Yes FLoReNce, I Will Do Better Next Time! Agentic Feedback Reasoning for Humorous Meme Detection

Olivia Shanhong Liu[1], Pai Chet Ng[2], De Wen Soh[1], Konstantinos N. Plataniotis[3]

[1]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
[2]Infocomm Technology Cluster, Singapore Institute of Technology, Singapore
[3]The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada
shanhong_liu@mymail.sutd.edu.sg, paichet.ng@singaporetech.edu.sg, dewen_soh@sutd.edu.sg, kostas@ece.utoronto.ca

Abstract

Humorous memes blend visual and textual cues to convey irony, satire, or social commentary, posing unique challenges for AI systems that must interpret intent rather than surface correlations. Existing multimodal or prompting-based models generate explanations for humor but operate in an open loop, lacking the ability to critique or refine their reasoning once a prediction is made. We propose FLoReNce, an agentic feedback reasoning framework that treats meme understanding as a closed-loop process during learning and an open-loop process during inference. In the closed loop, a reasoning agent is critiqued by a judge; the error and semantic feedback are converted into control signals and stored in a feedback-informed, non-parametric knowledge base. At inference, the model retrieves similar judged experiences from this KB and uses them to modulate its prompt, enabling better, self-aligned reasoning without finetuning. On the PrideMM dataset, FLoReNce improves both predictive performance and explanation quality over static multimodal baselines, showing that feedback-regulated prompting is a viable path to adaptive meme humor understanding.

Caution: This paper contains offensive content due to the nature of the topic, which may be disturbing or offensive to some readers. Reader discretion is advised.

Introduction

Humorous memes play a central role in online discourse, shaping opinions and spreading social commentary through visual–textual wit (Li et al. 2024b; Xi, Yu, and Wang 2025). Understanding their humor is not only crucial for applications such as content moderation, sentiment analysis, and cultural trend monitoring (Shifman 2013; Milner 2018; Vásquez and Aslan 2021; Rehman et al. 2025), but also for building AI systems capable of interpreting human intent and nuance (Băroiu and Trăușan-Matu 2022; Kalloniatis and Adamidis 2024). However, humor in memes rarely resides in explicit features, it often emerges from subtle semantic interactions between image and text, such as irony, contrast, or metaphor. Existing classifiers that solely rely on correlations between pixels and words therefore fail to capture the deeper incongruity that defines humor (Rahman et al. 2025; Singh et al. 2024a).

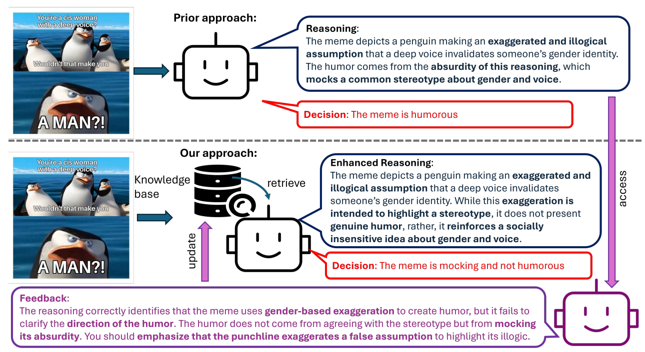


Figure 1: Comparison of our feedback-based reasoning correction approach (bottom) with prior approach (upper).

Recent studies have begun to incorporate reasoning into multimodal humor detection by prompting large vision–language models (VLMs) to generate textual explanations or multi-step justifications for their predictions. Approaches such as chain-of-thought prompting and multi-agent debate frameworks (Liu et al. 2023; Madaan et al. 2023; Zong et al. 2024; Li et al. 2024a; Schneider 2025; Liu et al. 2025a; Zhang et al. 2025; Lee et al. 2025) encourage models to articulate why an image–text pair may be humorous, improving interpretability over black-box classifiers like CLIP (Radford et al. 2021) or VL-BERT (Li et al. 2019). Other methods, including MemeCLIP (Shah et al. 2024) and TCRNet (Kasu et al. 2025), attempt to fuse textual and visual cues for irony or incongruity reasoning. While these methods mark progress toward explainable humor understanding, they remain fundamentally static: once a model produces an incorrect or shallow rationale, there is no mechanism for correction, reflection, or adaptation based on feedback, as shown in the upper panel in Fig. 1. In contrast, human humor comprehension is inherently dynamic, in which people refine their interpretations over time through critique, social feedback, and exposure to new cultural contexts, as illustrated in the lower panel of Fig. 1. This feedback process stabilizes understanding much like a control system regulates output to minimize error. Without such a loop, AI reasoning tends to oscillate between over- and under-predicting humor, lacking the self-corrective mechanism that enables humans to adaptively adjust their interpretations.

