Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection

Jinfa Huang[1]*, **Jinsheng Pan**[1]*, **Zhongwei Wan**[2], **Hanjia Lyu**[1], **Jiebo Luo**[1]
[1]University of Rochester [2]The Ohio State University
{jhuang90, jpan24, hlyu5}@ur.rochester.edu,
wan.512@osu.edu, jluo@cs.rochester.edu
`https://github.com/inFaaa/Evolver`

Abstract

Hateful memes continuously evolve as new ones emerge by blending progressive cultural ideas, rendering existing methods that rely on extensive training obsolete or ineffective. In this work, we propose *Evolver*, which incorporates Large Multimodal Models (LMMs) via Chain-of-Evolution (CoE) Prompting, by integrating the evolution attribute and in-context information of memes. Specifically, *Evolver* simulates the evolving and expressing process of memes and reasons through LMMs in a step-by-step manner using an evolutionary pair mining module, an evolutionary information extractor, and a contextual relevance amplifier. Extensive experiments on public FHM, MAMI, and HarM datasets show that CoE prompting can be incorporated into existing LMMs to improve their performance. More encouragingly, it can serve as an interpretive tool to promote the understanding of meme evolution.

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

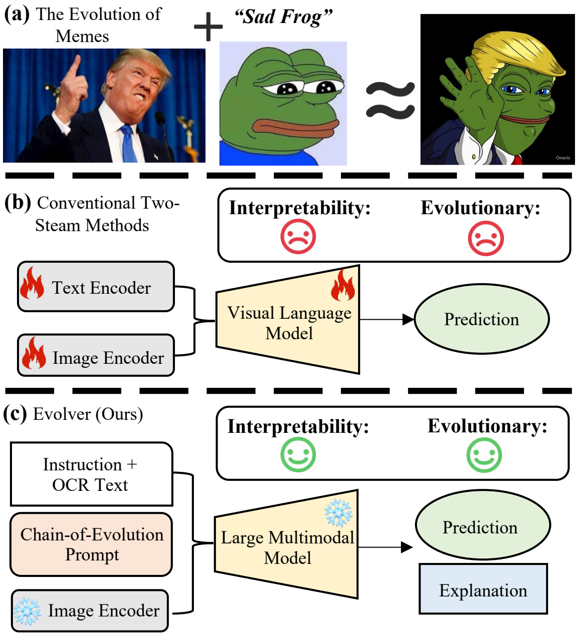


Figure 1: **The illustration of (a) the evolution of memes and comparison between (b) conventional two-stream methods, and (c) our *Evolver* method.** Memes evolve by fusing new cultural concepts. The meme of Trump is influenced by the meme of a sad frog in an image and text symbol, which creates a new hateful meme. Conventional hateful meme detection methods use trainable two-stream encoders and fusion for meme classification, with poor interpretability. In contrast, our *Evolver* captures the evolution and context of memes, utilizing them as prompts for large multimodal models to obtain a comprehensive understanding of memes.

1 Introduction

Hateful meme detection (Lippe et al., 2020; Kiela et al., 2020; Cao et al., 2020) is a crucial task in the field of multimodal research, aiming to identify content that combines text and images to propagate hate speech or offensive messages. Memes, as a widespread cultural phenomenon, proliferate on the Internet, blending images and texts to convey sophisticated meanings. Furthermore, the fusion of visual and textual elements in memes complicates the interpretation of their semantics and the identification of hateful undertones. The combination of text and images presents significant challenges in hateful meme detection, especially for detecting and moderating hateful content (Levine, 2013).

With the advances in image-text pre-training, efforts to leverage and fine-tune CLIP (Radford et al., 2021) for hateful meme detection have demonstrated notable success in achieving high accuracy (Kumar and Nandakumar, 2022; Arya et al., 2024; Hee et al., 2022). However, as shown in Figure 1(a), the extensive evolution of memes fusing together complicates the detection of hateful memes as they continuously evolve, bring new cultural elements to form new expressions. Traditional methods for hateful meme detection, as illustrated in Figure 1(b), suffer from limitations in interpretability and adaptability. These shortcom-

---

[*] Equal contribution.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7321–7330
January 19–24, 2025. ©2025 Association for Computational Linguistics

ings hinder the understanding of memes' evolving nature and their contextual nuances, leading to overfitting on training sets and diminished effectiveness in the dynamic and evolving meme landscape.

