Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, Bill Byrne
Department of Engineering
University of Cambridge
Cambridge, United Kingdom, CB2 1PZ
{jm2245, jc2124, gy266, wl356, wjb31}@cam.ac.uk

Abstract

Hateful memes have become a significant concern on the Internet, necessitating robust automated detection systems. While Large Multimodal Models (LMMs) have shown promise in hateful meme detection, they face notable challenges like sub-optimal performance and limited out-of-domain generalization capabilities. Recent studies further reveal the limitations of both supervised fine-tuning (SFT) and in-context learning when applied to LMMs in this setting. To address these issues, we propose a robust adaptation framework for hateful meme detection that enhances in-domain accuracy and cross-domain generalization while preserving the general vision-language capabilities of LMMs. Analysis reveals that our approach achieves improved robustness under adversarial attacks compared to SFT models. Experiments on six meme classification datasets show that our approach achieves state-of-the-art performance, outperforming larger agentic systems. Moreover, our method generates higher-quality rationales for explaining hateful content compared to standard SFT, enhancing model interpretability. Code available at https://github.com/JingbiaoMei/RGCL

This paper contains content for demonstration purposes that may be disturbing for some readers.

1 Introduction

The rise of social media has led to a surge in hateful content, notably in the form of memes. Manual detection is infeasible due to the vast amount of content and psychological risks for human moderators. Consequently, hateful meme detection systems have attracted considerable research interest (Kiela et al., 2021; Liu et al., 2022; Prakash et al., 2023; Shah et al., 2024).

Large Multimodal Models (LMMs) have emerged as a promising solution for this complex task (Hee et al., 2024b; Lin et al., 2025). Their strong capabilities across a range of general vision-language tasks provide a solid foundation for understanding the intricate interplay between text and image in memes (Zhu et al., 2023; Liu et al., 2023b). Furthermore, the generative nature of LMMs offers interpretability, allowing models to provide rationales for their detection decisions. Ideally, LMMs also bring improved generalizability, enabling them to adapt to the rapidly evolving landscape of online memes, making them well-suited for deployment in real-world content moderation systems. Despite this potential, current LMMs face the following challenges when applied to hateful meme detection.

1. **Sub-optimal performance.** LMMs struggle to learn the interplay of visual and textual cues inherent in hateful memes through standard supervised fine-tuning (SFT), as reported by Mei et al. (2024). We also report that SFT LMMs produce lower-quality rationales when explaining hateful content, possibly caused by overfitting and the scarcity of the training data.

2. **Limited out-of-domain generalization.** Memes constantly evolve with social trends and events, posing a generalization challenge (Cao et al., 2024; Mei et al., 2024). While in-context learning with retrieved examples from a dynamic meme database is a potential approach to generalize to unseen data for LMMs, Huang et al. (2024) found that this approach remains ineffective, highlighting the need for more effective methods to use few-shot meme examples.

3. **Degradation of general vision-language abilities can arise from fine-tuning for meme classification.** We observe that applying SFT for meme classification leads to overfitting, which degrades performance on

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23806–23828
November 4-9, 2025 ©2025 Association for Computational Linguistics

general multimodal benchmarks like MMMU (Yue et al., 2023). This undermines the rationale for choosing LMMs over single-purpose specialized models such as CLIP.

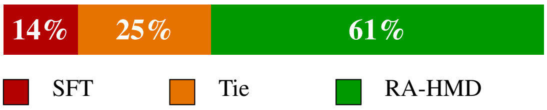


Figure 1: Comparison of rationales generated by SFT and RA-HMD Qwen2-VL-7B models on the HatefulMemes dataset. The bar chart shows the winning rate of rationale quality based on pairwise comparisons between the two models. A more detailed analysis is provided in Appendix L.

To address these challenges, we propose **RA-HMD** (**R**etrieval-**A**ugmented **H**ateful **M**eme **D**etection), a framework that incorporates architectural enhancements and a two-stage fine-tuning strategy to adapt LMMs for hateful meme detection without degradation in general vision-language ability.

We address the **three challenges** of applying LMMs to hateful meme classification through the following contributions:

1. We propose RA-HMD, a fine-tuning framework for adapting LMMs for hateful meme classification, achieving new state-of-the-art results on six widely used meme classification datasets. In addition, RA-HMD generates higher quality rationales compared to SFT models, thereby enhancing the interpretability of LMM predictions, as shown in Figure 1.

2. RA-HMD demonstrates more robust out-of-domain generalization compared to SFT models. Notably, when RA-HMD is combined with a retrieval-augmented KNN classifier, it demonstrates state-of-the-art performance for out-of-domain meme classification. Moreover, this setup enhances robustness against adversarial attacks and leverages few-shot meme examples more effectively than in-context learning, thereby addressing the challenge of adapting to rapidly evolving memes without the need for retraining.

3. RA-HMD expands LMMs ability to perform hateful meme classification and explaining hateful memes without compromising performance on other vision–language tasks, as shown by results in Section 4.4.

2 Related Work

2.1 Hateful Meme Detection

Most existing approaches to hateful meme detection rely on supervised learning, with the majority of research leveraging CLIP (Radford et al., 2021). Numerous studies have fine-tuned models based on CLIP using different modality fusion mechanisms (Pramanick et al., 2021b; Kumar and Nandakumar, 2022; Shah et al., 2024). Other works incorporate caption models into the CLIP-based feature fusion network to further enhance performance (Burbi et al., 2023; Cao et al., 2023; Ji et al., 2024). Additionally, contrastive learning techniques have been explored to address confounding factors in meme classification (Lippe et al., 2020; Mei et al., 2024).

With the emergence of LMMs, recent research has shifted toward using LMMs as generalist models, in contrast to the specialist nature of CLIP based models (Laurençon et al., 2023; Hu et al., 2024). Moreover, decoder-based LMMs offer an additional advantage: they can generate textual rationales to explain why a meme may be hateful (Lin et al., 2024; Hee et al., 2024b).

While LMMs such as Flamingo (Alayrac et al., 2022) have shown promise in hateful meme detection via SFT, fine-tuning strategies for LMMs remain underexplored. In fact, Mei et al. (2024) demonstrated that fine-tuned CLIP models can outperform much larger LMMs, highlighting the need for specialized methods. In this work, we address this gap by proposing LMM architecture refinement alongside a novel fine-tuning approach for LMMs that enhances their performance on hateful meme detection while preserving their general vision–language capabilities.

2.2 Low resource hateful meme detection

Low-resource hateful meme detection is critical for real-world applications that demand out-of-domain generalization. In this setting, an initially trained model is deployed to a new domain without gradient updates, relying only on demonstration examples for inference (Huang et al., 2024). Hee et al. (2024a) utilized retrieved few-shot examples to help LMMs generalize to unseen memes. Hu et al. (2024) and Huang et al. (2024) explored agent-based LMM systems with few-shot learning for out-of-domain settings. However, Huang et al. (2024) observed that in-context learning is less effective for meme classification compared to other tasks, highlighting the need for more effective strategies to use demonstration examples. In this work, we show that RA-HMD improves the in-context learning capabilities of LMMs. Furthermore, when combined with a retrieval-augmented KNN classifier, RA-HMD enables more effective use of demonstration examples than conventional in-context learning.

3 RA-HMD Methodology

3.1 Preliminaries

Problem Statement Hateful memes datasets are defined as $\{(I_i, T_i, y_i)\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{C\times H\times W}$ is the image portion of the meme in pixels; T_i is the caption overlaid on the meme; $y_i \in \{0, 1\}$ is the label, where 0 stands for benign, 1 for hateful.

Large Multimodal Models Some prior work has approached hateful meme detection via text generation with LMMs, where the LMM takes a meme (I_i, T_i) as an input to predict a single token label $\hat{y}_i^{LMH} \in$ {“benign”, “hateful”} (Lin et al., 2024). We refer to the final linear layer of the LMM as the LM Head (**LMH**), which maps hidden representations to a probability distribution over the vocabulary via a softmax function. For meme classification, the LMH decodes the hidden state of the last token and generates the output label. This contrasts with approaches based on CLIP, which train Logistic Regression Classifiers **(LRC)** on encoder CLS tokens (Kumar and Nandakumar, 2022).

3.2 RA-HMD Framework

Architecture enhancement Leveraging representations from large multimodal models (LMMs) for hateful meme classification is non-trivial, particularly when attempting to use LMM embeddings for classification while preserving the model’s original language generation capabilities.

Prior work has explored various strategies for adapting these representations to classification and retrieval tasks. In our study, we similarly experimented with multiple adaptation methods. Appendix D provides a comprehensive summary of these efforts, including failure cases of previous approaches and key insights that ultimately guided the design of RA-HMD.

A central takeaway is that earlier adaptation methods enabled LMMs to perform retrieval but failed to preserve their ability to generate text simultaneously. In contrast, our proposed architecture, combined with a two-stage training procedure, successfully addresses this limitation.

As illustrated in Figure 2, RA-HMD integrates an LMM with two additional trainable components: a Multilayer Perceptron (MLP) that projects the LMM final hidden state $\mathbf{h}_i$ into an embedding $\mathbf{g}_i$ for use in classification and retrieval; and an LRC operating on $\mathbf{g}_i$. Figure 2 shows how the architecture supports multiple fine-tuning and inference modes.

Retrieval During stage-2 training, FAISS-based (Johnson et al., 2021) nearest neighbor search retrieves contrastive learning examples from the encoded meme database $\mathbf{G}$. At inference, FAISS is used to retrieve neighbors for the Retrieval-augmented KNN Classifier (**RKC**).

Inference modes Figure 2 shows three different classifiers: LMH, LRC, and RKC. For pre-trained and SFT LMMs, we generate classification decisions using the LMH as described in Section 3.1. For RA-HMD models, we obtain meme classification decisions via the LRC, unless otherwise specified. Section 4.8 presents a detailed comparison of the three inference modes.

