

HarmEvoSkill: Skill-Memory Self-Evolution for Zero-shot Multimodal Harmful Meme Detection

Jianyu Zhou*
Hohai University
Nanjing, China
2361010312@hhu.edu.cn

Weihai Lu*†
Peking University
Beijing, China
luweihai@pku.edu.cn

Abstract

Multimodal harmful meme detection is challenging because harmful intent is often expressed through implicit image-text associations, cultural references, rhetorical framing, and context-dependent targets. Existing methods mainly rely on supervised multimodal models, MLLM prompting, or fixed agent workflows. However, they still lack experience accumulation and rely on static skill design, making it difficult to reuse prior judgments or adapt to emerging meme templates and long-tail harm patterns. To address these limitations, we propose HARMEVOSKILL (**H**armful Meme **E**volving **S**kill-Memory), a label-free self-evolving Skill-Memory framework for harmful meme detection. HARMEVOSKILL converts unlabeled reference memes into Case Memory, Insight Memory, and Risk Memory. It then automatically discovers candidate skills, abstracts recurring reasoning patterns, and induces executable skill cards without predefined skill labels. During inference, a Skill Planner retrieves relevant memories, selects task-specific skills, and generates evidence. A graph-assisted judge organizes the evidence into a harmful-harmless claim graph, makes the final prediction, and updates memory through graph diagnosis. Extensive benchmark evaluations show that HARMEVOSKILL achieves consistently strong performance among evaluated methods. On HarM, it improves Accuracy by 9.33% and Macro-F1 by 17.97% over the direct MLLM baseline.

1 Introduction

Multimodal harmful memes pose a major challenge to online safety because harmful intent is often conveyed implicitly through the interaction between images and text. Such intent may be encoded through humor, cultural references, rhetorical framing, evolving meme templates, or context-

dependent targets, rather than through explicit hateful words (Ma et al., 2025). A meme may look humorous, ironic, or politically critical, while implicitly attacking a protected group or normalizing harmful stereotypes. This makes harmful meme detection harder than standard multimodal classification: the model must jointly interpret visual and textual cues, account for cultural context, infer the target and stance, and distinguish harmful implication from benign humor, satire, or political criticism (Wang et al., 2026a; Kmainasi et al., 2026).

To address this challenge, existing methods mainly follow two directions. The first line develops supervised or prompt-based multimodal models for harmful and hateful meme classification, using textual, visual, and OCR signals to capture cross-modal harmful cues (Kiela et al., 2020; Pramanick et al., 2021; Cao et al., 2022). These methods learn useful task-specific patterns, but they often struggle with emerging meme templates, implicit cultural references, and long-tail rhetorical forms. The second line explores large multimodal models and agent-based reasoning for zero-shot or low-resource harmful meme understanding (Huang et al., 2025; Mei et al., 2025; Liu et al., 2025). These methods improve interpretability by using prompting, retrieval, or multi-agent deliberation, but their reasoning experience is still not accumulated as reusable memory, and their skills are usually designed as fixed prompts or predefined agents.

Despite this progress, current methods still face two key limitations:

(1) Lack of Experience Accumulation. Each meme is usually treated as an isolated inference case, so successful judgments, failure modes, and reusable insights are not preserved for future reasoning. This prevents the model from reusing prior harmfulness judgments, target-disambiguation experience, or hard-negative patterns when facing new memes.

*Equal contribution.

†Corresponding author.

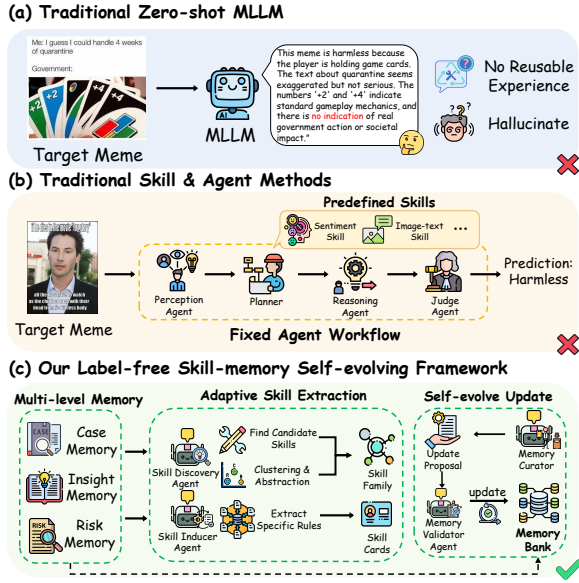


Figure 1: Motivation of HARMEVOSKILL, which constructs multi-level memory from unlabeled reference memes, automatically extracts executable skills, and updates the memory bank through self-evolving diagnosis.

(2) Static Skill Design. Existing agent or skill-based workflows mostly rely on predefined prompts, fixed agent roles, or manually designed reasoning procedures. As a result, they are difficult to adapt to new cultural templates, political events, coded references, or long-tail harm patterns. In harmful meme detection, useful skills should be discovered from prior cases and updated through later reasoning.

Figure 1 contrasts no-reference direct MLLM reasoning, fixed agent workflows, and our label-free inference setting with unlabeled reference memory. The first setting measures the raw reasoning ability of an MLLM without external reference memes, whereas our setting uses an unlabeled reference split but no gold labels, supervised task training or task-specific parameter updates.

Inspired by recent studies on agentic memory, self-evolving memory, and evolvable memory skills (Xu et al., 2025; Lang et al., 2026; Zhang et al., 2026), we propose HARMEVOSKILL, a label-free self-evolving Skill-Memory framework for multimodal harmful meme detection. We introduce a novel formulation of multimodal harmful meme detection as a validated skill-memory evolution problem, where memory construction, skill induction, graph-assisted reasoning, and validated updates form a task-specific self-evolving loop. Within this loop, HARMEVOSKILL integrates multi-level memory construction, automatic

skill induction, graph-assisted reasoning, and validated memory updates into a unified task-specific framework. Our code is available on GitHub¹.

Our contributions are summarized as follows:

- To the best of our knowledge, we introduce the first self-evolving Skill-Memory framework for multimodal harmful meme detection, unifying memory construction, skill induction, graph-assisted reasoning, and validated updates in a label-free loop.
- We design a multi-level memory architecture comprising Case, Insight, and Risk Memory to preserve instance-level cases, reusable cross-case insights, and ambiguous or failure-prone patterns.
- We develop an adaptive skill-evolution mechanism that discovers and induces executable skills without predefined labels, and combines memory-guided planning with harmful-harmless claim graphs and update validation.
- Extensive experiments across benchmark datasets show that HARMEVOSKILL consistently outperforms competitive baselines and achieves strong performance among evaluated methods. In particular, on HarM, HARMEVOSKILL improves Accuracy and Macro-F1 by 9.33% and 17.97% respectively.

2 Related Work

2.1 Harmful Meme Detection

Harmful meme detection has been widely studied as a multimodal safety problem. Early work mainly formulates it as supervised image-text classification. The Hateful Memes Challenge shows that unimodal cues are insufficient and introduces benign confounders to require joint visual-textual reasoning (Kiela et al., 2020). MOMENTA further studies harmful meme detection and target identification through global and local multimodal modeling (Pramanick et al., 2021). PromptHate exploits prompt-based language modeling and in-context examples to use implicit knowledge from pretrained language models for hateful meme classification (Cao et al., 2022). These methods establish strong task-specific baselines, but they usually rely on annotated data and may struggle when meme templates, cultural references, or harm mechanisms shift.

