SyRHM: Symbolic Reasoning with Retrieval for Zero-shot Harmful Meme Detection

Anonymous ACL submission

Abstract

Detecting harmful memes is critical for maintaining safe online communities. However, harmful intent is often implicit, arising from visual–textual incongruity and cultural stereotypes, which challenges existing multimodal detectors. We propose SyRHM, a framework that decomposes harmful meme detection into meaning-grounded retrieval and symbolic multi-stage reasoning. SyRHM retrieves semantically related memes by parsing multimodal content into textual elements and description, providing grounded context beyond surface-level similarity. Building on the retrieved context, SyRHM uses a translator stage to convert multimodal inputs into symbolic intermediate representations, and then performs multi-stage reasoning via planner and solver stages, enabling expressive and interpretable analysis of harmful intent. Experiments on FHM, HarM, and MultiOff show that SyRHM consistently outperforms multimodal and reasoning-based baselines, while providing reasoning traces for harmful content. The code is available at: `https://anonymous.4open.science/r/SyRHM-CC46/README.md`

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

1 Introduction

Memes have emerged as a powerful form of online expression, combining visual elements with text to convey complex, often culturally embedded messages. Beyond their humorous or entertaining purposes, memes can also serve as vehicles for covertly spreading harmful ideologies, posing a serious threat to inclusive online communities (Sharma et al., 2022; Pramanick et al., 2021a; Gomez et al., 2020). Detecting harmful memes (Lippe et al., 2020; Kiela et al., 2020; Cao et al., 2020) is particularly challenging because harmfulness often lies not in explicit surface features, but in implicit semantic associations, cultural references, or sarcasm. Therefore, effective detection requires both multimodal understanding and a reasoning process capable of inferring hostility from subtle, indirect cues.

Non-Harmful Non-Harmful Harmful

Figure 1: Similar memes can convey different meanings: the left examples are neutral, while the right example conveys harmful stereotypes targeting individuals with Down syndrome.

Early work on harmful meme detection primarily relies on supervised multimodal models (Pramanick et al., 2021a; Kiela et al., 2020; Suryawanshi et al., 2020) or fine-tuned vision–language models (Lippe et al., 2020; Muennighoff, 2020; Velioglu and Rose, 2020; Hee et al., 2022), which require large-scale high-quality annotations and struggle to generalize to evolving meme semantics. To reduce annotation dependence, recent studies explore few-shot in-context learning for harmful meme detection (Hee et al., 2024; Huang et al., 2024; Cao et al., 2024). More recently, large multimodal models (LMMs) have been applied in zero-shot settings, often augmented with retrieval of related memes to provide contextual grounding (Mei et al., 2024; Qu et al., 2023). Some approaches incorporate reasoning strategies, such as Chain-of-Evolution prompting in Evolver (Huang et al., 2025) and multi-agent debate in MinD (Liu et al., 2025), further improve decision robustness.

Despite these advances, retrieval-augmented methods suffer from two core limitations. First, global image–text retrieval favors surface similarity over semantic relevance, often returning visually or textually similar memes with opposite meanings (Figure 1).

Second, reasoning over retrieved evidence relies only on surface-level cues from images and captions, without capturing semantic relations, limiting understanding of implicit harmful content. To address these challenges, we propose SyRHM, a framework that integrates fine-grained associative retrieval with multi-stage symbolic reasoning for zero-shot harmful meme detection. Instead of relying on holistic image–caption embeddings, SyRHM first parses each meme into structured semantic elements (e.g., entities, attributes, and relations) and textual descriptions, enabling more precise associative retrieval grounded in underlying meaning rather than surface similarity. On top of the retrieved evidence, we introduce a multi-stage symbolic reasoning pipeline that translates natural language into symbolic representations and performs structured reasoning over semantic relations for interpretable decision-making. Through these strategies, SyRHM effectively uncovers context-sensitive hostility in memes, especially those using cultural knowledge, sarcasm, or intertextuality, and exhibit superiority across four zero-shot harmful meme detection benchmarks. Our main contributions are summarized as follows:

- We propose a associative retrieval module that treats treats each meme as a semantic unit and parses it into fine-grained elements, enabling more precise retrieval grounded in meaning rather than surface similarity.
- We introduce multi-stage symbolic reasoning for harmful meme detection, explicitly modeling semantic relations to enhance both interpretability and robustness.
- Extensive experiments on four harmful meme benchmarks demonstrate that SyRHM achieves state-of-the-art performance, outperforming prior reasoning frameworks such as MinD and Evolver by up to 9.0 accuracy points in the zero-shot setting.

2 Related Work

2.1 Harmful Meme Detection

Harmful memes pose a significant threat to online platforms, motivating the construction of several benchmark datasets. The Facebook Hateful Memes (FHM) dataset (Kiela et al., 2020) covers a broad range of hateful and non-hateful scenarios, while Multi-Off (Suryawanshi et al., 2020) and HarM (Pramanick et al., 2021a) focus on socio-political content (e.g., COVID-19 and U.S. politics). Early approaches to harmful meme detection primarily relied on two-stream or early-fusion architectures to integrate visual and textual information (Kiela et al., 2020; Suryawanshi et al., 2020; Radford et al., 2021; Pramanick et al., 2021b). With pretrained vision–language models (VLMs), fine-tuning became dominant, improving performance and offering post-hoc explanations via attention or attribution (Lippe et al., 2020; Hee et al., 2022; Muennighoff, 2020; Kumar and Nandakumar, 2022; Arya et al., 2024). However, their reliance on large-scale annotations constrains generalization to emerging events and unseen harmful patterns. More recently, large multimodal models (LMMs) have been explored in few-shot and zero-shot settings, often augmented with retrieved examples and in-context prompting to reduce task-specific supervision (Huang et al., 2024; Mei et al., 2024; Huang et al., 2025; Liu et al., 2025). While these methods improve generalization, they largely rely on global image–text retrieval and treat retrieved evidence as unstructured context, limiting their ability to model implicit cross-meme semantics and interpretability. We address this by introducing a fine-grained associative retrieval framework with multi-stage symbolic reasoning, which parses memes into semantic elements and performs structured inference over retrieved evidence using first-order logic.

2.2 LLM Reasoning

Recent advances in LLM-based reasoning show strong performance across diverse tasks (Huang and Chang, 2023), largely driven by prompting strategies such as Chain-of-Thought (CoT) (Wei et al., 2022; Zhang et al., 2022). By explicitly modeling intermediate reasoning steps or further emulating human cognitive patterns, CoT improves reasoning accuracy across multiple domains (Fei et al., 2023; Inaba et al., 2023; Fei et al., 2024; Zhou et al., 2022; Yao et al., 2023; Wang et al., 2023; Besta et al., 2024). However, their reliance on free-form natural language rationales can be suboptimal for tasks requiring precise logical manipulation. To address this limitation, prior work introduces structured intermediate steps, yielding improvements in code generation and mathematical reasoning (Li et al., 2025; Imani et al., 2023). Building on this, symbolic reasoning approaches first employ LLMs to translate natural language

