, and thus constraining both detection performance and interpretability. To address these challenges, we propose MemeLogos, a knowledge-driven retrieval-augmented reasoning framework for harmful meme detection. MemeLogos abstracts recurring harmfulness patterns and reasoning structures from diverse harmful cases into reusable knowledge units and organizes them into a dedicated star-shaped knowledge graph. By combining structured knowledge retrieval with an explicit multi-stage reasoning process inspired by cognitive mechanisms, MemeLogos leverages retrieved knowledge to guide harmfulness reasoning, improving both detection reliability and interpretability. Extensive experiments show that MemeLogos consistently outperforms existing harmful meme detection methods while generating more coherent and interpretable reasoning, demonstrating the effectiveness of structured harmfulness knowledge for multimodal content understanding.

Index Terms—Harmful meme detection, Knowledge graph, RAG

Warning: This paper contains potentially disturbing meme examples involving harmful content, presented solely for research purposes.

I. INTRODUCTION

UNLIKE unimodal harmful content, memes often convey implicit harmful meanings through subtle interactions between visual and textual cues, such as vulgarity, sarcasm, and violence [1]. For example, juxtaposing a victim’s photo with historical tragedy symbols lacks surface-level offensive words, yet infuses malicious historical irony. Since harmful semantics often depend on contextual knowledge, cross-modal compositionality and implicit logical relations, black-box multimodal alignment alone is insufficient for reliable

judgement [2]. Therefore, effective harmful meme detection requires not only cross-modal semantic understanding, but also the ability to organize contextual evidence into reusable harmfulness knowledge and reason over their structured semantic relations [3].

Harmful meme detection relies heavily on the mining of contextual knowledge and harmful judgement logic [4], [5]. Although current solutions have adopted retrieval augmentation, explanation and reasoning-driven detection, they still suffer from two core limitations: (1) **Insufficiency of performance**. Existing methods rely on similar case retrieval and explanatory insights from datasets, with knowledge organized at the instance or free-form text level, which rarely distill recurring underlying harmful patterns, making it hard to model stable logical relations across diverse meme scenarios, thereby restricting further performance improvement of the model. (2) **Limited quality of interpretability**. Existing explanation-oriented methods often justify predictions with retrieved cases or free-form rationales. However, such evidence is mostly instance-level and rarely reveals the essential harmfulness rule triggered by the meme, limiting a transparent reasoning path from multimodal cues to verdicts and weakening the traceability and auditability of interpretation.

Driven by the above observations, we propose MemeLogos, a reasoning framework for harmful meme detection, incorporating structured harmfulness knowledge with Retrieval-Augmented Generation (RAG) reasoning paradigm to establish interpretable and effective harmful judgement logics. We distill rule units from MemeMind [6], a large-scale harmful meme dataset with detailed Chain-of-Thought (CoT) annotations, to derive structured harmfulness knowledge, which is further organized into a harmfulness knowledge graph with a star-shaped topology. This graph-organized harmfulness knowledge provides two key advantages: (1) **Comprehensive and Reusable Harmfulness Rules**. The extracted rule units abstract recurring harmfulness patterns and underlying semantic mechanisms from diverse multimodal memes, capturing generalizable harmfulness knowledge beyond instance-specific visual-textual expressions. By preserving the essential semantic associations and judgment criteria shared across different scenarios, these rules reduce reliance on individual instances and support more consistent and reusable harmfulness reasoning. (2) **Fine-grained Logical Association**. By explicitly organizing the semantic attributes, supporting evidence, and verdicts of harmfulness rules, the knowledge graph establishes

Hexiang Gu, Min Ren, Huijia Wu, and Zhaofeng He are with Beijing University of Posts and Telecommunications, Beijing 100876, China (E-mail: guhexiang@bupt.edu.cn; minren@bupt.edu.cn; huijiawu@bupt.edu.cn; zhaofenghe@bupt.edu.cn).

Saihui Hou is with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. (Email: housaihui@bnu.edu.cn)

Changlan Liu is with the College of Electrical Engineering, Zhejiang University, Zhejiang 310027, China, and also with Zhongguancun Academy, Beijing, China. E-mail: liuchanglan@zju.edu.cn

Hongxun Gu is with China Telecommunications Corporation, Beijing 100033, China. E-mail: guhongxun@163.com

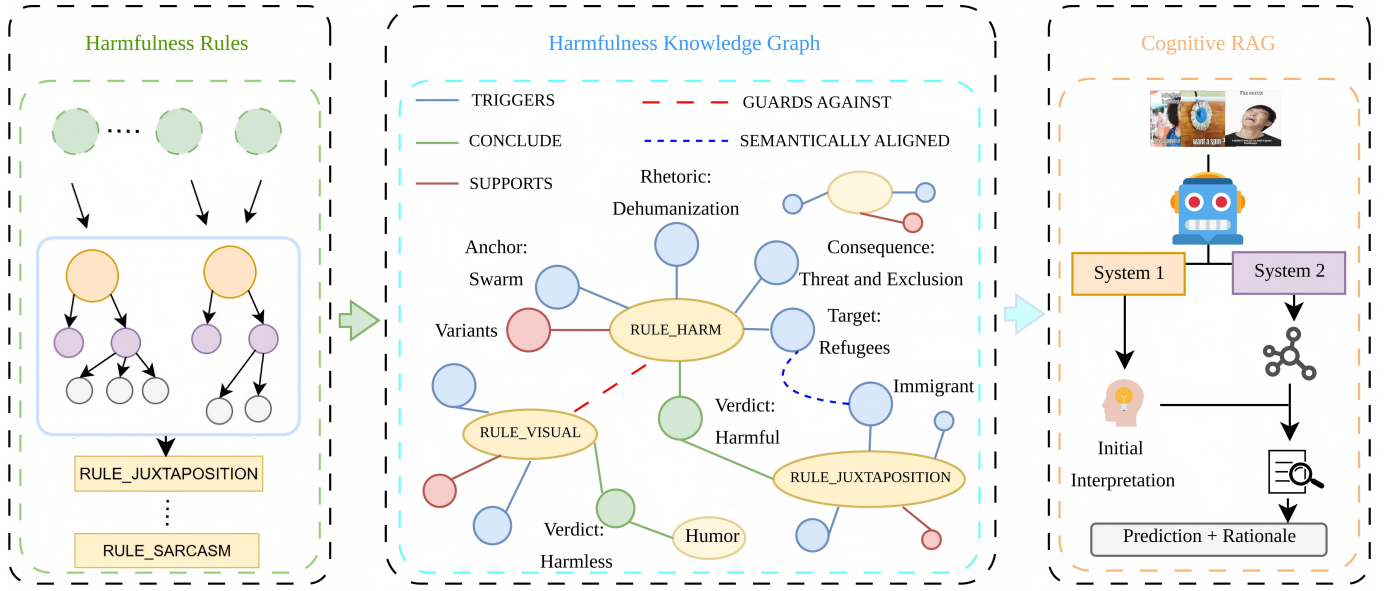


Fig. 1. Overview of MemeLogos. The framework abstracts reusable harmfulness rules from multimodal meme cases, organizes them into a structured harmfulness knowledge graph, and performs graph-grounded retrieval and cognitive reasoning for accurate and interpretable harmful meme detection.

traceable associations between harmfulness patterns and their judgments. Furthermore, cross-rule semantic relations facilitate knowledge expansion and contrastive retrieval between related rules with opposing verdicts, providing structured support for distinguishing subtle harmfulness boundaries and enabling more fine-grained harmfulness reasoning.

To fully leverage graph-organized harmfulness knowledge, inspired by Dual-System Theory of cognitive science [7], which posits that human cognition is driven by a fast, intuitive system and a slow, deliberative system, MemeLogos proposes a new RAG framework that can dynamically retrieve relevant rules from the knowledge graph and conduct contrastive reasoning, mitigating the risk of conventional RAG being misled by noisy or only superficially relevant retrieved evidence, which exhibits two distinct core advantages: (1) **Superior performance.** Such logical modeling significantly enhances generalizability across diverse domains and improves the overall performance of harmful meme detection. (2) **Deep interpretability.** The reasoning chain generated by MemeLogos transforms complex metaphor recognition into a traceable and structured reasoning process, thereby yielding a substantial improvement in interpretability.

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

- We construct structured harmfulness knowledge and organize it as a knowledge graph, which converts CoT annotations into reusable rule-level knowledge and provides a structured knowledge basis to deconstruct the underlying composition of harmful meaning.
- We propose MemeLogos, a RAG framework that formalizes harmful meme detection through knowledge reasoning. By bridging structured reasoning with RAG, MemeLogos significantly enhances detection performance while providing traceable and interpretable reasoning.
- We conduct extensive experiments on harmful meme

datasets, demonstrating that MemeLogos significantly outperforms existing approaches in performance.

II. RELATED WORK

A. Harmful Meme Detection

Harmful meme detection has become an important research topic for online content safety. Benchmark datasets such as Hateful Memes [1], HarMeme [8], MAMI [9], and ToxiCN_MM [10] have established diverse evaluation settings covering different harmfulness categories, languages, and social contexts. Early approaches primarily focused on multimodal feature extraction and fusion, while subsequent dual-stream architectures and vision-language pre-training models [11], [12] enabled more effective cross-modal representation learning. For instance, Hate-CLIPper [13] explicitly models interactions between CLIP-based visual and textual features to capture their complementary information for harmfulness classification.