¹<https://github.com/luoli-z/HarmEvoSkill>

Recent studies explore LLMs, MLLMs, retrieval, and agent-based reasoning for more flexible harmful meme understanding. LLM-based methods improve implicit harmfulness reasoning through distilled multimodal reasoning or harmful-harmless debate (Lin et al., 2023, 2024). Retrieval and agent-based methods further support low-resource or zero-shot adaptation by revising agent knowledge, retrieving related memes, or injecting contextual evidence into MLLMs (Huang et al., 2024; Liu et al., 2025; Mei et al., 2025). Related multimodal misinformation studies further explore multi-domain and domain-aware modeling to capture heterogeneous cross-modal patterns across different news domains (Tong et al., 2024; Lu et al., 2025). More recent work investigates domain-aware prompt tuning and retrieval-augmented multi-agent reasoning for more adaptive multimodal evidence integration (Zeng et al., 2026; Tong et al., 2025; Lu et al., 2026). Recent reasoning-oriented methods model meme evolution, social graph-of-thought inference, or thinking-based MLLM post-training (Huang et al., 2025; Wang et al., 2026b; Kmainasi et al., 2026). Closely related to our setting, ALARM performs label-free MLLM-agent self-improvement by distilling knowledge from pseudo-labeled explicit memes and contrastive meme pairs (Lang et al., 2026). However, ALARM represents acquired experience as revised agent knowledge rather than typed multi-level memory and does not explicitly induce executable skill cards. In contrast, HARMEVOSKILL constructs Case, Insight, and Risk Memory, induces executable skills, and combines memory-guided evidence planning, harmful-harmless claim graphs, and validator-controlled updates in a task-specific self-evolution loop.

2.2 Memory, Skill, and Self-Evolving Agents

Our work is also related to recent studies on self-evolving memory and skill-based agents. IMOL retrieves transferable cross-sample knowledge to improve multimodal reasoning under incomplete modalities (Zeng et al., 2025). A-MEM proposes an agentic memory system that dynamically organizes memories rather than relying on fixed memory operations (Xu et al., 2025). G-Memory introduces hierarchical agentic memory for multi-agent systems and organizes interaction histories through multi-level graph structures (Zhang et al., 2025). Evo-Memory formulates self-evolving memory as a test-time learning problem, where agents retrieve,

reuse, and update memory across streaming task sequences (Wei et al., 2025). These studies suggest that memory should not be treated as a static retrieval cache, but as an evolving structure that is updated as agents encounter new tasks.

Another line of work studies reusable skill acquisition for agents. EvolveR distills interaction trajectories into strategic principles for later task solving (Wu et al., 2025). MemSkill treats memory operations as learnable skills that can be selected and refined from hard cases (Zhang et al., 2026), while Memento-Skills stores reusable skills as structured external memory and updates the skill library through a read-write reflective loop (Zhou et al., 2026). These methods move beyond fixed prompting, but they mainly target general agent tasks, tool use, multi-turn interaction, or reasoning benchmarks. They do not address multimodal harmful meme detection, where the key difficulty lies in implicit visual-textual harm, social targets, cultural references, and ambiguous protected-group implications.

In contrast, HARMEVOSKILL adapts skill-memory evolution to multimodal harmful meme detection. It constructs Case, Insight, and Risk Memory from unlabeled reference memes, induces reusable reasoning skills, and updates memory through graph-assisted diagnosis. This makes memory and skill evolution serve fine-grained harmfulness judgment rather than generic task completion.

3 Methodology

We propose HARMEVOSKILL, a label-free self-evolving Skill-Memory framework for multimodal harmful meme detection. As shown in Figure 2, HARMEVOSKILL consists of three stages. First, it builds multi-level memory from unlabeled memes and adaptively induces executable skill cards. Second, it retrieves relevant memory to plan task-specific skills and generate structured evidence. Third, it integrates the evidence into a harmful-harmless claim graph for final judgment and self-evolving memory update.

3.1 Problem Definition

Given a multimodal meme $x = (I, T)$, where I denotes the image and T denotes the associated text, OCR content, or caption, the task is to predict a harmfulness label $y \in \{0, 1\}$. Here, $y = 1$ indicates harmful and $y = 0$ indicates harmless. In

our setting, the model has access to an unlabeled reference set:

$$\mathcal{R} = \{x_i\}_{i=1}^N, \quad x_i = (I_i, T_i), \quad (1)$$

where no gold labels are used at any stage and no task-specific model parameters are trained or updated. We refer to this protocol as label-free inference with unlabeled reference memory; unlike no-reference direct prompting, it permits access to $\mathcal{R}$ but not to its labels. Given a target meme x_q , the goal is to predict its harmfulness while producing structured evidence and reusable reasoning experience.

3.2 Label-free Skill-Memory Construction

The first stage constructs reusable memory from the unlabeled reference set. For each reference meme, an MLLM analyzer produces a structured analysis, including visual entities, OCR cues, image-text relations, target candidates, possible harm mechanisms, benign interpretations, and uncertainty signals. This follows the recent view that agent memory should store structured and dynamically organized observations rather than raw interaction traces alone (Xu et al., 2025; Zhang et al., 2025). These analyses are not used as pseudo labels. Instead, they serve as intermediate observations for memory construction.

We organize the resulting experience into a multi-level memory bank:

$$\mathcal{B} = \mathcal{M}^{case} \cup \mathcal{M}^{insight} \cup \mathcal{M}^{risk}, \quad (2)$$

where $\mathcal{B}$ denotes the full memory bank. $\mathcal{M}^{case}$ stores instance-level contexts, $\mathcal{M}^{insight}$ stores generalized reasoning patterns abstracted from similar cases, and $\mathcal{M}^{risk}$ stores ambiguous or error-prone patterns. This design is also related to recent memory-based agents that retain prior experiences, distill reusable strategies, and use successful or failed cases to guide later reasoning.

Case Memory keeps concrete meme-level information, including visual evidence, textual cues, target hypotheses, image-text relations, and competing harmful or harmless interpretations. It allows the framework to retrieve similar historical cases during inference.

Insight Memory abstracts reusable reasoning patterns from groups of related cases. For example, it may capture recurring visual metaphors, coded group references, rhetorical framing, target shifting, or cases where political criticism should be separated from protected-group attacks.

Risk Memory records cases with unstable or conflicting interpretations. These include low-confidence predictions, unclear targets, misleading humor, weak context, or cases where both harmful and harmless readings are plausible. Risk Memory is used to warn the planner and judge about likely failure modes.

Based on the typed memory bank, HARMEVOSKILL performs adaptive skill extraction. Unlike fixed agent pipelines, the skills are not manually predefined. A Skill Discovery Agent first mines case-aware skill proposals from Case, Insight, and Risk Memory. Similar proposals are then grouped and abstracted into seed skill families. A Skill Inducer further converts each seed family into an executable skill card. Each induced skill is represented as:

$$s_j = \{s_j^{tri}, s_j^{evi}, s_j^{pro}, s_j^{ver}, s_j^{fail}\}, \quad (3)$$

where s_j^{tri} denotes the triggering condition, s_j^{evi} specifies the required evidence, s_j^{pro} describes the reasoning procedure, s_j^{ver} defines the verification rule, and s_j^{fail} records known failure modes. A Memory Validator checks whether the generated skill is executable, non-redundant, and consistent with the memory bank. The validated skills form the induced skill pool:

$$\mathcal{S} = \{s_j\}_{j=1}^J, \quad (4)$$

where J is the number of validated skills.

3.3 Memory-guided Skill Planning and Evidence Generation

During inference, HARMEVOSKILL first extracts a query signature from the target meme. The signature summarizes visual content, OCR text, entities, target candidates, tone, image-text relation, possible harm mechanisms, and uncertainty cues. It is used to retrieve relevant Case, Insight, and Risk Memory.