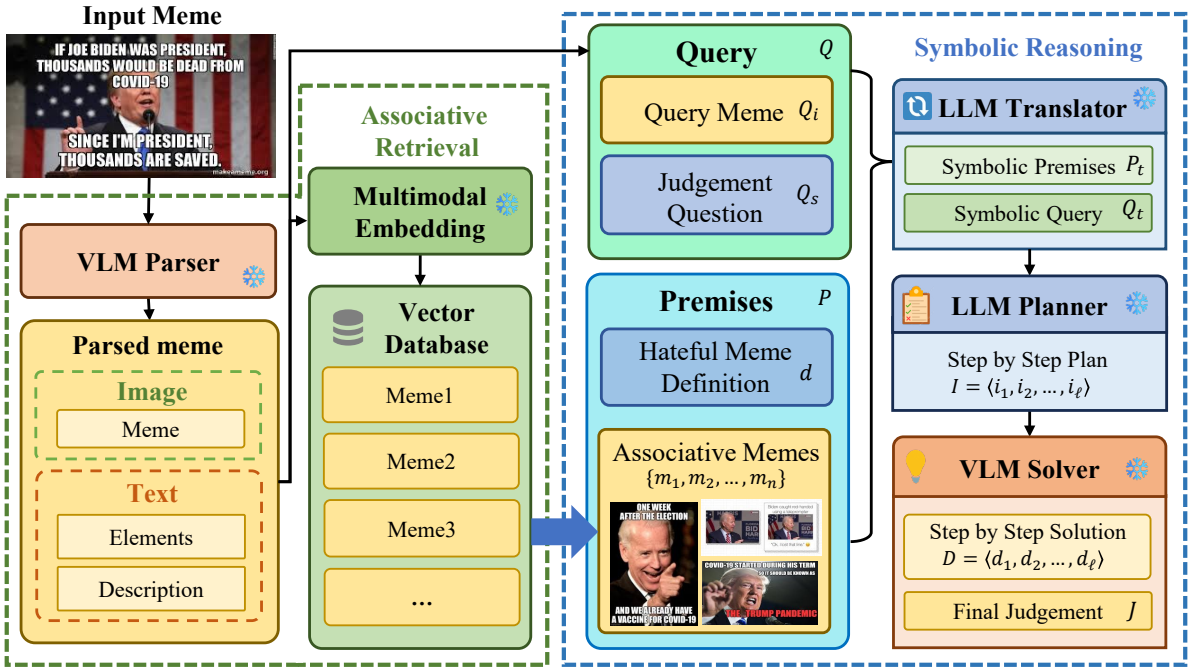

Figure 2: Framework of SYRHM

into formal logic processed by external solvers (Ye et al., 2023; Pan et al., 2023; Gaur and Saunshi, 2023; Olausson et al., 2023; Gao et al., 2023), while subsequent methods enable LLMs to jointly perform symbolic translation and logical inference, mitigating information loss and achieving stronger performance (Huang et al., 2025; Xu et al., 2024). However, existing symbolic CoT methods focus on text-only reasoning with explicit premises. We extend this line of work by integrating symbolic reasoning across multimodal content, enabling interpretable and context-aware assessment of harmful memes.

3 Method

3.1 Problem Statement

We define a harmful meme detection dataset as a set of memes where each meme is $M = (\mathcal{I}, y)$, $\mathcal{I}$ denotes the meme image, and $y \in \{harmful, harmless\}$ denotes the ground truth label. The exact label definitions may vary across datasets. Due to the increasing scarcity of high-quality labeled data as memes rapidly evolve (Sharma et al., 2022), we adopt a zero-shot setting in which the training set $\mathcal{S}_{\text{train}}$ is used solely as an unlabeled reference set $\mathcal{S}_{\text{ref}}$, and only its images are accessible to support harmful meme detection on a disjoint test set $\mathcal{S}_{\text{test}}$. Our core idea is to detect potentially harmful memes by jointly leveraging multimodal understanding, associative retrieval, and symbolic reasoning. As illustrated in Figure 2, the framework consists of two stages. First, an associative retrieval module retrieves semantically related memes from an unlabeled reference set and treats them as contextual Premises for the input meme. Second, a symbolic reasoning module operates over the retrieved Premises and the target Query to derive a final harmfulness decision along with an interpretable rationale.

3.2 Associative Retrieval

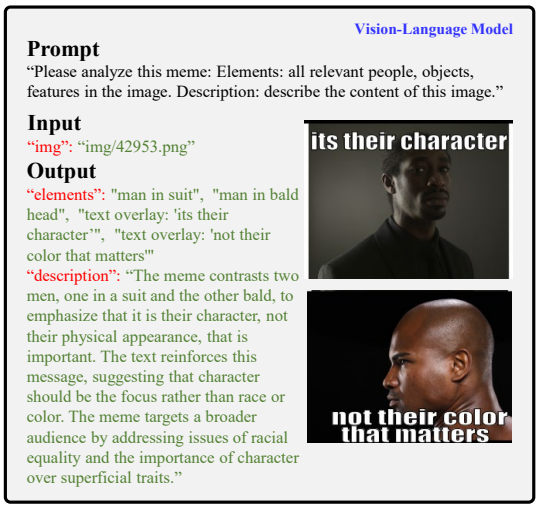

Figure 3: Visualization of the Parser step

As illustrated in Figure 3, our retrieval mech-

anism is built upon a structured abstraction of memes that bridges raw multimodal inputs and symbolic reasoning. We begin by converting each meme into a semantically grounded representation, which enables robust similarity matching beyond surface-level visual or textual overlap. Specifically, we employ a vision-language parser to jointly process the meme image and its overlaid text. The parser identifies salient visual entities (e.g., people, objects, symbols) and textual elements, and further generates a concise natural-language description summarizing the meme's overall semantics, serving as a high-level abstraction of its entities, interactions, and contextual cues across modalities. Formally, given a meme $M = (\mathcal{I}, y)$, we construct a parsed representation

$$M' = (\mathcal{I}, E, D), \tag{1}$$

where $\mathcal{I}$ denotes the raw image, E the extracted semantic elements, and D the generated description.

For each parsed meme M', we define a multimodal embedding function $f(\cdot)$ that maps it to a vector in a shared semantic space, so that conceptually similar memes are close even when their visual appearance or surface text differs:

$$\mathbf{v} = f(\mathcal{I}, E, D) = \lambda_0\, \phi_v(\mathcal{I}) + \lambda_1\, \phi_t(E) + \lambda_2\, \phi_t(D), \tag{2}$$

where $\phi_v(\cdot)$ and $\phi_t(\cdot)$ are frozen vision and text encoders from CLIP, respectively, and we fix $\lambda_0 + \lambda_1 + \lambda_2 = 1$.

To facilitate associative retrieval, we construct a reference memory from an unlabeled reference set $\mathcal{S}_{\text{ref}} = \{M_1, \ldots, M_N\}$. For each reference meme M_i, we obtain its parsed form $M_i' = (\mathcal{I}_i, E_i, D_i)$ and compute

$$\mathcal{V} = \{\mathbf{v}_i = f(\mathcal{I}_i, E_i, D_i)\}_{i=1}^{N}. \tag{3}$$

These embeddings are indexed in a vector database to support efficient similarity-based retrieval. .

