I Know What You Meme, Even If it Emerged Today: Understanding Evolving Memes through Open-World Knowledge Acquisition

Shanhong Liu[1], **Rui Cao**[1], **Pai Chet Ng**[2], **De Wen Soh**[1]
[1]Singapore University of Technology and Design
[2]Singapore Institute of Technology
shanhong_liu@mymail.sutd.edu.sg, {rui_cao,dewen_soh}@sutd.edu.sg,
paichet.ng@singaporetech.edu.sg

Abstract

Multimodal memes are dynamic and often require up-to-date background knowledge for interpretation. Existing methods often overlook such knowledge or rely on fixed parametric knowledge of pre-trained models that may be incomplete, outdated, or unavailable for emerging memes. We introduce *Query-Retrieve-Conclude*, a zero-shot framework that identifies missing knowledge, retrieves open-web evidence, and synthesizes evidence-grounded background knowledge for meme understanding and detection. We also introduce a curated meme understanding benchmark of recent memes from 2024–2026 with external background knowledge annotations. Experiments on three meme understanding datasets and five meme detection tasks show that our framework improves knowledge recovery, meme understanding and downstream detection over zero-shot baselines.

Disclaimer: *This paper contains discriminatory content that may be disturbing to some readers.*

1 Introduction

Internet multimodal memes are inherently dynamic and temporally evolving forms of online communication.[1] Their meanings often depend on recent news, viral templates, emerging public controversies, cultural references, or newly formed associations between public figures and events (Valensise et al., 2021). For example, the meme in Figure 1 requires more than recognizing the two depicted figures Donald Trump and Confucius. The phrase "CON-FUSE-US" evokes both "Confucius" and the English word"confuse us", creating a layered pun grounded in linguistic and cultural knowledge.

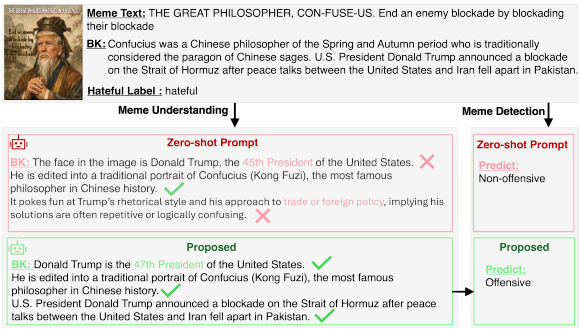


Figure 1: An example[2] illustrating the dynamic world knowledge gap in meme understanding. Direct zero-shot prompting misses the relevant context, whereas our framework recovers evidence-grounded background knowledge that supports correct downstream detection.

However, the point of the pun can only be understood in relation to time-evolving real-world events, including Iran's blockage of the Strait of Hormuz in February 2026 and the subsequent U.S. naval blockade after the failed Islamabad Talks.

Despite the inherently dynamic nature of memes, existing multimodal meme understanding methods remain limited in their use of background knowledge. Most methods built on vision-language models (VLMs) focus mainly on vision-language interactions (e.g., aligning regions with embedded text or capturing cross-modal incongruity) (Sharma et al., 2023; Nguyen et al., 2025), treating memes as closed multimodal reasoning problems. Moreover, they heavily rely on human annotated supervised data and the implicit parametric knowledge of VLMs (Hwang and Shwartz, 2023; Park et al., 2025, 2024; Nguyen et al., 2025), making them less adaptable when required knowledge is missing, outdated, or newly emerging. Although one prior work (Tripathi et al., 2026) incorporates external knowledge bases (e.g., ConceptNet[3] and Hatebase[4]), it relies on static symbolic knowledge, lim-

[1]https://en.wikipedia.org/wiki/Memetics
[2]Source of the meme used in this figure: https://knowyourmeme.com/
[3]https://conceptnet.io/
[4]https://hatebase.org/

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 30853–30868
October 24-29, 2026 ©2026 Association for Computational Linguistics

iting both updatability and human interpretability.

Meme detection builds on meme understanding, since reliable detection requires first interpreting the meaning and intention of the meme (Cao et al., 2022, 2023; Lin et al., 2024b, 2025; Liu et al., 2025; Rizwan et al., 2026; Liu et al., 2026a,b). However, existing detection methods share similar limitations: they mainly model multimodal interactions using supervised data (Lee et al., 2021; Agarwal et al., 2024), and rely on parametric knowledge of VLMs (Pan et al., 2025; Mei et al., 2025) or large language models (LLMs) (Lin et al., 2024a; Cao et al., 2022; Lin et al., 2025). These limitations make them less robust to emerging memes that require external commonsense, cultural, or time-sensitive background knowledge.

Our Approach. This motivates our zero-shot framework, *Query-Retrieve-Conclude*[5], which does not require task-specific supervised data. Rather than relying solely on VLMs for meme understanding or detection, it retrieves open-web evidence to acquire the essential and up-to-date world knowledge needed to address the dynamic nature of memes. The framework consists of three stages. The *Query* stage identifies missing knowledge by using reverse image search, caption generation, and question generation to produce search-oriented questions. The *Retrieval* stage retrieves external evidence from the open-web for each question and generates evidence-grounded answers. The *Conclude* stage synthesizes the resulting question-answer pairs into explicit background knowledge statements, following evidence-statement formulations in recent retrieval evaluation work (Cao et al., 2025; Akhtar et al., 2026), and uses these statements as contextual evidence for downstream task.

Beyond meme understanding, we further apply our framework to meme detection to further validate the utility and quality of generated background knowledge. The same pipeline produces evidence-grounded knowledge statements that help reveal the meme's intended meaning, which are then used as contextual evidence for detection. This allows detection decisions to be grounded not only in visual-textual interactions, but also in the external world knowledge needed to interpret dynamic events, cultural references, and evolving online discourse.

To validate our proposed method, we benchmark it on both meme understanding and meme detection datasets. For meme understanding, in addition to existing benchmarks, we introduce *KYM*, a highly curated diagnostic benchmark of recent memes from 2024 to 2026 that capture volatile internet phenomena. KYM further provides external background knowledge annotations, enabling evaluation under strict temporal distribution shifts and testing whether models can interpret memes grounded in rapidly evolving real-world events and online discourse. Experiments on **three meme understanding datasets** and **five meme detection datasets** demonstrate the effectiveness of our proposed method.

In summary, our contributions are as follows:

- We identify the dynamic world-knowledge gap in existing supervised and zero-shot meme understanding methods.
- We propose *Query-Retrieve-Conclude*, a zero-shot framework that retrieves open-web evidence and synthesizes grounded background knowledge for meme understanding and detection.
- We conduct extensive experiments on three meme understanding datasets, including *KYM*, a curated benchmark for recent memes tied to evolving real-world events, and five meme detection datasets.