For each memory item m_j in the memory bank, we compute a multimodal retrieval score:

$$s(x_q, m_j) = \lambda_t \cos(\mathbf{e}_q^t, \mathbf{e}_j^t) + \lambda_v \cos(\mathbf{e}_q^v, \mathbf{e}_j^v), \quad (5)$$

where $\mathbf{e}_q^t$ and $\mathbf{e}_q^v$ denote the textual and visual embeddings of the target meme, and $\mathbf{e}_j^t$ and $\mathbf{e}_j^v$ denote the corresponding embeddings of memory item m_j . λ_t and λ_v balance textual and visual similarity. The retrieved memory set is defined as:

$$\mathcal{B}_q = \arg \max_{B' \subseteq \mathcal{B}, |B'|=K} \sum_{m_j \in B'} s(x_q, m_j), \quad (6)$$

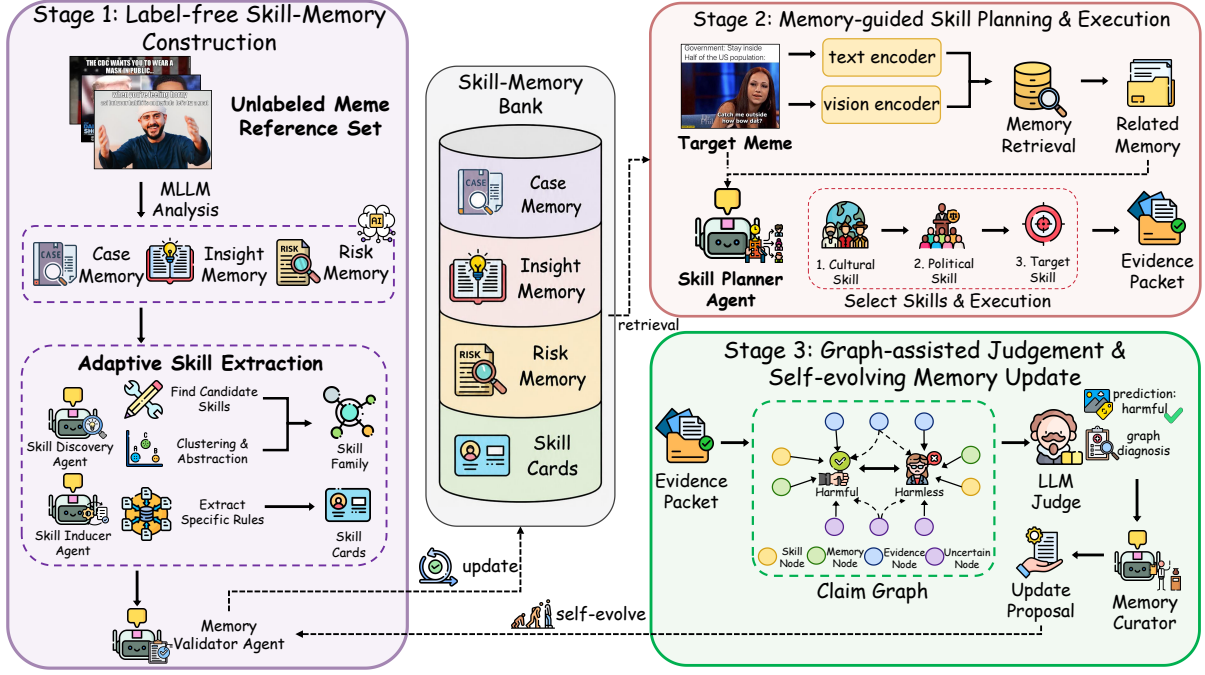


Figure 2: Overall framework of HARMEVOSKILL. Stage 1 builds multi-level memory from unlabeled reference memes and induces executable skill cards. Stage 2 retrieves relevant memories, selects task-specific skills, and generates structured evidence packets. Stage 3 builds harmful–harmless claim graphs for final judgment and graph-assisted memory update.

where K is the memory retrieval budget.

The retrieved memory provides two types of guidance for skill planning. Case and Insight Memory provide positive reasoning experience by showing how similar patterns were previously interpreted. Risk Memory provides cautionary signals by highlighting ambiguous targets, misleading humor, weak context, or harmful–harmless conflicts.

Based on $\mathcal{B}_q$, the Skill Planner selects L task-specific skills from the induced skill pool $\mathcal{S}$, forming $\mathcal{S}_q \subseteq \mathcal{S}$ with $|\mathcal{S}_q| = L$, where L is the skill planning budget. The planner does not execute all available skills. Instead, it selects skills whose triggers match the target meme and whose required evidence can be supported by the retrieved memory. This keeps the reasoning process focused and avoids irrelevant agent deliberation.

Each selected skill produces a structured evidence packet rather than a final label. An evidence packet contains the local claim, textual evidence, visual evidence, target hypothesis, harmful interpretation, harmless alternative, uncertainty signal, and a short rationale. This design separates local reasoning from final decision-making. It also allows later stages to inspect which skill contributes which evidence.

3.4 Graph-assisted Judgment

The third stage aggregates evidence packets through a claim graph:

$$\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q), \quad (7)$$

where $\mathcal{V}_q$ consists of claim nodes $\{v_{\text{harm}}, v_{\text{safe}}\}$, evidence nodes $\mathcal{V}_{\text{evi}}$, skill nodes $\mathcal{V}_{\text{skill}}$, memory nodes $\mathcal{V}_{\text{mem}}$, and uncertainty nodes $\mathcal{V}_{\text{unc}}$. Edges indicate whether an evidence node supports or weakens a claim, whether it is derived from a specific skill, whether it is supported by retrieved memory, and whether it conflicts with other evidence. Related multimodal misinformation work has also shown that graph-structured reasoning can capture event-level cross-modal relations and inconsistencies (Ma et al., 2024).

The LLM Judge receives the target meme, retrieved memories, evidence packets, and the serialized claim graph. It compares the harmful and harmless claims under the evidence structure and outputs the final prediction, supporting evidence, rejected alternatives, and a concise explanation. Compared with free-form direct reasoning, this graph-assisted design makes the decision process more explicit. It also helps separate strong harmful evidence from weak associative cues.

3.5 Self-evolving Memory Update

After judgment, HARMEVOSKILL performs graph-assisted diagnosis. The Memory Curator examines the claim graph, the final prediction, and unresolved uncertainty. If the case contains a reusable pattern, it proposes an update to Case Memory or Insight Memory. If the case exposes ambiguity, hallucinated evidence, misleading retrieval, or conflict between skills, it proposes an update to Risk Memory. If a skill produces useful or noisy evidence, the curator may revise its trigger, evidence requirement, verifier, or failure modes.

The separate Memory Validator then accepts, rejects, or requests revision of the proposal:

$$(\mathcal{B}_{t+1}, \mathcal{S}_{t+1}) = \text{Validate}(\mathcal{B}_t, \mathcal{S}_t, \Delta(\mathcal{G}_q, \hat{y}_q, r_q)) \quad (8)$$

where $\Delta(\mathcal{G}_q, \hat{y}_q, r_q)$ denotes the update proposal generated from the claim graph $\mathcal{G}_q$, the prediction $\hat{y}_q$, and the rationale r_q . A proposal is committed only if it is supported by explicit textual, visual, retrieved-memory, or verified image-text relational evidence; is non-redundant; and does not contain unresolved harmful-harmless conflict. Ambiguous proposals are rejected or stored as Risk Memory rather than confident Case or Insight Memory.