To overcome this dilemma and inspired by (Dawkins, 2016b,a) which shows that similar biological groups share some common traits, we introduce *Evolver*, a novel approach that uses Large Multimodal Models (LMMs) (Li et al., 2023a; Lyu et al., 2023; Yang et al., 2023) for hateful meme detection. By developing a benchmark specifically tailored for hateful meme detection based on LMMs, we aim to address the aforementioned limitations of traditional detection methods. Our approach not only enhances detection capabilities but also provides a more explainable and adaptable framework suitable for the evolving landscape of Internet memes. **Evolver** is an innovative approach in zero-shot hateful meme detection significantly. Extensive experiments have shown the effectiveness of our method across three widely recognized hateful meme detection datasets, demonstrating a superior ability to identify and interpret hateful memes.

Overall, the main contributions of this work are:

- We establish an LMM-based zero-shot hateful meme detection benchmark, which provides a comprehensive evaluation of the application of LMMs in social media.
- We propose a simple yet effective *Evolver* framework, which advances LMMs with Chain-of-Evolution prompting. It expands LMMs with an evolution reasoning ability while offering good interpretability.
- Extensive experiments on commonly used zero-shot hateful meme detection datasets with superior performance validate the efficacy and generalization of our method.

2 Approach

2.1 Problem Definition

We define the task of hateful meme detection as a binary classification task. Meme, in our case, consists of an image-text pair. This task can be illustrated as follows:

$$\hat{y}_j = g(\{X_v^j, X_t^j\}) \tag{1}$$

where $\hat{y}_j$ denotes the *j*-th prediction and the prediction $\hat{y}_j \in \{0, 1\}$ indicates the target image-text pair is hateful or not. $g(\cdot)$ is the multimodal model. $\{X_v^j, X_t^j\}$ is the *j*-th image-text paired input.

2.2 Large Multimodal Models

Vision Encoder serves as a translator to help language models understand visual content. It leverages frozen pretrained vision models such as CLIP (Radford et al., 2021) and ViT (Dosovitskiy et al., 2020) to encode visual content so that the language model can understand visual content:

$$h_v = \boldsymbol{W} \cdot \text{Enc}_{vis}(X_v) \tag{2}$$

where h_v is the language embedding tokens. $\boldsymbol{W}$ is the projection to transform visual features into language embedding tokens. $\text{Enc}_{vis}(X_v)$ denotes the visual feature extracted by pretrained model.

Large Language Model Decoder generates a sentence given tokens (Touvron et al., 2023). The process of generation can be represented as:

$$h_t = \text{Tokenize}(X_t) \tag{3}$$

$$p(w) = \prod_{i=1}^{n} p(w_i | w_{<i}, h_t, h_v) \tag{4}$$

where X_t is the input text, $\text{Tokenize}(\cdot)$ transform text into tokens, h_v is the image tokens, $p(w)$ is the probability of generating a sentence by a language model, $p(w_i|w_{<i}, h_t, h_v)$ is the probability of generating a token at the *i*-th position given the previously generated tokens, input text, and image tokens. The visual tokens are incorporated with textual tokens and then fed to the language model.

2.3 Evolver: Chain-of-Evolution Prompting

To improve LMMs' understanding of the online hateful memes which are evolving in nature, we design a novel Chain-of-Evolution (CoE) prompt which has three components: (a) an evolutionary pair mining module to identify most relevant candidate memes; (b) an evolution information extractor to extract key information from the candidate memes; and (c) a contextual relevance amplifier for more effective hatefulness detection.

Evolutionary Pair Mining It is unusual to find memes that have an evolutionary relationship with existing memes. Motivated by Qu et al. (2023), the evolution of a meme is defined as *new memes that emerge by fusing other memes or cultural ideas*. Therefore, the evolution of memes and old memes share similar textual and visual semantic regularities. We leverage this property to identify these old memes given the evolved meme. To this end, we can leverage the evolution of hateful memes to enhance LMMs' ability to understand hateful memes. In the implementation, we first generate the textual and visual embeddings from an external meme pool and target memes with CLIP. Specifically, the external meme pool should have two characteristics (1) do not overlap with the test set. (2) have enough evolutionary information. For simplicity, we use the training set as the carefully curated meme pools rather than any other dataset, where the memes follow the same definition of hatefulness/harm/misogyny. Then, we fuse the textual and visual embeddings with a fixed ratio. Ideally, meme evolution has temporal ability, which means that the memes change with time. However, it is difficult to accurately locate the upstream and downstream of meme evolution with existing technology. Therefore, we use multiple evolutionary memes to find the common characteristics of this evolution. Finally, given a target image-text pair, we retrieve the top-*K* similar memes using cosine similarity:

$$\text{memes} = \{A_i|\cos(\mathbf{A},\mathbf{B})_i \in \text{Top}_K(\cos(\mathbf{A},\mathbf{B}))\} \quad (5)$$

where $\mathbf{A} \in \mathcal{R}^{n\times d}$ is embedding of n candidate memes and $\mathbf{B} \in \mathcal{R}^{d}$ is the d-dimensional vector of the target meme. $\cos(\cdot)$ return a similarity vector and $\text{Top}_K(\cdot)$ returns K highest values given the input vector. For each evolution meme, we pair *K* memes that the evolution meme is derived from.