2 Related Work

Knowledge integration has proven to be effective in NLP tasks such as factual reasoning (Petroni et al., 2019; Cao et al., 2025; Akhtar et al., 2026). For multimodal meme understanding, recent studies have shown that external knowledge can enhance meme understanding and detection. For example, PromptHate (Cao et al., 2022) and Pro-Cap (Cao et al., 2023) utilize pretrained vision-language models to obtain contextual and cultural information for hateful meme analysis. More recent automated meme understanding studies, such as MemeIntent (Park et al., 2024) and MemeQA (Nguyen et al., 2025), collect background knowledge for existing datasets and fine-tune large multimodal models to support more holistic meme understanding. Additionally, (Tripathi et al., 2026) provides cultural grounding by integrating knowledge from existing meme datasets. Although these studies have enhanced understanding and detection, respectively, they remain limited when applied to

[5]Code: https://github.com/shanhongliuu/Query-Retrieve-Conclude

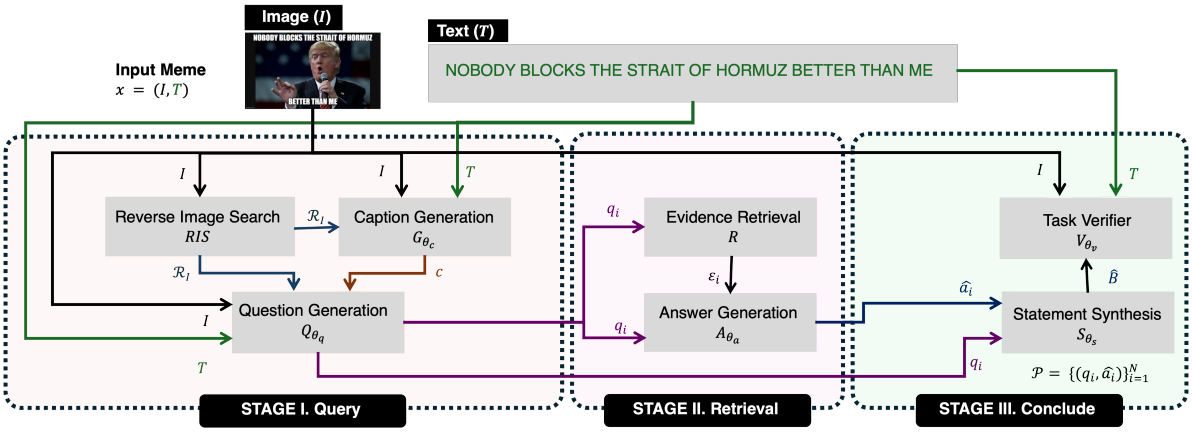


Figure 2: Overview of the proposed query-driven, retrieval-grounded framework for automated meme understanding and detection. Given a meme image I and text T, our proposed framework is organized into three stages. Stage I (*Query*) identifies missing knowledge through reverse image search, caption generation, and question generation. Stage II (*Retrieval*) retrieves external evidence for each generated question and produces evidence-grounded answers. Stage III (*Conclude*) synthesizes the question–answer pairs into explicit background knowledge statements $\hat{B}$ and uses them as contextual evidence for downstream meme detection.

recent or emerging memes whose meanings depend on timely events and evolving internet culture. More importantly, they do not explicitly model *what knowledge is missing* before attempting to explain or classify a meme. As a result, background knowledge is often treated as either an annotation, a retrieved resource, or a direct model output, rather than as the result of an explicit knowledge-gap identification and evidence acquisition process.

3 Methodology

We propose a query-driven, retrieval-grounded, and conclusion-oriented framework for meme understanding. This design is motivated by how humans interpret unfamiliar memes. When encountering a meme whose meaning is not immediately clear, a reader may first identify unfamiliar entities, phrases, or visual references; then search for relevant information; and finally combine the retrieved knowledge with the meme content to infer its intended meaning or social implication. As shown in Figure 2, the framework consists of three structured stages: *Query*, *Retrieve*, and *Conclude*. Rather than directly prompting a large VLM to explain or classify a meme, the novelty of our framework lies in treating background knowledge acquisition as a structured reasoning process: query–retrieve–conclude process. Rather than as a single-step generation problem or a direct classification prompt, we first identifying knowledge gaps[6], then retrieving external evidence, and finally synthesizing the retrieved information into task-relevant background knowledge for detection.

3.1 Stage I: Query

The first stage identifies the missing knowledge required to interpret the meme. To reduce the risk of prematurely fabricating background knowledge from parametric memory, this stage asks the model to formulate what needs to be known before the meme can be understood, thereby encouraging evidence-seeking behavior. Given the meme image I and text T, the goal is to generate a compact set of search-oriented questions $Q = \{q_1, \ldots, q_N\}$ that guide evidence retrieval. Unlike direct explanation-based prompting, this stage does not ask the model to immediately infer the meme's meaning. We first apply reverse image search (RIS) to the meme image to obtain visually associated web context, denoted as $\mathcal{R}_I = \mathrm{RIS}(I)$. Next, we generate a visually grounded caption c from the image, text, and reverse-image-search context. The caption generator uses $\mathcal{R}_I$ only as auxiliary visual grounding, helping resolve ambiguous entities, symbols, or scenes while preserving a literal description of the meme. Finally, the question generator produces a set of search-oriented questions,

$$Q = Q_{\theta_q}(I, T, c, \mathcal{R}_I). \tag{1}$$

[6]Following (Park et al., 2025), we consider missing knowledge to be the minimum amount of knowledge that is missing from the meme. It is the knowledge that needs to be combined with visual and textual cues from the meme in order to munderstand its meaning.

Each question targets a potential knowledge gap, such as the identity of a depicted person, the meaning of a phrase, the background of a public event, or the relationship between the image and the text. The output of this stage is therefore not a final interpretation, but a structured query plan for external knowledge acquisition.

3.2 Stage II: Retrieve

This retrieve stage is central to the open-world nature of our framework.[7] Because memes often refer to recent events, evolving cultural references, or platform-specific discourse, relying only on pretrained parametric knowledge can be insufficient (Valensise et al., 2021). Recognizing this gap, the second stage is designed to retrieve external evidence at inference time, such that the framework can adapt to emerging memes whose meanings may not be fully encoded in the model parameters. For each question $q_i \in Q$, the evidence retrieval module searches the web and returns a set of relevant passages, i.e., $\epsilon_i = R(q_i)$, where $R(\cdot)$ denotes the web search and evidence retrieval module. The full retrieved evidence set is $\epsilon = \{\epsilon_i\}_{i=1}^{N}$. Given each question q_i, its retrieved evidence ϵ_i, and the generated caption c, the answer generator produces a concise evidence-grounded answer:

$$\hat{a}_i = A_{\theta_a}(c, q_i, \epsilon_i). \tag{2}$$

The caption provides visual context, while the retrieved passages provide external textual grounding. The answer generator is constrained to use only the retrieved evidence, ensuring that the resulting answer is grounded in externally accessible information rather than unsupported model assumptions.

3.3 Stage III: Conclude

The conclusion stage serves two purposes: (i) to produce interpretable background knowledge statements that can be directly evaluated for meme understanding; and (ii) to use these statements as task-level evidence for downstream detection. This dual role allows the framework to connect meme understanding and meme detection within a single pipeline. The output from the previous two stages create a set of question-answer pairs, $\mathcal{P} = \{(q_i, \hat{a}_i)\}_{i=1}^{N}$, which are first transformed into declarative background knowledge statements:

$$\hat{B} = S_{\theta_s}(\mathcal{P}), \tag{3}$$

[7]The retrieval details are shown in Appendix C.3.

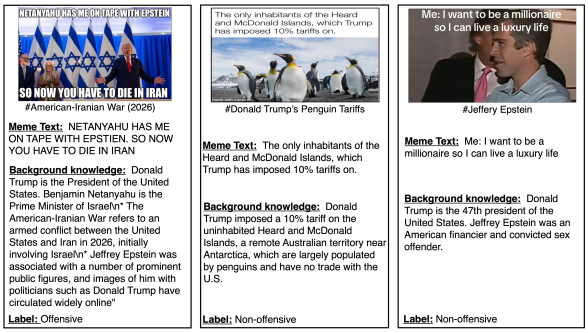


Figure 3: Example memes from the curated KYM dataset, collected from Know Your Meme, series on #2026 American-Iranian War; #Donald Trump's Penguin Tariffs; #Jeffery Epstein (from left to right).

where S_{θ_s} denotes the statement synthesis module and $\hat{B} = \{\hat{b}_1, \ldots, \hat{b}_M\}$ denotes the synthesized background knowledge. This step converts question-specific answers into compact statements that can be evaluated against human-annotated background knowledge and reused as contextual evidence for any task. For example, given the meme image I, text T, and synthesized background knowledge $\hat{B}$, we can design the verifier predicts the binary label for each task $s \in \mathcal{S}$:

$$\hat{y}^{(s)} = V_{\theta_v}^{(s)}(I, T, \hat{B}), \quad \hat{y}^{(s)} \in \{0, 1\}. \tag{4}$$

Unlike direct zero-shot detection, the verifier is provided with explicit background knowledge that explains relevant entities, events, phrases, or image-text associations. This enables the prediction to be grounded in retrieved evidence rather than relying only on the model's implicit knowledge.