Nevertheless, harmful meanings are frequently conveyed through implicit cross-modal interactions involving irony, metaphor, cultural references, and contextual incongruity, making them difficult to recognize through surface-level correlations alone. To address these challenges, recent studies have incorporated auxiliary semantic understanding and reasoning into harmful meme detection. CapAlign [14] employs question-driven informative caption generation to enrich visual semantics and improve cross-modal alignment. Other approaches leverage contextualized captioning and chain-of-thought reasoning to uncover implicit meanings and support interpretable predictions [5], [15]. Furthermore, Lin et al. [16] introduce multimodal reasoning distillation to transfer implicit semantic reasoning from large language models into harmful meme detectors. More recently, ExplainHM++ [17] integrates similar meme retrieval with multimodal debate to enhance

TABLE I
STATISTICS OF THE HARMFULNESS KNOWLEDGE GRAPH.

Nodes									Edges						
Type	Case	Rule	Anch.	Rhet.	Targ.	Cons.	Verd.	Total	Relation	Align.	Trig.	Guard	Supp.	Concl.	Total
Count	4477	2541	2195	2056	2002	1829	2541	17641	Count	15706	10164	8084	4477	2541	40972

harmfulness detection and interpretability through competing harmful and harmless interpretations.

Despite these advances, much of the existing reasoning and retrieval process remains centered on individual memes, free-form explanations, or retrieved examples. Although such information can enrich contextual understanding, recurring harmfulness patterns and their relationships across different scenarios are not always explicitly abstracted into reusable rule-level knowledge. Consequently, connecting multimodal semantic cues with generalizable harmfulness criteria and supporting evidence remains challenging, motivating more structured knowledge representation and reasoning for harmful meme detection.

B. Graph-Based Retrieval-Augmented Generation

Recent studies extend conventional Retrieval-Augmented Generation (RAG) by explicitly modeling structured relations among external knowledge units. GraphRAG [18] constructs entity-centric graph indexes and community summaries to support global reasoning over large document collections, while G-Retriever [19] formulates retrieval over textual graphs as a subgraph selection problem. HippoRAG [20] combines knowledge graphs with graph-based propagation to support multi-hop knowledge integration, and KG²RAG [21] exploits knowledge-graph relations to expand and organize semantically retrieved evidence. Beyond retrieval, Think-on-Graph [22] and Reasoning on Graphs [23] explore explicit graph traversal and relation-path reasoning over knowledge graphs, while Think-on-Graph 2.0 [24] further integrates structured graph retrieval with unstructured contextual evidence. StructRAG [25] additionally demonstrates the benefit of dynamically organizing retrieved knowledge into task-appropriate structured representations. These studies highlight the importance of relational structure for knowledge retrieval and reasoning. Different from general Graph-RAG methods that mainly operate over entity-, document-, or fact-level structures, MemeLogos abstracts multimodal harmful cases into reusable rule-level knowledge and explicitly models semantic and polarity-aware relations among harmfulness rules, enabling task-specific retrieval and contrastive reasoning for harmful meme detection.

III. METHODOLOGY

To boost the performance and interpretability of harmful meme detection, we propose MemeLogos, a reasoning-oriented detection framework, which integrates graph-based structured harmfulness knowledge and retrieval-augmented reasoning technology. The overall architecture is illustrated in Figure 1. The framework consists of two core stages: (1)

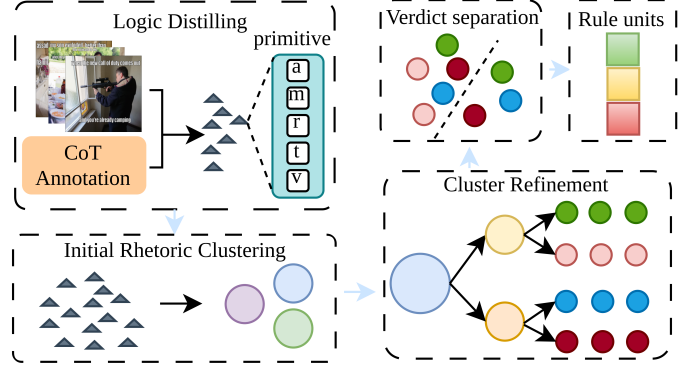


Fig. 2. Pipeline of Formalized Rule Extraction. MemeLogos first distills each meme into structured harmfulness primitives, then performs rhetoric-based clustering, verdict-aware separation, and semantic cluster refinement to aggregate recurring patterns, which are finally generalized into reusable harmfulness rule units.

Harmfulness Knowledge construction stage, we extract inherent harmfulness logic rules and build the knowledge graph for harmfulness detection; (2) **Graph-Grounded Cognitive RAG** inference stage, we combine adversarial graph retrieval and dual-system audit reasoning mechanism to achieve superior performance and auditable reasoning.

A. Harmfulness Knowledge

Existing methods often represent knowledge at the level of instances or unstructured textual explanations, making it difficult to capture inherent harmfulness patterns. Moreover, directly relying on models to generate knowledge from memes may introduce inherent semantic biases. To mitigate these issues, we ground harmfulness knowledge construction in meme instances and their detailed Chain-of-Thought (CoT) annotations. Specifically, we construct a harmfulness knowledge graph in two steps: 1) Formalized Rule Extraction, abstracting essential harmfulness patterns and rules from the memes with CoT annotations, and 2) Topological Graph Modeling, explicitly modeling harmfulness rules via star-shaped topology.

1) *Formalized Rule Extraction*: In this section, we leverage instance-level logic distillation and corpus-level rule generalization to map unstructured multimodal memes to formalized harmfulness rule units, as illustrated in Figure 2, so as to accurately extract and generalize the inherent harmful logic hidden within heterogeneous memes.

Instance-Level logic distillation. Given a sample, we decompose the instance into six logical primitives:

$$z_i = (a_i, m_i, r_i, t_i, c_i, v_i), \quad (1)$$

where a_i , m_i , r_i , t_i , c_i and v_i respectively denote the anchor, capturing the observable visual or textual carrier of the meme;

semantic mapping, represents the latent semantic interpretation assigned to the carrier; rhetorical mechanism, describing the compositional mechanism that connects the observable cue to the latent meaning; target, identifying the individual, group, institution, or social norm affected by the meme; consequence, recording the semantic or social effect and verdict, denoting the final harmfulness label. This logical decomposition reduces instance-specific multimodal noise while retaining the reasoning structure required for harmfulness assessment.

Corpus-level rule generalization. Based on the Instance-Level Logic Distillation, we progressively cluster and canonicalize the extracted primitives to obtain generalizable rule units. As rhetorical mechanisms provide an important semantic basis for harmfulness patterns, we first use it to construct coarse semantic communities. For each instance, the rhetorical primitive r_i is encoded by MPNet [26] as

$$\mathbf{h}_i^r = f_{\text{MPNet}}(r_i), \quad (2)$$

where f_{MPNet} denotes the MPNet semantic encoder. Then we apply initial clustering on $\mathbf{h}_i^r$ based on cosine similarity with a threshold of τ_r . This produces coarse clusters whose members share similar rhetorical mechanisms and establishes a foundation for fine-grained rule induction.

However, clusters formed solely according to rhetorical mechanisms may remain semantically heterogeneous, since the same rhetorical strategy can be instantiated with different visual anchors, semantic mappings, targets, and consequences. We therefore further refine each coarse cluster by incorporating these complementary primitives. For each instance, the anchor, mapping, target, and consequence texts are concatenated and encoded into a unified semantic representation $\mathbf{h}_i$:

$$\mathbf{h}_i = f_{\text{MPNet}}(a_i \oplus m_i \oplus t_i \oplus c_i), \quad (3)$$

where $\oplus$ denotes text-level concatenation. Because coarse clusters vary substantially in size, the refinement granularity is determined independently for each cluster. Let $\mathcal{G}_j$ denote the j -th coarse cluster and $n_j = |\mathcal{G}_j|$ its size. For each coarse cluster for refinement, we apply K-means to the unified representations and set the number of subclusters as

$$K_j = \max\left(2, \left\lfloor \frac{n_j}{100} \right\rfloor\right), \quad (4)$$

thereby targeting approximately 100 instances per refined cluster. Specifically, the refined subclusters are obtained by minimizing the objective

$$\min_{\{\mathcal{C}_{jk}, \boldsymbol{\mu}_{jk}\}_{k=1}^{K_j}} \sum_{k=1}^{K_j} \sum_{z_i \in \mathcal{C}_{jk}} \|\mathbf{h}_i - \boldsymbol{\mu}_{jk}\|_2^2, \quad (5)$$

where $\mathcal{C}_{jk}$ is the k -th refined subcluster of $\mathcal{G}_j$ and $\boldsymbol{\mu}_{jk}$ is its centroid. This adaptive refinement produces semantically more coherent candidate groups while avoiding a single global clustering granularity for coarse clusters of different sizes.