Evolution Information Extractor To extract the information we are interested in (*e.g.*, *hateful component*), we summarize paired memes with the help of a large multimodal model. The whole process can be expressed as follows:

$$\text{Info} = \text{LMM}([\text{memes}_{top_K}, X_{extract}]) \quad (6)$$

where Info stands for our evolutionary information and LMM indicates the large multimodal model. memes_{Top_K} are the Top_K memes retrieved in the previous step, and $X_{extract}$ is the instruction to guide the LMM to extract information. We present the detailed instruction of $X_{extract}$ as shown in Table 1. More definitions of hateful memes from different datases are shown in Appendix D.

Extract the common harmful feature of these image caption pairs based on the following hatefulness rules:

`Any attacks on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease should be considered hateful. 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. Mocking hate crime is considered hateful.`

Input: [image 0 : <image0>, caption 0 : texts[0], image 1 : <image1>, caption 1 : texts[1], image 2 : <image2>, caption 2 : texts[2], , image 3 : <image3>, caption 3 : texts[3], image 4 : <image4>, caption 4 : texts[4]']

Output: [Here is your response]

Table 1: The example of the instruction for extracting the evolutionary information.

Contextual Relevance Amplifier To enhance the in-context hatefulness information, we add a contextual relevance amplifier to the LMM during evolution information extraction and final prediction. The contextual relevance amplifier can help increase the search for hateful components. In our case, the contextual relevance amplifier is the definition of a hateful meme given by the dataset we use. Finally, we combine the extracted information and contextual relevance amplifier as the in-context enhancement and feed them to the model:

$$\hat{y} = \text{LMM}([\text{memes}_T, X_D, \text{Info}, \text{Amp}]) \quad (7)$$

where Info stands for the information extracted previously, $\hat{y}$ is the final prediction, memes_T is the memes we want to detect, and X_D is the instruction to ask the large multimodal model to detect hateful memes. Amp refers to contextual relevance amplifier. The example of the amplifier is the same as the `blue` part in Table 1.

The Principles of Prompts Our prompt strategy is simple yet effective, based on two key principles: (1) include the hateful meme definition and (2) limit the prompt to 30 words, addressing LMMs' challenges with long-text comprehension. We directly applied definitions from the FHM, MAMI, and HarM datasets, summarized to meet this length requirement with the whole prompt using GPT-4. To this end, we do not use any specialized prompt design, highlighting the robustness of our method.

3 Experiment

3.1 Experimental Setup

Datasets We evaluate *Evolver* on three widely-used public datasets, namely, Facebook Hateful Meme dataset (FHM) (Kiela et al., 2020), Harmful Meme dataset (HarM) (Pramanick et al., 2021), and Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022).

Implementation We implement our **Evolver** based on the MMICL and LLaVA-1.5. For

Methods	Model Size	Dataset: FHM AUC ↑	Dataset: FHM ACC ↑	Dataset: MAMI AUC ↑	Dataset: MAMI ACC ↑	Dataset: HarM AUC ↑	Dataset: HarM ACC ↑
Typical Models (full-Supervised)							
CLIP BERT (Pramanick et al., 2021)	<1B	67.0	58.3	77.7	68.4	82.6	80.8
Text BERT (Kiela et al., 2020)	<1B	66.1	57.1	74.5	67.4	81.4	78.7
Image-Region (Kiela et al., 2020)	<1B	56.7	52.3	70.2	64.2	74.5	73.1
API-based LMM (Zero-shot)							
Gemini-Pro-V (Team et al., 2023)	-	66.0	65.7	74.5	74.5	71.3	76.2
GPT-4V (OpenAI, 2023)	-	70.5	70.3	-	-	-	-
Open-source LMM (Zero-shot)							
Openflamingo (Awadalla et al., 2023)	7B	57.0	56.4	56.8	56.8	51.7	55.8
LLaVA-1.5 (Liu et al., 2023a)	13B	61.8	61.4	57.4	57.4	55.0	54.5
MMICL (Zhao et al., 2023)	11B	59.9	60.4	67.3	67.3	52.1	63.8
MiniGPT-v2 (Zhu et al., 2023)	7B	58.8	59.1	62.3	62.3	57.1	60.3
BLIP-2 (Li et al., 2023b)	11B	56.4	55.8	59.4	59.4	56.8	60.6
InstructBLIP (Dai et al., 2023)	13B	59.6	60.1	64.1	64.1	55.7	60.1
Evolver (Ours)	11B	**63.5**	**63.6**	**68.6**	**68.6**	**67.7**	**65.5**
Evolver† (Ours)	13B	62.3	62.5	59.9	59.9	59.3	57.3