4 Experiment

4.1 Implementation Details

Datasets. For meme understanding, we use the existing MemeIntent (Park et al., 2024) and MemeInterpret (Park et al., 2025) datasets, and further curate a new KYM dataset to evaluate whether our framework can handle newly emerging memes. As shown by the examples in Figure 3, each meme in the KYM dataset is annotated with background knowledge, intent and an offensiveness label. Overall we annotated 100 memes scraped from *Know Your Meme*[8], covering meme series from 2024 to 2026. Specifically, these memes are centered on recent and culturally situated topics, such as the *American–Iranian War*, *Jeffrey Epstein*, *Donald*

[8]https://knowyourmeme.com/

Task	Source	Label Distribution		Total
		Label	Count	
Hatefulness	MemeInterpret	Hateful Non-hateful	490 510	1000
Misogyny	MAMI	Misogynistic Non-misogynistic	500 500	1000
Offensiveness	MultiOFF	Offensive Non-offensive	303 440	743
Sarcasm	MSD	Sarcastic Non-sarcastic	910 910	1820
Harmfulness	Harm-C, Harm-P	Harmful Non-harmful	444 619	1063

Table 1: Dataset statistics for meme detection tasks.

Model		Datasets		
QA-pairs	BKS	KYM	MemeIntent	MemeInterpret
Zero-shot Generated BKS				
–	Qwen3	0.46	0.66	0.73
–	Gemma3	0.59	0.65	0.66
****Proposed**				
Gemma3	Gemma3	0.65	0.70	0.74
Qwen3	Qwen3	0.78	0.76	0.79
Gemma3	Qwen3	0.68	0.73	0.76
Qwen3	Gemma3	0.78	0.75	0.78

Table 2: Recall scores of zero-shot generated background knowledge and QA-pair-converted knowledge statements against ground-truth background knowledge. Qwen3-VL-32B (Qwen3) and Gemma-3-27B-IT (Gemma3) are used interchangeably for QA-pair generation and knowledge-statement generation.

Trump's interest in purchasing Greenland, and *the United States of America*. Which often depend on evolving political events, public discourse, and cultural references, thus they provide a challenging testbed for evaluating whether our framework can recover up-to-date background knowledge for meme interpretation.

For downstream meme detection task, inspired by GOAT-Bench (Lin et al., 2024b), we evaluate our framework across five tasks related to social abuse and nuanced meme interpretation: hatefulness, misogyny, offensiveness, sarcasm, and harmfulness. Table 1 summarizes the label distribution and dataset size for each task.

Evaluation. We evaluate both evidence recovery quality and downstream meme detection performance. Recent studies (Cao et al., 2025; Akhtar et al., 2026) show that reference-based evidence evaluation with large language models (LLMs) aligns best with human judgments. We extend this idea to our multimodal meme understanding setting, where retrieved web evidence may include both text and images. Recall that each generated QA pair is converted into an evidence statement. We then perform reference-based evaluation by comparing these generated evidence statements with the ground-truth evidence annotations. We exploit Gemini-3.1-Flash [9] as the scoring model for the this evaluation inspired by its power (Cao et al., 2025; Akhtar et al., 2026). We then report evidence recall, defined as the percentage of ground-truth evidence instances successfully retrieved. The prompts used for QA-pair-to-statement conversion and evidence evaluation are provided in Appendix D.4 and Appendix D.5, respectively. Additionally, to assess the reliability of the reference-based evaluation, we compare the resulting scores with human judgments obtained from independent raters (details in Appendix A.1).

For downstream meme detection, we evaluate how well the recovered evidence supports prediction across the target tasks, using Accuracy and Macro-averaged F1 Score as the main evaluation metrics. We additionally conduct a human-alignment check, referred to as *Decision Support*, to examine whether the predicted background knowledge contains sufficient information to justify the annotated detection labels (details in Appendix A.2).

Models. We use different models for meme understanding and meme detection. For meme understanding, **Qwen3-VL-32B-Instruct** (Bai et al., 2025) and **Gemma-3-27B-it** (Gemma Team et al., 2025) are used. For reference-based evaluation, we use Gemini-3.1-Flash. For meme detection, we evaluate three open-source LVLMs: **llava-1.5-7b-hf** (Liu et al., 2023), **Qwen3-VL-8B-Instruct** (Bai et al., 2025), and **Gemma-3-12B-it** (Kamath et al., 2025).

4.2 Baselines

To contextualize the performance of our proposed framework, we compare against zero-shot baselines that with and without background knowledge, which are:

- **Zero-shot inference without background knowledge.** We evaluate meme detection using only the image and embedded text, without any generated background knowledge. This setting includes vanilla zero-shot inference, MemeAgent (Lin et al., 2025), an agent-

[9] https://deepmind.google/models/model-cards/gemini-3-1-flash-image/

	Model	Detection Task											
	Detection	Hatefulness		Misogyny		Offensiveness		Sarcasm		Harmfulness		Overall	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
	Zero-shot without BKS												
Vanilla	LLaVA-1.5-7B	0.54	0.45	0.59	0.52	0.58	0.57	0.50	0.34	0.61	0.54	0.56	0.48
	Qwen3-VL-8B	0.63	0.61	0.64	0.62	0.60	0.59	0.61	0.60	0.63	**0.61**	0.62	0.61
	Gemma3-12B	0.67	0.67	0.73	0.72	0.62	0.61	0.63	0.62	0.65	0.64	0.67	0.65
MemeAgent	LLaVA-1.5-7B	0.55	0.43	0.60	0.45	0.60	0.45	0.56	0.42	0.58	0.46	0.58	0.44
	Qwen3-VL-8B	0.64	0.60	0.67	**0.65**	0.62	0.41	0.64	0.60	0.62	0.59	0.64	0.57
	Gemma3-12B	0.66	0.63	0.61	0.49	0.60	0.56	0.61	0.65	0.62	0.58	0.62	0.59
MiND	LLaVA-1.5-7B	0.54	**0.60**	0.57	0.50	0.59	0.56	0.55	0.47	0.60	0.53	0.58	0.44
	Qwen3-VL-8B	0.63	0.60	**0.68**	0.64	0.61	0.58	0.63	0.59	0.64	0.60	0.64	0.57
	Gemma3-12B	0.66	0.65	0.70	0.68	0.61	0.59	0.62	0.61	0.64	0.62	0.65	0.63
	With Zero-shot Generated BKS												
	LLaVA-1.5-7B	0.56	0.42	0.61	0.55	**0.62**	0.60	0.61	0.56	0.63	0.55	0.61	0.54
	Qwen3-VL-8B	0.65	0.60	0.65	0.61	0.62	0.61	0.63	**0.63**	0.65	0.60	0.64	0.61
	Gemma3-12B	0.69	0.66	0.75	0.74	0.66	0.64	0.65	0.64	0.67	0.66	0.68	0.67
	****Proposed**												
	LLaVA-1.5-7B	**0.61**	0.59	**0.63**	**0.60**	**0.62**	**0.61**	**0.62**	**0.59**	**0.64**	**0.60**	**0.62**	**0.60**
	Qwen3-VL-8B	**0.66**	**0.62**	0.67	0.64	**0.64**	**0.62**	**0.65**	0.62	**0.66**	0.61	**0.66**	**0.62**
	Gemma3-12B	**0.71**	**0.70**	**0.80**	**0.79**	**0.67**	**0.67**	**0.69**	**0.68**	**0.70**	**0.70**	**0.71**	**0.71**

Table 3: Experimental results (Accuracy and F1) across five meme detection tasks. Results evaluate zero-shot detection without background knowledge, dedicated baselines (MemeAgent, MiND), zero-shot generated BKS, and our proposed *Query-Retrieve-Conclude*. Best results for each model backbone are in **bold**; second-best are underlined.

based zero-shot baseline that conducts a discussion over the meme before producing the final prediction, and MiND (Liu et al., 2025), a retrieval-augmented zero-shot baseline that does not update model parameters but uses similar memes from a retrieval pool to derive task-level insights for final inference.