Next, although memes may exhibit similar surface features or rhetorical patterns, their harmfulness can differ depending on the underlying semantic context. To prevent such samples with conflicting verdicts from being generalized into the same rule, we further partition each cluster according to the sample

verdict v_i , yielding separate harmful and harmless candidate rules. Finally, to balance rule generalization and long-tail risk coverage, we introduce an asymmetric support-based filtering strategy with $\tau_s = 3$, where s_j denotes the number of supporting instances for candidate rule:

$$\text{Keep}(\mathcal{R}_j) = \begin{cases} 1, & s_j \geq \tau_s, \\ 1, & s_j < \tau_s \wedge v_j = \text{Harmful}, \\ 0, & s_j < \tau_s \wedge v_j = \text{Harmless}. \end{cases} \quad (6)$$

Specifically, harmless rules are required to exhibit sufficient cross-instance recurrence to reduce overly instance-specific abstractions, whereas low-support harmful rules are retained to preserve potentially rare but important harmfulness patterns. This asymmetric design limits redundant rule proliferation while maintaining coverage of long-tail harmful scenarios.

Harmfulness rule units. Although multimodal memes vary considerably in content and presentation, recurring harmfulness patterns often share a stable logical structure. We therefore canonicalize each retained cluster into a formalized rule unit:

$$\mathcal{R}_j = \langle \mathcal{M}_j, \mathcal{P}_j, \mathcal{S}_j \rangle, \quad \mathcal{P}_j = (A_j, Rh_j, T_j, C_j, V_j), \quad (7)$$

where $\mathcal{M}_j$ is a high-level semantic description of the harmfulness pattern, and $\mathcal{P}_j$ represents the stable attributes, including Anchor, Rhetoric, Target, Consequence and Verdict, which are obtained by abstracting logical primitives from the Corpus-Level Rule. $\mathcal{S}_j$ contains representative cases. Rather than relying on unconstrained knowledge generation, we employ a multimodal large language model to synthesize $\mathcal{M}_j$ and $\mathcal{P}_j$ based on the extracted logical primitives, supporting instances and harmfulness label within each cluster. This evidence-grounded generalization aims to capture recurring semantic patterns while reducing unsupported interpretations and preserving consistency with the underlying instances. Furthermore, each induced rule retains explicit associations with its supporting cases, facilitating knowledge traceability and subsequent retrieval-augmented reasoning.

2) Topological Graph Modeling: Harmfulness rule units extracted in the Formalized Rule Extraction stage are inherently high-order relational objects composed of multiple logical primitives. Conventional entity-relation graphs only encode binary facts as independent triples, which cannot natively match the structure of rule units. Such representation fragments the internal semantics of rules. Therefore, to fully retain the semantic integrity of each rule, realize explicit structured modeling, and facilitate retrieval and evidence tracing, we reconstruct plain-text harmfulness rule units into a structured harmfulness knowledge graph embedded with star-shaped topology, as shown in Harmfulness Knowledge Graph of Figure 1, which contains star-shaped rule subgraphs and cross-rule relations.

Star-shaped rule subgraphs. For each harmfulness rule unit, we construct a rule-centered local star subgraph to explicitly organize its semantic attributes, supporting evidence, and associated verdict. The central Rule node represents a generalized harmfulness pattern and is connected to six types of peripheral nodes: Anchor (Anch.), Rhetoric (Rhet.),

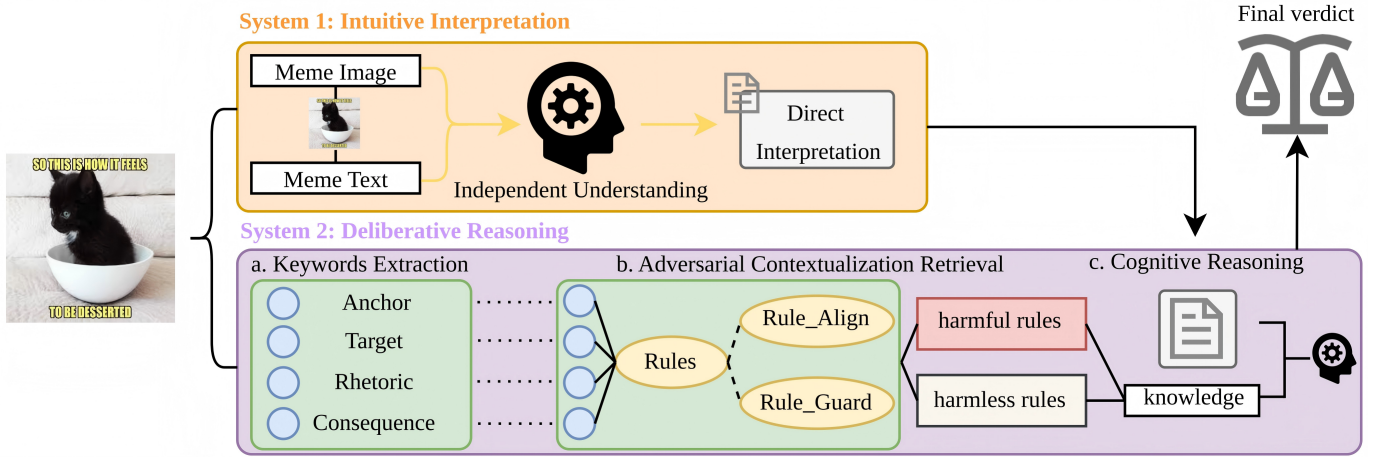


Fig. 3. Overview of Cognitive RAG. System 1 independently interprets the input meme without retrieved knowledge, while System 2 extracts preliminary semantic cues, retrieves adversarial harmful and harmless rules from the knowledge graph, and integrates the initial interpretation with retrieved evidence for knowledge-grounded reasoning and final harmfulness judgment.

Target (Targ.), Consequence (Cons.), Support_cases (Case), and Verdict (Verd.). The first four capture the key semantic characteristics underlying the rule, while Case nodes preserve supporting instances as empirical evidence and the Verd. node specifies the corresponding harmfulness judgment. The Mapping primitive is not explicitly included in the graph, as it mainly captures instance-specific cross-modal correspondences rather than reusable rule-level attributes. This rule-centered organization separates generalized harmfulness knowledge from its supporting instances while preserving their explicit associations.

We further define three types of structural relations to organize these dependencies. TRIGGERS edges (Trig.) connect Anch., Rhet., Targ., and Cons. nodes to the Rule node, representing semantic conditions relevant to rule applicability. SUPPORTS edges (Supp.) connect Case nodes to their corresponding Rule node, linking generalized knowledge to its supporting instances. CONCLUDE edges (Concl.) connect each Rule node to its Verd. node, associating the rule with its harmfulness judgment. Together, these relations establish explicit topological paths, providing a structured foundation for rule matching, evidence tracing, and knowledge retrieval.

Cross-rule relations. Although rule-centered star subgraphs explicitly organize the semantic attributes and supporting evidence of individual rules, they lack direct associations across different rule units, limiting the exploration of related knowledge beyond local structures. To address this limitation, we introduce two types of cross-rule semantic relations based on semantic similarity and verdict polarity: (1) SEMANTICALLY ALIGNED edges (Align.) connect attribute nodes of the same semantic type when their similarity exceeds a threshold τ_a . These relations capture semantic associations between attributes belonging to different rules, allowing retrieval to expand beyond directly matched attributes and identify relevant rules across different meme scenarios. This facilitates knowledge sharing and improves the accessibility of semantically related rule-level knowledge. (2) GUARDS AGAINST edges (Guard.) connect Rule nodes that exhibit

high semantic similarity above a threshold τ_g but have opposite verdicts. These relations capture potentially contrasting harmfulness interpretations under similar semantic conditions, highlighting that semantically related patterns may lead to different judgments depending on their context. By linking rules with opposing verdicts, these edges facilitate the retrieval of contrastive evidence for examining rule applicability and distinguishing harmful from harmless interpretations.

Together, these cross-rule relations complement the rule-centered structural edges by enabling semantic retrieval expansion and polarity-aware knowledge exploration. The resulting graph integrates local rule-level evidence with cross-rule semantic associations, providing a structured knowledge foundation for subsequent Cognitive RAG. Graph statistics are reported in Table I.

B. Graph-Grounded Cognitive RAG

The harmfulness knowledge graph provides structured support for harmful meme detection. Building on this graph, we propose Cognitive RAG, as shown in Figure 3, which supports Training-free inference to produce harmfulness analysis and enables Training-Alignment enhancement for meme feature extraction and reasoning.

1) *Training-free Cognitive RAG:* Inspired by Dual-System Theory, where a fast system relies on instinct and experience while a slow system drives conscious reasoning, we design Cognitive RAG with two complementary systems: intuitive interpretation (System 1) and deliberative reasoning (System 2). System 1 independently interprets the input meme without relying on retrieved knowledge, mitigating potential retrieval-induced bias. System 2 then performs deliberative reasoning through three stages: (1) core attribute extraction, which identifies preliminary semantic cues as retrieval hypotheses. (2) adversarial contextualization, which retrieves relevant rules to construct a polarity-balanced context containing competing interpretations. (3) cognitive reasoning, which integrates the initial interpretation with retrieved context to perform reason-

TABLE II

PERFORMANCE COMPARISON ON THE MEMEMIND DATASET. WE COMPARE MEMELOGOS WITH REPRESENTATIVE BASELINES. MEMELOGOS DENOTES THE TRAINING-FREE FRAMEWORK, WHILE MEMELOGOS_ALIGN FURTHER INCORPORATES KEYWORD AND REASONING ALIGNMENT. $\diamond$ INDICATES DIRECT EVALUATION OF VISION-LANGUAGE MODELS WITHOUT TASK-SPECIFIC FINE-TUNING. **BOLD** INDICATES THE BEST RESULT, WHILE UNDERLINING INDICATES THE SECOND-BEST RESULT.