Table 2: **Comparison among different LMMs on zero-shot hateful meme detection benchmarks.** Evolver and Evolver† are MMICL and LLaVA 1.5 with CoE prompting. The API-based models have ethical considerations, we can not directly apply our CoE prompt above them. Best results in the open-source setting are highlighted in **bold**.

EPM	EIE	CRA	Datatset: FHM ACC↑
			55.0
✓			57.3 (+2.3)
	✓		61.7 (+6.7)
	✓	✓	60.6 (+5.6)
✓	✓		63.0 (+8.0)
✓	✓	✓	**63.5** (+8.5)

Table 3: **Ablation study of the three components** of our method on FHM dataset.

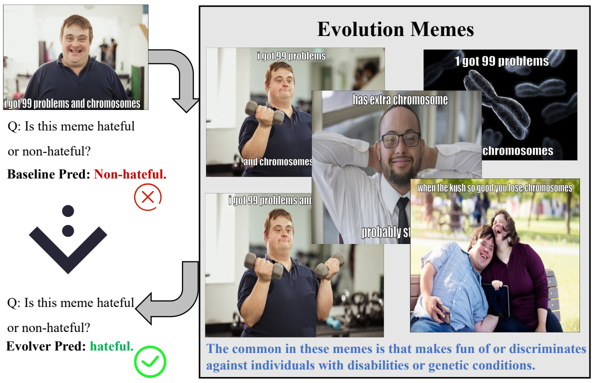

i got 99 problems and chromosomes
Q: Is this meme hateful
or non-hateful?
Baseline Pred: Non-hateful.
Q: Is this meme hateful
or non-hateful?
Evolver Pred: hateful.
Evolution Memes
i got 99 problems
and chromosomes
has extra chromosome
probably st
1 got 99 problems
chromosomes
i got 99 problems and
when the kush so good you lose chromosomes
The common in these memes is that makes fun of or discriminates against individuals with disabilities or genetic conditions.

Figure 2: **Example results of the Evolver (Ours) and the baseline model (MMICL).** For more examples refer to the Appendix.

MMICL, we set the minimum length for the generation to 50, the maximum length for the generation to 80 during the Evolution Info Extractor, and the `temperature` to 0.2. During the final prediction, we set the minimum length of generation to 1, and the maximum length for generation to 50. For LLaVA-1.5, we set `temperature` to 0.2 and maximum generated tokens to 1024 for both stages. The embedding size of textual and visual embeddings is $N \times 768$, and N is the number of memes. We fuse textual and visual embeddings with a fixed ratio of 4:1 by element-wise add in practice.

Baselines The detailed description of the baseline models including API-based (*e.g.*, GPT-4V) and open-source LMMs (*e.g.*, LLaVA) can be found in Appendix A.

Metrics We adopt ACC (accuracy) and AUC (area under the ROC curve) as evaluation metrics.

LMM Backbones We implement our *Evolver* based on MMICL (Zhao et al., 2023) and LLaVA-1.5 (Liu et al., 2023a). Please refer to the Appendix B for more details.

3.2 Main Result

Table 2 highlights the effectiveness of our Chain-of-Evolution Prompting strategy across three datasets. We compare zero-shot results among closed-source LMMs, open-source LMMs, and open-source LMMs with the CoE Prompting strategy. First, we observe that with more evolutionary context, LMMs exhibit a greater ability to recognize hateful content. Notably, both MMICL and LLaVA-1.5 with CoE Prompting outperform their zero-shot baselines, with LLaVA-1.5 achieving a 1.1% improvement and MMICL a 3.2% improvement in accuracy on the FHM dataset. Additionally, within open-source LMMs, model size does not necessarily lead to a better understanding of hateful memes. For instance, on MAMI, MiniGPT-v2 (7B) outperforms LLaVA-1.5 (13B). Furthermore, despite some closed-source LMMs being unavailable for evaluation on certain datasets due to model safety updates, these closed-source models still outperform open-source LMMs in zero-shot results, suggesting a significant gap for open-source models to close. The typical existing training-based models are fully supervised settings, mainly based on CLIP (Radford et al., 2021) and BERT (Devlin, 2018), which can not provide interpretability. However, our setting is a zero-shot setting, mainly including the LMM-based methods, which can explicitly reflect the evolution of memes and training-free. The superior performance of Gemini-Pro-V and GPT-4V is due to their larger parameter count and extensive training data. Furthermore, without security restrictions, our method could be seamlessly integrated into these models.