3.3 Stage 1: Logistic Regression Augmented Supervised Fine-tuning

In stage 1, the LMM is fine-tuned via Low-Rank Adaptation (Hu et al., 2022), which applies trainable low-rank matrices to the model while freezing its original weights. The MLP and LRC are updated simultaneously. We optimize the joint loss for each training example i:

$$\mathcal{L}_i^{\text{Stage1}} = \mathcal{L}_i^{LM} + \mathcal{L}_i^{LR}, \tag{1}$$

where $\mathcal{L}_i^{LM}$ is the language modeling objective used in SFT. In the context of meme classification, the model is trained to predict a single target token $s(y_i)$:

$$s(y_i) = \begin{cases} \text{“benign”} & \text{if } y_i = 0 \\ \text{“hateful”} & \text{if } y_i = 1 \end{cases}. \tag{2}$$

$\mathcal{L}_i^{LM}$ is computed as the negative log-likelihood of generating the correct target token, conditioned on the input image and text:

$$\mathcal{L}_i^{LM} = -\log p(\hat{y}_i^{LMH} = s(y_i) \mid I_i, T_i) \tag{3}$$

The $\mathcal{L}_i^{LR}$ is the binary cross-entropy loss applied to the LRC prediction $\hat{y}_i^{LRC}$:

$$\mathcal{L}_i^{LR} = -y_i \log \hat{y}_i^{LRC} - (1 - y_i) \log(1 - \hat{y}_i^{LRC}) \tag{4}$$

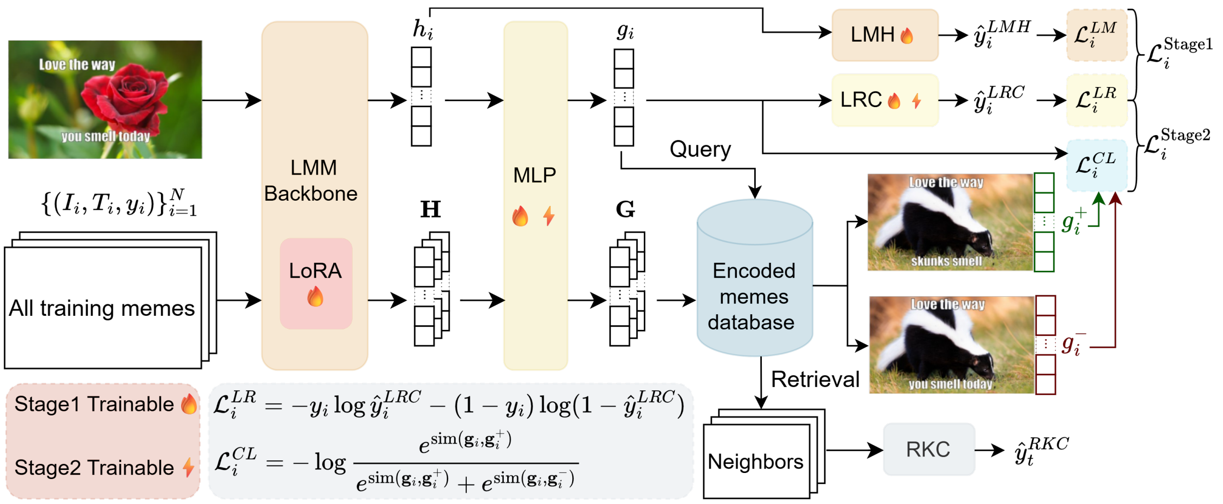


Figure 2: Architecture of RA-HMD. We decompose the LMM into two components: the LMM Backbone and the LM Head (LMH). For each training example i, the last hidden state $\mathbf{h}_i$ is fed to the LMH to obtain the LM loss $\mathcal{L}_i^{LM}$. $\mathbf{h}_i$ is also fed to a trainable multilayer perceptron (MLP) to generate an embedding $\mathbf{g}_i$ for use as a retrieval query and as a feature for the Logistic Regression Classifier (LRC) to compute the cross entropy loss $\mathcal{L}_i^{LR}$. During training, contrastive learning examples are retrieved from the encoded meme database $\mathbf{G}$ for computing the contrastive loss $\mathcal{L}_i^{CL}$. At inference, the same process retrieves the K nearest neighbors for Retrieval-augmented KNN Classification (RKC), which predicts the label $\hat{y}_t^{RKC}$ for an inference example t.

Jointly optimizing the language modeling loss $\mathcal{L}_i^{LM}$ with the cross-entropy loss $\mathcal{L}_i^{LR}$ allows the LMM to rapidly adapt to the hateful meme detection task.

3.4 Stage 2: LMM Contrastive Fine-tuning

In stage 2, the LMM is frozen; only the MLP and LRC are fine-tuned to refine retrieval-aligned representations. Stage 2 jointly optimizes:

$$\mathcal{L}_i^{\text{Stage2}} = \mathcal{L}_i^{CL} + \mathcal{L}_i^{LR}, \tag{5}$$

where $\mathcal{L}_i^{LR}$ is defined in Eq. 4, and $\mathcal{L}_i^{CL}$ is the Contrastive Learning Loss.

To compute $\mathcal{L}_i^{CL}$, we retrieve pseudo-gold positive examples similar to RGCL (Mei et al., 2024) and hard negative examples (Schroff et al., 2015) from the training set. Specifically, for a given sample i with embedding $\mathbf{g}_i$, we use FAISS (Johnson et al., 2021) to perform the nearest neighbor search between $\mathbf{g}_i$ and every other target embedding $\mathbf{g}_j \in \mathbf{G}$ from the training set.

Pseudo-gold positive examples are same-label examples that have high similarity scores with $\mathbf{g}_i$, while hard negative examples are opposite-label examples that have high similarity scores. We denote the embedding of the pseudo-gold positive example and hard negative example as $\mathbf{g}_i^+$ and $\mathbf{g}_i^-$, respectively. $\mathcal{L}_i^{CL}$ is then computed as:

$$\begin{aligned}\mathcal{L}_i^{CL} &= L(\mathbf{g}_i, \mathbf{g}_i^+, \mathbf{g}_i^-)\\ &= -\log \frac{e^{\text{sim}(\mathbf{g}_i, \mathbf{g}_i^+)}}{e^{\text{sim}(\mathbf{g}_i, \mathbf{g}_i^+)} + e^{\text{sim}(\mathbf{g}_i, \mathbf{g}_i^-)}},\end{aligned} \tag{6}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. Stage 2 fine-tuning explicitly aligns the representations of semantically similar meme pairs, thereby improving the generalization of LMMs to distribution shifts in unseen datasets.

3.5 Retrieval Augmented KNN Classification

In addition to the LMH and LRC, RKC is used specifically for out-of-domain meme classification. For a test meme t, we retrieve K similar memes within the embedding space from the meme database $\mathbf{G}$. We perform similarity-weighted majority voting to obtain the prediction:

$$\hat{y}_t^{RKC} = \sigma(\sum_{k=1}^{K} \bar{y}_k \cdot \text{sim}(g_k, g_t)), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and

$$\bar{y}_k := \begin{cases} 1 & \text{if } y_k = 1 \\ -1 & \text{if } y_k = 0 \end{cases}. \tag{8}$$

Additionally, to enable RKC on pretrained or SFT LMMs that do not incorporate an MLP, we use the last hidden state $\mathbf{h}_i$ for the nearest neighbor search. The results are provided in Appendix F.

4 Experiments

We evaluate on six meme classification datasets: **HatefulMemes** (Kiela et al., 2021), **HarMeme** (Pramanick et al., 2021a), **MAMI** (Fersini et al., 2022), **Harm-P** (Pramanick et al., 2021b), **MultiOFF** (Suryawanshi et al., 2020) and **PrideMM** (Shah et al., 2024). A detailed description and statistics are in Appendix A. Implementation details are described in Appendix B.

4.1 Comparing RA-HMD to Baseline Systems under Supervised Settings

Table 1 presents the performance of baseline systems under supervised fine-tuning settings. We compare RA-HMD against a range of strong baselines: the best prior models for each dataset[1]; supervised fine-tuned CLIP-based classifiers; and Large Multimodal Models (LMMs). All models are fine-tuned and evaluated for each dataset separately.

CLIP-based Classifiers We compare the performance of fine-tuned CLIP (Radford et al., 2021) model with two other fine-tuning methods for CLIP-based systems: HateCLIPper (Kumar and Nandakumar, 2022) and RGCL (Mei et al., 2024).

Large Multimodal Models We experiment with three LMMs from two model families: LLaVA-1.5-7B (Liu et al., 2023a), Qwen2-VL-2B and Qwen2-VL-7B (Wang et al., 2024b). We report the performance of these LMMs in the following settings: pre-trained models with zero-shot and few-shot prompts using the LMH; SFT LMMs using the LMH; and classification using LRC under the RA-HMD fine-tuning framework. We further include the results with GPT-4o (OpenAI, 2024a) with optimized prompting for each dataset for reference. For GPT-4o, the token likelihood is not accessible to compute the AUC score.

Best Prior Models Visual Program Distillation (VPD) (Hu et al., 2024) and ExplainHM (Lin et al., 2024) are LLM agent-based systems. The remaining state-of-the-art models, including ISSUES (Burbi et al., 2023), Pro-Cap (Cao et al., 2023), RGCL (Mei et al., 2024) and MemeCLIP (Shah et al., 2024), are based on fine-tuning CLIP-based vision and language models. Detailed descriptions of these methods are provided in Appendix N.

Observation 1: Fine-tuned CLIP-based classifiers outperform baseline LMMs. As shown in Table 1, RGCL (#3) achieves the highest performance among CLIP-based classifiers, surpassing standard fine-tuned CLIP (#1) by approximately 10% across multiple datasets. On 5 out of 6 datasets, RGCL performs better than, or on par with, all three SFT LMMs (#7, #11, #15).

Observation 2: In-context learning exhibits limited efficacy for meme classification. We compare the zero-shot (#5, #9, #13) and few-shot (#6, #10, #14) performance of the pre-trained LMMs. Our findings indicate that, in-context learning does not benefit meme classification, which is consistent with previous results (Hee et al., 2024a; Huang et al., 2024). HarMeme is the only dataset where few-shot systems consistently outperform zero-shot systems. On Harm-P and MultiOFF, although the accuracies of zero-shot and few-shot remain comparable, the few-shot experiments yield a significant gain in F1 score. This improvement is due to a more balanced precision and recall after providing demonstration examples to the system.