- **Detection with zero-shot generated background knowledge.** Following MemeInterpret (Park et al., 2025), we prompt Qwen3-VL-32B to generate background knowledge directly from the meme image and text, and then use the generated knowledge as additional context for downstream detection. [10]

4.3 Results and Discussion

The evaluation results for automated meme understanding and downstream detection are presented in Table 2 and Table 3, respectively. Our analysis focuses on how open-world knowledge acquisition fills critical reasoning gaps where closed parametric baselines fail.

Meme Knowledge Recovery Analysis. As reported in Table 2, zero-shot models that rely entirely on internal parametric parameters struggle to reconstruct accurate background knowledge statements (BKS), particularly on newly emerging contexts. Under the zero-shot baseline, Qwen3 achieves an evidence recall score of only 0.46 on the curated KYM dataset, demonstrating that even large-scale VLMs fail to encapsulate highly dynamic, temporally shifting sociopolitical events (2024–2026).

In contrast, our proposed multi-stage framework consistently improves evidence recall across all evaluated benchmarks. When pairing Qwen3 as both the QA-pair generator and the BKS synthesizer, our framework improves recall on the KYM dataset from 0.46 to 0.78 (+32% absolute increase). This performance gain is similarly reflected across established benchmarks, with recall advancing from 0.66 to 0.76 on MemeIntent, and from 0.73 to 0.79 on MemeInterpret. Crucially, the cross-model evaluation (e.g., using Qwen3 for query generation and Gemma3 for statement synthesis, yielding 0.78 on KYM) reveals that the performance bottleneck in meme understanding lies not in the text synthesis stage, but in the precision of the initial search query.

Downstream Detection Performance. The empirical results across the five downstream detection tasks (Table 3) confirm that the quality of recovered background knowledge translates directly into

[10] The prompts used for zero-shot knowledge generation and detection are shown in Appendix D.6 and Appendix D.7 respectively.

Framework Configuration	Hatefulness		Sarcasm		Overall	
	Acc.	F1	Acc.	F1	Acc.	F1
Ours (Full Framework)	**0.75**	**0.74**	**0.72**	**0.71**	**0.74**	**0.74**
w/o Stage I: (Direct Retrieval)	0.69	0.66	0.67	0.65	0.69	0.67
w/o Stage II: (Parametric Only)	0.66	0.62	0.65	0.62	0.66	0.62
w/o Stage III: (Raw QA)	0.72	0.71	0.69	0.68	0.71	0.70
w/o Visual Grounding Context (c & $\mathcal{R}_I$)	0.65	0.61	0.63	0.60	0.64	0.61

Table 4: Ablation analysis across representative tasks using the Qwen3-VL-32B backbone. Configurations isolate down-stream performance shifts when structural blocks are systematically removed from our query-driven, open-world knowledge acquisition framework.

reasoning accuracy. Under the zero-shot without BKS baseline, vanilla models exhibit a severe performance ceiling; Gemma3-12B yields an overall F1 of 0.65, while smaller architectures like LLaVA-1.5-7B achieve just 0.48.

Iterative multi-agent reasoning (MemeAgent) and standard instance-level retrieval (MiND) offer marginal adjustments but do not break this ceiling. For instance, on the highly nuanced task of Sarcasm detection, MemeAgent decreases LLaVA-1.5's performance or plateaus Gemma3's performance (0.65 F1) because multi-agent debates over unencoded open-world context inevitably amplify parametric hallucinations. Similarly, while MiND utilizes visual similarity to anchor its predictions, it yields zero architectural mechanism to capture rapid temporal distribution shifts, keeping Gemma3's overall F1 constrained to 0.63.

Augmenting models with standard zero-Shot Generated BKS yields modest baseline improvements (e.g., raising Gemma3's overall F1 to 0.67), but it introduces severe structural noise. As observed in the LLaVA-1.5 configuration, zero-shot background knowledge actually degrades performance on Hatefulness (F1 drops from 0.45 to 0.42) because direct contextual generation induces severe semantic anchoring traps (further analyzed in Section 4.4).

Our proposed framework completely bypasses these degradation risks. Grounding downstream classifiers with our open-world knowledge framework pushes Gemma3-12B to a state-of-the-art accuracy of 0.71 and an F1 score of 0.71. Most notably, our framework secures massive, absolute F1 leaps on culturally dense fields: +0.07 in Misogyny (0.79 vs. 0.72) and +0.06 in Sarcasm (0.68 vs. 0.62) compared to vanilla execution.

Ablation Study. To isolate the exact architectural contributions of our *Query-Retrieve-Conclude* framework, we systematically remove key components using the primary Qwen3-VL-32B backbone. Eliminating the structured Question Generation phase (*w/o Stage I*) forces the system to pull open-world web evidence based purely on surface-level text or visual keywords. This configuration causes the overall F1 score to drop significantly from 0.74 to 0.67, confirming that direct query scraping fails to capture the implicit, multi-layered ironies embedded in evolving memes. Bypassing external web searches entirely and forcing a reliance on internal parametric parameters (*w/o Stage II*) results in an even more severe performance collapse, dropping the Hatefulness F1 score to 0.62. This drop highlights that even a large-scale 32B foundation model cannot reliably infer shifting or highly volatile real-world contexts without live information retrieval. Finally, feeding uncurated, raw question-answer fragments directly to the downstream discriminator rather than using our statement synthesis module (*w/o Stage III*) limits contextual efficiency and introduces structural token noise, which decreases the F1 score by a distinct 0.04 points. This systematic degradation across all ablation states demonstrates that each step in our framework is essential for accurately identifying knowledge gaps and protecting downstream models from semantic anchoring errors.

4.4 Case Study

To explicitly demystify why open-world knowledge decomposition outpaces direct generation, Table 5 presents a qualitative comparison between our framework and the zero-shot baseline on a sample from the MemeIntent (Park et al., 2024) dataset. [11]

The meme features an image of Chelsea Clinton's wedding with a red circle highlighting an unexpected guest, captioned with morbid options regarding an impending death. The zero-shot baseline immediately falls into a *semantic anchoring trap*. Because the parametric weights of the model strongly couple "Bill Clinton" and "scandal" with the 1990s White House accounts, the model misidentifies the circled figure as Monica Lewinsky and hallucinates that the meme's humor is derived from that historic affair. Consequently, the baseline completely misses the actual temporal target, yielding a totally broken context statement.

Our proposed framework systematically evades

[11] The purpose of this case study is not to replace the quantitative evaluation, but to illustrate how question-guided knowledge acquisition changes the type of background knowledge produced compared with direct zero-shot generation.