	Method	Accuracy	Precision	Recall	F1-score
Single-Modality	CLIP Text-Only [12]	74.14	75.54	56.92	64.92
	CLIP Image-Only [12]	80.25	79.47	71.74	75.41
Multi-Modality	CLIP Image+Text [12]	80.81	80.81	71.06	75.62
	VisualBERT_COYO [27]	63.21	63.95	63.48	63.72
	Pro-Cap [28]	75.58	78.09	78.76	78.44
	ISSUES [29]	82.54	83.09	74.80	78.72
	MemeCLIP [30]	79.18	80.48	78.98	79.72
	RGCL [31]	81.65	78.38	79.21	78.79
	RA-HMD [32]	81.23	80.41	74.53	77.36
	Qwen2.5-VL-32B $\diamond$ [33]	78.44	81.22	63.25	71.11
	Qwen3.6-plus $\diamond$ [34]	77.35	92.26	50.98	65.67
	GPT-5.5 $\diamond$ [35]	77.91	86.97	57.37	69.14
Our Method	MemeLogos	<u>86.96</u>	83.45	<u>87.07</u>	<u>85.22</u>
	MemeLogos_Align	89.50	<u>87.01</u>	89.08	88.03

ing and hypothesis falsification, deriving the reasoning process and final verdict.

System 1: Intuitive Interpretation. Previous studies have shown that irrelevant or conflicting retrieved knowledge can mislead model predictions and interfere with the utilization of intrinsic knowledge [36], [37]. To mitigate this issue, System 1 independently interprets the input meme without accessing external knowledge, forming an initial rationale $h_i^{(1)}$ based solely on its multimodal content:

$$h_i^{(1)} \sim p_\theta(\cdot | x_i), \quad (8)$$

where x_i denotes the input meme image and text. This independent interpretation provides a complementary perspective for subsequent cognitive reasoning, helping the model integrate its intrinsic understanding of the meme with retrieved knowledge.

System 2: Deliberative Reasoning. Building upon the System 1, System 2 incorporates retrieval knowledge to support harmfulness reasoning and verification. The deliberative process is organized into three stages: keyword extraction, adversarial contextualization retrieval, and cognitive reasoning.

a. Keywords Extraction. To reduce retrieval interference from redundant meme details, we extract a compact set of logical probes $\mathcal{K}_i = \{\hat{a}_i, \hat{r}_i, \hat{t}_i, \hat{c}_i\}$, corresponding to the anchor, rhetorical mechanism, target, and consequence of the input meme. These probes represent preliminary semantic hypotheses, providing structured cues for identifying potentially relevant rules. We then encode them into a retrieval query set $\mathbf{Q}$ for schema-aligned matching with the corresponding attribute nodes in the harmfulness knowledge graph.

b. Adversarial Contextualization Retrieval. Since semantically similar patterns may correspond to different harmfulness verdicts, relying solely on the most relevant rules may overlook competing interpretations. We therefore construct a polarity-balanced context through graph-grounded retrieval. Specifically, we first compute the semantic relevance between $\mathbf{Q}$ and rule attribute nodes. We obtain directly matched rules $\mathcal{R}_{\text{dir}}$ through TRIGGERS edges, expand semantically related rules $\mathcal{R}_{\text{rel}}$ through ALIGNED edges, and identify contrastive rules $\mathcal{R}_{\text{risk}}$ through GUARDS edges. Candidate rules are then lexicographically reranked according to the priority $\mathcal{R}_{\text{dir}} > \mathcal{R}_{\text{rel}} > \mathcal{R}_{\text{risk}}$. Finally, we retrieve K harmful and K harmless rules to construct the adversarial context Σ , providing competing rule-level knowledge for subsequent reasoning.

c. Cognitive Reasoning. Given the initial interpretation $h_i^{(1)}$ and the retrieved adversarial context Σ , System 2 performs knowledge-grounded reasoning to critically examine against harmful and harmless rules. By jointly considering the initial interpretation and retrieved knowledge, the model conduct comprehensive reasoning to produce a refined rationale $h_i^{(2)}$ and the final verdict $\hat{y}_i$:

$$(\hat{y}_i, h_i^{(2)}) \sim p_\theta(\cdot | x_i, h_i^{(1)}, \Sigma), \quad (9)$$

This dual-system design combines independent multimodal interpretation with contrastive knowledge-grounded reasoning, mitigating potential retrieval-induced bias while supporting more reliable and traceable harmfulness judgments.

2) Training-Alignment Cognitive RAG: To further improve the quality of knowledge retrieval and reasoning, we develop MemeLogos_Align, a training-aligned variant of Cognitive

RAG through two optimization strategies: keywords alignment and reasoning alignment.

Keywords Alignment. To enhance the consistency of core semantic attribute extraction, we perform supervised fine-tuning using logical primitives derived from annotated rules as supervision targets. This process guides the model to map multimodal meme inputs into structured semantic representations aligned with the attribute schema of the harmfulness knowledge graph, rather than unconstrained textual descriptions. By reducing the semantic mismatch between extracted queries and graph attributes, keyword alignment aims to improve retrieval precision and the relevance of retrieved rules.

Reasoning Alignment. To further improve the utilization of retrieved knowledge during cognitive reasoning, we employ Group Relative Policy Optimization (GRPO) [38] to optimize the model through outcome-based feedback. For each meme, the model samples a group of candidate reasoning trajectories, each producing a final harmfulness verdict. We assign an accuracy-based reward according to the correctness of the predicted verdict:

$$r_{i,g} = \mathbb{I}[\hat{y}_{i,g} = y_i], \quad (10)$$

where $\hat{y}_{i,g}$ denotes the verdict produced by the g -th trajectory and y_i is the ground-truth label. Through group-relative optimization, this reward encourages reasoning trajectories that lead to accurate harmfulness judgments, facilitating more effective integration of independent semantic interpretations and retrieved rule-level knowledge.

IV. EXPERIMENT

A. Experimental Setup

Baselines and Datasets. The experiments are divided into two groups. The single-modality group adopts the CLIP [12] to process text and image modalities independently, while the multimodality group includes representative multimodal methods, such as CLIP, Pro-Cap [28], ISSUES [29], MemeCLIP [30], VisualBERT-COCO [27], U-CoT+ [5], RGCL [31], RA-HMD [32], ExPO-HM [39], and vision-language models. Except for the vision-language models baselines, all methods in Table II are uniformly trained and evaluated on MemeMind. We also conduct cross domain experiments in PrideMM [30].

Implementation details. We adopt Qwen3-VL-8B [40] as the backbone for Cognitive RAG inference and Qwen3-VL-32B [40] for knowledge extraction and rule generalization. All-Mpnet-Base-V2 as the semantic encoder for knowledge construction and retrieval. For formalized rule extraction, the extracted logical primitives are encoded into 768-dimensional embeddings with a batch size of 128. We first perform coarse clustering based on rhetorical mechanisms using a cosine similarity threshold of $\tau_r = 0.7$. Within each coarse cluster, the Anchor, Mapping, Target, and Consequence primitives are concatenated and encoded for K-Means refinement, with the number of clusters adaptively determined to contain approximately 100 samples per refined cluster. Harmful and harmless samples are subsequently separated according to their verdicts to avoid merging contradictory patterns. For

refined clusters containing more than 10 samples, K-Means-based diversity sampling is employed to select representative supporting cases. For topological graph modeling, we set the similarity thresholds of ALIGNED and GUARDS edges to $\tau_a = 0.9$ and $\tau_g = 0.75$, respectively, and construct the resulting harmfulness knowledge graph using NetworkX. During Cognitive RAG inference, we retrieve the top-3 rules from each polarity for contrastive reasoning. To prevent data leakage, all rule extraction and knowledge graph construction are performed exclusively on the training split of MemeMind. We report Accuracy, Precision, Recall, and F1-score as the primary evaluation metrics.

B. Experimental Results

Table II reports the main results on MemeMind. Among single-modality baselines, the image-only CLIP model substantially outperforms its text-only counterpart, achieving an F1-score of 75.41 compared with 64.92, which indicates that visual information provides important cues for harmful meme detection. However, combining image and text with CLIP yields only a marginal improvement to 75.62 F1, suggesting that simple multimodal fusion is insufficient to capture the implicit semantic interactions underlying harmfulness. More advanced multimodal approaches, including Pro-Cap, ISSUES, MemeCLIP, RGCL, and RA-HMD, achieve stronger overall performance, confirming the effectiveness of cross-modal modeling and task-specific semantic enhancement. Nevertheless, their performance remains limited when harmfulness depends on implicit contextual associations and reasoning beyond surface-level visual-textual correspondence. In particular, the strongest existing baseline achieves 82.54% accuracy, leaving considerable room for further improvement.

In addition, stronger general vision-language models do not necessarily achieve superior performance on harmful meme detection under direct zero-shot evaluation. As shown in Table II, Qwen2.5-VL-32B, Qwen3.6-plus, and GPT-5.5 obtain F1-scores of 71.11, 65.67, and 69.14, respectively. In particular, Qwen3.6-plus and GPT-5.5 achieve relatively high precision but substantially lower recall, indicating a conservative prediction tendency that misses a considerable proportion of harmful memes. These results suggest that general multimodal capability alone is insufficient to reliably capture the implicit contextual associations and compositional harmfulness semantics required by this task. Task-specific harmfulness knowledge and structured reasoning therefore remain important for improving both recognition coverage and decision reliability.