3.3 Ablation Study

In Table 3, evolutionary pair mining (EPM) boosts the baseline with an improvement of up to 2.3%. Moreover, the evolutionary information extractor (EIE) and contextual relevance amplifier (CRA) significantly improve the generalization ability. Our full model achieves the best performance and outperforms the baseline by 8.5% in accuracy. This demonstrates that three parts of *Evolver* are beneficial by integrating with the evolution information.

3.4 Qualitative Analysis

We visualize the paired evolution memes retrieved by the evolutionary pair mining modules in Figure 2, showing how the evolution information influences the final prediction. The origin model (MMICL) without considering the meme's evolution, predicts non-hateful because it is not easy to detect the boy's disability or genetic conditions. After obtaining the evolution information like "discriminates against individuals with disabilities or genetic conditions.", our *Evolver* rectify the prediction, which supports the rationale of our method.

3.5 Impact of Number of Evolutionary Memes

As shown in Figure 3, we show the effect of the number K of the evolutionary memes. We find that our method significantly improves the baseline under all evolutionary meme number settings. Specifically, we set the $K = 5$ to achieve the best performance in practice.

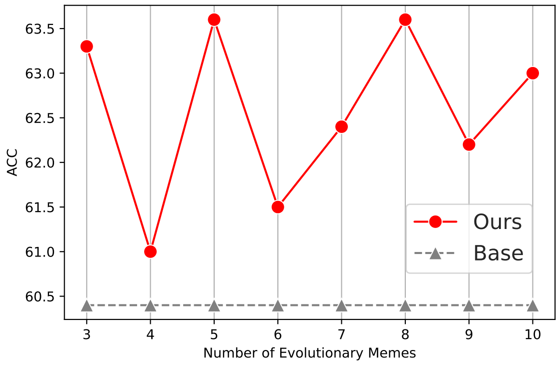


Figure 3: Effect of the number of evolutionary memes.

4 Conclusion

We present *Evolver* to seamlessly boost LMMs for hateful meme detection via Chain-of-Evolution prompting. By integrating evolution of memes, our method can adapt to unseen memes. Experimental results demonstrate the effectiveness of *Evolver*.

Impact Statements

Ethics Statement. While our objective is to mitigate the spread and impact of hate speech online, we recognize the potential for misuse or unintended consequences of this technology. Our framework is designed exclusively for detecting hateful memes, with its use strictly limited to academic or approved research environments. It is not intended for content generation. Insights from this research can support public awareness campaigns, encouraging informed digital citizenship and empowering users to actively participate in recognizing and reporting harmful content. We are committed to ensuring that our technology is used ethically and responsibly.

Reproducibility Statement. We have clarified inference details including hyper-parameters, and the chain-of-evolution pipeline in Section 2.3 and Appendix. In addition, all the datasets used in this paper are open-source and can be accessed online.

Limitation. While *Evolver* demonstrates significant advances in detecting hateful memes by leveraging the evolution of memes and large multimodal models (LMMs), it has some limitations. First, the effectiveness of our approach relies heavily on the quality and diversity of the curated meme pool used for seeking evolutionary memes. Furthermore, biases inherent in these datasets could potentially affect the model's ability to generalize across different cultural contexts and meme evolution patterns not represented in the related data.

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*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Richard Dawkins. 2016a. *The extended selfish gene*. Oxford University Press.

Richard Dawkins. 2016b. *The selfish gene*. Oxford university press.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

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.

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. *arXiv preprint arXiv:2210.05916*.

Marne Levine. 2013. Controversial, harmful and hateful speech on facebook. *Internet: https://www. facebook. com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054 (24.3. 2014)*.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023a. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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 preprint arXiv:2012.12871*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. GPT-4v(ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

OpenAI. 2023. GPT-4v(ision) system card.

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

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 *IEEE Symposium on Security and Privacy (S&P)*. IEEE.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of LMMs: Preliminary explorations with GPT-4v(ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Baseline Details

To ensure a fair comparison, we create the LMM-based zero-shot hateful speech detection benchmark. (a) For API-based LMMs, we select GPT-4V and Gemini-Pro-V, which are among the most popular models. (b) For open-source LMMs, our comparison extends to include six widely used models: Openflamingo, LLaVA-1.5, MMICL, MiniGPTv2, BLIP-2, and InstructBLIP.

GPT-4V (OpenAI, 2023) first creates a GPT-based LMM on the massive volume of datasets on the internet. They trained the language decoder, then aligned it with the vision encoder.