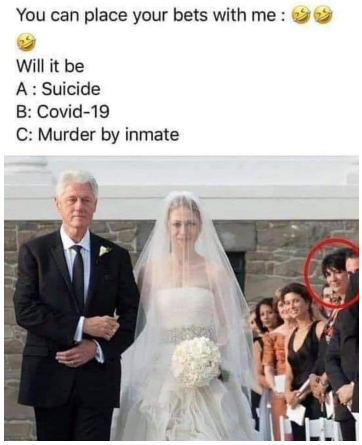
Meme	Proposed	Zero-shot Baseline
 **Ground Truth BKS:** 1. Bill and Hillary Clinton are married. 2. People generally invite people they are associated with to their wedding. 3. Ghislaine Maxwell was a known associate of Jeffrey Epstein. 4. Jeffrey Epstein allegedly killed himself in prison. 5. There is a conspiracy theory that he was murdered. 6. Murders are often covered up by claiming the deceased died from a natural cause.	**[Q1]:** Who is Ghislaine Maxwell and what is her connection to Bill Clinton and Chelsea Clinton? **[Q2]:** What is the significance of the wedding photograph featuring Bill Clinton, Chelsea Clinton, and Ghislaine Maxwell? **[Q3]:** What does the phrase 'Will it be A. suicide B. COVID-19 C. Murder by inmate' refer to in the context of Ghislaine Maxwell? --- **[Our QA-pairs Converted BKS]:** 1. **Ghislaine Maxwell** is a former girlfriend and accomplice of **Jeffrey Epstein** who has claimed to have a friendship with Bill Clinton and attended Chelsea Clinton's wedding as a guest. 2. The wedding photograph featuring Bill Clinton, Chelsea Clinton, and Ghislaine Maxwell is significant **because it shows Maxwell in attendance**, sparking scrutiny and speculation about the relationship between the Clintons and Maxwell. 3. The phrase *"Will it be A. suicide B. COVID-19 C. Murder by inmate"* refers to **speculation about the cause of death of Jeffrey Epstein,** not Ghislaine Maxwell; it reflects conspiracy theories surrounding Epstein's death in prison, with some suggesting murder rather than suicide.	**[Zero-shot Generated BKS]:** 1. **Bill Clinton**, the former President of the United States, was involved in a **scandal** in the 1990s related to an affair with **Monica Lewinsky**, a White House intern, which led to his impeachment. This scandal is widely known and often referenced in political and cultural discourse. 2. The woman in the wedding dress is Chelsea Clinton, Bill Clinton's daughter, and the image is from her wedding in 2010. The red circle highlights a woman in the background who resembles **Monica Lewinsky**, which is the source of the meme's humor and controversy. 3. The meme plays on the public's awareness.

Table 5: Example background knowledge statements generated by our proposed method and the zero-shot baseline. Important information is highlighted in **bold**.

this failure mode. By refusing to directly guess the contextual premise, the model instead flags the visual anomaly as an explicit knowledge gap, generating targeted queries: *"Who is the woman... what is the significance of the wedding photograph?"* The resulting web-grounded synthesis accurately surfaces *Ghislaine Maxwell*, her real-world association with Jeffrey Epstein, and the underlying conspiracy theories surrounding mysterious inmate deaths.

5 Conclusion

Our work highlights that meme understanding is not only multimodal reasoning problems, but also open-world knowledge acquisition problems, since many memes depend on recent events, evolving cultural references, and implicit social associations that may be absent from both the meme content and a model's parametric memory. Our results show that explicitly acquiring and grounding missing background knowledge improves meme understanding by recovering more reference-aligned contextual knowledge, and improves meme detection by providing stronger evidence for downstream prediction. This suggests that robust meme analysis requires models to recognize what they do not know, seek external evidence, and use the acquired knowledge to support interpretable and reliable decisions. Future work can extend this direction by evaluating larger and more diverse multilingual meme collections, improving retrieval and evidence verification, reducing inference cost, and incorporating uncertainty estimation or human feedback to build more adaptive, reliable, and culturally aware meme understanding and detection systems.

6 Limitations

Dependence on External Evidence. Our framework relies on the quality of retrieved evidence. While inference-time retrieval helps address dynamic and emerging memes, web and reverse image search may return incomplete, noisy, outdated, or biased results. Errors from captions, entity recognition, or question generation may further propagate into misleading background knowledge.

Limited Evaluation Scope. Our experiments mainly focus on English-language memes and social-abuse-related detection tasks, with a relatively small curated KYM dataset. Further validation is needed for multilingual, culturally localized, and low-resource meme communities. LLM-based evidence evaluation may also introduce judgment bias.

Retrieval Cost and Adaptive Planning. The framework always retrieves from the external web, which increases inference latency and search cost even when parametric knowledge or static knowledge bases may suffice. Future work could develop

agentic pipelines that dynamically decide whether to use parametric knowledge, curated knowledge bases, or open-web search for each meme.

Potential Risk. Because our framework explicitly retrieves and synthesizes background knowledge, it may surface sensitive attributes, or politically and culturally biased associations that are only indirectly implied by a meme. Such knowledge can improve interpretation, but it may also amplify harmful framing or expose sensitive information if not carefully controlled. Future work should incorporate bias-aware filtering, sensitive-content safeguards, and uncertainty signals when using retrieved background knowledge for meme understanding.

Acknowledgments This research/project is supported by the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 2021_09_01). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. Mememqa: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5042–5078.

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2026. Ev2r: Evaluating evidence retrieval in automated fact-checking. *Transactions of the Association for Computational Linguistics*, 14:530–561.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web. *arXiv preprint arXiv:2505.17978*.

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.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 193 others. 2025. Gemma 3 Technical Report. *arXiv preprint*.

Eunjeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1433–1445. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, S Perrin, T Matejovicova, A Ramé, M Rivière, L Rouillard, and 1 others. 2025. Gemma 3 technical report. corr abs/2503.19786 (2025). *arXiv preprint arXiv:2503.19786*.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

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

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024b. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Transactions on Intelligent Systems and Technology*.

Xuanrui Lin, Chao Jia, Junhui Ji, Hui Han, and Usman Naseem. 2025. Ask, acquire, understand: A multimodal agent-based framework for social abuse detection in memes. In *Proceedings of the ACM on Web Conference 2025*, pages 4734–4744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Olivia Shanhong Liu, Pai Chet Ng, De Wen Soh, and Konstantinos N Plataniotis. 2026a. Yes florence, i will do better next time! agentic feedback reasoning for humorous meme detection. *arXiv preprint arXiv:2601.07232*.

Shanhong Liu, Pai Chet Ng, De Wen Soh, Malika Meghjani, and Konstantinos N Plataniotis. 2026b. Beyond a joke: Multi-angle reasoning for detecting and explaining harmful humor in memes. *arXiv preprint arXiv:2607.15442*.

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.

Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Pengda Qin, Yuhong Li, Da Chen, and Bill Byrne. 2025. Expo-hm: Learning to explain-then-detect for hateful meme detection. *CoRR*, abs/2510.08630.

Khoi PN Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, and 1 others. 2025. Memeqa: Holistic evaluation for meme understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18926–18946.

Fengjun Pan, Anh Tuan Luu, and Xiaobao Wu. 2025. Detecting harmful memes with decoupled understanding and guided cot reasoning. *CoRR*, abs/2506.08477.

Jeongsik Park, Khoi PN Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. 2024. Memeintent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643.

Jeongsik Park, Khoi PN Nguyen, Jihyung Park, Minseok Kim, Jaeheon Lee, Jae Won Choi, Kalyani Ganta, Phalgun Ashrit Kasu, Rohan Sarakinti, Sanjana Vipperla, and 1 others. 2025. Memeinterpret: Towards an all-in-one dataset for meme understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16073–16087.

Karl Pearson. 1896. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.

Naquee Rizwan, Subhankar Swain, Paramananda Bhaskar, Gagan Aryan, Shehryaar Shah Khan, and Animesh Mukherjee. 2026. See, explain, and intervene: A few-shot multimodal agent framework for hateful meme moderation. *arXiv preprint arXiv:2601.04692*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, pages 9763–9771. AAAI Press.

Charles Spearman. 1961. The proof and measurement of association between two things.

Sahil Tripathi, Gautam Siddharth Kashyap, Mehwish Nasim, Jian Yang, Jiechao Gao, and Usman Naseem. 2026. They said memes were harmless—we found the ones that hurt: Decoding jokes, symbols, and cultural references. In *Proceedings of the ACM Web Conference 2026*, pages 4910–4918.

Carlo M Valensise, Alessandra Serra, Alessandro Galeazzi, Gabriele Etta, Matteo Cinelli, and Walter Quattrociocchi. 2021. Entropy and complexity unveil the landscape of memes evolution. *Scientific Reports*, 11(1):20022.