Given a query meme M_q, we obtain its embedding $\mathbf{v}_q$ using the same parsing and embedding pipeline, compute its cosine similarity to each reference embedding $\mathbf{v}_i$, and retrieve the top-K most similar memes as follows:

$$s_i = \text{sim}(\mathbf{v}_q, \mathbf{v}_i), \tag{4}$$

$$\mathcal{R} = \{m_i \mid i \in \text{Top-}K(\{s_1, \ldots, s_N\})\}. \tag{5}$$

Finally, for each retrieved meme $m_i \in \mathcal{R}$, we discard the raw image and retain only its symbolic textual components (E_i, D_i) and serve them as contextual premises for the subsequent symbolic reasoning stage.

3.3 Symbolic Reasoning

Building on the associative retrieval stage, this phase integrates the query meme with the retrieved contextual examples to perform structured symbolic reasoning, yielding an interpretable judgment of whether the parsed query meme M' is harmful. Given a formulated question F (e.g., "Is this image harmful?"), we define the combined query as

$$Q = \{M_q', F\}, \tag{6}$$

which is then evaluated over a set of premises

$$P = \{d, m_1, m_2, \ldots, m_n\}, \tag{7}$$

where d denotes the dataset-provided definitions of hateful content, and each m_i corresponds to a retrieved meme from the associative retrieval stage.

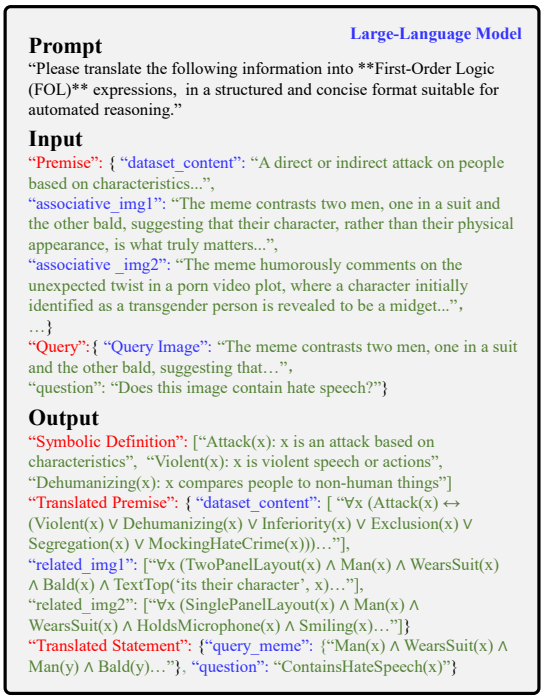

Figure 4: Illustration of the Translator process. Visualizations of the other reasoning steps are provided in the Appendix.

The reasoning pipeline unfolds in three stages: **(1) Translator:** converts multimodal meme representations into structured FOL-style symbolic forms that explicitly capture entities, relations, and contextual cues; **(2) Planner:** constructs a logical reasoning chain that connects the premises P with the query Q, ensuring interpretability at each step; **(3) Solver:** incorporates the original visual modality and executes the reasoning plan using a VLM, preserving visual evidence and mitigating information loss from symbolic abstraction. By

coupling symbolic reasoning with multimodal cues, the framework produces judgments that are both logically consistent and faithful to the original multimodal content.

3.3.1 Translator

As shown in Figure 4, the Translator converts the natural-language premises P and query Q into symbolic forms suitable for logical reasoning, ensuring structural representation across visual, textual, and contextual content for the downstream modules. Formally, the premises are translated into FOL as:

$$P_t = \{d_t, m_{1t}, m_{2t}, \ldots, m_{nt}\}, \tag{8}$$

where the subscript t indicates the translated symbolic form. Similarly, the query is represented as $Q_t = \{M'_t, F_t\}$, which serves as input of Planner.

3.3.2 Planner

The Planner generates a human-interpretable sequence of reasoning steps that connect the premises P to the query Q. To reduce information loss from pure symbolic abstraction, the Planner operates on merged representations of natural language and symbolic forms:

$$P_c = P \oplus P_t, \qquad Q_c = Q \oplus Q_t. \tag{9}$$

where $\oplus$ denotes a merge operator.

Conditioned on (P_c, Q_c), the Planner produces a sequence of intermediate reasoning steps:

$$I = \langle i_1, i_2, \ldots, i_\ell \rangle, \tag{10}$$

where each i_j specifies a instruction that applies premise rules to relevant entities, relations, or textual cues (e.g., "Check entity X for ContainsDiscriminatoryLanguage using rule Y").

The sequence ends with a logical synthesis step, aggregating intermediate results into a final decision according to formal definitions, e.g., a meme is harmful if Offensive $\Leftrightarrow$ ContainsHateSpeech $\vee$ ContainsDiscriminaton $\vee$ ContainsVulgarity. where the atomic predicates follow dataset definitions, if any of the previous checks is true, the meme is labeled harmful. This structured plan ensures reasoning is interpretable, traceable, and executable by the Solver.

3.3.3 Solver

The Solver executes the reasoning plan I under multimodal grounding. Given the merged representations (P_c, Q_c), the plan I, and the target meme image Q_i, each step $i_j \in I$ is processed sequentially, producing a reasoning trace:

$$D = \langle d_1, d_2, \ldots, d_\ell \rangle, \tag{11}$$

where each d_j records the applied premise rule, relevant entities or contexts, and intermediate decisions with supporting evidence. Finally, the trace D is aggregated to issue the overall judgment $J \in \{0, 1\}$, where $J = 1$ indicates that the query meme M_q is harmful. For datasets with multi-class or multi-label annotations, the framework can be extended to predict multiple labels.

By combining symbolic reasoning with multimodal cues, the final judgments are both logically consistent and faithful to the original meme.

4 Experiment

4.1 Experimental Setup

4.1.1 Datasets

We evaluate SYRHM on four widely used harmful meme detection benchmarks. Specifically, we use the Facebook Hateful Memes (FHM) dataset (Kiela et al., 2020) and MultiOff (Suryawanshi et al., 2020) for binary harmfulness classification. We further include the HarM datasets (Pramanick et al., 2021a), comprising HarM-c (COVID-19) and HarM-p (U.S. politics), which support both binary prediction and two fine-grained classification tasks: harmfulness level ({Not Harmful, Somewhat Harmful, Very Harmful}) and target type ({Individual, Organization, Community, Society}). Dataset statistics and prompt details are provided in Appendix B.

4.1.2 Baseline

We benchmark several open-source VLMs in a zero-shot setting, including `LLaVA-1.5-13B` (Liu et al., 2023), `Qwen2.5-VL-7B` (Wang et al., 2024), `LLaVA-v1.6-Vicuna-13B` (Liu et al., 2024), `LLaVA-OneVision-Qwen2-7B`(Li et al., 2024), `InternVL3.5-8B` (Wang et al., 2025) and `DeepSeek-VL-7B` (Lu et al., 2024). We additionally include two recent reasoning-based methods, MinD (Liu et al., 2025) and Evolver (Huang et al., 2025).