This update is not unconditional self-training. The framework does not add every prediction back into memory. Instead, it uses graph diagnosis to identify reusable reasoning patterns, uncertainty sources, and skill-level failure modes. Through this process, HARMEVOSKILL forms a self-evolving loop: unlabeled memes are transformed into memory, memory induces skills, skills generate evidence, graph judgment produces diagnosis, and diagnosis further refines memory and skills.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate HARMEVOSKILL on five multimodal harmful meme benchmarks: HarM, FHM, Harm-P, MultiOFF, and PrideMM (Kiel et al., 2020; Pramanick et al., 2021; Suryawanshi et al., 2020; Shah et al., 2024). These datasets cover diverse harmfulness scenarios, including hateful implication, offensive humor, political harmfulness, protected-group targeting, and LGBTQ+-related harmful content. Detailed dataset statistics are provided in Appendix A.

Baselines. We compare HARMEVOSKILL with three categories of baselines. **1): Direct MLLM**

Prompting includes representative MLLMs under direct or CoT-style prompting, including LLaVA-NeXT-34B (Li et al., 2024), Qwen2.5-VL-32B (Bai et al., 2025), Gemini-2.0-Flash (Pichai et al., 2024), GPT-4o, GPT-4o-mini (OpenAI, 2024b,a), and GPT-5.4-mini. **2): Harmful Meme Methods** includes supervised, prompt-based, retrieval-based, and agent-style harmful meme detection methods, including PromptHate, MR.HARM, ProCap, LoReHM, Evolver, and MIND (Cao et al., 2022; Lin et al., 2023; Cao et al., 2023; Huang et al., 2024, 2025; Liu et al., 2025). **3): Skill-Memory and Self-evolving Methods** includes skill-memory or self-evolving harmful meme detection methods, including ALARM and our skill-memory variants (Lang et al., 2026). All baselines are evaluated under the same evaluation sets and metrics. For prior methods without reported results on all five datasets, we reproduce or adapt their core strategies under our unified evaluation protocol.

Metrics and Implementation. Following prior studies in harmful meme detection (Liu et al., 2025), we report Accuracy and Macro-F1. Macro-F1 is emphasized because harmful meme datasets often contain label imbalance, and both false positives on benign humor and false negatives on implicit harmful content are important. Unless otherwise specified, HARMEVOSKILL uses GPT-5.4-mini as the default MLLM backbone. During inference, the Skill Planner retrieves K relevant memory items and selects L task-specific skills for evidence generation. We set $K = 5$ and $L = 3$ in the main experiments. More details are given in Appendix B.

4.2 Main Results

Table 1 reports the main results on five benchmarks. HARMEVOSKILL achieves strong overall performance across all datasets. Compared with GPT-5.4-mini Direct, HARMEVOSKILL improves Accuracy and Macro-F1 by 9.33% and 17.97% on HarM, and by 4.40% and 5.12% on FHM, respectively. Compared with the strongest non-ours baseline, HARMEVOSKILL further improves Accuracy/Macro-F1 by 2.02%/2.43% on MultiOFF and 2.76%/2.58% on PrideMM. These results show that the gains do not simply come from a stronger backbone, but from the proposed skill-memory mechanism. By organizing reusable experience into Case, Insight, and Risk Memory, inducing executable skills from memory,

Method	FHM		HarM		Harm-P		MultiOFF		PrideMM	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
<i>Direct and CoT-style MLLM Baselines</i>										
LLaVA-NeXT-34B	64.80	64.63	66.67	62.57	87.53	86.94	72.48	71.81	75.15	74.48
Qwen2.5-VL-32B	68.70	68.49	65.54	62.93	88.59	87.55	74.50	72.53	78.30	77.74
Gemini-2.0-Flash (CoT)	72.80	72.55	72.60	70.94	88.99	88.31	75.17	73.92	79.29	78.46
GPT-4o (CoT)	72.90	72.68	74.29	73.52	89.85	88.71	77.18	76.58	80.28	79.86
GPT-4o-mini	68.50	67.92	68.93	66.47	88.20	88.14	74.50	73.88	78.90	78.42
GPT-5.4-mini (CoT)	70.20	69.84	72.03	65.58	89.39	88.55	75.17	74.69	79.49	79.43
GPT-5.4-mini	68.80	68.08	68.64	59.39	88.73	88.13	74.50	73.91	79.09	78.64
<i>Harmful Meme Detection and Agent-style Reasoning Methods</i>										
PromptHate	68.90	68.69	71.47	69.52	89.06	88.44	75.84	74.58	78.50	77.41
MR.HARM	69.50	69.01	72.03	70.62	89.66	88.76	75.17	74.13	80.67	80.01
Pro-Cap	69.10	68.52	72.32	70.77	89.46	88.92	76.51	75.68	79.88	78.79
LoReHM	70.60	70.03	73.45	72.23	90.19	89.18	77.18	76.55	81.07	80.96
Evolver	70.10	68.18	73.73	72.91	89.99	89.45	77.85	77.26	81.66	81.52
MIND	71.00	70.41	74.58	73.86	90.52	89.82	77.85	77.13	80.87	79.98
<i>Skill-Memory and Self-evolving Methods</i>										
ALARM	70.90	70.79	75.14	74.52	90.72	89.91	78.52	77.96	82.64	82.38
HARMEVOSKILL w/o Evolution	71.50	71.02	75.42	74.81	90.85	90.17	77.85	76.64	82.25	81.58
HARMEVOSKILL	73.20	73.20	77.97	77.36	91.71	90.75	80.54	80.39	85.40	84.96

Table 1: Main results on multimodal harmful meme benchmarks.

Variant	HarM		FHM	
	Acc.	Macro-F1	Acc.	Macro-F1
Full HARMEVOSKILL	77.97	77.36	73.20	73.20
w/o Case Memory	75.42	74.91	71.60	71.16
w/o Insight Memory	74.86	74.28	70.40	70.31
w/o Risk Memory	75.14	74.63	71.20	70.84
w/o Adaptive Skill Extraction	73.45	72.76	70.10	70.02
w/o Skill Planner	74.58	73.65	70.80	70.47
w/o Claim Graph	75.71	75.02	71.30	71.04
w/o Memory Update	76.27	75.68	71.80	71.42
w/o Reference Memory	70.34	68.90	69.40	69.08

Table 2: Ablation study on HarM and FHM.

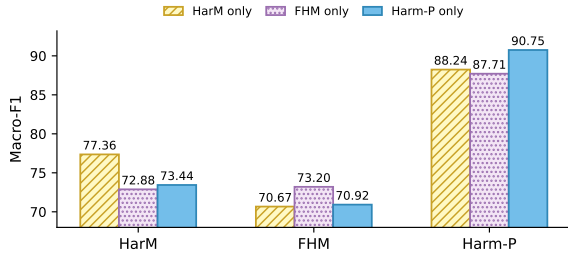


Figure 3: Memory source analysis measured by Macro-F1.

and updating them through graph-assisted diagnosis, HARMEVOSKILL improves implicit target reasoning, image-text incongruity understanding, protected-group attribution, and harmful-harmless evidence comparison.