Contents of Appendix

A Human Alignment of Evaluation Metrics 11
- A.1 Alignment Check on Evidence Evaluation 11
- A.2 Alignment Check on Decision Support For Detection Task 11

B Additional Experimental Results 12
- B.1 Precision Scores on QA-pairs Converted Statements 12
- B.2 More Evidence Evaluation Results 12
- B.3 More Detection Results 12

C Details of Experiments 13
- C.1 Computation Resources and Costs 13
- C.2 Experiment Environment and Packages 13
- C.3 Retrieval Details 13

D Prompts in Use 14
- D.1 Prompts for Caption Generation . 14
- D.2 Prompts for Question Generation . 14
- D.3 Prompts for Answer Generation . 14
- D.4 Prompts for QA Conversion to Evidence Statement 14
- D.5 Prompts for Evidence Evaluation. 15
- D.6 Prompts for Zero-shot Background Knowledge Generation 15
- D.7 Prompts for Detection 15

E Declaration on AI Assistance 16

A Human Alignment of Evaluation Metrics

A.1 Alignment Check on Evidence Evaluation

Human Alignment Check. To further ensure alignment with human judgment in our setting, we conducted a small-scale human evaluation to validate our reference-based evaluation of generated background knowledge. Specifically, we randomly selected 10 memes from each of the three meme understanding datasets, including our curated KYM dataset, MemeIntent (Park et al., 2024), and MemeInterpret (Park et al., 2025). The alignment checks were done by four university students and two young professionals based in English speaking countries.[12]

As shown in Figure 4, for each example, annotators were presented with the meme image, meme text, model-generated background knowledge, and the corresponding human-annotated reference background knowledge. They were asked to compare the generated background knowledge against the reference background knowledge based on two criteria: **Coverage**, which measures how well the generated background knowledge captures the information contained in the reference background knowledge, and **Usefulness**, which measures whether the generated background knowledge helps users understand the intended meaning of the meme. As shown in Table 6, the automatic recall-based evaluation achieves Spearman(Spearman, 1961) correlations of $\rho_{\text{KYM}} = 0.78$, $\rho_{\text{MemeIntent}} = 0.74$, and $\rho_{\text{MemeInterpret}} = 0.77$, and Pearson(Pearson, 1896) correlations of $r_{\text{KYM}} = 0.76$, $r_{\text{MemeIntent}} = 0.73$, and $r_{\text{MemeInterpret}} = 0.75$. These results indicate a strong alignment between human judgments and the reference-based evaluation scores on the evidence.

Metric	KYM	MemeIntent	MemeInterpret
Spearman (ρ)	0.78	0.74	0.77
Pearson (r)	0.76	0.73	0.75

Table 6: Correlation between human evaluation and automatic recall-based evaluation of generated background knowledge. We report Spearman's rank correlation coefficient (ρ) and Pearson's correlation coefficient (r) for KYM, MemeIntent, and MemeInterpret.

A.2 Alignment Check on Decision Support For Detection Task

To further assess alignment with human judgment, we conduct a *Decision Support* evaluation on MemeInterpret (Park et al., 2025), examining whether the predicted background knowledge provides sufficient information to justify the annotated detection label. Specifically, we treat the detection outputs produced from the predicted evidence, as described in Appendix A.1, as model predictions, and the raters rate how well the predicted background knowledge supports the given classification label. With a mean Decision Support rating of **4.28** out of 5.0 (with a standard deviation of ± 0.64), this strong average confirms that our query-driven open-world evidence discovery does not merely collect generic, superficially fluent paragraphs; rather, it successfully isolates and delivers the exact high-fidelity empirical facts required to systematically support and justify downstream classifications.

[12]The evaluators were properly instructed and made aware of the potential risk of disturbing content.

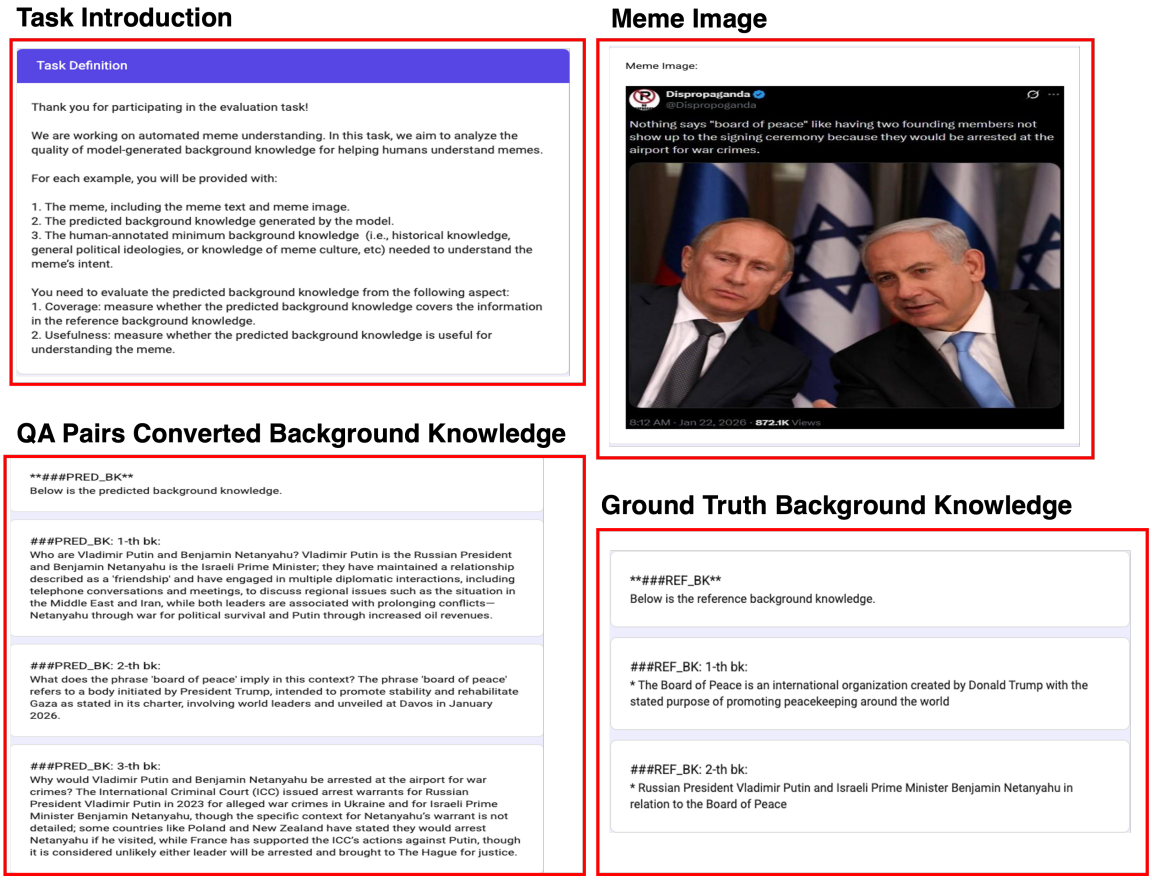


Figure 4: Instructions for Alignment Check on Evidence Evaluation

B Additional Experimental Results

B.1 Precision Scores on QA-pairs Converted Statements

We computed the precision scores (i.e., the percentage of generated questions result in necessary knowledge annotated by human) of our reference-based evaluation to further examine whether the generated questions lead to relevant evidence rather than excessive or noisy evidence. As shown below, the precision scores for KYM, MemeIntent, and MemeInterpret are 0.71, 0.67, and 0.72, respectively. These results demonstrate that our method does not simply produce many questions and evidence statements; instead, it provides evidence that is largely aligned with the required human-annotated background knowledge.