4.1.3 Implementation

We implement SYRHM with `Qwen2.5-VL-7B` as the default VLM backbone, which we found to be stable and competitive across all datasets in pilot runs. The *Parser* and *Solver* modules therefore

Method	FHM			Multioff		
	ACC	BACC	MCC	ACC	BACC	MCC
Open-source LMM (Zero-shot)						
LLaVA-1.5-13B	55.01	55.13	10.65	48.15	56.23	14.41
Qwen2.5VL-7B	63.20	63.01	26.53	64.43	59.94	21.70
LLaVA-v1.6-Vicuna-13B	56.90	56.66	13.72	61.07	**65.94**	33.46
LLaVA-OneVision-Qwen2-7B	56.60	56.25	13.37	59.73	60.16	19.81
InternVL3.5-8B	60.40	60.36	20.81	62.42	62.04	23.57
DeepSeek-VL-7B	54.35	54.31	8.63	61.49	61.80	22.97
Reasoning Based Methods						
MinD	63.80	64.21	31.13	61.74	64.30	28.30
Evolver	59.60	59.08	21.32	65.10	58.92	21.70
SyRHM	**66.10**	**65.84**	**32.83**	**69.80**	65.58	**34.03**

Table 1: Results on the binary classification datasets FHM and MultiOff.

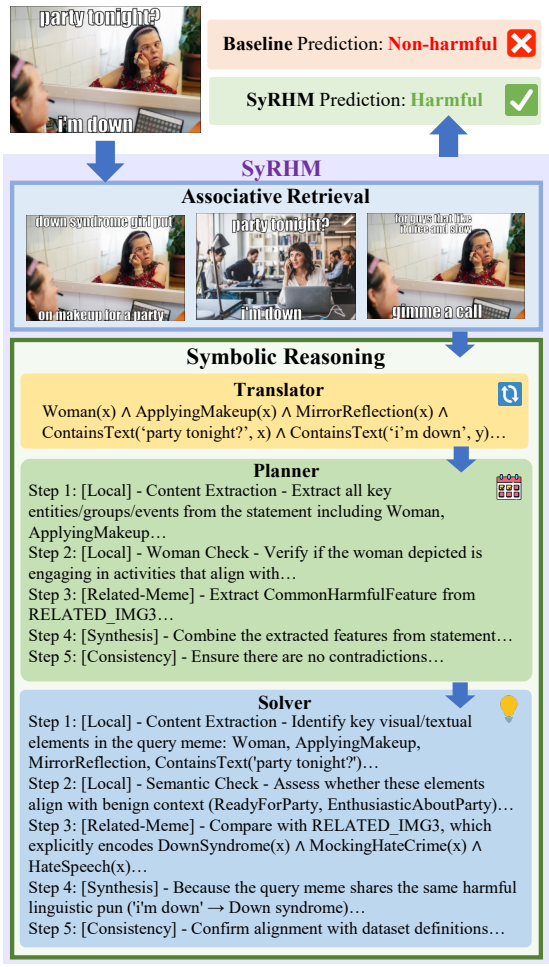

Figure 5: Qualitative examples of SyRHM and the baseline. Additional cases are provided in the Appendix.

use `Qwen2.5-VL-7B` for multimodal understanding. To ensure a fair comparison, we instantiate all reasoning-based baselines with the same VLM backbone. For the LLM-only stages, the *Translator* and *Planner* adopt `Qwen2.5-14B-Instruct` to support symbolic abstraction and structured planning. Unless otherwise specified, all generation stages use a context length of 4096 and temperature 0.2. For associative retrieval, we embed images/elements/descriptions with `CLIP-ViT-B/32` (Radford et al., 2021) in equal ratio and index them in `Milvus` (Guo et al., 2022); we retrieve top-k neighbors with k=3 using equal weights. For reproducibility, we fix random seeds and repeat each experiment three times, reporting the mean performance.

4.1.4 Evaluation Metrics

We report **Accuracy (ACC)**, **Balanced Accuracy (B-ACC)**, and **Matthews Correlation Coefficient (MCC)** as primary metrics to account for overall performance and class imbalance. For the multi-label HarM dataset, we additionally report **Macro-F1** for multi-lable tasks.

4.2 Main Results

To evaluate SyRHM's effectiveness in harmful meme detection, we conduct experiments on the binary-label datasets FHM and MultiOff. As shown in Table 1, SyRHM outperforms prior reasoning-based methods (`MinD` and `Evolver`) across all metrics, achieving 66.10 ACC on FHM and 69.80 ACC on MultiOff. In FHM, it achieves the highest ACC, BACC, and MCC, while in MultiOff it attains the best ACC and MCC and ranks second in BACC, slightly behind LLaVA-v1.6-Vicuna-13B.

We further evaluate SyRHM on the multi-label HarM-c and HarM-p benchmarks. As shown in Table 2, SyRHM achieves the best performance across all tasks, reaching 60.28 ACC, 60.65 BACC, and 21.70 MCC for binary harmful classification

Method	Harmful or Not			Harmfulness Level	Target Type
	ACC	BACC	MCC	M-F1	M-F1
Open-source LMM (Zero-shot)					
LLaVA-1.5-13B	42.94	54.04	11.08	25.83	2.19
Qwen2.5VL-7B	67.23	53.97	18.23	35.92	3.82
LLaVA-v1.6-Vicuna-13B	54.80	62.06	25.26	35.03	2.95
LLaVA-OneVision-Qwen2-7B	68.93	58.99	24.63	36.54	4.52
InternVL3.5-8B	64.41	69.82	38.99	42.86	12.25
DeepSeek-VL-7B	54.80	61.13	22.73	37.90	3.81
Reasoning Based Methods					
MinD	41.53	54.44	15.46	25.31	12.25
Evolver	65.82	52.51	11.20	37.86	4.60
SyRHM	**70.34**	**71.78**	**41.59**	**51.90**	**15.94**

(a) Performance on HarM-c.

Method	Harmful or Not			Harmfulness Level	Target Type
	ACC	BACC	MCC	M-F1	M-F1
Open-source LMM (Zero-shot)					
LLaVA-1.5-13B	51.55	52.85	8.05	24.22	4.88
Qwen2.5VL-7B	54.65	53.13	11.25	35.73	4.75
LLaVA-v1.6-Vicuna-13B	55.77	56.49	14.06	39.77	10.50
LLaVA-OneVision-Qwen2-7B	56.62	55.16	17.29	35.57	5.26
InternVL3.5-8B	55.49	56.61	16.60	35.09	30.36
DeepSeek-VL-7B	53.52	54.27	9.36	33.67	8.24
Reasoning Based Methods					
MinD	50.42	51.99	7.63	27.77	13.92
Evolver	55.21	53.69	13.27	32.59	11.84
SyRHM	**60.28**	**60.65**	**21.70**	**42.88**	**35.96**

(b) Performance on HarM-p.

Table 2: Performance comparison on HarM datasets: (a) HarM-c with binary harmful classification, harmfulness level, and target type tasks; (b) HarM-p with the same tasks.

on the HarM-p benchmark, outperforming prior reasoning-based methods (`MinD` and `Evolver`) by substantial margins. For harmfulness level prediction on the HarM-p benchmark, it attains 42.88 Macro-F1, an improvement of 10.29 points over `Evolver`, and for target type prediction, it achieves 35.96 Macro-F1, far surpassing previous reasoning-based approaches and open-source LMMs. Figure 5 further illustrates how associative retrieval and symbolic reasoning enable our framework to capture implied harmful intent that backbone VLMs often miss. Overall, these results demonstrate that SyRHM consistently excels in both binary and fine-grained multi-label tasks, highlighting its superior reasoning capabilities.