Observation 3: RA-HMD outperforms all strong baseline systems across six datasets Across six datasets and three LMMs, fine-tuning with RA-HMD significantly improves performance over SFT (Table 1: #7, #8; #11, #12; #15, #16). Statistical significance tests comparing RA-HMD and SFT further validate these results, with all p-values below 0.05 (see Appendix C). Notably, as indicated in #16, Qwen2-VL-7B fine-tuned with RA-HMD outperforms VPD-PaLI-X-55B on HatefulMemes. Moreover, RA-HMD improves upon RGCL with gains of over 4% in AUC and 3% in accuracy on HatefulMemes. These gains show RA-HMD's effectiveness in improving LMMs for meme classification over SFT.

4.2 Comparing RA-HMD with Baseline Systems under Low-Resource Settings

Online hate speech is constantly evolving, posing a challenge to systems as the distribution of memes encountered in the wild departs from that of the training data. To simulate real-world deployment constraints, we evaluate systems on out-of-domain examples under low-resource settings where gradient updates are prohibited and only demonstration examples are available (Huang et al., 2024; Hee et al., 2024a; Cao et al., 2024).

We adopt a cross-dataset evaluation protocol similar to Mei et al. (2024): models fine-tuned on **HarMeme** are evaluated on **HatefulMemes**, while models trained on **HatefulMemes** are evaluated on

[1] From a recent paper (Nguyen and Ng, 2024); some datasets have been updated with the new best results.

		HatefulMemes		HarMeme		MAMI		Harm-P		MultiOFF		PrideMM	
	Model	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Best prior results		VPD-55B		ISSUES		Pro-Cap		ExplainHM		RGCL		MemeCLIP	
		89.2	80.8	92.8	81.6	83.8	73.6	90.7	90.7	67.1	58.1	76.1	75.1
Supervised fine-tuned CLIP-based Classifiers													
1	CLIP	79.8	72.0	82.6	76.8	77.7	68.4	80.6	80.3	62.4	48.1	72.4	72.3
2	HateCLIPper	85.5	76.1	89.7	84.8	87.2	74.8	87.6	86.9	62.4	54.8	75.5	74.1
3	RGCL	87.0	78.8	91.8	87.0	89.4	78.4	89.9	89.5	67.1	58.1	76.3	76.5
Large Multimodal Models													
4	GPT-4o	-	71.3	-	72.9	-	79.4	63.1	64.5	58.3	58.1	75.3	73.7
	LLaVA-1.5-7B												
5	*w/ zero-shot*	63.7	57.6	71.4	48.6	67.6	58.3	61.6	46.4	59.6	51.7	63.4	65.6
6	*w/ few-shot*	63.4	57.2	73.4	59.6	68.1	62.7	53.5	52.2	38.9	56.0	62.1	64.0
7	*w/ SFT*	85.2	78.7	91.4	79.1	86.0	73.9	82.8	82.8	67.8	57.8	73.2	76.0
8	*w/ RA-HMD*	89.7	80.9	**93.5**	**88.2**	**91.2**	79.7	89.6	89.3	70.9	63.6	**78.1**	**78.7**
	Qwen2-VL-2B												
9	*w/ zero-shot*	64.8	54.2	61.1	56.7	67.2	51.0	53.9	21.8	63.1	36.3	57.8	53.3
10	*w/ few-shot*	61.7	59.1	62.1	65.8	64.8	58.8	53.0	51.6	67.1	44.9	55.4	54.3
11	*w/ SFT*	84.0	76.2	90.2	82.5	77.7	68.6	80.3	79.7	66.4	54.5	73.7	74.2
12	*w/ RA-HMD*	88.4	79.1	92.9	87.7	89.3	79.4	88.9	88.7	68.5	61.8	76.0	76.7
	Qwen2-VL-7B												
13	*w/ zero-shot*	71.9	63.2	64.8	64.1	76.2	58.5	55.5	22.9	63.4	35.9	65.3	62.9
14	*w/ few-shot*	71.5	63.8	71.5	67.2	73.4	66.1	55.6	65.2	64.4	54.7	69.1	56.6
15	*w/ SFT*	86.3	78.6	91.8	85.9	82.6	72.4	85.9	86.3	67.8	55.5	75.1	74.9
16	*w/ RA-HMD*	**91.1**	**82.1**	93.2	88.1	90.4	**79.9**	**91.6**	**91.1**	**71.1**	**64.8**	**78.1**	78.4

Table 1: Comparison with baseline systems under supervised settings. For large multimodal models, we report the pre-trained models zero-shot and few-shot performance (using 4-shot evaluation), along with a comparison between SFT and RA-HMD. Best performance is highlighted in **bold**; second-best is underlined.

all other datasets. This protocol simulates a scenario in which a trained meme classification system is deployed to evaluate trending memes. Few-shot and RKC examples are drawn from the training split of each of the target evaluation datasets to avoid test set contamination.

We compare RA-HMD fine-tuned LMM with the RKC against the following systems: SFT LMMs with zero-shot and few-shot prompting using LMH; GPT-4o (OpenAI, 2024a); specialized low-resource systems (LOREHM (Huang et al., 2024), Mod-hate (Cao et al., 2024)). For GPT-4o, we report results without prompt optimization for each dataset, as this setting assumes the hate type is not known in advance. Further discussion and comparison of GPT-4o results can be found in Appendix K.

Observation 1: Fine-tuning on one memes classification dataset does not help LMMs to improve generalization on other meme classification datasets

Cross-domain fine-tuned LMMs show no consistent improvements over pre-trained LMMs for either zero-shot or few-shot prompting. Qwen2-VL-7B zero-shot (#11 in Table 2) matches its SFT model performance (#13 in Table 1) on HatefulMemes and PrideMM but has performance degradation on the remaining four datasets.

Observation 2: SFT LMMs with in-context learning is ineffective

As shown in Table 2 #6, #9 and #12, the few-shot approach remains similarly ineffective after LMMs are fine-tuned on different domains of hateful meme datasets, offering no significant gains over the SFT zero-shot models (Table 2 #5, #8, #11). In Section 4.6, we further analyze the effect of the number of shots in in-context learning and find that increasing the number of shots does not improve performance. Moreover, in Section 4.7, we show that in-context learning with RA-HMD consistently outperforms SFT models, demonstrating RA-HMD's effectiveness in enhancing out-of-domain generalization.

Observation 3: RA-HMD fine-tuned LMMs

	Evaluated on	HatefulMemes		HarMeme		MAMI		Harm-P		MultiOFF		PrideMM	
	Model	AUC	Acc.	AUC	Acc.	AUC	Acc.	Acc.	F1	Acc.	F1	Acc.	F1
	Low resourced systems												
1	GPT-4o	-	66.4	-	68.4	-	72.9	55.4	55.1	61.1	51.1	63.8	62.3
2	Mod-Hate	64.5	58.0	73.4	69.5	67.4	61.0	-	-	-	-	-	-
3	LOREHM	-	65.6	-	73.7	-	75.4	-	-	-	-	-	-
	Systems fine-tuned under cross-dataset settings												
	Fine-tuning set	HarMeme		HatefulMemes									
4	RGCL	69.9	66.9	64.3	61.1	67.8	62.4	56.4	57.1	53.7	45.1	59.8	61.5
	LLaVA-1.5-7B												
5	*SFT + zero-shot*	63.8	59.4	61.9	48.5	69.1	61.1	55.2	28.7	62.8	32.5	58.1	53.3
6	*SFT + few-shot*	63.1	56.4	69.9	52.8	65.5	50.1	55.6	49.6	56.0	38.9	48.5	55.3
7	*RA-HMD + RKC*	74.2	65.2	**89.5**	**81.9**	80.0	74.5	**67.3**	**67.8**	62.4	51.7	68.8	67.7
	Qwen2-VL-2B												
8	*SFT + zero-shot*	64.1	59.7	61.3	52.2	66.4	57.3	53.5	20.3	62.3	29.3	56.4	59.0
9	*SFT + few-shot*	61.3	53.8	57.4	65.0	73.8	66.0	56.9	55.5	53.7	42.4	55.4	60.9
10	*RA-HMD + RKC*	70.9	62.8	86.0	78.4	74.8	72.3	63.4	66.0	63.4	53.4	69.0	69.1
	Qwen2-VL-7B												
11	*SFT + zero-shot*	71.1	64.1	63.0	55.2	71.1	61.9	54.7	21.5	63.1	29.7	64.5	63.6
12	*SFT + few-shot*	72.3	60.6	67.2	62.4	73.4	66.0	56.4	64.9	62.0	53.7	55.4	60.9
13	*RA-HMD + RKC*	**77.1**	**69.3**	88.8	81.7	**81.4**	**75.6**	64.5	66.4	**63.8**	**55.6**	**69.3**	**69.3**

Table 2: Comparing out-of-domain meme classification performance under low-resource settings. For systems fine-tuned under cross-dataset settings, models are fine-tuned on HarMeme and evaluated on the HatefulMemes dataset. For the remaining evaluation datasets, models are fine-tuned on the HatefulMemes dataset. For LMMs, we compare the SFT models using zero-shot and in-context learning with the RA-HMD fine-tuned models using RKC. Few-shot examples (4-shot) and RKC examples are drawn from the training split of each evaluation dataset. #2 and #3 are taken from the original paper. Best performance is highlighted in **bold**; second-best is underlined.

with RKC inference mode outperforms baseline methods

RA-HMD fine-tuned LMMs using RKC outperform the baseline SFT LMMs in both zero-shot and few-shot settings under the same cross-dataset fine-tuning settings. Notably, RA-HMD trained Qwen2-VL-7B with RKC improves over the baseline SFT few-shot model by 21.6% in AUC and 19.3% in accuracy on HarMeme (Table 2 #11-13). The ablation study in Section 4.6, which varies the number of top k for RKC, further demonstrates that RKC uses demonstration examples more effectively than the few-shot in-context learning framework. Moreover, as shown in Appendix F, applying RKC to both pretrained and SFT LMMs results in worse performance compared to RA-HMD, underscoring the effectiveness of our fine-tuning strategy.