In contrast, MemeLogos achieves 86.96% accuracy and 85.22% F1, outperforming the strongest existing baselines. These improvements demonstrate the effectiveness of jointly leveraging structured harmfulness knowledge and Cognitive RAG. The harmfulness knowledge graph provides reusable and organized task-specific knowledge, while Cognitive RAG enables the model to effectively incorporate such knowledge into the reasoning process, thereby improving its ability to capture implicit harmful semantics and make more reliable predictions. Furthermore, MemeLogos_Align improves the

TABLE III
COMPUTATIONAL EFFICIENCY OF DIFFERENT KNOWLEDGE REPRESENTATIONS. ACCESS DENOTES KNOWLEDGE RETRIEVAL, WHILE INFERENCE DENOTES DOWNSTREAM PREDICTION WITHOUT FULL REASONING.

Method	Access (s)	Inference (s)	Total (s)
Rule-RAG	0.076	0.092	0.168
Graph-RAG	0.170	0.107	0.277
Instance-RAG	0.139	0.278	0.417

TABLE IV
ABLATION ON GRAPH STRUCTURE.

Category	Model Variant	Accuracy	F1
Representation	Instance-RAG	82.96	80.44
	Rule-RAG	82.76	80.41
Variants	w/o ALIGNED edges	86.13	84.47
	w/o GUARDS edges	85.32	83.51
Our Method	MemeLogos	86.96	85.22

performance to 89.50% accuracy and 88.03% F1, yielding additional gains of 2.54 and 2.81 percentage points over the training-free MemeLogos, respectively. This further demonstrates that aligning both semantic keyword extraction and reasoning behavior with the structured harmfulness knowledge can enhance the utilization of retrieved evidence and improve the overall detection capability.

C. Ablation study

In this section, we evaluate the contribution of each component design in MemeLogos. For all ablation experiments, we maintain the same experimental configurations described in previous sections.

Ablation on knowledge graph. To independently analyze the effects of knowledge representation and graph structure, we conduct ablation experiments with all other experimental configurations fixed. We first compare the controlled computational overhead of different knowledge access strategies in Table III, while the actual inference overhead of the complete MemeLogos framework is detailed in Appendix ?? . Specifically, Instance-RAG directly retrieves similar meme instances, Rule-RAG retrieves induced rules based on plain-text semantic similarity, and Graph-RAG adopts the knowledge graph-based rule retrieval strategy used in MemeLogos. To isolate the computational cost of knowledge access, all variants use the same direct prediction setting rather than the full end-to-end MemeLogos pipeline. Under this controlled setting, the total latency is 0.417s, 0.168s, and 0.277s for Instance-RAG, Rule-RAG, and Graph-RAG, respectively.

Then, we further evaluate the variants and MemeLogos under the full end-to-end inference setting as shown in Table IV. The results show that retrieving induced rules in the plain-text space achieves performance comparable to direct retrieval of similar meme instances. This indicates that the induced rule-level knowledge can effectively preserve harmfulness-relevant information from meme instances. However, Rule-

TABLE V
ABLATION OF COMPONENTS IN COGNITIVE RAG.

Variant	Accuracy	F1
w/o Keywords Extraction	83.49	83.13
w/o Adversarial Context	85.11	82.64
w/o Cognitive Reasoning	64.55	39.47
MemeLogos	86.96	85.22

TABLE VI
PERFORMANCE OF MEMELOGOS ACROSS DIFFERENT MULTIMODAL BACKBONES.

Backbone	Variant	Accuracy	F1
Qwen3-VL-8B	<i>Zero-shot</i>	73.24	66.32
	<i>MemeLogos</i>	86.96	85.22
InternVL3.5-8B	<i>Zero-shot</i>	69.75	63.81
	<i>MemeLogos</i>	85.43	84.76
MiniCPM-V-4.5	<i>Zero-shot</i>	75.62	75.67
	<i>MemeLogos</i>	81.95	82.54

RAG still underperforms MemeLogos with the star-shaped graph structure, demonstrating the benefit of structured knowledge organization for retrieval and reasoning.

Moreover, removing either ALIGNED or GUARDS edges leads to a consistent yet relatively limited performance degradation. This indicates that the semantic cross-rule relations do provide useful signals for expanding relevant knowledge and characterizing semantic associations and decision boundaries among related rules. However, their contribution is smaller than that of the structural relations directly induced by the rule schema. This suggests that the primary benefit of the graph comes from its explicit rule-centered structural organization, while semantic relation edges mainly serve as complementary cues for relational expansion and contrastive retrieval.

Ablation on component in Cognitive RAG. To independently evaluate the contribution of core components in cognitive RAG, we conduct ablation experiments on MemeLogos with the training-free settings and other configurations consistent. As shown in Table V, we first replace the keywords extraction module with a free-form caption or query automatically generated by the model for retrieval. The results show that replacing the structured logical keywords with such free-form queries leads to a performance degradation. This suggests that explicitly aligning retrieval queries with the graph schema provides more discriminative and task-relevant cues than unconstrained textual descriptions, thereby improving rule retrieval.

We then replace adversarial contextualization with conventional Top- K rule retrieval, where rules are solely according to retrieval relevance. The resulting performance degradation suggests that polarity-aware contrastive context helps the model distinguish semantically similar but decisionally different cases, leading to more reliable harmfulness judgments.

In addition, with other settings fixed, we replace Cognitive Reasoning with a standard retrieval-augmented inference

TABLE VII
ABLATION OF TRAINING ALIGNMENT.

Strategy	Accuracy	F1
MemeLogos	86.96	85.22
Keywords Alignment	87.20	85.68
Reasoning Alignment	88.57	86.88
MemeLogos_Align	89.50	88.03

scheme. The results show that removing Cognitive Reasoning substantially degrades performance, showing that simply exposing the model to the conflict retrieved knowledge may introduce redundant or distracting information and consequently impair prediction. In contrast, the dual-system cognitive reasoning process explicitly verifies, contrasts, and refines the retrieved evidence before making the final decision, enabling more reliable knowledge utilization.

Ablation on Backbones. To evaluate the extensibility of MemeLogos, we conduct comprehensive experiments across multiple backbones, with other settings consistent. To avoid architectural overlap, we select MiniCPM-V-4.5 [41] and InternVL3.5-8B [42] as backbones and use the same zero-shot CoT prompt. As illustrated in Table VI, MemeLogos consistently improves the performance of other backbones, validating the robustness of MemeLogos as a model-agnostic framework for harmful meme reasoning and detection.

Ablation on Training Alignment. To isolate and evaluate the impact of Keyword Alignment and Reasoning Alignment, we conduct ablation experiments by keeping other components fixed. As shown in Table VII, MemeLogos denotes the training-free setting, while other strategy represents its trained variant, the inclusion of Keyword Alignment yields substantial improvements across multiple core metrics, demonstrating that Keyword Alignment is crucial for enhancing the precision of semantic retrieval within the multimodal context. Furthermore, the performance uplift from Reasoning Alignment validates the efficacy of reasoning refinement via GRPO. By facilitating reasoning exploration, this process enables the model to better utilize the retrieved structured knowledge, thereby improving prediction reliability in complex cases.

D. Quality of Knowledge Graph

To assess the quality of the constructed harmfulness knowledge graph, we employ GPT-5.5, an advanced multimodal language model from a different model family than the knowledge construction model, as an independent evaluator. This cross-model assessment examines the graph-wide knowledge quality along three complementary dimensions: Support Traceability, Cross-scenario Coverage, and Rhetorical Coherence. Specifically, we provide GPT-5.5 with batches of harmfulness rule units and the corresponding graph structures, and instruct it to assess the overall knowledge quality according to predefined evaluation criteria. The resulting statistics provide an assessment of semantic grounding, cross-scenario reusability, and intra-cluster consistency.

TABLE VIII
HUMAN EVALUATION RESULTS ON INTERPRETABILITY.

Method	Faith.	Suff.	Rel.
Qwen3-VL-8B	3.73	3.42	—
MemeLogos	4.43	4.19	4.61

Support Traceability. We assess whether supporting variants maintain clear semantic connections to their source anchors. A variant is considered traceable when its core semantics and harmfulness-relevant attributes remain consistent with the source instance without introducing contradictory or unsupported information. The GPT-based evaluation reports a traceability rate of 90.30%, suggesting that most supporting evidence preserves identifiable semantic associations with its source instances.

Cross-scenario Coverage. We examine whether harmfulness rules capture recurring patterns across different scenarios rather than being restricted to individual templates or contexts. A rule is considered cross-scenario when its supporting variants originate from at least two distinct scenarios. The evaluation reports that 97.09% of multi-variant rules satisfy this criterion, suggesting that the knowledge exhibits considerable cross-scenario recurrence and potential reusability.

Rhetorical Coherence. We further assess whether supporting variants within the same rule cluster exhibit consistent rhetorical mechanisms. The assessment reports an overall rhetorical consistency of 78.19%, suggesting that most rule clusters preserve coherent rhetorical patterns, while the inherent ambiguity and overlap of rhetorical mechanisms may introduce uncertainty into evaluation.

Overall, these results indicate that the constructed harmfulness knowledge and graph structure largely preserve semantic consistency and evidence traceability while capturing recurring harmfulness patterns across diverse meme scenarios, providing a reliable structured knowledge foundation for subsequent harmful meme detection.