4.3 Ablation Study

4.3.1 Associative retrieval

We study the effect of retrieval size k on SyRHM's performance. As shown in Figure 6, accuracy follows a non-monotonic trend: it drops slightly for small k ($k \leq 2$), peaks at $k=3$ (70.90), remains stable for $k=4$–5, and then gradually declines as k increases further. This suggests that moderate retrieval provides sufficient associative evidence for abstraction, while excessive retrieval introduces noisy or weakly related items that hinder reasoning. Notably, even without retrieval ($k=0$), SyRHM outperforms the `Qwen2.5-VL-7B` baseline by 1.67 points, highlighting the benefit of structured reasoning alone.

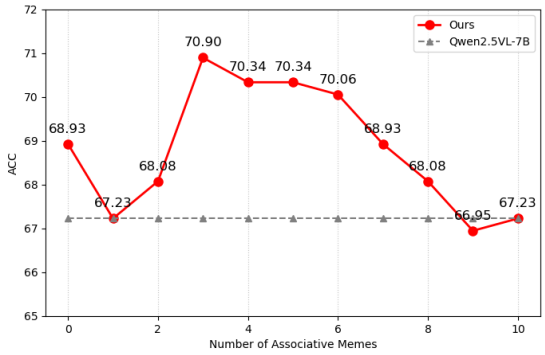

Figure 6: Performance variation with the number of retrieved memes k on HarM.

We further ablate the retrieval weighting across modalities. Table 3 compares five representative settings. Vision-only retrieval (**S1**) performs worst, while combining semantic cues consistently improves accuracy. The best result is achieved by relying solely on non-visual semantics (**S5**), followed by balanced multimodal weighting (**S4**). This suggests that meme harmfulness is primarily conveyed through semantic content, such as targets, relations, and contextual cues, which are better captured by symbolic parses and generated descriptions than by raw visual appearance, highlighting that while visual information can contribute, semantic grounding plays the dominant role in associating memes with harmfulness.

Table 3: Parameter settings for associative retrieval with three modalities. λ_0, λ_1, and λ_2 denote the weights of image, elements, and generated descriptions, respectively.

Setting ID	λ_0	λ_1	λ_2	ACC
S1	1.0	0.0	0.0	67.80
S2	0.5	0.5	0.0	68.93
S3	0.5	0.0	0.5	69.77
S4	0.34	0.33	0.33	<u>70.34</u>
S5	0	0.5	0.5	**71.75**

4.3.2 Effectiveness of Symbolic Reasoning Stages

To assess the contribution of each stage of our reasoning pipeline, we conduct an ablation study by removing one component at a time while keeping all other settings fixed. As shown in Figure 7, the complete model achieves an accuracy of 70.90. Removing the Translator reduces the accuracy to 68.36, while removing the Planner results in an even larger drop to 67.23, indicating that these two modules are the most influential. In contrast, removing the Solver only slightly decreases performance to 70.47. These results suggest that most performance gains come from structured abstraction and explicit planning rather than additional verification. The Planner contributes the largest improvement (+3.67), followed by the Translator (+2.54), while the Solver provides a smaller yet consistent gain by grounding symbolic decisions back to the multimodal input. Overall, combining symbolic representations with explicit planning improves both reasoning completeness and interpretability, leading to more accurate harmfulness judgments.

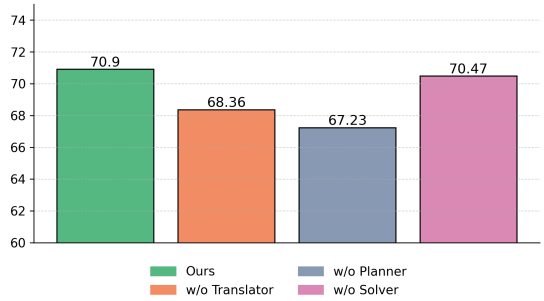

Figure 7: Ablation study: impact of removing individual reasoning stages on model accuracy.

5 Conclusion

We propose SyRHM, a unified framework that combines fine-grained associative retrieval with multi-stage symbolic reasoning to enable context-aware and interpretable harmful meme detection. Unlike existing methods that rely on holistic image–text similarity, SyRHM parses memes into structured semantic elements and textual descriptions, enabling precise retrieval grounded in underlying semantic associations. Building on the retrieved evidence, the framework introduces symbolic intermediate representations and a planner–solver reasoning pipeline to explicitly model cross-meme semantic relations and implicit hostility, providing clear and logically grounded explanations for model decisions. Extensive experiments on multiple standard harmful meme benchmarks demonstrate strong performance of SyRHM in zero-shot detection settings. Beyond harmful meme detection, this paradigm of integrating retrieval, symbolic reasoning, and multimodal understanding offers a general and extensible solution for broader multimodal reasoning and safety-oriented understanding tasks.

6 Limitations

Our paper has the following potential limitations:

1. Symbolic representations may omit nuanced semantic cues such as cultural context or affective intensity. To mitigate this, our framework jointly reasons over symbolic representations and the original natural language descriptions, which increases reasoning context length and inference cost.

2. SyRHM relies on both the availability of semantically related retrieved memes for effective grounding and the base LMM's understanding of social and cultural nuances. Retrieval quality may degrade for novel or emerging meme content, and the model may struggle to identify subtly encoded harmful content or evolving memes that require deep cultural knowledge readily recognized by human moderators.

3. Our evaluation is mainly based on models with 7B parameters. Extending our analysis to other large models may require more computational resources and is left for future work.

7 Ethics Statement

This work aims to mitigate the spread of harmful memes and reduce exposure to abusive or discriminatory content, contributing to safer and more inclusive online environments. The types of harmful content studied in this work, including hate speech and offensive stereotypes, are well-documented concerns in social media research. We acknowledge the potential risk that malicious users may attempt to reverse-engineer or adapt harmful memes to evade automated detection systems based on the proposed framework. Such misuse is strongly discouraged and condemned, as it contradicts the intended purpose of this research. We emphasize that the proposed method is designed to assist, rather than replace human moderation efforts. All experiments in this study are conducted using publicly available research datasets and follow their respective usage agreements. The datasets consist solely of meme images and associated annotations and do not contain personally identifiable user information. The opinions and content expressed in the meme samples do not reflect the views of the authors.

8 Reproducibility Statement

We provide detailed descriptions of the inference process, including all hyperparameters, in Section 4.1.3. The overall pipeline design and prompt formulations are described in Section 3.3 and further elaborated in the Appendix B.2 and C. All datasets used in this work are publicly available and can be accessed online. In addition, we release our code to facilitate reproducibility at: `https://anonymous.4open.science/r/SyRHM-CC46/README.md`.

References

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEe Access*, 12:22359–22375.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, pages 4575–4584.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. *arXiv preprint arXiv:2308.01906*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Rentong Guo, Xiaofan Luan, Long Xiang, Xiao Yan, Xiaomeng Yi, Jigao Luo, Qianya Cheng, Weizhi Xu, Jiarui Luo, Frank Liu, and 1 others. 2022. Manu: a cloud native vector database management system. *Proceedings of the VLDB Endowment*, 15(12):3548–3561.