Observation 4: RA-HMD trained LMMs with RKC inference outperform other low resource methods

When compared to other low-resource methods, our RA-HMD fine-tuned LLaVA-1.5-7B with RKC matches the performance of LOREHM on the HatefulMemes dataset. Notably, LOREHM uses a newer and larger LLaVA-1.6-34B within an agent-based framework. Furthermore, our method outperforms LOREHM by 8.2% in accuracy on HarMeme, highlighting our methods' effectiveness under low-resource settings.

4.3 Effects of Two-Stage Fine-tuning

We assess the contribution of each stage within our two-stage RA-HMD fine-tuning process. As shown in Tables 3, omitting either stage leads to performance degradation in both supervised and cross-dataset settings, with Stage 1 contributing more substantial gains.

When only Stage 1 is applied and Stage 2 is omitted, the performance loss is less severe under supervised settings than in cross-dataset evaluations. We attribute this to the contrastive loss in Stage 2, which explicitly optimizes retrieval by aligning representations of semantically similar meme pairs, thereby enhancing robustness to distribution shifts in unseen datasets.

We also compared a variant where we jointly optimize the losses from both stages in a single training phase.

$$\mathcal{L}_i^{\text{Combined}} = \mathcal{L}_i^{CL} + \mathcal{L}_i^{LR} + \mathcal{L}_i^{LM}. \qquad (9)$$

This combined training yields suboptimal results, demonstrating that the two-stage fine-tuning effectively resolves the optimization conflict between task adaptation (Stage 1) and representation alignment (Stage 2). Furthermore, since the LMM remains trainable throughout combined training, updating the encoded meme database incurs significantly higher computational costs compared to the two-stage fine-tuning approach, where the LMM is frozen in stage 2. This staged separation thus enables more efficient training while obtaining stronger performance.

In Appendix H, we also conduct ablation studies by removing individual loss terms within each stage and found that every loss component is essential.

	HatefulMemes		HarMeme	
Mode	AUC	Acc.	AUC	Acc.
RA-HMD	**91.1**	**82.1**	**93.2**	**88.1**
w/ Stage 1 only	90.2	81.4	92.0	86.2
w/ Stage 2 only	84.4	74.2	90.1	85.6
w/ $\mathcal{L}_i^{\mathit{Combined}}$	88.9	77.8	90.2	83.4

(a) Supervised settings, see Table 1 for detailed settings

	HatefulMemes		HarMeme	
Mode	AUC	Acc.	AUC	Acc.
RA-HMD	**77.1**	**69.3**	**88.8**	**81.7**
w/ Stage 1 only	74.4	66.7	86.3	78.7
w/ Stage 2 only	72.0	62.1	84.9	78.1
w/ $\mathcal{L}_i^{\mathit{Combined}}$	72.2	65.3	87.5	80.2

(b) Cross-dataset settings, see Table 2 for detailed settings

Table 3: Ablation study of RA-HMD two-stage fine-tuning framework on Qwen2-VL-7B, evaluating the impact of Stage 1 and Stage 2 Fine-tuning. For $\mathcal{L}_i^{\text{Combined}}$, we jointly optimize the three loss objectives from both stages in a single training process as shown in Eq. 9.

4.4 Performance on General Vision-Language Benchmarks

Table 4 compares the pretrained Qwen2-VL-2B with its SFT and RA-HMD variants, both fine-tuned on the HatefulMemes dataset, across three general vision-language benchmarks: MMMU (Yue et al., 2023), SEED-Bench (Li et al., 2023a), and GQA (Hudson and Manning, 2019). Evaluation settings and additional results are provided in Appendix I. The SFT model shows performance degradation across all three benchmarks, while RA-HMD maintains performance comparable to the pretrained model. These results indicate that RA-HMD robustly preserves the general vision-language capabilities of LMMs.

Model	MMMU	SEEDBench	GQA
Qwen2-VL-2B	40.2	72.7	60.4
+*SFT*	39.1	72.1	57.0
+*RA-HMD*	40.4	72.7	60.1

Table 4: Comparison of the pretrained, SFT, and RA-HMD Qwen2-VL-2B models on three general vision-language benchmarks. The SFT and RA-HMD models are fine-tuned on the HatefulMemes dataset.

	HatefulMemes	
Model	**AUC**	**Acc.**
Baseline (From Table 1)		
SFT	86.3	78.6
RA-HMD + *LRC*	91.1	82.1
RA-HMD + *RKC*	90.8	81.8
Under Adversarial Attack		
SFT	80.5 (-5.8)	72.3 (-6.3)
RA-HMD + *LRC*	84.4 (-6.7)	75.5 (-6.6)
RA-HMD + *RKC*	86.8 (-4.0)	76.6 (-5.2)
w/ Augmented DB	**88.4** (-2.4)	**78.4** (-3.4)

Table 5: Comparison of the SFT and RA-HMD Qwen2-VL-7B models on HatefulMemes under adversarial attack. Values in parentheses denote performance drop compared to non-attack.

4.5 Robustness Under Adversarial Attack

We follow Aggarwal et al. 2023 to assess the system's robustness under adversarial attack. Specifically, we adopt the SaltPepper-I-High attack described in their work, which injects white and black pixels across the image in a manner that does not compromise the overall perception of semantic content.

We compare the SFT and the RA-HMD tuned Qwen2-VL-7B systems (corresponding to #15 and #16 in Table 1). As shown in Table 5, RA-HMD consistently outperforms SFT on SaltPepper-I-High–perturbed data, exhibiting less severe performance degradation. Moreover, when the perturbed examples are incorporated into the retrieval database alongside the original ones, robustness improves further. This shows that our RA-HMD retrieval-guided approach offers a simple yet effective way to enhance system robustness against adversarial attacks as soon as these attacks are detected.

4.6 Numbers of Shots and Neighbors

In Appendix E, we ablate the effects of varying the number of shots for few-shot in-context learning and varying the number of top K nearest neighbors for RKC. We find that merely adding more shots does not necessarily improve performance for in-context learning, aligning with the findings of Huang et al. (2024). In contrast, increasing K in RKC leads to steady performance gains, with improvements plateauing around $K = 20$. These results suggest that the RKC inference mode of RA-HMD makes more effective use of demonstration examples compared to in-context learning.

4.7 Comparing Out-of-Domain Generalization Across Model Variants

To further evaluate out-of-domain generalization across different model variants, we compare RA-HMD against both pretrained and SFT models in Appendix F, following the same protocol as in Table 2. Our results show that RA-HMD consistently outperforms both baselines under both the in-context learning framework and the retrieval-augmented KNN classifier (RKC), highlighting its robustness and effectiveness for generalizing to unseen meme distributions.

4.8 Comparing Different Inference Modes

We compare RA-HMD using three classifiers: LMH, LRC and RKC in Appendix G. Under supervised settings, the performance is similar. However, in cross-dataset scenarios, RKC outperforms both LMH and LRC, underscoring its superior effectiveness in handling out-of-domain examples.

4.9 Comparing Rationales Generated by Models

We compare meme explanations generated by the SFT and RA-HMD fine-tuned Qwen2-VL-7B models (Table 1 rows 15 and 16) on the validation set of the HatefulMemes dataset, where human-annotated rationales are available as ground-truth references (Hee et al., 2023). RA-HMD is evaluated after Stage-1 fine-tuning, since Stage-2 does not fine-tune the Language Model Head for language generation.

Following prior work (Yang et al., 2023), we evaluate explanation quality using two approaches based on LLM-as-judge:

- pair-wise comparison,
- rubric-based evaluation.

The LLM judge measures how closely model-generated explanations align with human rationales. Full details of the evaluation protocol are provided in Appendix L.

The pair-wise comparison results are:

- RA-HMD beats SFT: 61.5
- RA-HMD ties SFT: 13.8
- SFT beats RA-HMD: 24.7

and these results are visualized in Figure 1.

For the rubric-based evaluation (scored on a scale of 0–10), the SFT baseline achieved an average score of 4.9, while RA-HMD obtained a higher score of 5.6, further indicating stronger alignment with human rationales.

Based on the analysis of the generated explanations, we find that improvements in classification accuracy are supported by a deeper semantic understanding of memes. For challenging examples, where comprehension of background events or fine-grained details from the image is required, the RA–HMD fine-tuned system generates explanations that are both more accurate and more informative. Nevertheless, with an average score of 5.6, the performance remains far from perfect. Further discussion of this aspect is provided in Limitations.

4.10 Demonstration examples

In Appendix M, we present a case analysis comparing the classification results of RA-HMD and SFT. Appendix L provides example rationales generated by each model.

5 Conclusion

We propose RA-HMD, a robust adaptation framework for LMMs tailored for hateful meme classification. Our approach effectively improves both in-domain accuracy and out-of-domain generalization, achieving state-of-the-art results across six meme classification datasets while preserving the general vision-language capabilities of the underlying models.

Limitations

Insufficient Definition of Hate Speech: Hate speech is described using various terminologies, including online harassment, online aggression, cyberbullying, and harmful speech. The United Nations Strategy and Plan of Action on Hate Speech acknowledges that definitions of hate speech can be controversial and subject to debate (Nderitu, 2020).

Similarly, the UK Online Harms White Paper highlights that certain harms may be insufficiently defined (Woodhouse, 2022).

Variation in the Definition of Hate Speech: We acknowledge that the definition of hate speech can be subjective and varies across different cultural and legal contexts. To this end, we evaluate our methods on six widely used meme classification datasets, allowing for generalization across different definitions of hate speech. As the discourse on defining hate speech evolves, we align our research with this ongoing process and plan to incorporate new datasets as they become available.

Fine-grained Vision Understanding: In our error analysis, we find that the system is unable to recognize subtle visual details in memes. Enhancing image understanding through a more powerful vision encoder could further improve performance, which we leave for future work.

Preference and RL-based Tuning Methods for Reasoning: In this work, our analysis and methodology primarily build on SFT, which may partly explain the limitations in reasoning quality. Future research could benefit from integrating reasoning capabilities from emerging models such as OpenAI-o1 (OpenAI, 2024b) and DeepSeek-R1 (DeepSeek-AI, 2025). A deeper investigation into preference-based and RL-based tuning methods, including DPO and GRPO, may prove valuable for incentivizing stronger reasoning abilities in LMMs.