Gemini-Pro-V (Team et al., 2023) curates a large multimodal model with various image-text data on the web and then trains the model with image-text data from scratch.

Openflamingo (Awadalla et al., 2023) builds a LMM using CLIP as vision encoder and MPT/RedPajama as a language decoder. The model is trained on open-source image-text datasets.

LLaVA 1.5 (Liu et al., 2023a) improves the performance of the model they propose previously (Liu et al., 2023b) where CLIP is used as vision encoder and Vicuna as language decoder, with MLP projection to bridge the gap between vision encoder and language decoder.

MiniGPTv2 (Zhu et al., 2023) connects the vision and language space with Q-former and an MLP projection. The model is pre-trained with a large collection of aligned image-text pairs and then instruction fine-tuned on curated high-quality image-text pairs.

BLIP-2 (Li et al., 2023b) propose a model architecture using only Q-former to connect the vision encoder and language decoder to perform a series of multi-modal tasks. The pre-training of model has two stages.

InstructBLIP (Dai et al., 2023) further performs instruction fine-tuning based on BLIP-2 various converted multi-model datasets.

MMICL (Zhao et al., 2023) propose a framework using Q-former and an MLP layer to connect a vision encoder and language decoder.

B Data Details

For more clarity, we show the detailed hateful meme detection dataset analysis in Table 4. Here, we give a detailed description of each dataset.

Dataset	# of Train Samples	# of Test Samples
FHM	8,500	1,000
MAMI	9,948	1,000
HarM	3,013	354

Table 4: Summary of hateful meme detection datasets.

FHM (Kiela et al., 2020) curate a dataset of various hateful memes collected on the web to help build models to detect hateful memes. As shown in Table 4, it has 8,500 training examples and 1,000 test examples. The definition of a hateful meme is:

> "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. Mocking hate crime is also considered hate speech."

Table 5: The definition of hatefulness in the FHM.

MAMI (Fersini et al., 2022) propose a dataset of Misogyny memes, which contains 9,948 training examples as well as 1,000 test examples. The definition of Misogyny given by **MAMI** is:

> "meme is misogynous if it conceptually describes an offensive, sexist, or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group of women. Misogyny can be expressed in the form of shaming, stereotype, objectification, and/or violence."

Table 6: The definition of hatefulness in the MAMI.

HarM (Pramanick et al., 2021) builds a dataset of harmful meme related to Covid. As shown in Table 4, this dataset has 3,013 training samples and 354 test examples. Additionally, **HarM** gives the definition of a harmful meme as follows:

> "Multi-modal unit consisting of an image and an embedded text that has the potential to cause harm to an individual, an organization, a community, or society"

Table 7: The definition of hatefulness in the HarM.

C Computing Resource Requirements

Our Evolver is seamlessly integrated into the LMMs in the inference state, requiring minimal computational resources. For the main results, we conduct experiments on one GTX-3090-24G or one A100-40G. For the ablation study, we measure the comparison on one A100-40G.

D Prompt Engineering

In this section, we show the detailed prompt of Evolution Information Extractor (EIE) for **MAMI** and **HarM** datasets in Table 8 and Table 9.

> Extract the common misogynous feature of these image caption pairs based on the following misogynous rules:
> `meme is misogynous if it conceptually describes an offensive, sexist, or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group of women. Misogyny can be expressed in the form of shaming, stereotype, objectification, and/or violence.`

Table 8: The prompt design for EIE in the MAMI.

> Extract the common harmful feature of these image caption pairs based on the following misogynous rules:
> `meme is harmful if Multi-modal unit consisting of an image and an embedded text that has the potential to cause harm to an individual, an organization, a community, or society.`

Table 9: The prompt design for EIE in the HarM.

Furthermore, we show the detailed prompt of our Evolver in the final prediction for **MAMI** and **HarM** datasets in Table 10 and Table 11.

> Determine if an image <image0> with its caption: ocr_{text} is misogynous or not misogynous. Below is the misogynous analysis of 5 evolutional image text pairs, where our target image text pair is derived from those evolution image text pairs. Use the misogynous of evolution to help you make decision.
> Evolution: <Info>
> Requirement:
> 1. The output should be a word of either ’misogynous’ or ’not misogynous’. 2. meme is misogynous if it conceptually describes an offensive, sexist, or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group of women. Misogyny can be expressed in the form of shaming, stereotype, objectification, and/or violence.
> [Here is your expert response]

Table 10: Prompt design for final prediction of MAMI.