Ming Shan Hee, Aditi Kumaresan, and Roy Ka-Wei Lee. 2024. Bridging modalities: Enhancing cross-modality hate speech detection with few-shot in-context learning. *arXiv preprint arXiv:2410.05600*.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM web conference 2022*, pages 3651–3655.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, 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.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, pages 1049–1065.

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.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1522–1532.

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. *Advances in neural information processing systems*, 33:2611–2624.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv e-prints*, page arXiv:2012.12871.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

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

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.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347.

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

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the evolution of (hateful) memes by means of multimodal contrastive learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 293–310. IEEE.

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

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael 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.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv e-prints*, page arXiv:2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv e-prints*, page arXiv:2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. 2024. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. *arXiv e-prints*, page arXiv:2412.16953.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Jiacheng Ye, Chengzu Li, Lingpeng Kong, and Tao Yu. 2023. Generating data for symbolic language with large language models. *arXiv preprint arXiv:2305.13917*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The eleventh international conference on learning representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A The Use of LLM

We used a large language model (ChatGPT, OpenAI) solely for English copyediting, including grammar correction, wording and minor stylistic re-writes, and occasional LaTeX formatting help. The model was not used for idea generation, literature search, data collection/annotation, coding, analysis, or producing results. All scientific claims and contributions were written and verified by the authors, and no non-public data were shared with the model. The authors assume full responsibility for the content of the paper.

B Data Details

B.1 Dataset Overview

We study four datasets covering three evaluation tasks. **FHM** and **MultiOff** are annotated only for **Task 1**, whereas **HarM-c** and **HarM-p** additionally support **Task 2** and **Task 3**. Dataset sizes (train/test) are listed in Table 4.

Dataset	Train Samples	Test Samples
FHM	8,500	1,000
MultiOff	445	149
HarM-c	3,013	354
HarM-p	2,938	355

Table 4: Dataset statistics (train/test) used in our experiments.

Per-task label distributions are shown in Table 5 and definitions are as follows:

Task 1 (harmfulness, binary): Decide whether a meme conveys harmful content considering image–text interaction (e.g., slurs, dehumanizing metaphors, threats or disparagement). Labels: 0=*not harmful*, 1=*harmful*.

Task 2 (harmfulness level, ordinal): Grade the severity of harmfulness when present; 1 reflects mild/implicit disparagement or sarcasm, while 2 indicates explicit or strongly implied hostility or incitement. Labels: 0=*not harmful*, 1=*somewhat harmful*, 2=*very harmful*.

Task 3 (influence target type): Identify the primary target of the harmful implication of the meme conditioned on harmfulness. If Task 1 = 0 or Task 2 = 0, set label of Task 3 to 0 (*none*); Labels: 0=*none*, 1=*individual*, 2=*organization*, 3=*community*, 4=*society*.

Dataset	Task 1		Task 2			Task 3				
	0	1	0	1	2	0	1	2	3	4
FHM	510	490	–	–	–	–	–	–	–	–
MultiOff	91	58	–	–	–	–	–	–	–	–
HarM-c	230	124	230	21	103	230	119	0	3	2
HarM-p	184	171	184	159	12	184	81	69	15	6

Table 5: Label distribution for each dataset across the three tasks.

B.2 Dataset Descriptions and Prompt Design

FHM provides a large-scale set of memes. It has 8,500 training examples and 1,000 test examples. and its goal is a binary classification of whether a meme is hateful.

Definition
A meme is considered hateful if it contains a direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. The mocking of hate crimes is also considered hate speech.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"is_hateful": <boolean>}
"is_hateful": Set to true if the image is harmful according to the definition, otherwise false.

Figure 8: Definition and instructions for FHM.

MultiOff contains 445 train and 149 test memes collected from Kaggle (2016 US election) and social media platforms, reflecting real-world political memes and emphasizing both image and text as context.

Definition
A meme is considered offensive if it contains content that is likely to offend, insult, or upset people based on factors such as race, ethnicity, religion, gender, sexual orientation, political views, or other personal characteristics. Offensive content may include offensive speech, discriminatory language, vulgarity, or content that mocks or belittles specific groups or individuals.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"is_offensive": <boolean>}
"is_offensive": Must be `true` if the image is offensive according to the definition, otherwise `false`.

Figure 9: Definition and instructions for MultiOff.

HarM includes two datasets: Harm-C (3,544 memes) and Harm-P (3,552 memes) are real-world meme datasets on COVID-19 and US politics, collected from web and social media, deduplicated and filtered, and annotated by experts for harm intensity and target type (individual, organization, community, society).

Definition
A multimodal unit consisting of an image and embedded text that has the potential to cause harm to an individual, an organization, a community, or society.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"harm_level": <string>, "target_type": <string>}
"harm_level": Must be one of ["not harmful", "somewhat harmful", "very harmful"].
"target_type": Must be one of ["individual", "organization", "community", "society"].
If "harm_level" is "not harmful", then "target_type" should be "none".

Figure 10: Definitions and instructions shared by HarM-c and HarM-p.

Prompt — **Large-Language Model**
"Please derive a step-by-step plan using the First-OrderLogic rule for determining the conclusion based on the context."

Input
"Symbolic Definition": ["Attack(x): x is an attack based on characteristics", "Violent(x): x is violent speech or actions", "Dehumanizing(x): x compares people to non-human things"]
"Translated Premise": { "dataset_content": ["∀x (Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x) ∨ MockingHateCrime(x)))…"],
"related_img1": ["∀x (TwoPanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ Bald(x) ∧ TextTop('its their character', x)…"],
"related_img2": ["∀x (SinglePanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ HoldsMicrophone(x) ∧ Smiling(x)…"]}
"Translated Statement": {"query_meme": {"Man(x) ∧ WearsSuit(x) ∧ Man(y) ∧ Bald(y)…"},
"question": "ContainsHateSpeech(x)"}

Output
"Step by Step Plan": ["Step 1: [Local] - Content Extraction - Parse symbols from input: two male portraits; TextTop('its their character')…",
"Step 2: [Local] - Symbol Mapping - From Symbolic Definition and Premise: Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x)…",
"Step 3: [Local] - Protected Attribute - Detect MentionsProtectedAttr(x, color) from 'color'. Using: Modus Ponens.",
"Step 4: [Related-Meme] - Retrieve related_img* by cues {'character', 'color', 'skin', 'race'}; compute CommonHarmfulFeature as intersection/majority over…",
"Step 5: [Synthesis] - If CommonHarmfulFeature = ∅ and no harmful predicates hold locally, infer ¬Attack(x); else Attack(x)…",
"Step 6: [Decision] - Define ContainsHateSpeech(x) ↔ (Attack(x) ∧ MentionsProtectedAttr(x, _)); output Harm Level and Target Type. Using: Modus Ponens + Consistency Check."]

Figure 11: Illustration of the Planner Stage

C Symbolic Reasoning Module Design

Section 3.3 describes our symbolic reasoning pipeline. Due to space limitations, the main text illustrates only the Translator. Here we present visualizations of the remaining two components: the *Planner* as shown in Figure 11 and the *Solver* as shown in Figure 12.

Prompt — **Vision-Language Model**
"Your task is to carry out a logical solution based on the Step-By-Step Plan and input image."