Ethical Statement

Reproducibility. Detailed experimental setups, implementation specifics, and hyperparameter settings are provided in Appendix B to ensure reproducibility. The source code can be accessed from GitHub.

Usage of Datasets. The datasets used in this study—HatefulMemes, HarMeme, MAMI, Harm-P, MultiOFF, and PrideMM—were curated for research purposes to combat online hate speech. We strictly adhere to the terms of use established by the dataset authors.

Societal benefits. Hateful meme detection systems, like RA-HMD, can be used to automatically detect hateful content online, contributing significantly to reducing online hate speech. By reducing hate speech, fostering safer digital environments, and supporting human content moderators, these systems can make a significant impact on online communication and safety. We believe these benefits are both substantial and essential in the broader effort to create a more secure and respectful digital space.

Intended use. We intend to enforce strict access controls for model release. The model will be available only to researchers who agree to our terms of use, which explicitly state that the system is designed solely for the detection and prevention of hateful speech. Its use for any purposes that promote, condone, or encourage hate speech or harmful content is strictly prohibited.

Misuse Potential. Although our system is not inherently designed to induce bias, training on existing datasets such as HatefulMemes may inadvertently propagate existing biases towards certain individuals, groups, or entities (Pramanick et al., 2021b). To mitigate the risk of unfair moderation resulting from these dataset-induced biases, it is essential to incorporate human oversight into the moderation process if deployed.

Deployment consideration. Cultural differences and subjective topics introduce biases in moderating online hate speech. Expressions that may seem benign to some can be deeply offensive to others. Our RKC inference mode relies on retrieving examples that generalize well across various domains, allowing the creation of multiple retrieval sets tailored to diverse cultural sensitivities without requiring retraining. However, before deploying such systems, it is crucial to carefully evaluate dataset annotations, particularly when addressing cultural differences and subjective interpretations. Key factors include data curation guidelines, potential annotator biases, and the inherently context-dependent definitions of hate speech. These considerations are essential to ensuring the system is deployed responsibly and effectively across varied cultural contexts.

Environmental Impact Training large-scale models is computationally intensive and contributes to global warming due to heavy GPU usage. However, our approach mitigates this issue by fine-tuning LMMs using quantized LoRA, a parameter-efficient method. As a result, our system can be trained in under four hours on a single consumer-grade GPU RTX 3090, costing less than 1 USD, significantly reducing both training time and computational cost compared to full-scale LMM fine-tuning. Furthermore, since our method generalizes across different domains without requiring retraining, it further minimizes computational overhead.

Acknowledgments

Jingbiao Mei is supported by Cambridge Commonwealth, European and International Trust for the undertaking of the PhD in Engineering at the University of Cambridge.

Jinghong Chen is supported by the Warwick Postgraduate Studentship from Christ's College and the Huawei Hisilicon Studentship for the undertaking of the PhD in Engineering at the University of Cambridge.

Guangyu Yang is supported by Cambridge Commonwealth, European and International Trust for the undertaking of the PhD in Engineering at the University of Cambridge.

Weizhe Lin is supported by a Research Studentship funded by Toyota Motor Europe (RG92562(24020)) for the undertaking of the PhD in Engineering at the University of Cambridge.

Prof. Bill Byrne holds concurrent appointments as a Professor of Information Engineering at Cambridge University and as an Amazon Scholar. This publication describes work performed at Cambridge University and is not associated with Amazon.

We would also like to thank all the reviewers for their knowledgeable reviews that helped us strengthen the contribution.

References

Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023*, page 3734–3743. ArXiv:2302.05703 [cs].

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2824–2828.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5244–5252. Association for Computing Machinery.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, WWW '24, page 4575–4584. Association for Computing Machinery.

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. (arXiv:2305.06500). ArXiv:2305.06500 [cs].

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. (arXiv:2501.12948). ArXiv:2501.12948 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. (arXiv:2305.14314). ArXiv:2305.14314 [cs].

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

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)*, page 533–549, Seattle, United States. Association for Computational Linguistics.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. (arXiv:2305.17678). ArXiv:2305.17678 [cs].

Ming Shan Hee, Aditi Kumaresan, and Roy Ka-Wei Lee. 2024a. Bridging modalities: Enhancing cross-modality hate speech detection with few-shot in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 7785–7799, Miami, Florida, USA. Association for Computational Linguistics.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024b. Recent advances in hate speech moderation: Multimodality and the role of large models. (arXiv:2401.16727). ArXiv:2401.16727 [cs].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9590–9601, Seattle, WA, USA. IEEE.

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*, page 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506.

Junhui Ji, Xuanrui Lin, and Usman Naseem. 2024. Capalign: Improving cross modal alignment via informative captioning for harmful meme detection. In *Proceedings of the ACM Web Conference 2024*, page 4585–4594, Singapore Singapore. ACM.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. (arXiv:2005.04790). ArXiv:2005.04790 [cs].

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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. (arXiv:2306.16527). ArXiv:2306.16527 [cs].

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. (arXiv:2307.16125). ArXiv:2307.16125 [cs].

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2359–2370, New York, NY, USA. Association for Computing Machinery.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2025. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. (arXiv:2401.01523). ArXiv:2401.01523 [cs].

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:2012.12871). ArXiv:2012.12871 [cs].

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. (arXiv:2310.03744). ArXiv:2310.03744 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. (arXiv:2304.08485). ArXiv:2304.08485 [cs].

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, Bangkok, Thailand. Association for Computational Linguistics.

Wairimu Nderitu. 2020. United nations strategy and plan of action on hate speech.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o system card. (arXiv:2410.21276). ArXiv:2410.21276 [cs].

OpenAI. 2024b. Openai o1 system card. (arXiv:2412.16720). ArXiv:2412.16720 [cs].

Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. Totaldefmeme: A multi-attribute meme dataset on total defence in singapore. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, MMSys '23, page 369–375, New York, NY, USA. Association for Computing Machinery.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

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*, page 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, page 8748–8763. PMLR.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. ArXiv:1503.03832 [cs].

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

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*, page 32–41, Marseille, France. European Language Resources Association (ELRA).

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. (arXiv:2409.12191). ArXiv:2409.12191 [cs].

John Woodhouse. 2022. Regulating online harms. *UK Parliament*.

Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. 2024. Grounding language models for visual entity recognition. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XI*, page 393–411, Berlin, Heidelberg. Springer-Verlag.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 5490–5505, Singapore. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. (arXiv:2304.10592). ArXiv:2304.10592 [cs].

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *arXiv preprint arXiv:2403.13164*.

A Dataset details and statistics

Table 6 shows the data split for our evaluation datasets.

HatefulMemes (Kiela et al., 2021) Released by Meta in 2020, HatefulMemes contains 12,000 memes annotated as hateful or benign by trained experts. This benchmark dataset synthesizes memes targeting religion, race, disability, and gender. It includes confounder examples where the benign memes are generated by altering either the image or text to challenge models' ability in multimodal reasoning.

HarMeme and Harm-P HarMeme is a dataset containing approximately 3,000 memes centered on COVID-19 related political memes. A companion dataset, Harm-P (Pramanick et al., 2021b), contains around 3,000 memes related to US politics. Although the original HarMeme was later renamed Harm-C in subsequent work, we adhere to its original name following previous studies (Cao et al., 2022). In HarMeme, memes are annotated into three classes: very harmful, partially harmful, and harmless. Consistent with prior work (Cao et al., 2022; Pramanick et al., 2021b), we merge the very harmful and partially harmful categories into a single hateful class, while treating harmless memes as benign.

MAMI (Fersini et al., 2022) The MAMI dataset focuses on detecting misogynistic memes sourced from various social media platforms, including Twitter and Reddit, as well as meme creation and sharing websites, and even anti-women websites and forums. It contains annotation for two tasks: (1) binary classification of misogyny and (2) categorization of misogyny types. In this work, we address the binary task of identifying whether a meme is misogynistic.

MultiOFF (Suryawanshi et al., 2020) MultiOFF consists of memes gathered from Reddit, Facebook, Twitter, and Instagram, curated specifically for the detection of offensive content. Notably, the training set is extremely small, containing fewer than 500 meme examples. We use this dataset to evaluate the applicability of our methods under ultra low-resource conditions.

PrideMM (Shah et al., 2024) PrideMM contains LGBTQ+-themed memes annotated for four tasks: hate speech detection, hate target identification, topical stance classification, and humor detection. In this work, we use the hate speech classification annotations for hateful meme detection.

Datasets	Train		Test	
	#Benign	#Hate	#Benign	#Hate
HatefulMemes	5450	3050	500	500
HarMeme	1949	1064	230	124
MAMI	4500	4500	500	500
Harm-P	1534	1486	173	182
MultiOFF	258	187	58	91
PrideMM	2581	2482	260	247

Table 6: Statistical summary of HatefulMemes and HarMeme datasets

For HatefulMemes, HarMeme, and MAMI, we report the Area Under the Receiver Operating Characteristic Curve (AUC) and Accuracy (Acc) in line with previous studies (Kumar and Nandakumar, 2022; Cao et al., 2023; Mei et al., 2024; Cao et al., 2024). For Harm-P, MultiOFF, and PrideMM, we report Accuracy and F1 score consistent with the literature (Pramanick et al., 2021b; Mei et al., 2024; Shah et al., 2024; Lin et al., 2024).

To access the Facebook HatefulMemes dataset, one must follow the license from Facebook[2]. HarMeme and Harm-P are distributed for research purposes only, without a license for commercial use. MultiOFF is licensed under CC-BY-NC. MAMI is under Apache License 2.0. There is no specified license for PrideMM.

B Experiment Setup and Implementation Details

Environment. `PyTorch 2.5.1`, `CUDA 12.4`, Huggingface `Transformer 4.45.0` and `Python 3.10.12` were used for implementing the experiments. FAISS (Johnson et al., 2021) vector similarity search library with version `faiss-gpu 1.7.2` was used to perform dense retrieval. All the reported metrics were computed by `TorchMetrics 1.0.1`.