B.2 More Evidence Evaluation Results

Recall that we have shown Gemini-3.1-Flash 8 as the scoring model for the reference-based evaluation where we compare the QA-pairs converted statements with the ground-truth evidence annotation[13]; we conducted an additional recall evaluation using GPT-5-mini.[14] As shown in Table 7, we observe a similar trend to the Gemini-based results. Specifically, zero-shot generated BKS struggles to reconstruct accurate background knowledge, particularly for newly emerging memes. In contrast, our proposed multi-stage framework consistently improves evidence recall across the evaluated benchmarks.

Method	KYM	MemeIntent	MemeInterpret
Zero-shot Generated BKS	0.63	0.59	0.61
Proposed	**0.68**	**0.61**	**0.64**

Table 7: Performance comparison with zero-shot generated background knowledge statements (BKS).

B.3 More Detection Results

We conducted additional experiments using Qwen3-VL-32B (Bai et al., 2025) as the detector for hateful meme and sarcasm detection. As

[13] Details in Section 4.

[14] https://developers.openai.com/api/docs/models/gpt-5-mini

Method	Hatefulness (Acc., F1)	Sarcasm (Acc., F1)	Overall (Acc., F1)
Vanilla zero-shot (w/o BKS)	(0.68, 0.66)	(0.65, 0.64)	(0.66, 0.65)
MemeAgent (w/o BKS)	(0.69, 0.67)	(0.68, 0.64)	(0.69, 0.66)
MIND (w/o BKS)	(0.71, 0.71)	(0.68, 0.63)	(0.70, 0.67)
With Zero-shot Generated BKS	(0.72, 0.71)	(0.70, 0.72)	(0.71, 0.72)
Proposed	**(0.74, 0.72)**	**(0.75, 0.77)**	**(0.75, 0.75)**

Table 8: Performance comparison across Hatefulness and Sarcasm tasks.

Dataset	ImgBB Upload	RIS	Caption	Question (Qwen3)	Retrieve	Answer (Qwen3)	Statement (Gemma3)	Detection (LLaVA)	Total	Avg. Runtime / Meme
KYM	∼30 s	∼5 min	∼2–4 min	∼2–5 min	∼20 min	∼4–5 min	∼2–4 min	–	∼0.59–0.73 h	**∼22–27 s**
MemeInterpret	∼1.5 min	∼60 min	∼20–30 min	∼20–30 min	∼250 min	∼40–50 min	∼20–40 min	∼3 min	∼6.9–7.7 h	**∼25–28 s**
MSD	∼3 min	∼110 min	∼40–50 min	∼40–50 min	∼520–560 min	∼90–100 min	∼70–80 min	∼6 min	∼14.7–16.0 h	**∼29–32 s**

Table 9: Runtime Costs Across Stages

shown in Table 8, the proposed method improves over vanilla zero-shot prompting, zero-shot baselines, and zero-shot generated BKS when using a stronger 32B detector. Overall, we observe that our proposed method achieves an average score of 0.75 accuracy and 0.75 F1 across the two tasks. This improves over vanilla zero-shot prompting by +0.09 accuracy and +0.10 F1, MemeAgent (Lin et al., 2025) by +0.06 accuracy and +0.09 F1, MIND (Liu et al., 2025) by +0.05 accuracy and +0.08 F1,and zero-shot generated BKS by +0.04 accuracy and +0.03 F1. Which suggest that the background knowledge generated by our proposed framework improves downstream meme detection not only for small VLMs[15], but also when using a stronger detector.

C Details of Experiments

C.1 Computation Resources and Costs

All experiments are conducted on four NVIDIA L40S GPUs, each with 48 GB of dedicated GPU memory.

Table 9 summarize the runtime cost across all stages, from upload the images to statement generation, and downstream detection tasks (we select two datasets). Since some logs do not contain explicit timestamps, we report approximate runtime ranges for the model-generation stages based on execution observation. The measured local retrieval/ranking time is approximately 4 seconds per image/question, while the full open-web retrieval latency can be higher and depends on network conditions, search API latency, and page-fetching time, as discussed in the Limitations section.

[15]Details in Appendix C.2.

C.2 Experiment Environment and Packages

In this section, we introduce the experiment environment and packages in use. All experiments are implemented in PyTorch 2.10.0+cu128 with CUDA 12.8. For meme understanding, which involves caption generation, question generation, answer generation, and background knowledge statement generation, we use open-source models loaded through the HuggingFace ecosystem, including *Qwen/Qwen3-VL-32B-Instruct* and *google/gemma-3-27b-it*. For meme detection, we use *llava-hf/llava-1.5-7b-hf*, *Qwen/Qwen3-VL-8B-Instruct*, and *google/gemma-3-12b-it*. The version of *huggingface_hub* used in our experiments is 1.12.0. For proprietary-model experiments, we access Gemini through the API and use *gemini-3.1-flash-lite*.

C.3 Retrieval Details

Reverse Image Search For retrieving the related information for the meme images, we first upload them to obtain an accessible image URL, we then use Google reverse image search to obtain the information related to the images. The returned results include page titles, URLs, and text snippets associated with visually similar or contextually related images. Following the evidence retrieval strategy used in (Cao et al., 2025), we rank the retrieved textual snippets using BM25 (Robertson and Zaragoza, 2009) and retain only the top five chunks. We save the top five ranked text snippets from the retrieved web pages, rather than raw page contents or an unconstrained LLM-generated summary. These snippets are used as auxiliary visual grounding for caption and question generation, helping resolve ambiguous entities, symbols, or scenes before the framework proceeds to evidence-grounded retrieval

and QA.

Retrieve In the Retrieve stage. To improve the validity of the retrieved information, we follow the evidence retrieval strategy used in (Cao et al., 2025): we rank the retrieved textual snippets using BM25 (Robertson and Zaragoza, 2009) and retain only the top five chunks. We then apply lightweight source filtering by prioritizing informative sources such as Wikipedia and BBC websites, while excluding shopping and advertising domains such as Amazon and eBay. We will add these retrieval details in the revised version.

D Prompts in Use

In this section, we provide the exact prompts in use for our proposed pipeline and baselines.

D.1 Prompts for Caption Generation

```
You are an Image Captioner for memes. You are given a
meme image, its embedded text, and RIS visual hints. Your
task is to generate a literal description of what is
visibly shown. Use RIS only to help identify ambiguous
visible words, people, symbols, or known scenes.

### INPUTS
Meme text:{text}
RIS visual hints: {ris_summary}

STRICT OUTPUT:
Return JSON ONLY in this format:
{{
  "image_caption": "...",
  "recognized_people": ["..."]
}}

Please generate captions:
```

D.2 Prompts for Question Generation

```
You are a Knowledge Gap Analyst for meme understanding.
You are given a meme image, its embedded text, and
Recognized people list (may be empty). Your task is to
generate a a small set of useful background knowledge
questions that a curious human viewer might ask to better
understand the meme.

Please note: The goal is not to explain the meme directly
or to fill a fixed number of slots, but to ask the most
useful questions that a person would likely search for in
order to understand the meme.

### INPUTS
Meme text: {text}
Recognized people: {recognized_people}
Generated image caption: {gen_caption}

### OUTPUT (JSON ONLY)
{{
  "question_types_used": ["entity", "phrase",
"association"],
  "questions": ["...", "...", "..."]
}}

Please generate your question:
```

D.3 Prompts for Answer Generation

```
You are answering a question using only the provided
retrieved documents. The retrieved documents were
specifically retrieved for this question. Your goal is to
produce the strongest evidence-grounded answer possible.

Important:
- Use ONLY the provided retrieved documents.
- Do NOT use outside knowledge, memory, assumptions, or
unsupported claims.
- If the documents support a full answer, give a concise
full answer.
- If the documents support only part of the question,
give the best partial evidence-grounded answer.

### INPUTS
Question: {question}
Retrieved documents: {evidence_block}

Return JSON ONLY in this format:
{{
  "answer": "...",
  "used_chunk_ids": [1, 2]
}}

Please generate your answer:
```