4.3 Ablation Study

We remove one major component at a time. Table 2 shows that all modules contribute to the final performance. Removing Case, Insight, or Risk Memory weakens instance-level analogy, abstract

Backbone	HarM Acc.			FHM Acc.		
	Direct	Ours	Δ	Direct	Ours	Δ
Qwen2.5-VL-7B	63.28	68.08	+4.80	62.90	68.20	+5.30
GPT-4o-mini	68.93	74.58	+5.65	68.50	71.20	+2.70
Qwen2.5-VL-32B	65.54	71.19	+5.65	68.70	72.00	+3.30
GPT-4o	71.75	75.99	+4.24	71.60	74.10	+2.50

Table 3: Backbone generalization on HarM and FHM.

pattern reuse, or ambiguity control, respectively. Removing adaptive skill extraction, the Skill Planner, or the claim graph also causes clear drops, showing that executable skills, focused planning, and harmful-harmless evidence comparison are all important.

4.4 Generalization Analysis

Figure 3 reports memory-source transfer results measured by Macro-F1. Each group denotes a target dataset, and each bar denotes the memory source. The matched source usually performs best, while the off-source results remain competitive, suggesting that the induced skills capture reusable reasoning patterns beyond dataset-specific shortcuts. Table 3 further shows that HARMEVOSKILL consistently improves different MLLM backbones, indicating that the proposed skill-memory mechanism is not tied to a single model.

4.5 Skill-Memory Evolution Analysis

Memory Scale Analysis. Table 4 studies the effect of reference memory scale and memory quality. Performance improves as the reference set increases from 25% to 100%. This indicates that

Setting	HarM		FHM	
	Acc.	Macro-F1	Acc.	Macro-F1
w/o reference memory	70.34	68.90	69.40	69.08
25% reference set	73.73	73.08	71.20	70.74
50% reference set	75.71	75.12	71.90	71.86
100% reference set	77.97	77.36	73.20	73.20
random memory	72.88	72.13	70.10	70.02
noisy memory	74.58	73.97	71.50	71.03

Table 4: Memory scale analysis. Noisy memory injects weakly related memory items into the memory bank.

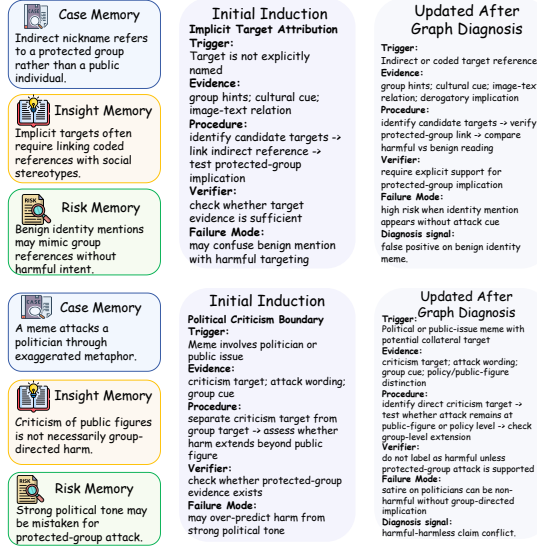


Figure 4: Skill evolution trajectory.

broader historical experience helps the framework retrieve more relevant cases and abstract more useful skills. However, random memory and noisy memory are less effective than structured memory. This shows that the improvement does not come from simply adding more context, but from organizing reference memes into Case Memory, Insight Memory, and Risk Memory.

Skill Evolution Trajectory. To verify that the skills are not manually predefined, we trace how skills are discovered and updated from memory. As shown in Figure 4, a seed skill family is first mined from recurring memory patterns. It is then induced into an executable skill card with triggering conditions, evidence requirements, reasoning procedures, verification rules, and failure modes. After graph-assisted diagnosis, the skill card is updated by narrowing unreliable triggers, adding missing evidence requirements, or recording new failure modes. This process shows that HARMEVOSKILL learns reusable skills from historical cases rather than relying on a fixed agent workflow.

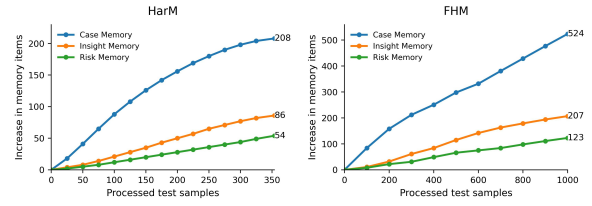


Figure 5: Memory growth across self-evolution iterations on HarM and FHM.



Figure 6: Case studies of HARMEVOSKILL.

Memory Growth Analysis. Figure 5 compares the growth of Case Memory, Insight Memory, and Risk Memory on HarM and FHM. Case Memory grows fastest because it stores instance-level observations, while Insight Memory increases more slowly because it requires abstraction across related cases. Risk Memory grows when the framework encounters ambiguous targets, weak context, misleading retrieval, or harmful-harmless conflicts. The distinct growth patterns across the two datasets indicate that the three memory types play complementary roles in the self-evolving process.

4.6 Case Study

Figure 6 presents three representative cases: implicit harmfulness, political or activist content, and hard negative humor. Direct prompting tends to rely on surface text or negative tone, while HARMEVOSKILL retrieves relevant memory, activates task-specific skills, and compares harmful and harmless interpretations before prediction. These cases show that memory-guided skill planning helps reduce both missed implicit harm and false positives on benign memes.

5 Conclusion

We propose HARMEVOSKILL, a self-evolving Skill-Memory framework for label-free multimodal harmful meme detection. HARMEVOSKILL constructs Case, Insight, and Risk Memory from unlabeled reference memes, automatically induces executable skills, and updates memory through graph-assisted diagnosis. Experiments on five benchmarks show that HARMEVOSKILL consistently outperforms existing baselines. Further analyses demonstrate the contributions of multi-level memory, adaptive skill induction, and harmful-harmless claim comparison. These results suggest that persistent skill-memory evolution is a promising direction for multimodal safety reasoning.

Limitations

HARMEVOSKILL has several limitations. First, it introduces additional inference cost because memory retrieval, skill planning, evidence generation, claim-graph judgment, and memory update are performed before the final prediction. This improves reliability but makes the framework slower than direct MLLM prompting. Second, the memory update process is label-free and may introduce noise when graph diagnosis is based on incomplete or ambiguous evidence. Third, some harmful memes require external socio-cultural or historical knowledge that is not always covered by the reference memory. Finally, this work focuses on image-text memes and does not yet address video memes, audio cues, or multi-turn social media contexts.

Ethics Statement

This work studies multimodal harmful meme detection for online safety. The datasets used in our experiments are existing research benchmarks, and we do not collect private user data or attempt to identify individuals. Some examples may contain offensive, hateful, or harmful content because such content is necessary for evaluating harmful meme detection systems. We use these examples only for research purposes and avoid amplifying harmful messages beyond the minimum needed for analysis.

A potential risk of this work is misuse in automated content moderation. Although HARMEVOSKILL improves evidence-grounded harmfulness detection, it should not be used as a fully automated moderation system without human

oversight. False positives may suppress benign humor, political criticism, or activist speech, while false negatives may fail to detect implicit harmful content. We therefore recommend using the framework as a decision-support tool with human review, especially in culturally sensitive or politically contested contexts.