Implementation Details. We use QLoRA (Dettmers et al., 2023) to fine-tune all LMMs, as our experiments show that LoRA and QLoRA perform similarly on this task while significantly outperforming full-parameter fine-tuning. The details for fine-tuning are covered in Appendix B.1. All reported metrics were based on the mean of five runs with different seeds. For statistical

[2] https://hatefulmemeschallenge.com/#download

significance testing, each model is run five times with different random seeds. For baseline models, we strictly follow the settings specified in their original papers.

Implementation environment. We conducted our experiments on a workstation equipped with an NVIDIA RTX 3090.

Run time The run time for RA-HMD two-stage fine-tuning on the HatefulMemes dataset is approximately 4 hours on a single NVIDIA RTX 3090 GPU, and costs around 1 USD.

To optimize efficiency in stage 2, we pre-extract the final hidden states from the frozen LMM and store them on disk before training, avoiding redundant LMM computations. This reduces the stage 2 training time to approximately 10 minutes.

In our ablation study, we examine the performance impact of merging the two-stage loss into a single fine-tuning stage. Since the LMM remains trainable in this setting, we cannot precompute and store the frozen LMM features, leading to significantly higher computational costs. This approach requires approximately 12 hours to complete fine-tuning on a single RTX 3090.

B.1 LLaVA and Qwen2-VL experiments

We freeze the vision module throughout fine-tuning, following the standard LMM fine-tuning protocol. For prompt formatting, we adhere to InstructBLIP (Dai et al., 2023). For LLaVA few-shot experiments, since LLaVA is not explicitly trained to support in-context learning, we follow the procedure outlined by Zong et al. (2024) to enable few-shot learning on LLaVA. For fine-tuning LLaVA (Liu et al., 2023b,a), we follow the original hyperparameters setting[3] for fine-tuning on downstream tasks for both the SFT and RA-HMD stage 1 fine-tuning.

For Qwen2-VL fine-tuning, we employ the officially recommended fine-tuning library `LLaMA-Factory 0.9.1`[4] with official hyperparameter settings for downstream tasks in both the SFT and RA-HMD stage 1 fine-tuning. The only modifications are the LoRA hyperparameters: we employ a larger rank (64) and alpha (128), which are kept fixed throughout all experiments. We fix these LoRA hyperparameters throughout.

Note that in the open-source version on GitHub [5], we update the LLaMA-Factory version to support Qwen2.5-VL fine-tuning for later study, but the reported results in this paper are based on the version noted above. We further provide scripts to run Qwen2.5-VL-7B and Qwen2.5-VL-32B experiments. Surprisingly, both these two models achieve almost identical results compared to the Qwen2-VL-7B model reported in this paper. Given that, we do not further scale up the parameter to 72B for this study.

For few-shot learning with Qwen2-VL, we follow the official multi-round conversation prompt format to ensure consistency with the model's intended usage.

B.2 Hyperparameters for MLP and Stage 2 Fine-tuning

The default hyperparameters for the MLP and the stage 2 contrastive fine-tuning are shown in Table 7. With this configuration of hyperparameters, the number of trainable parameters is about 5 million.

Modelling Hyperparameter	Value
Projection dimension of MLP	1024
Number of layers in the MLP	2
Optimizer	AdamW
Maximum epochs	30
Batch size	64
Learning rate	0.0001
Weight decay	0.0001
Gradient clip value	0.1
Stage 2 Hyperparameter	**Value**
# hard negative examples	1
# pseudo-gold positive examples	1
Similarity metric	Cosine similarity
Loss function	NLL
Top-K for RKC	20

Table 7: Default hyperparameter values

C Statistical Significance Test

We conduct statistical significance tests comparing the performance of SFT and RA-HMD fine-tuned LMMs, as reported in Table 8.

D Insights for using LMMs representation for Meme Classification

There has been substantial interest in adapting decoder-only language models for retrieval and classification tasks. In this section, we summarize the novelty of our approach in comparison to previous efforts that attempt to repurpose decoder-only

[3] https://github.com/haotian-liu/LLaVA

[4] https://github.com/hiyouga/LLaMA-Factory

[5] https://github.com/JingbiaoMei/RGCL

	HatefulMemes		HarMeme		MAMI		Harm-P		MultiOFF		PrideMM	
Model	AUC	Acc.	AUC	Acc.	AUC	Acc.	Acc.	F1	Acc.	F1	Acc.	F1
LLaVA-1.5-7B												
p-value	9.8e^{-3}	3.5e^{-3}	1.2e^{-2}	8.5e^{-3}	4.4e^{-3}	6.2e^{-3}	2.5e^{-3}	1.6e^{-3}	6.1e^{-3}	4.6e^{-3}	5.6e^{-3}	8.9e^{-3}
Qwen2-VL-2B												
p-value	4.2e^{-3}	4.8e^{-3}	9.1e^{-3}	7.6e^{-3}	6.5e^{-4}	1.9e^{-4}	9.3e^{-4}	1.1e^{-3}	2.0e^{-2}	7.7e^{-3}	7.1e^{-3}	8.3e^{-3}
Qwen2-VL-7B												
p-value	8.7e^{-4}	2.4e^{-3}	3.5e^{-2}	1.6e^{-2}	2.5e^{-3}	2.6e^{-3}	6.3e^{-3}	8.6e^{-3}	1.3e^{-2}	5.9e^{-3}	9.2e^{-3}	7.2e^{-3}

Table 8: For each LMM, we provide the p-value from significance testing between SFT and RA-HMD.

large language models (LLMs) or large multimodal models (LMMs) for such tasks.

We categorize our adaptation attempts into two groups: (1) those that entirely failed to work for hateful meme classification, and (2) those that showed some promise in improving classification performance but significantly compromised the model's language generation capabilities. Finally, we explain the rationale that led to the design of our current RA-HMD architecture.

Attempts That Do Not Work

Some prior approaches aim to modify the model architecture with contrastive training objective to enable bidirectional representations from decoder-only models.

For instance, LLM2Vec (BehnamGhader et al., 2024) proposes to introduce bidirectional attention and masked next-token prediction to make decoder-only models suitable for retrieval and general-purpose text embeddings.

AutoVER (Xiao et al., 2024), on the other hand, introduces a special [RET] token and learns its embedding for visual entity retrieval.

However, neither approach yielded meaningful improvements in our setting. When applied to hateful meme classification, both methods achieved classification accuracies only slightly above 60%. We attribute this failure to the limited availability of hateful meme training data, which is insufficient for training these complex adaptations effectively.

Approaches That Partially Work

Other works have explored using decoder-only LLM or LMM embeddings by either pooling the output token representations or using the final-layer embedding of the last token (Zhang et al., 2024; Li et al., 2023b; Wang et al., 2024a), often in combination with contrastive learning.

In our experiments, we found that mean pooling generally underperformed compared to using the final token embedding. While this approach achieved a reasonable classification accuracy of about 77% on the HatefulMemes dataset using Qwen2-VL-7B (RA-HMS Qwen2-VL-7B has an accuracy of 82%), it had a major drawback: the language generation ability of the model was completely compromised. In practice, the model became unable to generate coherent text, suggesting that the learned representations were no longer aligned with the original language modeling objective.

Our approach

To address these limitations, we propose two key improvements that underpin the RA-HMD architecture:

- **MLP Projection Head**: We introduce a lightweight, trainable multilayer perceptron (MLP) on top of the final token embedding from the LMM. While the language model head (LMH) continues to use the original last-token embedding for language generation, the classification and retrieval heads operate on the MLP-projected embedding. This separation enables the model to retain its language generation capability while learning representations that are better suited for classification and retrieval.

- **Preserving the Language Modeling Objective**: During Stage 1 of RA-HMD fine-tuning, we retain the original language modeling loss in the training objective. This encourages the base embeddings to remain useful for text generation, avoiding overfitting solely to the classification task and preserving the model's general-purpose functionality.

E Numbers of Shots and Neighbors

We ablate the effects of varying the number of shots for few-shot in-context learning and varying the number of top K nearest neighbors for RKC.

Figure 3 demonstrates that increasing the number of in-context examples for LMMs does not consistently yield performance improvements over the zero-shot setting, and in some cases even causes loss. These findings suggest that merely adding more shots does not necessarily improve performance, which is consistent with findings from Huang et al. (2024).

Figure 3 shows that as the number of nearest neighbors K for RKC increases, the performance continues to increase for both AUC and accuracy, plateauing at around $K = 20$. The consistent improvement in performance indicates that RKC trained with RA-HMD utilizes demonstration examples more effectively than the standard in-context learning framework.

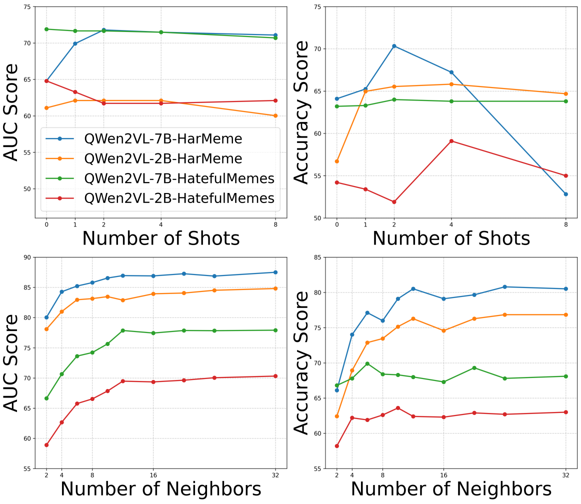


Figure 3: Effects of increasing number of shots for in-context learning with pre-trained LMM and effects of increasing top K nearest neighbors for RKC trained with RA-HMD

F Comparing Out-of-Domain Generalization Across Model Variants with In-Context Learning and RKC

In this section, we compare the performance of the RKC inference mode against few-shot in-context learning for pre-trained LMMs, SFT LMMs, and LMMs fine-tuned using our proposed RA-HMD framework under the cross-dataset setting in Table 9. We observe three things:

- RA-HMD-trained Qwen2-VL-7B exhibits more robust generalization in the cross-dataset setting. Its in-context learning performance in few-shot scenarios consistently surpasses that of other models.
- RKC consistently outperforms the in-context learning approach across all LMM variants, demonstrating its superior ability to leverage demonstration examples.
- Moreover, RA-HMD fine-tuned LMMs with RKC outperform SFT LMMs with RKC, highlighting the effectiveness of our fine-tuning strategy.