E Related Work

Hateful Meme Detection. Most works of hateful meme detection are based on pretrained vision-language models and fine-tune them on the hateful meme detection data. For example, Kiela et al. (2020) introduced the Hateful Memes Challenge and proposed a multimodal model that combines visual and textual features using a transformer-based architecture. Hate-CLIPper proposed a multimodal contrastive learning approach to improve the representation learning of memes by aligning the visual and textual modalities. These works demonstrate the effectiveness of using pretrained vision-language models. Differently, we discuss the potential of LMMs for hateful meme detection and introduce a knowledge-enhance LMM.

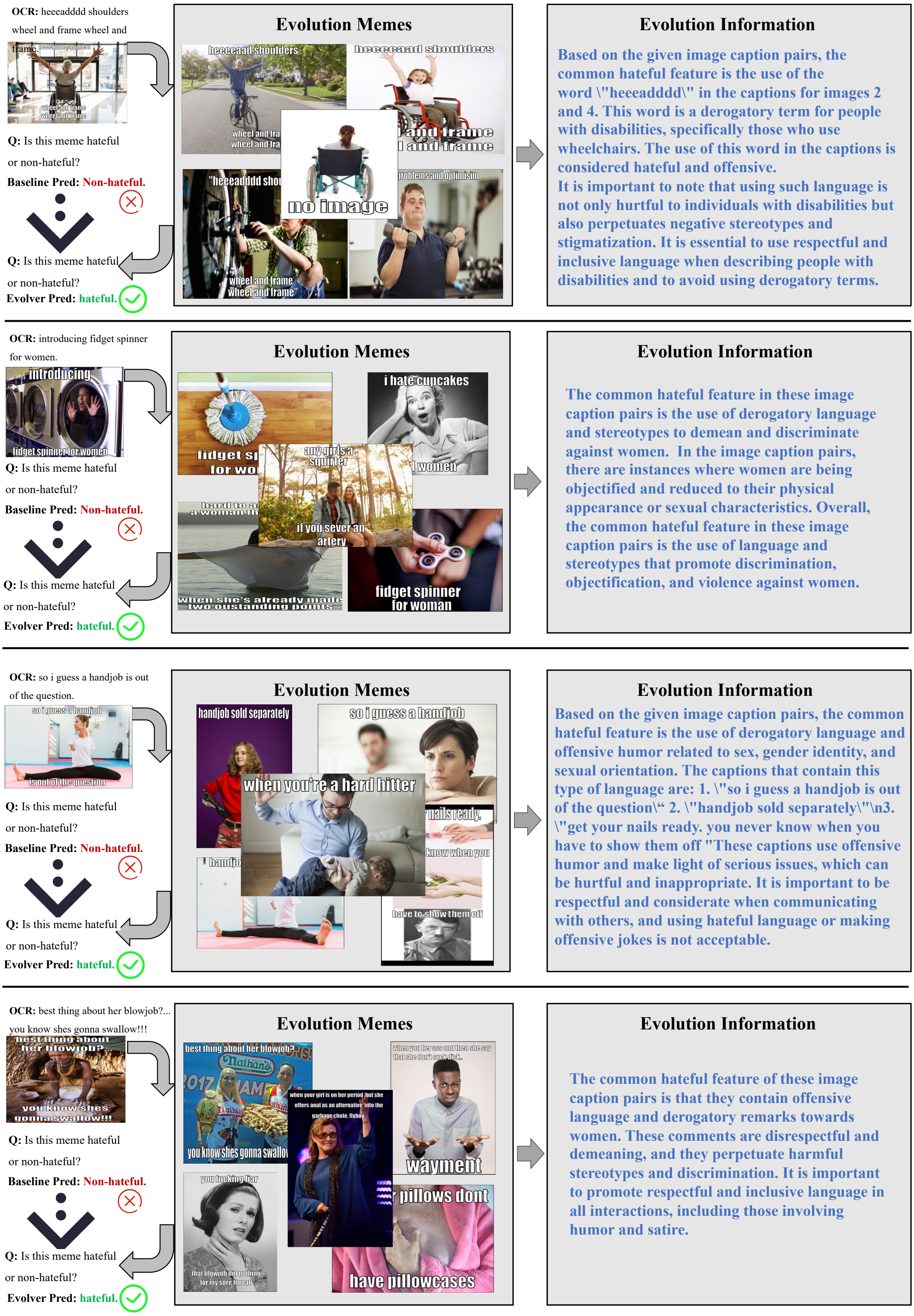
> Determine if an image <image0> with its caption: ocr_{text} is harmful or not harmful. Below is the harmfulness analysis of 5 evolutional image text pairs, where our target image text pair is derived from those evolution image text pairs. Use the harmfulness of evolution to help you make decision.
> Evolution: <Info>
> Requirement:
> 1. The output should be a word of either ’harmful’ or ’not harmful’.
> 2. meme is misogynous if Multi-modal unit consisting of an image and an embedded text that has the potential to cause harm to an individual, an organization, a community, or society.
> [Here is your expert response]

Table 11: Prompt design for final prediction of HarM.