Acknowledgments

We use ChatGPT for language polishing, LaTeX formatting, and code assistance. All technical content, experimental results, and references are reviewed and verified by the authors.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. [ProCap: Leveraging a frozen vision-language model for hateful meme detection](#). In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252. Association for Computing Machinery.
- Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. [Prompting for multimodal hateful meme classification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jianzhao Huang, Hongzhan Lin, Ziyang Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. [Towards low-resource harmful meme detection with LMM agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.
- Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2025. [Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7321–7330, Abu Dhabi, UAE. Association for Computational Linguistics.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

- Mohamed Bayan Kmainasi, Mucahid Kutlu, Ali Ezzat Shahroor, Abul Hasnat, and Firoj Alam. 2026. [Can thinking models think to detect hateful memes?](#) In *Companion Proceedings of the ACM Web Conference 2026*, pages 935–944. Association for Computing Machinery.
- Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. 2026. [From shallow humor to metaphor: Towards label-free harmful meme detection via LMM agent self-improvement](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 588–599. Association for Computing Machinery.
- Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. [LLaVA-NeXT: Stronger llms supercharge multimodal capabilities in the wild](#). Blog post.
- Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. [Towards explainable harmful meme detection through multimodal debate between large language models](#). In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370, New York, NY, USA. Association for Computing Machinery.
- Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. [Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.
- Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. [MIND: A multi-agent framework for zero-shot harmful meme detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria. Association for Computational Linguistics.
- Weihai Lu, Yu Tong, and Zhiqiu Ye. 2025. [DAMMFND: Domain-aware multimodal multi-view fake news detection](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 559–567.
- Weihai Lu, Zhejun Zhao, Yanshu Li, and Huan He. 2026. [MM-StanceDet: Retrieval-augmented multimodal multi-agent stance detection](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31449–31464, San Diego, California, United States. Association for Computational Linguistics.
- Yihan Ma, Xinyue Shen, Yiting Qu, Ning Yu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. [From meme to threat: On the hateful meme understanding and induced hateful content generation in Open-Source vision language models](#). In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4703–4722, Seattle, WA. USENIX Association.
- Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. [Event-radar: Event-driven multi-view learning for multimodal fake news detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, Bangkok, Thailand. Association for Computational Linguistics.
- Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. [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*, pages 23806–23828, Suzhou, China. Association for Computational Linguistics.
- OpenAI. 2024a. [GPT-4o mini: Advancing cost-efficient intelligence](#). OpenAI Blog.
- OpenAI. 2024b. [GPT-4o system card](#). OpenAI.
- Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. [Introducing Gemini 2.0: Our new AI model for the agentic era](#). Google Blog.
- Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. [MOMENTA: A multimodal framework for detecting harmful memes and their targets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. [MemeCLIP: Leveraging CLIP representations for multimodal meme classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Michael Arcan, and Paul Buitelaar. 2020. [Multimodal meme dataset \(MultiOFF\) for identifying offensive content in image and text](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association.
- Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. [DAPT: Domain-aware prompt-tuning for multimodal fake news detection](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911. Association for Computing Machinery.
- Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. [MMDFND: Multi-modal multi-domain fake news detection](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186. Association for Computing Machinery.
- Mo Wang, Kaixuan Ren, Pratik Jalan, Ahmed Ashraf, Tuong Vy Vu, Rahul Seetharaman, Shah Nawaz, and Usman Naseem. 2026a. [From native memes](#)

- to global moderation: Cross-cultural evaluation of vision–language models for hateful meme detection. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pages 9788–9799, New York, NY, USA. Association for Computing Machinery.
- Xiuxian Wang, Yuting Su, Wenhui Li, Xiaowen Wang, Zhuojun Li, and Anan Liu. 2026b. *SGoT-R1: Social graph of thought reasoning-enhanced multimodal large language model for harmful meme detection*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26597–26605.
- Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. 2025. *Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory*. *arXiv preprint arXiv:2511.20857*.
- Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and Botian Shi. 2025. *EvolveR: Self-evolving llm agents through an experience-driven lifecycle*. *arXiv preprint arXiv:2510.16079*.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. *A-Mem: Agentic memory for LLM agents*. In *Advances in Neural Information Processing Systems*, volume 38, pages 20004–20031.
- Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. *IMOL: Incomplete-modality-tolerant learning for multi-domain fake news video detection*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933, Vienna, Austria. Association for Computational Linguistics.
- Zhi Zeng, Jiaying Wu, Minnan Luo, Di Zhang, Yifei Yang, Xiangzheng Kong, Herun Wan, and Zihan Ma. 2026. *From detection to understanding: Multi-turn reasoning for video misinformation analysis*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37009–37027, San Diego, California, United States. Association for Computational Linguistics.
- Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. 2025. *G-Memory: Tracing hierarchical memory for multi-agent systems*. In *Advances in Neural Information Processing Systems*, volume 38, pages 14587–14617.
- Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026. *Memskill: Learning and evolving memory skills for self-evolving agents*. *arXiv preprint arXiv:2602.02474*.
- Huichi Zhou, Siyuan Guo, Anjie Liu, Zhongwei Yu, Ziqin Gong, Bowen Zhao, Zhixun Chen, Menglong Zhang, Yihang Chen, Jinsong Li, Runyu Yang, Qiangbin Liu, Xinlei Yu, Jianmin Zhou, Na Wang, Chunyang Sun, and Jun Wang. 2026. *Memento-skills: Let agents design agents*. *arXiv preprint arXiv:2603.18743*.

A Dataset Details

We evaluate HARMEVOSKILL on five multimodal harmful meme benchmarks: HarM, FHM, Harm-P, MultiOFF, and PrideMM. These datasets cover different forms of multimodal harm, including hateful implication, political harmfulness, offensive humor, and LGBTQ+-related harmful content. Dataset statistics are reported in Table 5.

HarM contains multimodal memes annotated with harmfulness labels. Following our binary classification setting, we merge the original *harmful* and *somewhat harmful* labels into the harmful class, and treat the remaining samples as harmless. FHM is a widely used hateful meme benchmark that requires joint reasoning over image and text; we use its original binary hateful/non-hateful labels. Harm-P is the politics-related subset of HarMeme used in MOMENTA, focusing on politically harmful memes (Pramanick et al., 2021). Following the same binary setting, we merge harmful and somewhat harmful samples into the harmful class and treat the remaining samples as harmless. For Harm-P, we use a 1,508-sample evaluation set under our unified evaluation protocol rather than the original 355-sample MOMENTA test split; all compared methods are evaluated on this same set. MultiOFF contains offensive memes with both visual and textual cues, where offensive samples are mapped to harmful and non-offensive samples to harmless. PrideMM consists of LGBTQ+-related multimodal memes and is used to evaluate protected-group-related harmfulness detection.

Dataset	Domain	Samples
HarM	Harmful memes	354
FHM	Hateful memes	1,000
Harm-P	Political harmful memes	1,508
MultiOFF	Offensive memes	149
PrideMM	LGBTQ+-related memes	507

Table 5: Dataset statistics used in our experiments.

B Implementation and Reproduction Details

Unified Evaluation Protocol. All methods are evaluated on the same evaluation sets and reported with Accuracy and Macro-F1. Accuracy measures overall classification correctness, while Macro-F1 gives equal weight to the harmful and harmless classes. We report Macro-F1 because harmful meme datasets may contain label imbalance, and performance on the minority class is important for safety-sensitive evaluation.

Hyper-parameters. The main experiments use $K = 5$ retrieved memory items and $L = 3$ selected skills. Additional sensitivity analysis for K and L is reported in Appendix E. Unless otherwise specified, all experiments use the same binary label mapping and the same evaluation script.

Artifact Licenses and Access. We use public research datasets and released models or model APIs according to their original access conditions. Raw images are not redistributed when redistribution is restricted. Our released package contains code, prompts, scripts, and metadata for reproduction, while users should obtain restricted datasets from the original providers.

C Additional Discussion of Core Modules

C.1 Discussion of Claim Graph

The claim graph is designed to reduce unstructured free-form judgment in multimodal harmful meme detection. Direct MLLM prompting often merges visual evidence, textual cues, target hypotheses, and social context into a single explanation. This may cause the model to over-rely on surface humor, political tone, or isolated negative words. In contrast, our claim graph separates harmful and harmless interpretations into competing claim nodes, and links each local evidence packet to the claim it supports or weakens.