E. Interpretability Analysis

We conduct a human evaluation on 1,000 randomly sampled memes using a 5-point Likert scale. Three independent annotators assess the generated reasoning along three complementary dimensions: **Faithfulness** (Faith.), which measures the consistency between the generated reasoning and the observable multimodal evidence in the input meme. A high score requires that the explanation correctly reflects the visual and textual content without introducing unsupported entities, intentions, or semantic relations. We emphasize that this metric evaluates input-grounded consistency rather than causal faithfulness between the explanation and the model's internal decision process; **Sufficiency** (Suff.), which evaluates whether the explanation covers the key evidence and reasoning steps required to understand the final prediction. A sufficient explanation should identify the critical multimodal cues, clarify their semantic interaction, and explicitly connect them to the harmfulness judgment. Explanations that merely restate the meme content or provide a verdict without intermediate

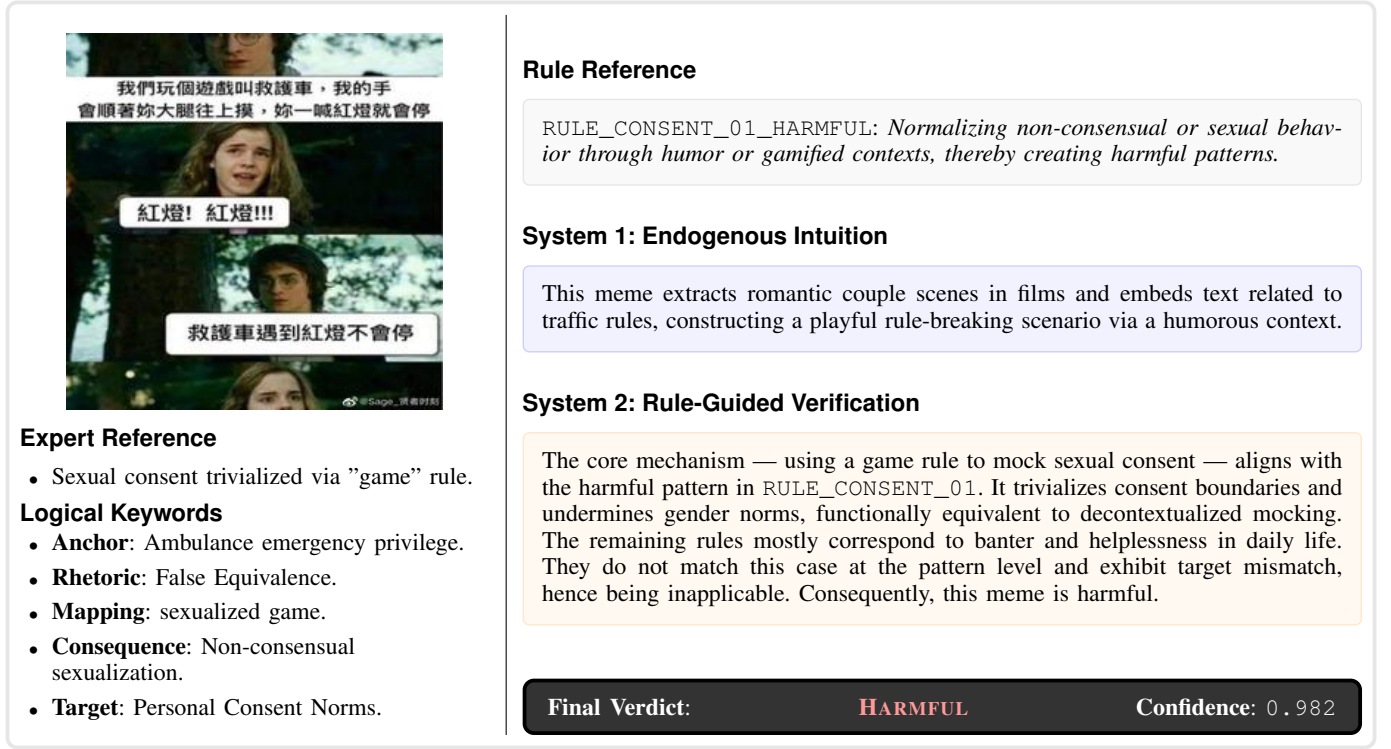


Fig. 4. Qualitative illustration of the reasoning process in MemeLogos. The framework extracts key semantic cues, retrieves relevant harmfulness rules, and performs dual-system reasoning to progressively verify the evidence and derive the final harmfulness verdict.

justification receive lower scores; and **Rule Relevance** (Rel.), which evaluates whether the retrieved rule-level knowledge is semantically applicable to the current meme and whether the generated reasoning uses such knowledge appropriately. A high score requires that the retrieved rules capture the central harmfulness mechanism of the meme rather than only matching superficial visual or lexical features. This metric is only applicable to MemeLogos, since the vanilla backbone does not perform rule retrieval. These metrics respectively evaluate whether the reasoning remains grounded in the input meme, whether it provides sufficient evidence for understanding the final judgment, and whether the retrieved harmfulness rules are applicable to the current case. Detailed information of the three dimensions is shown in Appendix ??

Table VIII shows that MemeLogos achieves higher Faithfulness and Sufficiency scores than the Qwen3-VL-8B backbone with CoT prompting, while also obtaining a high Rule Relevance score. The three annotators exhibit substantial agreement, with ordinal Krippendorff's α values of 0.79, 0.83, and 0.74 for Faithfulness, Sufficiency, and Rule Relevance, respectively. These results indicate that the retrieved rule-level knowledge provides effective support for generating more grounded and sufficiently justified reasoning, thereby enhancing the interpretability of harmfulness judgments.

To further illustrate how this knowledge-grounded reasoning process operates in practice, we provide a representative qualitative example in Figure 4. In this example, the meme combines the visual context of an ambulance-related emergency privilege with a sexualized "game-rule" interpretation, where the humorous framing implicitly trivializes personal

consent. MemeLogos first identifies the key multimodal cues and abstracts them into structured logical keywords, including the anchor, false-equivalence rhetoric, sexualized mapping, target, and potential consequence. Based on these cues, the framework retrieves a relevant harmfulness rule concerning the normalization of non-consensual or sexual behavior through humorous or gamified contexts. System 1 then forms an initial interpretation of the meme, while System 2 contrasts the retrieved rule with alternative candidates and verifies its applicability before producing the final harmful verdict.

Despite the generally reliable reasoning demonstrated above, MemeLogos may still fail when implicit slang semantics or misleading surface-level similarities dominate the retrieval process. Figure 5 presents a representative failure case in which MemeLogos incorrectly classifies a harmful meme as harmless. The partially illuminated MILFORD PLAZA sign appears as MILF PLAZA, and together with the caption Bro stop the fucking car, conveys a sexualized interpretation associated with objectification. However, the model fails to fully recognize the slang meaning of MILF and instead emphasizes the surface-level visual incongruity, retrieving a harmless rule related to humorous visual juxtaposition. This error mainly stems from two factors: insufficient understanding of colloquial and metaphorical expressions, and retrieval that overweights rhetorical similarity while underweighting the target and harmful consequence. The case therefore highlights the need for stronger slang-aware semantic understanding and more balanced retrieval over different harmfulness attributes.