Model	Mode	HatefulMemes AUC	HatefulMemes Acc.	HarMeme AUC	HarMeme Acc.
Pre-trained	Few-shot	71.5	63.8	71.5	67.2
Pre-trained	RKC	74.5	64.5	80.1	72.4
SFT	Few-shot	72.3	60.6	67.2	62.4
SFT	RKC	75.8	67.1	84.5	75.4
RA-HMD	Few-shot	74.3	63.5	73.2	68.1
RA-HMD	RKC	77.1	69.3	88.8	81.7

Table 9: Comparing Pre-trained, SFT and RA-HMD systems with few-shot learning and RKC with Qwen2-VL-7B under cross-dataset settings. See Table 2

G Comparing Different Inference Modes

Table 10 compares Qwen2-VL-7B fine-tuned with RA-HMD using the three classifiers. Our results indicate that under supervised settings, the differences among the three inference modes are minimal. However, under cross-dataset settings, there is a significant disparity in generalization performance. Notably, RKC outperforms both LMH and LRC, underscoring its superior effectiveness in handling out-of-domain examples.

Inference Mode	HatefulMemes AUC	HatefulMemes Acc.	HarMeme AUC	HarMeme Acc.
LMH	90.2	81.9	92.8	88.0
LRC	91.1	82.1	93.2	88.1
RKC	90.8	81.8	93.2	88.0

(a) Supervised settings, see Table 1 for detailed settings

Inference Mode	HatefulMemes AUC	HatefulMemes Acc.	HarMeme AUC	HarMeme Acc.
LMH	74.2	64.3	64.5	60.3
LRC	59.5	55.4	57.9	52.2
RKC	77.1	69.3	88.8	81.7

(b) Cross-dataset settings, see Table 2 for detailed settings

Table 10: Comparing different inference modes using RA-HMD fine-tuned Qwen2-VL-7B. RKC shows much better out-of-domain generalization compared to other inference modes.

H Ablation study on the loss function

Table 11 shows the results when each loss objective is removed from different stages of fine-tuning.

Notably, when the cross-entropy loss is removed in stage 1 for the logistic regression component, the LRC fails to train properly via backpropagation, resulting in performance that is equivalent to random guessing. Consequently, we exclude this case from our comparison. Overall, we observe that removing any loss function from the fine-tuning objective leads to a significant drop in performance, highlighting the importance of each loss term in optimizing the model.

Furthermore, Mei et al. (2024) utilize in-batch negative examples alongside hard negative examples during training. However, we find that incorporating in-batch negatives in Stage 2 of RA-HMD's contrastive fine-tuning introduces noise and leads to a slight degradation in performance.

	HatefulMemes		HarMeme	
Mode	AUC	Acc.	AUC	Acc.
RA-HMD	**91.1**	**82.1**	**93.2**	**88.1**
w/o $\mathcal{L}^{LM}$ in stage 1	88.4	79.6	90.9	85.1
w/o $\mathcal{L}^{CL}$ in stage 2	90.2	81.2	91.9	86.4
w/o $\mathcal{L}^{LR}$ in stage 2	89.2	80.6	91.6	87.2

(a) Supervised settings, see Table 1

	HatefulMemes		HarMeme	
Mode	AUC	Acc.	AUC	Acc.
RA-HMD	**77.1**	**69.3**	**88.8**	**81.7**
w/o $\mathcal{L}^{LM}$ in stage 1	75.4	66.6	87.3	81.1
w/o $\mathcal{L}^{CL}$ in stage 2	73.8	64.3	82.9	76.5
w/o $\mathcal{L}^{LR}$ in stage 2	76.4	67.9	86.9	80.6

(b) Cross-dataset settings, see Table 2

Table 11: Ablation study of RA-HMD two-stage fine-tuning framework on Qwen2-VL-7B, evaluating the impact of removing any of the loss objectives.

I Evaluation on General Vision-Language Benchmarks

To evaluate general vision-language capabilities, we use models fine-tuned on the HatefulMemes dataset via both supervised fine-tuning (SFT) and our proposed RA-HMD approach. We conduct the evaluation using the `VLMEvalKit` package[6] (Duan et al., 2024). For MMMU, we report accuracy based on exact match using the `MMMU_DEV_VAL` split to ensure reproducibility. For SeedBench, we use the `SeedBench_IMG` subset, also evaluated with exact match. For GQA, we report results on the `GQA_TestDev_Balanced` split. The full results are provided in Table 12.

Model	MMMU	SEEDBench	GQA
Qwen2-VL-2B	40.2	72.7	60.4
+*SFT*	39.1	72.1	57.0
+*RA-HMD*	40.4	72.7	60.1
Qwen2-VL-7B	49.3	76.4	62.4
+*SFT*	48.0	75.2	61.2
+*RA-HMD*	49.2	76.4	62.3

Table 12: Comparison of the pretrained, SFT, and RA-HMD Qwen2-VL-2B models on three general vision-language benchmarks. The SFT and RA-HMD models are fine-tuned on the HatefulMemes dataset.

J Task-specific fine-tuning and Multi-task fine-tuning

Here, we compare RA-HMD–tuned Qwen2-VL-7B models under two settings: (i) task-specific fine-tuning, where separate models are trained for each dataset, and (ii) multi-task fine-tuning, where a single model is jointly trained on all datasets. As shown in Table 13, multi-task fine-tuning results on Qwen2-VL-7B with RA-HMD suggest that a single system fine-tuned on all available datasets performs comparably to domain-specific systems.

K GPT-4o Results and Prompts

We adopt the `gpt-4o-2024-05-13` model for our evaluations. We explore two prompting strategies for GPT-4o. In Table 1, we present results using optimized prompts, where the specific type of hate (e.g., hatefulness, harmfulness, offensiveness, misogyny, etc.) is explicitly provided based on the characteristics of each dataset. Below, we include the detailed prompt formulations used for each dataset:

- **HatefulMemes**:
  You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains hateful content or not. Please answer with 'yes' or 'no'.

- **Harmeme**:
  You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains harmful content or not. Please answer with 'yes' or 'no'.

- **Harmp**:

[6] https://github.com/open-compass/VLMEvalKit

Model	HatefulMemes	HarMeme	MAMI	Harm-P	MultiOFF	PrideMM
Single-task fine-tuning	82.1	88.1	79.9	91.6	71.1	78.1
Multi-task fine-tuning	82.0	87.8	80.4	91.6	70.4	78.5

Table 13: Accuracy comparison between single-task and multi-task fine-tuning across six datasets.

You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains harmful content or not. Please answer with 'yes' or 'no'.

- **MultiOFF**:
  You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains offensive content or not. Please answer with 'yes' or 'no'.

- **PrideMM**:
  You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains hateful content related to LGBTQ+ Pride movement or not. Please answer with 'yes' or 'no'.

- **MAMI**:
  You are an expert in identifying meme-based social abuse. Given the meme, with the Text [T] accompanied by the image. The task is to classify whether the meme contains misogyny or not. Please answer with 'yes' or 'no'.

The low-resource comparison in Table 2 is designed to reflect real-world scenarios to detect the evolving harmful memes on the internet, where the specific type of hate is often unknown. Accordingly, we use the general term "hate" across all six datasets in this setting. Below, we include the detailed prompt formulations:

- Given the meme, with the Text [T] accompanied by the image. Does the meme contain any hateful content or any social abuse?

We directly present the comparison between the two sets of results in Table 14. For reference, Goat-Bench (Lin et al., 2025) published GPT-4o results on similar datasets using task-specific prompts. However, their evaluation is based on different data splits, making the results not directly comparable. Below, we summarize key differences in performance:

- **Hatefulness**: This benchmark corresponds to our HatefulMemes dataset. While we use the `test_seen` split, Goat-Bench uses the `test_unseen` split. Despite this difference, the results are comparable: they report an accuracy of 71.7, while ours is 71.3.

- **Harmfulness**: This benchmark aligns with our combined evaluation on HarMeme and Harm-P. Goat-Bench reports an accuracy of 66.01, whereas we achieve 72.9 on HarMeme and 63.1 on Harm-P.

- **Offensiveness**: Due to the small size of the MultiOFF dataset, Goat-Bench evaluates performance using the combined training, validation, and test sets. In contrast, we report results based solely on the test set. They report an accuracy of 62.13 and an F1 score of 61.16, compared to our accuracy of 58.3 and F1 score of 58.1.

L Rationale Generated by LMMs

We compare meme explanations generated by the SFT and RA-HMD fine-tuned Qwen2-VL-7B models (Table 1 rows 15 and 16). RA-HMD is evaluated after Stage-1 fine-tuning, as Stage-2 does not fine-tune the Language Model Head for language generation. For evaluation, we prompt the above two models with:

"Does the meme contain any hate speech or offensive content? Please provide a detailed explanation."

Following prior work (Yang et al., 2023), we assess explanation quality using an LLM judge. Specifically, we provide GPT-4o-mini (`gpt-4o-mini-2024-07-18`) with reference explanations from (Hee et al., 2023) and use the comparison template from (Yang et al., 2023) for pairwise evaluation. For reference, the prompt we use is:

	HatefulMemes		HarMeme		MAMI		Harm-P		MultiOFF		PrideMM	
Model	AUC	Acc.	AUC	Acc.	AUC	Acc.	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o (Specific Prompt)	-	71.3	-	72.9	-	79.4	63.1	64.5	58.3	58.1	75.3	73.7
GPT-4o (General Prompt)	-	66.4	-	68.4	-	72.9	55.4	55.1	61.1	51.1	63.8	62.3

Table 14: Comparing the performance with GPT-4o with different prompts

```
Please act as an impartial judge and
evaluate the quality of the model-
generated reasoning provided by two AI
assistants. You will compare the two
model-generated reasoning with the
reference human reasoning for a hateful
meme.

Your evaluation should consider which
response is more similar to the true
answers. Begin your evaluation by
comparing the two responses and provide
a short explanation. After providing
your explanation, output your final
verdict by strictly following this
format: "[[A]]" if assistant A is more
accurate, "[[B]]" if assistant B is more
 accurate, and "[[C]]" for a tie.


[The Start of Assistant A's Answer]
{assistant_a["reasoning"]}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{assistant_b["reasoning"]}
[The End of Assistant B's Answer]}

[The Start of Reference Answer]
{reference}
[The End of Reference Answer]
```

Results are as follows:

- RA-HMD beats SFT: 61.5
- RA-HMD ties SFT: 13.8
- SFT beats RA-HMD: 24.7

Based on the analysis of the generated explanations, we find that improvements in classification accuracy are supported by deeper semantic understanding of memes. For challenging examples, where comprehension of background events or fine-grained details from the image is required, the RA-HMD fine-tuned system generates better explanations.