The same structure is used to control label-free memory update. The update is not triggered by the final prediction alone. The Memory Curator validates each update through three checks. First, the proposed memory item or skill revision must be supported by explicit textual evidence, visual evidence, retrieved memory, or a verified image-text relation. Second, if harmful and harmless claims are both strongly supported, the case is not added as confident Case or Insight Memory, but is rejected

or stored as Risk Memory. Third, redundant updates, sentiment-only updates, or updates without a clear target and harm mechanism are rejected. These checks prevent the framework from simply adding every prediction back into memory.

Different diagnosis outcomes lead to different update actions. Stable case-level evidence updates Case Memory, recurring patterns update Insight Memory, and ambiguous targets, weak context, misleading retrieval, or harmful-harmless conflicts update Risk Memory. Skill cards are revised only when graph diagnosis identifies a specific failure mode, such as an over-broad trigger, missing evidence requirement, or unreliable verifier. Thus, the label-free update mechanism is evidence-based memory curation rather than unconditional self-training.

This design makes the decision and update process more inspectable and helps control false positives. For example, a meme may contain strong criticism or activist language, but the harmless claim can still be supported if the target is a public institution rather than a protected group. The claim graph therefore provides a structured space for resolving harmful-harmless evidence conflict and filtering noisy memory updates.

C.2 Discussion of Adaptive Skill Extraction

Adaptive skill extraction is introduced to avoid manually predefined skill inventories. Harmful memes evolve quickly, and fixed skills may not cover new meme templates, coded references, or emerging social events. Instead of hand-designing all reasoning skills, HARMEVOSKILL mines case-aware skill proposals from Case Memory, Insight Memory, and Risk Memory. Similar proposals are clustered and abstracted into seed skill families, which are then induced into executable skill cards.

Each skill card contains triggering conditions, required evidence, reasoning procedure, verification rule, and failure modes. This format makes the skill executable during inference and also allows later updates. When graph diagnosis reveals that a skill produces unsupported or noisy evidence, the skill can be revised by narrowing its trigger, adding missing evidence requirements, or recording a new failure mode. This process distinguishes adaptive skill extraction from fixed agent roles or static prompt templates.

C.3 Discussion of Memory-guided Skill Planner

The Skill Planner selects a small number of task-specific skills according to the target meme and retrieved memory. This is necessary because harmful meme detection does not require the same reasoning path for every sample. Some memes mainly require target disambiguation, while others require political criticism boundary checking, image-text incongruity reasoning, cultural-template interpretation, or benign-humor filtering.

The planner uses retrieved Case Memory and Insight Memory to identify potentially useful reasoning directions, and uses Risk Memory to avoid known failure modes. It then selects skills whose triggers match the target meme and whose evidence requirements can be supported by retrieved memory. This reduces irrelevant agent deliberation and keeps inference focused on the most plausible reasoning dimensions.

D Prompt Templates

This section provides representative prompt templates for the main agents in HARMEVOSKILL. All agents are required to return structured outputs and support their judgments with textual, visual, or memory-based evidence.

Prompt for Skill Discovery Agent

Role. You are a skill discovery agent for multimodal harmful meme detection.

Input. Image description: {image_description}

OCR text: {ocr_text}

Retrieved context: {context}

Preliminary interpretation: {preliminary_interpretation}

Task. Analyze the meme and identify reusable reasoning patterns that may help future harmful meme detection. Do not assign a final label. Focus on visual-textual relations, implicit targets, protected-group cues, rhetorical strategies, cultural references, and possible false-positive or false-negative risks.

Output format. Return a JSON object with: case_pattern, possible_skill_family, trigger_cues, required_evidence, risk_notes, and short_rationale.

Prompt for Skill Inducer Agent

Role. You are a skill inducer agent.

Input. Candidate skill family: {skill_family}

Supporting Case Memory: {case_memory}

Supporting Insight Memory: {insight_memory}

Supporting Risk Memory: {risk_memory}

Task. Induce an executable skill card from recurring memory patterns. The skill must be supported by the provided memory rather than manually assumed. Each skill should specify when it should be used, what evidence it requires, how it reasons, how its local conclusion is verified, and when it may fail.

Output format. Return a JSON object with: skill_name, trigger, required_evidence, procedure, verifier, and failure_modes.

Prompt for Memory-guided Skill Planner

Role. You are a memory-guided skill planner for multimodal harmful meme detection.

Input. Target image description: {image_description}

OCR text: {ocr_text}

Retrieved Case Memory: {case_memory}

Retrieved Insight Memory: {insight_memory}

Retrieved Risk Memory: {risk_memory}

Available skill cards: {skill_cards}

Task. Select a small set of task-specific skills for the target meme. Do not select all available skills. Choose only skills whose triggers match the meme and whose required evidence can be checked. For each selected skill, generate a structured evidence packet instead of a final prediction.

Output format. Return a JSON object with selected_skills. For each selected skill, include: skill_name, why_selected, local_claim, textual_evidence, visual_evidence, target_hypothesis, harmful_interpretation, harmless_alternative, uncertainty, and short_rationale.

Prompt for Graph-assisted Judge

Role. You are a graph-assisted judge for multimodal harmful meme detection.

Input. Target meme: {target_meme}

Structured evidence packets: {evi-

dence_packets}
Retrieved risk notes: {risk_notes}
Task. Build two competing interpretations: a harmful claim and a harmless claim. For each claim, list supporting and weakening evidence. Compare the two claims and make the final prediction. The decision must be based on evidence rather than surface offensiveness alone.
Output format. Return a JSON object with: harmful_claim, harmful_supporting_evidence, harmful_weakening_evidence, harmless_claim, harmless_supporting_evidence, harmless_weakening_evidence, conflict_summary, final_prediction, confidence, and decision_rationale.

Prompt for Memory Curator Agent

Role. You are a memory curator agent for a self-evolving harmful meme detection system.
Input. Target meme: {target_meme}
Final prediction: {prediction}
Claim graph diagnosis: {graph_diagnosis}
Selected skills: {selected_skills}
Existing memory summary: {existing_memory}
Task. Generate update proposals for Case Memory, Insight Memory, Risk Memory, and skill cards. Do not use gold labels. Update memory only when the claim graph provides sufficient evidence. Reject weak, redundant, or unsupported updates.
Output format. Return a JSON object with: case_memory_update, insight_memory_update, risk_memory_update, skill_update, update_type, update_content, and validator_note.

E Additional Experiments

E.1 Additional Ablation Results

Table 6 reports additional ablation results on Harm-P, MultiOFF, and PrideMM. The results are consistent with the main ablation study: adaptive skill extraction, multi-level memory, skill planning, and claim-graph judgment all contribute to performance. The larger drops on MultiOFF and PrideMM further show the importance of Risk

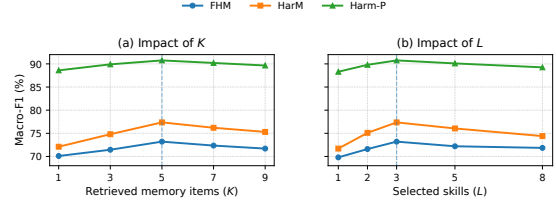


Figure 7: Hyper-parameter analysis of memory retrieval budget K and skill planning budget L .