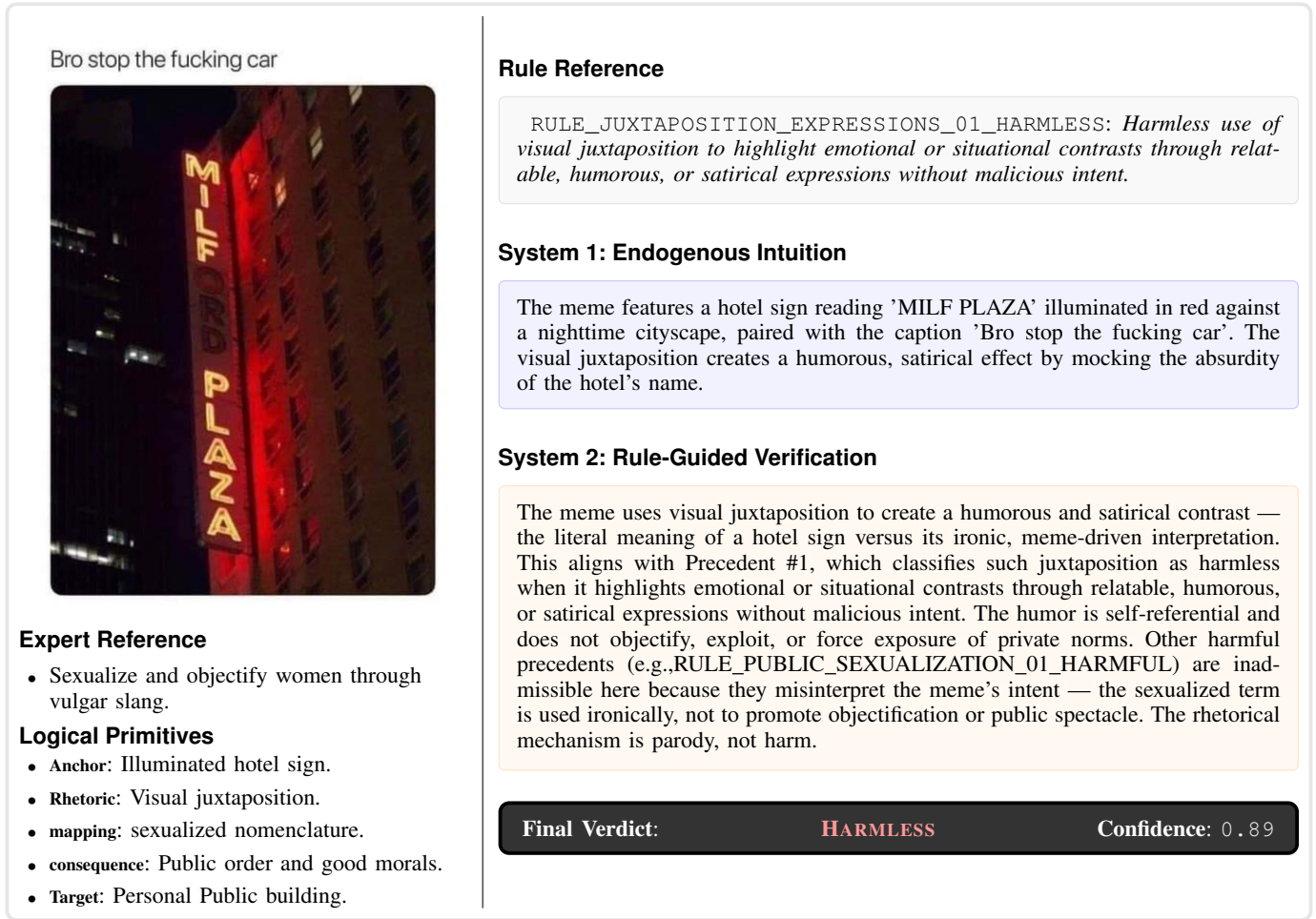


Fig. 5. Failure case of MemeLogos. The model captures the surface-level visual juxtaposition but retrieves and over-relies on an inappropriate harmless rule, failing to recognize the implicit sexualized meaning conveyed by the slang expression and consequently producing an incorrect verdict.

TABLE IX
CROSS-DOMAIN ZERO-SHOT EVALUATION OF MEMELOGOS ON PRIDEMM.

Method	Accuracy	F1
In-domain Supervised Evaluation		
CLIP	72.39	72.33
MemeCLIP	76.06	75.09
U-CoT+	71.60	71.37
RGCL	76.30	76.50
RA-HMD	78.10	78.40
ExPO-HM	—	78.70
Cross-domain Zero-shot Evaluation		
Qwen3-VL-8B	64.69	69.61
MemeLogos	77.51	79.50

F. Cross-domain Experiment

To evaluate the cross-domain generalization ability of MemeLogos, we conduct zero-shot transfer experiments on PrideMM using only the harmfulness knowledge constructed from the MemeMind training set, without target-specific fine-tuning. As shown in Table IX, we include the reported

results of existing in-domain supervised methods as references. MemeLogos still maintains stable detection performance on PrideMM, indicating that the learned rule-level harmfulness knowledge captures reusable patterns that generalize across different data distributions.

G. Parameter Sensitivity Analysis

In this section, we perform parameter sensitivity analyses for the harmfulness knowledge graph and Cognitive RAG in MemeLogos on the MemeMind dataset. The remaining training hyperparameters were fixed without tuning based on the default configurations of the corresponding backbone and optimization framework.

1) *Graph Sensitivity Analysis:* We investigate the sensitivity of MemeLogos to the key parameters involved in rule induction and graph construction. Specifically, we analyze the coarse clustering threshold τ_r , the Semantically Aligned edge threshold τ_a , and the Guards Against edge threshold τ_g . All configurations are evaluated using the same fixed set of 5,000 training instances and 200 validation instances.

For τ_r , we vary the threshold from 0.60 to 0.90 while keeping the graph parameters fixed. As shown in Fig 6, the performance exhibits a non-monotonic but relatively stable

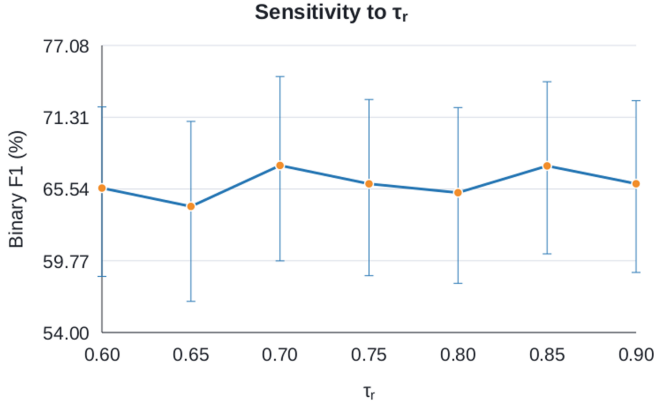


Fig. 6. The sensitivity experiments of MemeLogos to the rule-clustering threshold τ_r , conducted on a fixed 5,000/200 diagnostic subset. Error bars denote 95% confidence intervals estimated using 5,000 class-stratified bootstrap resamples.

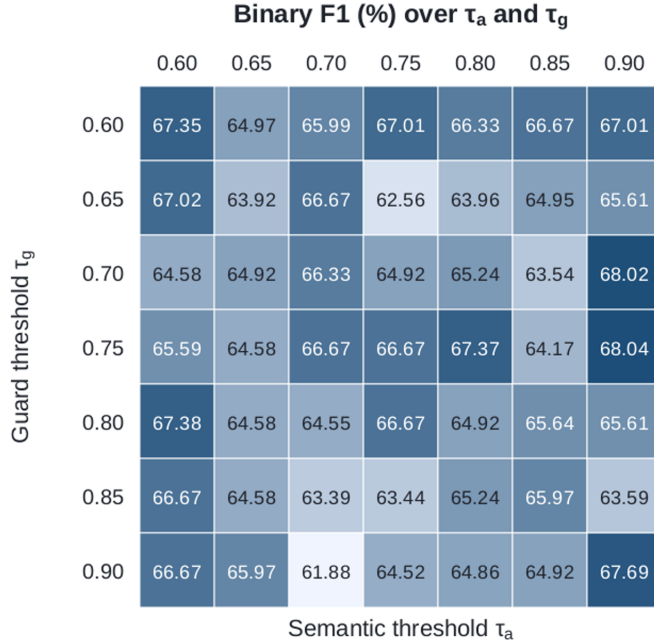


Fig. 7. The sensitivity experiments of MemeLogos to the combinations of the Semantically Aligned edge threshold τ_a and Guards Against edge threshold τ_g conducted on a fixed 5,000/200 diagnostic subset.

trend across different values of τ_r . The best observed overall performance is obtained at $\tau_r = 0.70$, achieving an Accuracy of 71.50% and an F1-score of 67.40%. Comparable performance is also observed at several other settings, indicating that the method is not highly sensitive to moderate changes in the rule-clustering threshold.

We further jointly vary τ_a and τ_g from 0.60 to 0.90, as shown in Fig 7. The highest observed F1-score of 68.04% is achieved at $(\tau_a, \tau_g) = (0.90, 0.75)$, while several other parameter combinations obtain similar performance. The absence of a single sharp performance peak suggests that MemeLogos does not rely on a narrowly tuned graph-construction setting and maintains relatively stable performance over a broad range of threshold combinations.

TABLE X
SENSITIVITY ANALYSIS OF THE NUMBER OF RETRIEVED RULES FOR EACH POLARITY.

Rules per Polarity (k)	Accuracy	F1
1	86.65	84.93
3	86.96	85.22
5	86.62	84.91

2) *Retrieval count Sensitivity Analysis:* We further analyze the sensitivity of Cognitive RAG to the number of retrieved rules. Specifically, we vary the number of rules retrieved for each polarity, denoted by K , while keeping all other settings unchanged. The results are reported in Table X.

As shown in Table X, the model achieves comparable performance across different values of K , with an accuracy variation of only 0.34 percentage points. Retrieving too few rules may limit the availability of relevant evidence, whereas retrieving more rules may introduce additional irrelevant or redundant information. The best result is obtained at $K = 3$, suggesting that a moderate retrieval size provides a suitable balance between evidence coverage and retrieval noise. Nevertheless, the limited performance variation demonstrates that Cognitive RAG is robust to the number of rules retrieved for each polarity.

V. CONCLUSION

In this paper, we propose MemeLogos, a knowledge-driven retrieval-augmented reasoning framework for harmful meme detection. Unlike conventional approaches relying on instance-level retrieval or unstructured explanations, MemeLogos abstracts recurring harmfulness patterns from CoT-annotated memes into reusable rule-level knowledge and organizes them into a structured harmfulness knowledge graph with explicit semantic and verdict-polarity-aware relations. Building upon this knowledge representation, Cognitive RAG integrates independent meme interpretation, graph-grounded contrastive retrieval, and deliberative reasoning to enable more comprehensive harmfulness analysis and judgment. Extensive experiments demonstrate that MemeLogos effectively improves harmful meme detection performance. Furthermore, ablation studies and interpretability evaluations highlight the contributions of structured knowledge organization and cognitive reasoning to more accurate predictions and evidence-grounded explanations. Overall, this work demonstrates the value of abstracting instance-level semantic knowledge into reusable, structured reasoning units, providing an effective approach to knowledge-enhanced multimodal harmful content detection.