Input
"img": "img/42953.png"
"Symbolic Definition": ["Attack(x): x is an attack based on characteristics", "Violent(x): x is violent speech or actions", "Dehumanizing(x): x compares people to non-human things"]
"Translated Premise": { "dataset_content": ["∀x (Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x) ∨ MockingHateCrime(x)))…"],
"related_img1": ["∀x (TwoPanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ Bald(x) ∧ TextTop('its their character', x)…"],
"related_img2": ["∀x (SinglePanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ HoldsMicrophone(x) ∧ Smiling(x)…"]}
"Translated Statement": {"query_meme": {"Man(x) ∧ WearsSuit(x) ∧ Man(y) ∧ Bald(y)…"},
"question": "ContainsHateSpeech(x)"}
"Step by Step Plan": ["Step 1: [Local] - Content Extraction - Parse symbols from input: two male portraits…
"Step 2: [Local] - Symbol Mapping - From Symbolic Definition…

Output
"label": 0,
"solution": ["Step 1: [Local] - Content Extraction - TwoPanel; Man(x), Man(y); TextTop('its their character')…",
"Step 2: [Local] - Protected Attribute - MentionsProtectedAttr(m, color) detected (skin color)…",
"Step 3: [Local] - Harm Predicates - EqualityMessage(m) ⇒ ¬Violent(m) ∧ ¬Dehumanizing(m) ∧ ¬Inferiority(m) ∧ ¬Exclusion(m)…",
"Step 4: [Related-Meme] - Majority over related_img* shows equality/anti-discrimination…",
"Step 5: [Synthesis] - Attack(m) ↔ (Violent ∨ Dehumanizing ∨ Inferiority ∨ Exclusion ∨ Segregation ∨ MockingHateCrime)…",
"Step 6: [Decision] - ContainsHateSpeech(m) = Attack(m) ∧ MentionsProtectedAttr(m, color) ⇒ false…"]

Figure 12: Process Visualization of covid_memes_5560 form HarM

D Scaling Language Models

We conducted a controlled study to examine how scaling LLM affects performance within our framework. Here, the VLM is kept fixed as the perceptual backbone, while the LLM responsible for reasoning is varied in size. To further test the generality of our approach, in addition to Qwen2.5-VL-7B used in the main experiments, we also include InternVL3.5-8B, another VLM that performs well in baseline evaluations. Table 6 reports the results on the HarM dataset across all three tasks.

We observe that introducing even a 7B LLM substantially improves performance over the VLM-only baseline, highlighting the effectiveness of symbolic reasoning. While model size has limited impact on the relatively easier binary harmful classification task, larger LLMs consistently yield gains on more fine-grained tasks. In particular, the 14B and 32B models provide notable improvements on harmfulness level prediction and target type identi-

Method		Harmful or Not			Harmfulness Level	Target type
VLM	LLM	ACC	BACC	MCC	M-F1	M-F1
Qwen2.5VL-7B	–	67.23	53.97	18.23	35.92	3.82
Qwen2.5VL-7B	Qwen2.5-7B	67.80	67.78	34.13	47.68	11.02
Qwen2.5VL-7B	Qwen2.5-14B	**70.34**	**71.78**	**41.59**	**51.90**	15.94
Qwen2.5VL-7B	Qwen2.5-32B	68.08	70.79	39.73	46.98	**16.01**
InternVL3.5-8B	–	64.41	69.82	38.99	42.86	12.25
InternVL3.5-8B	Qwen2.5-7B	**68.93**	69.21	36.83	47.63	12.67
InternVL3.5-8B	Qwen2.5-14B	67.80	**72.62**	**43.93**	**50.82**	16.70
InternVL3.5-8B	Qwen2.5-32B	68.36	71.19	40.52	48.63	**17.27**

Table 6: Performance of API-based models on the HarM dataset across three tasks.

fication, indicating that stronger reasoning capacity better supports nuanced inference.

E Cross-dataset Generalization of Associative Retrieval

Table 7: Cross-dataset retrieval results on three tasks.

Dataset	Task1 ACC	Task2 M-F1	Task3 M-F1
FHM	68.93	51.57	12.98
MultiOff	75.14	50.89	14.55
HarM-c	70.34	51.90	15.94
HarM-p	70.34	55.22	12.64

To assess whether associative retrieval generalizes beyond the meme pool used for evaluation, we fix **HarM-c** as the target test set and vary the **source dataset** used to construct the retrieval vector database (FHM, MultiOff, HarM-c, and HarM-p). For each source dataset, we build the vector DB from its training split and retrieve top-k nearest neighbors for each HarM-c test query using the same retrieval pipeline. The results in Table 7 show that performance remains competitive when retrieval is performed from external datasets, and in some cases even surpasses the in-domain retrieval baseline (HarM-c $\rightarrow$ HarM-c), suggesting that the proposed associative retrieval does not rely on accessing the exact same meme pool as the evaluation set and exhibits cross-dataset generalization. We attribute the stronger results of MultiOff and HarM (HarM-c/HarM-p) sources to domain overlap: these datasets contain a substantial content related to politics, which aligns well with HarM-c, whereas political memes constitute only a relatively small portion of FHM, leading to less targeted retrieval evidence under cross-dataset search.

F Case Study

To provide a more concrete understanding of how our SyRHM framework operates, we present several representative examples that illustrate the step-by-step process of meme analysis, including both the associative retrieval and symbolic reasoning stages. These case studies are intended to complement the quantitative results by showing how the model leverages context, prior examples, and logical reasoning to reach accurate classifications.

Figure 13 shows a meme depicting a man shearing a sheep with the text “aye, tone! why would dey delouse ’em if dey was just gonna whack ’em?!”, alongside retrieved memes featuring sheep in humorous or absurd situations. The baseline model misclassifies it as harmful, likely focusing on violent imagery, but SyRHM correctly predicts it as non-harmful. By extracting entities (man, sheep, shear tool) and reasoning about the humorous context, local content, and related meme patterns, SyRHM identifies that the text and imagery do not constitute hate speech, leading to a consistent and correct non-harmful classification.

Figure 14 presents a political meme showing Trump with the text “TRUMP’S ADMINISTRATION ANNOUNCED 5 MILLION TEST KITS TO BE DELIVERED IN A MONTH — A MONTH TOO LATE,” alongside related images highlighting public frustration and sarcasm about COVID-19 testing. The baseline model mislabels it as non-harmful, while SyRHM correctly classifies it as harmful. By extracting entities (Trump, administration, test kits, COVID-19), analyzing the critique and sarcasm, and cross-referencing related memes for accusatory or blame-framing signals, SyRHM determines that the meme constitutes harmful content targeting an individual, producing an accurate and interpretable result.