Additionally, we conduct a rubric-based analysis, following the procedure in Yang et al. 2023. Specifically, we prompted the LLM judge to assign a score from 0 to 10 to each system's rationale, based on how well it aligns with the human-written explanations.

For reference, the prompt we use is:

```
Compare the model-generated reasoning
with the reference human reasoning for
this hateful meme.

Reference: {reference_reasoning}
Model: {model_reasoning}

Rate how well the model reasoning aligns
 with the reference on a scale of 0-10:
- 9-10: Excellent alignment, captures
all key points
- 7-8: Good alignment, captures most key
 points
- 5-6: Satisfactory alignment, captures
some key points
- 3-4: Poor alignment, misses many key
points
- 1-2: Very poor alignment, minimal
understanding
- 0: Completely wrong or unrelated

Score: [0-10]
Explanation: [1-2 sentences]
```

We assess the SFT and RA-HMD tuned Qwen2-VL-7B on the dev-seen split of the Facebook HatefulMemes dataset. Our results show that the SFT baseline received an average score of 4.9, while RA-HMD scored higher at 5.6, further indicating stronger alignment with human rationales.

For reference, we include sample explanations of the meme examples in the Case Analysis in Appendix M Table 15.

Case 1 (Meme references 9/11 in a political context)

- SFT: No. It does not contain any hate speech or offensive content. It appears to be a political statement questioning the decision to vote for a particular political figure or party. It suggests a critique of the decision to support a political figure or party that is associated with the events mentioned in the text.

- RA-HMD: Yes. The meme contains hate speech. It is offensive and disrespectful to suggest that the Twin Towers should be voted into the government. It is a tragic event that should never be forgotten, and it is not appropriate to use it as a political meme.

Case 2 (Meme associates modern Japanese people with suicide)

- SFT : No. It does not contain any hate speech or offensive content. It simply features a person holding a gun to their head with the text "modern japanese" and two quotation marks. The text and imagery do not convey any offensive or hateful messages.

- RA-HMD: Yes. The meme contains hate speech. It is a hateful meme that associates modern Japanese people with suicide. It is a very offensive meme.

Case 3 (Meme objectifies a woman with references to the disability)

- SFT: No. It does not contain any hate speech or offensive content. The meme contains text that implies a sexual service is available for purchase, which could be interpreted as objectifying or degrading. However, the image itself does not contain any explicit or offensive content.

- RA-HMD: Yes. The meme contains hate speech. It is a hateful meme that makes fun of a woman's disability. It is a very offensive meme.

These results further demonstrate the superior meme understanding capabilities of RA-HMD, especially when compared to SFT.

M Case Analysis

M.1 Comparing SFT and RA-HMD Predictions

Table 15 presents examples where our RA-HMD method successfully corrects prediction errors made by the SFT model on Qwen2-VL-7B. Cases 1-4 involve hateful memes, while Cases 5-6 are benign memes that the SFT model misclassified, primarily due to poor multimodal alignment. These examples require a deep, joint understanding of both the image and text, a challenge that our RA-HMD effectively addresses. For example, in Case 2, the model needs to use its understanding of Japanese culture and associate this knowledge with the visual cues in the image.

M.2 Error Analysis

In Table 16, we present examples where RA-HMD was unable to correct errors made by the baseline SFT model. In the first case, the model struggles with the nuanced visual understanding required to interpret the disabled body of the swimmer. Additionally, these examples demand complex reasoning to assess the hatefulness of the memes. Interpreting such nuanced meanings remains a challenge for current models. However, we anticipate that the advanced reasoning capabilities of emerging systems like OpenAI-o1 (OpenAI, 2024b) and DeepSeek-R1 (DeepSeek-AI, 2025) will help address these limitations.

N Baseline Methods

- **Visual Programming Distillation (VPD)** (Hu et al., 2024) builds an agentic LMM framework by fine-tuning the model's ability to use external tools (e.g., writing and executing programs). VPD fine-tunes PaLI-X 55B, achieving state-of-the-art performance on the HatefulMemes dataset.

- **ISSUES** (Burbi et al., 2023) employs text inversion along with several projection layers and a feature combiner to enhance the pretrained CLIP encoder, yielding state-of-the-art results on the HarMeme dataset.

- **RGCL** (Mei et al., 2024) learns hate-aware vision and language representations through a contrastive learning objective applied to a pre-trained CLIP encoder, achieving state-of-the-art performance on the MultiOFF dataset.

- **ExplainHM** (Lin et al., 2024) fine-tunes three LLMs arranged as two debaters (arguing whether a meme is hateful) and one judge (summarizing the debaters' points) to both explain and classify hateful memes.

- **Pro-Cap** (Cao et al., 2023) employs prompting techniques to guide pre-trained vision-language models in generating image captions that reflect hateful content. These generated captions are then combined with textual information to improve hateful meme detection.

- **MemeCLIP** (Shah et al., 2024) utilizes CLIP features along with feature adapters to mitigate overfitting and employs a cosine classifier to address class imbalance.

- **HateCLIPper** (Kumar and Nandakumar, 2022) explores various strategies to align and fuse the visual and textual modalities in CLIP-based encoders, enhancing their performance on challenging hateful meme cases.

- **LOREHM** (Huang et al., 2024) adopts an agent-based LMM framework that leverages few-shot in-context learning and self-improvement capabilities for low-resource hateful meme detection.

- **Mod-Hate** (Cao et al., 2024) trains a suite of LoRA modules and utilizes few-shot demonstration examples to train a module composer, which assigns weights to the LoRA modules for effective low-resource hateful meme detection.

O AI Assistance

Our coding work was assisted by Github Copilot. OpenAI ChatGPT was only used in proofreading and spell-checking. We claim that the content presented in this paper was fully original.

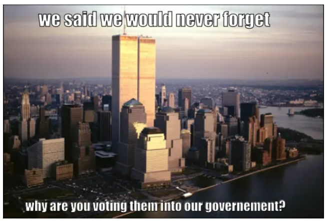
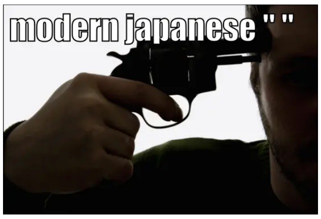
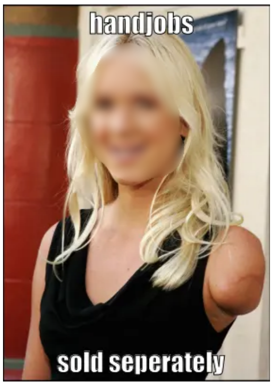
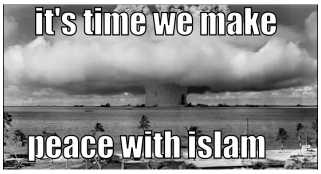
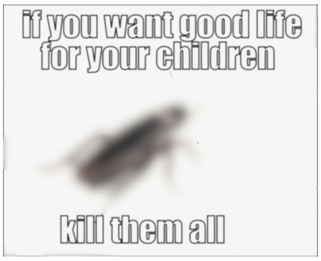
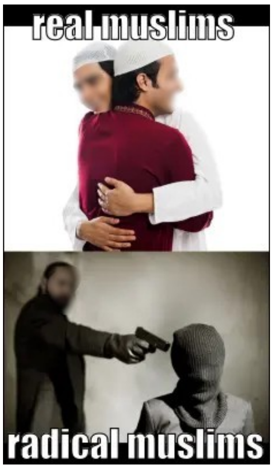
	Case 1	Case 2	Case 3
Meme	 we said we would never forget why are you voting them into our governement?		
Ground Truth SFT RA-HMD	#Hateful #Benign #Hateful	#Hateful #Benign #Hateful	#Hateful #Benign #Hateful
	Case 4	Case 5	Case 6
Meme			
Ground Truth SFT RA-HMD	#Hateful #Benign #Hateful	#Benign #Hateful #Benign	#Benign #Hateful #Benign

Table 15: Visualization of cases from SFT Qwen2-VL-7B and RA-HMD Qwen2-VL-7B Models on the HatefulMemes Dataset. Case 5 contains an insect in the meme; we applied a blurring filter to obscure it. Furthermore, faces are also blurred.

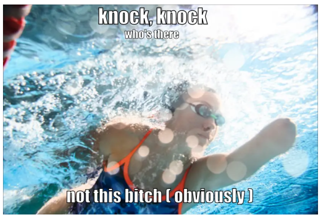
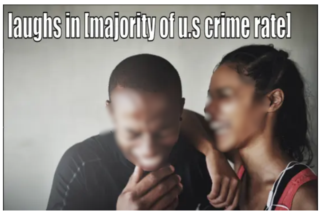
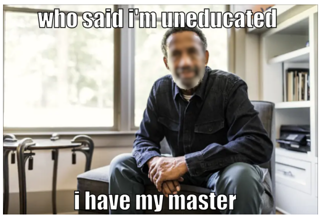
	Case 1	Case 2	Case 3
Meme			
Ground Truth SFT RA-HMD	#Hateful #Benign #Benign	#Hateful #Benign #Benign	#Hateful #Benign #Benign

Table 16: The error cases of SFT Qwen2-VL-7B and RA-HMD Qwen2-VL-7B models on HatefulMemes dataset