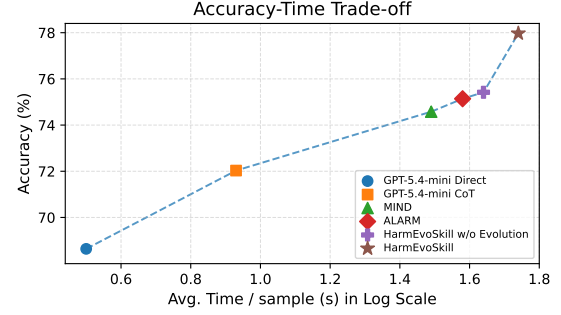


Figure 8: Accuracy-time trade-off on HarM. HARMEVOSKILL improves performance at the cost of longer inference time.

Memory and harmful-harmless claim comparison in hard-negative and protected-group-related cases.

E.2 Additional Backbone Results

Table 7 reports additional backbone results on Harm-P, MultiOFF, and PrideMM. HARMEVOSKILL improves all tested backbones, but the gain varies across models and datasets. This suggests that the proposed skill-memory mechanism is complementary to backbone capacity rather than being a prompt tuned for one specific MLLM.

E.3 Hyper-parameter Analysis

We analyze the sensitivity of the memory retrieval budget K and the skill planning budget L . Figure 7 shows that performance first improves as more relevant memory items and skills are used. However, excessive retrieval or too many selected skills may introduce irrelevant memory and redundant reasoning paths. We therefore set $K = 5$ and $L = 3$ in the main experiments.

E.4 Efficiency and Cost Analysis

Figure 8 visualizes the accuracy-time trade-off on HarM. HARMEVOSKILL is slower than direct MLLM prompting because it performs memory retrieval, skill planning, evidence generation, claim-graph judgment, and memory update. However,

Variant	Harm-P		MultiOFF		PrideMM	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Full HARMEVOSKILL	91.71	90.75	80.54	80.39	85.40	84.96
w/o Case Memory	90.32	89.30	77.85	76.98	83.43	82.91
w/o Insight Memory	89.79	88.72	76.51	75.44	82.45	82.36
w/o Risk Memory	90.05	89.17	75.84	74.65	84.02	83.52
w/o Adaptive Skill Extraction	88.99	87.86	75.17	74.20	82.05	81.79
w/o Skill Planner	89.52	88.55	77.18	76.11	83.04	82.58
w/o Claim Graph	90.25	89.48	77.85	77.02	83.63	83.14
w/o Memory Update	90.72	89.91	79.19	78.46	84.22	83.82

Table 6: Additional ablation results on Harm-P, MultiOFF, and PrideMM.

Backbone	Harm-P Acc.			MultiOFF Acc.			PrideMM Acc.		
	Direct	Ours	Δ	Direct	Ours	Δ	Direct	Ours	Δ
Qwen2.5-VL-7B	87.86	90.58	+2.72	71.81	77.85	+6.04	77.32	83.63	+6.31
GPT-5.4-mini	88.73	91.71	+2.98	74.50	80.54	+6.04	79.09	85.40	+6.31
Qwen2.5-VL-32B	88.59	91.18	+2.59	74.50	79.19	+4.69	78.30	84.22	+5.92
GPT-4o (CoT)	89.85	91.98	+2.13	77.18	81.21	+4.03	80.28	85.80	+5.52

Table 7: Additional backbone results on Harm-P, MultiOFF, and PrideMM.

the additional computation brings substantial performance gains.

F Additional Skill-Memory Evolution Examples

F.1 Additional Skill Extraction Examples

Table 8 shows additional examples of automatically extracted skills. Each skill is induced from recurring memory patterns rather than manually predefined categories.

F.2 Additional Before-after Skill Cards

Table 9 provides another before-after skill update. The update is generated after graph diagnosis identifies that the initial skill over-relies on surface visual incongruity.

G Error Analysis and Future Work

Figure 9 summarizes the remaining error types of HARMEVOSKILL. Although HARMEVOSKILL reduces many errors caused by implicit targets, image-text incongruity, and benign hard negatives, several challenging cases remain.

First, cultural-template failures occur when a meme depends on platform-specific or community-specific background knowledge. Second, historical harm trivialization remains difficult because it often requires external knowledge about historical events

and their social meaning. Third, misogynistic framing may be expressed through subtle objectification or coded language, making the harmful implication hard to verify from image-text evidence alone. Fourth, target ambiguity remains difficult when a meme simultaneously references public figures, institutions, and demographic groups.

These errors suggest several directions for future work. Future versions of HARMEVOSKILL can incorporate stronger external context verification, culturally grounded memory construction, and more fine-grained target taxonomies. Another important direction is to develop stronger validators for label-free memory update, so that noisy or weakly supported memory items are less likely to affect later reasoning.

Induced Skill	Memory Evidence	Trigger	Failure Mode
Implicit Target Attribution	Repeated cases where the target is expressed through coded references or indirect group cues.	The target is not explicitly named but can be inferred from image-text association.	May confuse benign identity mention with group-directed attack.
Political Criticism Boundary	Cases where strong political tone does not necessarily imply protected-group harm.	A meme involves a politician, public institution, or policy issue.	May over-predict harm from negative political sentiment.
Benign Humor Filtering	Risk Memory records humorous or absurd memes that were falsely predicted as harmful.	The meme contains unusual imagery, informal language, or exaggerated humor.	May under-detect implicit harm when humor masks a protected-group attack.
Historical Harm Trivialization	Insight Memory captures cases where historical trauma is minimized or turned into ridicule.	The meme references war, disaster, genocide, slavery, or historical oppression.	Requires external context and may miss subtle trivialization.

Table 8: Additional examples of automatically induced skills.

Field	Before Update	After Update
Skill	Image-text Incongruity Reasoning	Image-text Incongruity Reasoning
Trigger	Image and text appear semantically mismatched.	Image and text mismatch creates a possible harmful implication.
Evidence Requirement	Visual object and caption mismatch.	Visual object, caption, target candidate, and harm mechanism.
Procedure	Detect whether the image contradicts the text.	Determine whether the mismatch changes the target or creates group-directed harm.
Verifier	Check if there is a contradiction.	Check whether the contradiction supports harmfulness rather than general humor.
Failure Mode	May over-predict harmfulness for absurd jokes.	Treat absurd humor as risky only when target and harm evidence are present.

Table 9: Additional before-after skill card update.

(a) Case 1: Cultural-template Failure

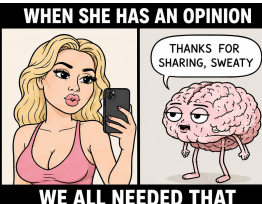


HarmEvoSkill

I recognize that the meme uses an unusual community-specific template and emotionally charged wording. The image-text pairing appears intentionally mocking, and the phrase "NPC meltdown" may imply hostility toward the depicted target. Since the template meaning is unclear from the available context, I treat the post as harmful.

Input Meme

(b) Case 2: Misogynistic Framing



HarmEvoSkill

I see a humorous or exaggerated depiction involving gender. The post appears to make a joke about behavior, appearance, and social interaction. Since the attack is **not stated directly** and the harmful implication is expressed through stereotype-based framing, I treat it as non-harmful humor.

Input Meme

(c) Case 3: Historical Harm Trivialization



HarmEvoSkill

I notice that the meme refers to a well-known historical event and presents it with a joking tone. The image combines a disaster reference with entertainment framing, but the post does not contain an **explicit target** or direct attack. I therefore interpret it mainly as dark humor rather than harmful content.

Input Meme

(d) Case 4: Target Ambiguity



HarmEvoSkill

I detect criticism and negative language, but the target can be read as a politician, a policy issue, or a broader social group. The strongest wording appears to criticize "you people," so I interpret the meme as **group-directed hostility** rather than public-figure or policy criticism.

Input Meme

Figure 9: Additional error analysis of HARMEVOSKILL.