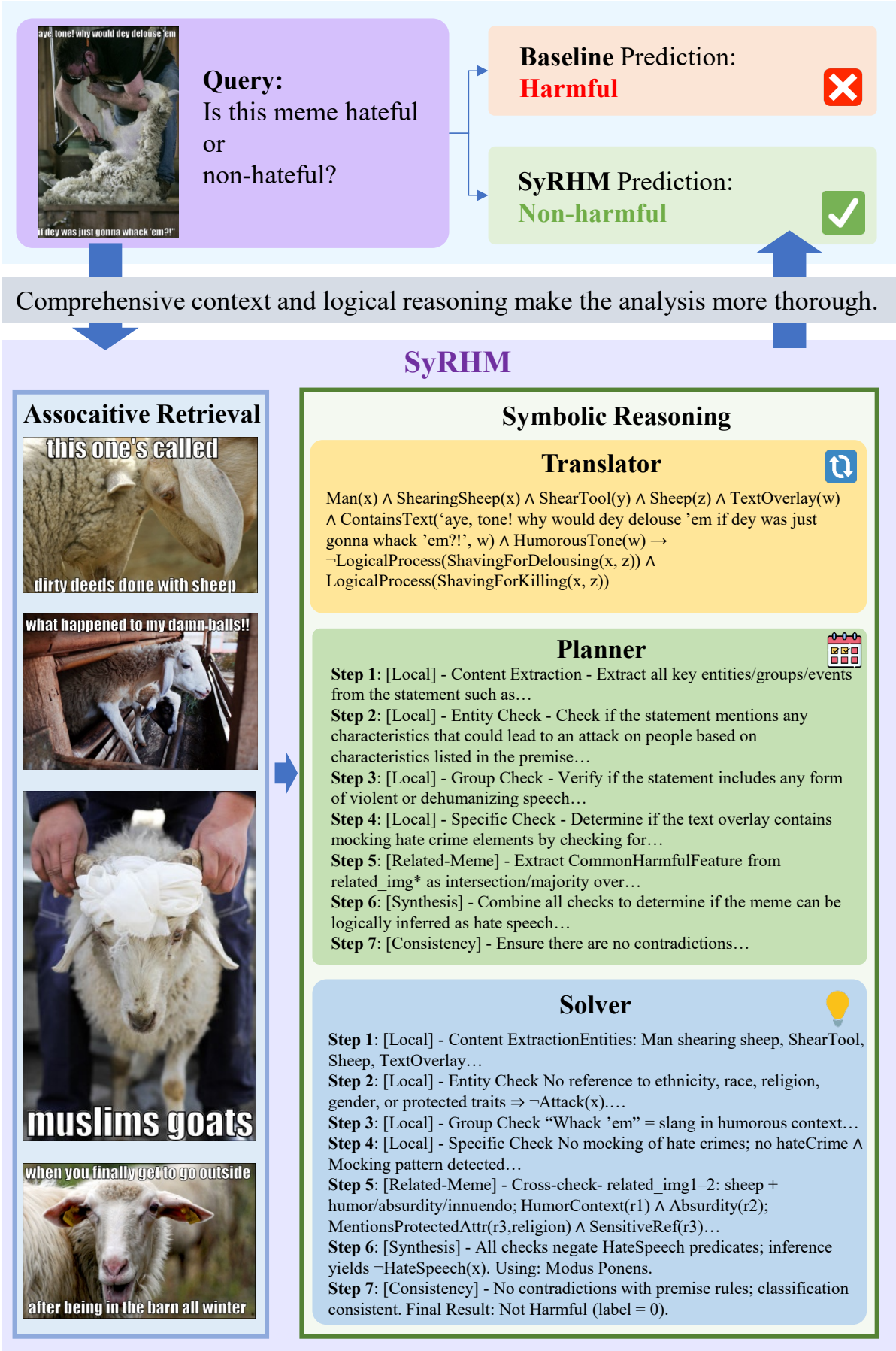

Figure 13: Process Visualization of 01634 form FHM

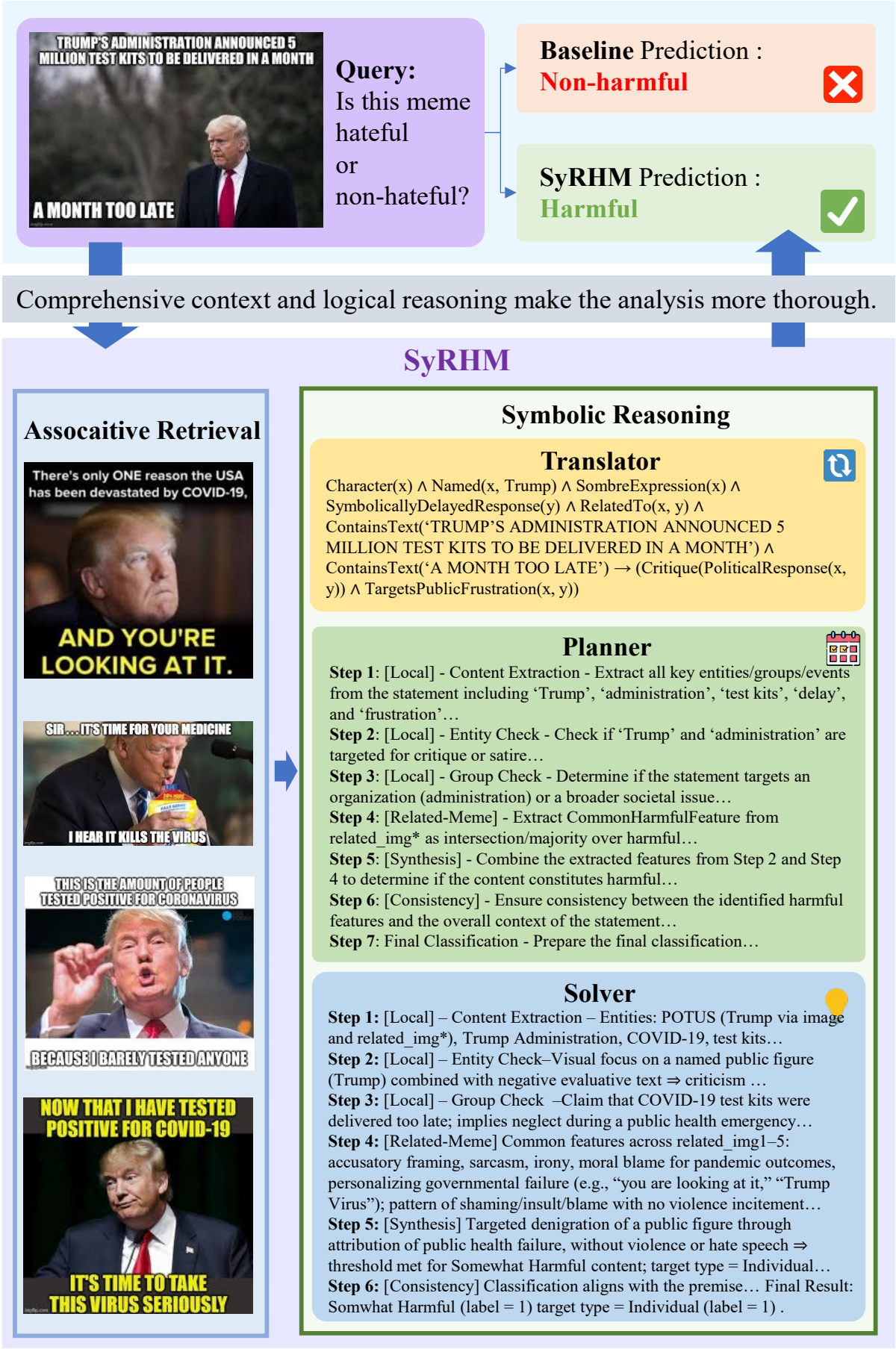

Figure 14: Process Visualization of covid_memes_5560 form HarM