D.4 Prompts for QA Conversion to Evidence Statement

Following (Cao et al., 2025; Akhtar et al., 2026), we convert QA pairs to evidence statement, for both evidence evaluation and maintaining the evidence history.

```
You are a expert writer. Given a question ([QUES]) and
its answer [ANS], your goal is to convert the caption and
the QA pair into a statement [STAT].
Below are some examples:

[QUES]: What event or incident is referenced by the image
of the Statue of Liberty with dark smoke rising behind a
city skyline?
[ANS]: The image of the Statue of Liberty with dark smoke
rising behind a city skyline refers to the events
following the 9/11 terrorist attacks in Lower Manhattan.
[STAT]: The event referenced by the image of the Statue
of Liberty with dark smoke rising behind a city skyline
is the 9/11 terrorist attacks in Lower Manhattan.

[QUES]: What is the significance of \"MN\" in the context
of this meme?
[ANS]: In the context of the meme, \"MN\" refers to
Minnesota, as evidenced by references to Minneapolis,
Minnesota.
[STAT]: \"MN\" refers to Minnesota, a U.S. state in the
context of this meme.

[QUES]: Who is George Floyd, and how is him relevant to
Minnesota in this meme?
[ANS]: George Floyd was an African American man. He was
killed during an arrest by Minneapolis police in May
2020. His death led to large protests in Minneapolis.
[STAT]: George Floyd  was an African man. He is relevant
to Minnesota in the meme because he died during a police
arrest in Minneapolis, and the incident became widely
associated with protests against police violence.

Please convert the QA pair below into its statement:
[QUES]: {ques}
[ANS]: {ans}
[STAT]:
```

D.5 Prompts for Evidence Evaluation.

As described in Section 4, we compare the textual of retrieved evidence and ground-truth evidence.

```
You will get as input a reference evidence ([REF]) and a
predicted evidence ([PRED]).
The predicted evidence contains generated statements. The
reference evidence contains the reference background
knowledge statements. Please verify the correctness of
the predicted evidence by comparing it to the reference
evidence, following these steps: 1. Evaluate each fact in
the predicted evidence individually: is the fact
supported by the REFERENCE evidence? Do not use
additional sources or background knowledge. 2. Evaluate
each fact in the reference evidence individually: is the
fact supported by the PREDICTED evidence? Do not use
additional sources or background knowledge. 3. Finally
summarise (1.) how many predicted facts are supported by
the reference evidence and explanations ([PRED in REF]
and [PRED in REF Exp]), (2.) how many reference facts are
supported by the predicted evidence and explanations
([REF in PRED] and [REF in PRED Exp]). Generate the
output as shown in the examples below:

[PRED]: 1. Donald Trump is an American politician, media
personality, and businessman who is the 47th president of
the United States. 2. The phrase "It affects virtually
nobody. It's an amazing thing." refers to the coronavirus
(COVID-19), as stated by Donald Trump while downplaying
the severity of the virus.

[REF]: 1. Trump has been accused of downplaying the
severity of COVID. 2. Many people have been negatively
affected by COVID.
[PRED in REF]: 1
[PRED in REF Exp]: 1. The first predicted statement
provides general biographical information about Donald
Trump, which is not directly supported by the reference
evidence. 2. The second predicted statement is supported
by the reference evidence because the reference states
that Trump has been accused of downplaying the severity
of COVID, and that many people have been negatively
affected by it.
[REF in PRED]: 2
[REF in PRED Exp]: 1. The first reference fact is
supported by the second predicted statement, which states
that Trump downplayed the severity of COVID. 2. The
second reference fact is supported by the second
predicted statement, which refers to COVID and implies
its serious impact.

[PRED]: 1. Ilhan Omar is an American politician and
member of the U.S. House of Representatives. 2. Ilhan
Omar's statement "I hate Trump" is being paired with the
response "Most Terrorists do" as a rhetorical criticism
that associates her statement with extremist views.
[REF]: 1. Ilhan Omar is a U.S. congresswoman from
Minnesota. 2. The meme frames Ilhan Omar negatively by
linking her anti-Trump statement to terrorism.
[PRED in REF]: 2
[PRED in REF Exp]: 1. The first predicted statement is
supported by the first reference fact, since both state
that Ilhan Omar is a U.S. politician/congresswoman. 2.
The second predicted statement is supported by the second
reference fact, since both describe the meme as linking
Omar's anti-Trump statement with terrorism in a negative
framing.
[REF in PRED]: 2
[REF in PRED Exp]: 1. The first reference fact is
supported by the first predicted statement. 2. The second
reference fact is supported by the second predicted
statement.

[PRED]: 1. The phrase "THE PARTY OF DIVERSITY" refers to
the Democratic Party. 2. The labels "GALAXY DUST" and
"MOSTLY GAS" are used in the meme to suggest exaggerated
or mocking identity labels.
[REF]: 1. The meme critiques the Democratic Party's claim
to diversity. 2. The meme uses unusual labels to mock
identity-based politics.
[PRED in REF]: 2
[PRED in REF Exp]: 1. The first predicted statement is
supported by the first reference fact because both refer
to the Democratic Party and its claim to diversity. 2.
The second predicted statement is supported by the second
reference fact because both describe the unusual labels
as part of the meme's mockery.
[REF in PRED]: 2
[REF in PRED Exp]: 1. The first reference fact is
supported by the first predicted statement. 2. The second
reference fact is supported by the second predicted
statement.

Return the output in the exact format as specified in the
examples, do not generate any additional output:
[PRED]: [PRED_EVID]
[REF]: [REF_EVID]
```

D.6 Prompts for Zero-shot Background Knowledge Generation

We adopt the zero-shot background knowledge generation prompt from (Park et al., 2025).

```
You will be provided with a meme and its embedded text.
Your task is to infer the background knowledge that a
reader of the meme needs to possess before they can
understand the ultimate intent behind the creation or
sharing of a meme, as perceived by its audience.
Background knowledge is the minimum amount of knowledge
that is missing from the meme. It is the knowledge that
needs to be combined with visual and textual cues from
the meme in  order to munderstand its meaning. Give me
background knowledge in the form of a list. For example:
'1. Soccer is the sport that  children like a lot. 2.
There are two main political parties in the US:
Democratic and Republican.' Each background knowledge
item must be in one to three sentences.
### Background knowledge:
```

D.7 Prompts for Detection

Recall that we have defined the detection task as:

$$\mathcal{S} \in \left\{ \begin{array}{l} \text{hatefulness, misogyny, offensiveness,} \\ \text{sarcasm, harmfulness} \end{array} \right\},$$

Inspired by (Lin et al., 2024b) , for each task $* \in \mathcal{S}$, we standardize the detection prompt by conditioning on the meme content as follows:

- For Vanilla Zero-shot Detection:
  "Given the meme, with the text [$\mathcal{T}$] accompanied by the image [$\mathcal{I}$], is this meme [$*$?]"

- For Detection with Background Knowledge:
  "Given the meme, with the text [$\mathcal{T}$] accompanied by the image [$\mathcal{I}$], the background knowledge [$\mathcal{B}$],[16] *is this meme [$*$?]".*

Below are the prompts that are used:

```
Given the meme, with the text [{text}] accompanied by the
image [{img_name}], is this meme {task_word}?
Return exactly one character:
1 if yes
0 if no
Do not output any other word.
```

[16] For detection with zero-shot generated background knowledge, $\mathcal{B}$ is generated by D.6. In our proposed pipeline, $\mathcal{B}$ is the background-knowledge statements converted from the QA pairs.

```
Given the meme, with the text [{text}] accompanied by the
image [{img_name}], the background knowledge of the meme
[{background_knowledge}], is this meme {task_word}?
Return exactly one character:
1 if yes
0 if no
Do not output any other word.
```

E Declaration on AI Assistance

AI-assisted writing tools were used to polish language, correct grammar, and improve the organization and clarity of the manuscript. The research design, experiments, analysis, and conclusions were conducted and verified by the authors. The authors reviewed all AI-assisted edits and take full responsibility for the final content.