REFERENCES

- [1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html>

- [2] D. Bandyopadhyay, A. Ganguly, B. Gain, and A. Ekbal, "SEMANTIFY: unveiling memes with robust interpretability beyond input attribution," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 6189–6197. [Online]. Available: <https://www.ijcai.org/proceedings/2024/684>
- [3] R. Garg, T. Padhi, H. Jain, U. Kursuncu, and P. Kumaraguru, "Just kiddin' : Knowledge infusion and distillation for detection of indecent memes," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, ser. Findings of ACL, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., vol. ACL 2025. Association for Computational Linguistics, 2025, pp. 23 067–23 086. [Online]. Available: <https://doi.org/10.18653/v1/2025.findings-acl.1184>
- [4] Y. Li, L. Zhang, L. Zhang, and Q. Yan, "Kid: Knowledge-injected dual-head learning for knowledge-grounded harmful meme detection," 2026.
- [5] F. Pan, X. Wu, T. Quan, and A. T. Luu, "Read as you see: Guiding unimodal llms for low-resource explainable harmful meme detection," in *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, H. Hacid, Y. Maarek, F. Bonchi, I. Guy, and E. Yilmaz, Eds. ACM, 2026, pp. 1672–1683. [Online]. Available: <https://doi.org/10.1145/3774904.3792570>
- [6] H. Gu, Q. Yu, S. Hou, Z. Fang, H. Wu, and Z. He, "Mememind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection," 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280000554>
- [7] D. Kahneman, *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
- [8] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2783–2796. [Online]. Available: <https://doi.org/10.18653/v1/2021.findings-acl.246>
- [9] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen, "SemEval-2022 task 5: Multimedia automatic misogyny identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022.
- [10] J. Lu, B. Xu, X. Zhang, H. Wang, H. Zhu, D. Zhang, L. Yang, and H. Lin, "Towards comprehensive detection of chinese harmful memes," 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273098367>
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:225039882>
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:231591445>
- [13] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252846273>
- [14] J. Ji, X. Lin, and U. Naseem, "Capalign: Improving cross modal alignment via informative captioning for harmful meme detection," *Proceedings of the ACM Web Conference 2024*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:269639711>
- [15] R. Sood, A. Anaissi, W. Huang, and A. Braytee, "Memhatecaptioning: Enhancing hate speech detection in memes with context-aware captioning and chain-of-thought," in *Companion Proceedings of the ACM on Web Conference 2025*. New York, NY, USA: Association for Computing Machinery, 2025, p. 2034–2041.
- [16] H. Lin, Z. Luo, J. Ma, and L. Chen, "Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023.
- [17] H. Lin, W. Gao, J. Ma, Y. Deng, Z. Luo, B. Wang, R. Yang, and T.-S. Chua, "Explainm++: Explainable harmful meme detection with retrieval-augmented debate between large multimodal models," *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- [18] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.
- [19] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi, "G-retriever: Retrieval-augmented generation for textual graph understanding and question answering," *Advances in Neural Information Processing Systems*, vol. 37, pp. 132 876–132 907, 2024.
- [20] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su, "Hipporag: Neurobiologically inspired long-term memory for large language models," *Advances in neural information processing systems*, vol. 37, pp. 59 532–59 569, 2024.
- [21] X. Zhu, Y. Xie, Y. Liu, Y. Li, and W. Hu, "Knowledge graph-guided retrieval augmented generation," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 8912–8924.
- [22] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 3868–3898.
- [23] L. Luo, Y.-F. Li, R. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 14 400–14 423.
- [24] S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, C. Yang, J. Mao, and J. Guo, "Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 52 782–52 806.
- [25] Z. Li, X. Chen, H. Yu, H. Lin, Y. Lu, Q. Tang, F. Huang, X. Han, L. Sun, and Y. Li, "Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 36 107–36 124.
- [26] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnnet: Masked and permuted pre-training for language understanding," 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:215827489>
- [27] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:199528533>
- [28] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 5244–5252.
- [29] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Mapping memes to words for multimodal hateful meme classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2832–2836.
- [30] S. B. Shah, S. Shiwakoti, M. Chaudhary, and H. Wang, "Memeclip: Leveraging clip representations for multimodal meme classification," 2024.
- [31] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin, "Improving hateful meme detection through retrieval-guided contrastive learning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024.
- [32] J. Mei, J. Chen, G. Yang, W. Lin, and B. Byrne, "Robust adaptation of large multimodal models for retrieval augmented hateful meme detection," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025.
- [33] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *ArXiv*, vol. abs/2502.13923, 2025.
- [34] Qwen Team, "Qwen3.6-Plus: Towards real world agents," Qwen Blog, Apr. 2026, accessed: Aug. 4, 2026. [Online]. Available: <https://qwen.ai/blog?id=qwen3.6>

- [35] OpenAI, “GPT-5.5 system card,” OpenAI, Apr. 2026, accessed: Aug. 4, 2026. [Online]. Available: <https://openai.com/index/gpt-5-5-system-card/>
- [36] W. Yu, H. Zhang, X. Pan, P. Cao, K. Ma, J. Li, H. Wang, and D. Yu, “Chain-of-note: Enhancing robustness in retrieval-augmented language models,” in *Proceedings of the 2024 conference on empirical methods in natural language processing*, 2024, pp. 14 672–14 685.
- [37] C. Amiraz, F. Cuconasu, S. Filice, and Z. Karnin, “The distracting effect: Understanding irrelevant passages in rag,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 18 228–18 258.
- [38] Z. Shao, P. Wang, Q. Zhu, R. Xu, J.-M. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” 2024.
- [39] J. Mei, M. Sun, J. Chen, P. Qin, Y. Li, D. Chen, and B. Byrne, “Expohm: Learning to explain-then-detect for hateful meme detection,” 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:282055901>
- [40] S. Bai, Y. Cai, R. Chen, K. Chen, X.-H. Chen, Z. Cheng, L. Deng, W. Ding, R. Fang, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, Q. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Y. Meng, X. Ren, X. yi Ren, S. Song, Y. chen Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, “Qwen3-vl technical report,” 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:283262018>
- [41] T. Yu, Z. Wang, C. Wang, F. Huang, W. Ma, Z. He, T. Cai, W. Chen, Y. Huang, Y. Zhao, B. Xu, J. Cui, Y. Xu, L. Ruan, L. Zhang, H. Liu, J. Tang, H. Liu, Q. Guo, W. Hu, B. S. He, J. Zhou, J. Cai, J. Qi, Z. Guo, C. Chen, G. Zeng, Y. Li, G. Cui, N. Ding, X. Han, Y. Yao, Z. Liu, and M. Sun, “Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe,” 2025.
- [42] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, Z. Wang, Z. Chen, H. Zhang, G. Yang, H. Wang, Q. Wei, J. Yin, W. Li, E. Cui, G. Chen, Z. Ding, C. Tian, Z. Wu, J. Xie, Z. Li, B. Yang, Y. Duan, X. Wang, H. Hao, S. Li, X. Zhao, H. Duan, N. Deng, B. Fu, Y. He, Y. Wang, C. He, B. Shi, J. He, Y. Xiong, H. Lv, L. Wu, W. Shao, K. Zhang, H. Deng, B. Qi, B. Qi, Q. Guo, W. Zhang, Y. Gu, W. Ouyang, L. Wang, M. Dou, X. Zhu, T. Lu, D. Lin, J. Dai, B. Zhou, W. Su, K. Chen, Y. Qiao, W. Wang, and G. Luo, “Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280710824>

Hexiang Gu received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2025. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include multimodal large language model, and multimodal content safety.

Saihui Hou received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2014 and 2019, respectively. He is currently an Associate Professor at the School of Artificial Intelligence, Beijing Normal University. His research interests include computer vision and machine learning, with a focus on gait recognition, which aims to identify individuals based on their walking patterns.

Changlan Liu received the B.S. degree from Northeast Agricultural University in 2026. He is currently pursuing the direct Ph.D. degree with the College of Electrical Engineering, Zhejiang University. His research interests include LLM security.

Min Ren received the B.E. and Ph.D. degrees from the National University of Defense Technology and University of Chinese Academy of Sciences in 2013 and 2023, respectively. He is currently an Associate Professor at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include embodied intelligence and computer vision.

Huijia Wu received the B.Sc. degree in e-commerce from Dalian Maritime University, China in 2007, the M.Sc. degree in e-commerce from Shanghai University of Finance and Economics, China in 2011, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), China. He worked as a Senior Researcher at the WeChat Pattern Recognition Center and as an Advertising Algorithm Expert at Alibaba’s Alimama, with 8 years of experience and achievements in both academia and industry. He has published over 10 papers in leading international conferences such as ACL, AAAI, EMNLP, and COLING. His research interests include natural language processing, personalized recommendation systems and machine learning.

Hongxun Gu received the BSc and MSc degrees from Huazhong University of Science and Technology, Wuhan, China, in 1994 and 1997 respectively, and the MSc degree in Management from The Hong Kong Polytechnic University in 2016. He is currently the General Manager of the Network and Information Security Management Department in China Telecommunications Corporation. He is an Adjunct Professor at Huazhong University of Science and Technology. He is a Board Member of the China Cyber Security Society, a Fellow of the China Institute of Communications. His research interests cover cyberspace security, artificial intelligence security, and related fields.

Zhaofeng He received the B.S. degree in Electronic Engineering and Information Science from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree in Computer Applied Technology from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a Professor at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing. His research interests include computer vision, biometrics, and intelligent decision-making in games.