Large Multimodal Models. With the release of the GPT-4V (OpenAI, 2023), many researchers focus on the development and application of LMMs in processing and understanding multimodal data. Discuss groundbreaking models like GPT-4V and LLAVA-1.5 (Liu et al., 2023a), highlighting their capabilities in interpreting complex datasets that include a combination of text and images. Address the challenges these models face, such as sensitivity to noise and typographical errors, and the ongoing efforts to improve their accuracy and robustness in real-world applications. In this work, we adopt the LMMs to hateful meme detection, where set an LMM-based zero-shot hateful meme detection benchmark. Furthermore, we show that our *Evolver* further boosts the LMM’s performance.

Prompt-based Learning. Prompt-based learning has evolved significantly from its inception with large language models (LLMs) to its current application in various tasks. The concept of prompt engineering emerged with the release of GPT-3, which demonstrated that LLMs could be guided to generate desired outputs through carefully crafted prompts. Inheriting the merit from LLM, prompt-based learning in large multimodal models (LLMs) has also gained extensive research attention. In our work, we present a novel chain-of-evolution prompting strategy to seamlessly boost the existing LMM for hateful meme detection.

More Cases Analysis. To show the effectiveness of our method and align a better understanding of hatefulness in FHM dataset, we show more examples generated by *Evolver* and baseline in Fig. 4.



OCR: heeeadddd shoulders wheel and frame wheel and frame.

Q: Is this meme hateful or non-hateful?

Baseline Pred: Non-hateful. ⊗

Q: Is this meme hateful or non-hateful?

Evolver Pred: hateful. ✓

Based on the given image caption pairs, the common hateful feature is the use of the word \"heeeadddd\" in the captions for images 2 and 4. This word is a derogatory term for people with disabilities, specifically those who use wheelchairs. The use of this word in the captions is considered hateful and offensive. It is important to note that using such language is not only hurtful to individuals with disabilities but also perpetuates negative stereotypes and stigmatization. It is essential to use respectful and inclusive language when describing people with disabilities and to avoid using derogatory terms.

OCR: introducing fidget spinner for women.

Q: Is this meme hateful or non-hateful?

Baseline Pred: Non-hateful. ⊗

Q: Is this meme hateful or non-hateful?

Evolver Pred: hateful. ✓

The common hateful feature in these image caption pairs is the use of derogatory language and stereotypes to demean and discriminate against women. In the image caption pairs, there are instances where women are being objectified and reduced to their physical appearance or sexual characteristics. Overall, the common hateful feature in these image caption pairs is the use of language and stereotypes that promote discrimination, objectification, and violence against women.

OCR: so i guess a handjob is out of the question.

Q: Is this meme hateful or non-hateful?

Baseline Pred: Non-hateful. ⊗

Q: Is this meme hateful or non-hateful?

Evolver Pred: hateful. ✓

Based on the given image caption pairs, the common hateful feature is the use of derogatory language and offensive humor related to sex, gender identity, and sexual orientation. The captions that contain this type of language are: 1. \"so i guess a handjob is out of the question\“ 2. \"handjob sold separately\"\n3. \"get your nails ready. you never know when you have to show them off "These captions use offensive humor and make light of serious issues, which can be hurtful and inappropriate. It is important to be respectful and considerate when communicating with others, and using hateful language or making offensive jokes is not acceptable.

OCR: best thing about her blowjob?... you know shes gonna swallow!!!

Q: Is this meme hateful or non-hateful?

Baseline Pred: Non-hateful. ⊗

Q: Is this meme hateful or non-hateful?

Evolver Pred: hateful. ✓

best thing about her blowjob?
you know shes gonna swallow!!!
Evolution Memes
best thing about her blowjob?...
you know shes gonna swallow
when you her ass and then she say that she don't suck dick..
when your girl is on her period but she offers anal as an alternative into the garbage chute, flyboy
wayment
you fucking liar
that blowjob did nothing for my sore throat
pillows dont
have pillowcases
Evolution Information

The common hateful feature of these image caption pairs is that they contain offensive language and derogatory remarks towards women. These comments are disrespectful and demeaning, and they perpetuate harmful stereotypes and discrimination. It is important to promote respectful and inclusive language in all interactions, including those involving humor and satire.

Figure 4: Example results of the Evolver (Ours) and the baseline model (MMICL).