In this paper, we propose **FLoReNce**, **F**eedback-**Lo**op **Re**asoner with **N**on-parametric Experien**ce** (FLoReNce), a novel agentic framework that addresses these limitations by introducing feedback-driven adaptation into multimodal humor understanding. Our FLoReNce models the reasoning process as a closed-loop system that continuously refines its interpretation through structured critique and control. It integrates four key modules: 1) a Vision–Language Reasoning Agent that interprets the meme and produces an initial rationale; 2) a Judge Agent that evaluates this reasoning against ground truth and issues semantic feedback; 3) a PID controller that transforms the prediction error and feedback into quantitative control signals; and 4) a Knowledge Base that accumulates these experiences as non-parametric memory. Together, these components enable FLoReNce to dynamically adjust its reasoning strategy at inference time, by retrieving similar past cases from memory and modulating its prompts based on prior feedback, without requiring any parameter updates or retraining. In doing so, FLoReNce bridges the gap between static reasoning and human-like adaptive understanding, allowing large vision–language models to learn from their own interpretive history and progressively stabilize their perception of humor.

Our key contributions are as follows:

- We formulate humor reasoning as a closed-loop state-space system, where prediction errors and semantic feedback are treated as control signals regulating the reasoning process. This formulation enables control-driven adaptive prompting that dynamically adjusts the reasoning behavior of a frozen vision–language model for humor classification.
- We construct a feedback-informed Knowledge Base (KB) that encodes both the model's own reasoning and the Judge's critique, rather than storing raw training examples as in prior retrieval-based approaches. This non-parametric memory evolves through feedback, capturing interpretive refinements that improve inference-time reasoning.
- We demonstrate the effectiveness of FLoReNce on the PrideMM dataset, achieving strong performance with enhanced interpretability and adaptability. Even with a minimal retrieval setting (top-$K = 1$), FLoReNce attains an F1-score of 0.7708, showing that dynamic feedback and control substantially improve humor understanding over static reasoning baselines.

Related Works

Humorous Meme Detection

Early research on humorous meme detection treated the task as a shallow multimodal classification problem. Vlad et al. (2020) and Guo et al. (2020) explored fusing image and OCR text using handcrafted or CNN-based features. Later, transformer-based encoders such as BERT and ALBERT (Devlin et al. 2019; Lan et al. 2019) were paired with deep visual backbones like VGG and DenseNet (Simonyan and Zisserman 2014; Huang et al. 2017), forming parallel branches or late-fusion pipelines (Gupta et al. 2020). Subsequent work introduced cross-modal attention to align visual and textual cues for contextual humor understanding—e.g., Pramanick, Akhtar, and Chakraborty (2021) used joint attention to capture image–text incongruity, and Phan et al. (2022); Kumari, Adak, and Ekbal (2024); Singh et al. (2024b); Kasu et al. (2025) extended this with transformer-based multimodal fusion and multitask setups.

Despite these advances, most existing models treat humor detection as a surface-level classification task, lacking the ability to reason about how or why a meme is humorous (Kumari et al. 2024; Liu et al. 2024). These models often conflate different humor mechanisms (e.g., irony, absurdity, wordplay) without modeling them explicitly.

LLM-Based Multi-Agent Frameworks

Large Language Models (LLMs) have been increasingly adopted as autonomous agents capable of planning and deliberation across diverse domains such as embodied decision-making (Smit et al. 2023; Wang et al. 2023; Pham et al. 2023), collaborative problem solving (Liu et al. 2023; Madaan et al. 2023; Zong et al. 2024; Li et al. 2024a), and social simulation (Schneider 2025; Zhang et al. 2025). These efforts demonstrate that LLMs can reason beyond static prompting when embedded in multi-agent environments. Frameworks such as ChatEval (Chan et al. 2023), Debate (Liang et al. 2024; Nguyen, Childress, and Yin 2025), and CoT-SelfConsistency (Wang et al. 2022) simulates interactive and consensus-driven reasoning, while stance and meme-understanding frameworks like COLA (Lan et al. 2024), LoReHM (Huang et al. 2024), and MiND (Liu et al. 2025b) extend such agentic reasoning to multimodal social media analysis.

However, these systems remain fundamentally open-loop: they exchange or aggregate textual responses but lack a principled feedback mechanism that regulates reasoning trajectories. Their refinement typically depends on oracle supervision or external evaluators, conditions infeasible in zero-shot or subjective classification settings such as humor or hate detection. FLoReNce departs from these paradigms by introducing a closed-loop control formulation in which judge feedback is numerically encoded, stored in a Knowledge Base, and reused to modulate future reasoning prompts. This bridges discrete linguistic critique with continuous control dynamics, transforming reactive multi-agent dialogue into a stable feedback-regulated reasoning process. Prior works in multimodal humor understanding and LLM-based multi-agent reasoning either rely on static fusion or unregulated dialogue. FLoReNce unifies both strands under a control-theoretic perspective, where reasoning, judgment, and retrieval interact through feedback loops, offering a principled path toward interpretable and self-regulating humor understanding.

FLoReNce Framework

Problem Statement. We define a humorous meme detection dataset as a set of memes where each meme is denoted $m = (x^{\text{img}}, x^{\text{text}}, y)$, with image x^{img}, OCR text x^{text}, and humor label $y \in \{0, 1\}$. The *Reasoning Agent* is a frozen

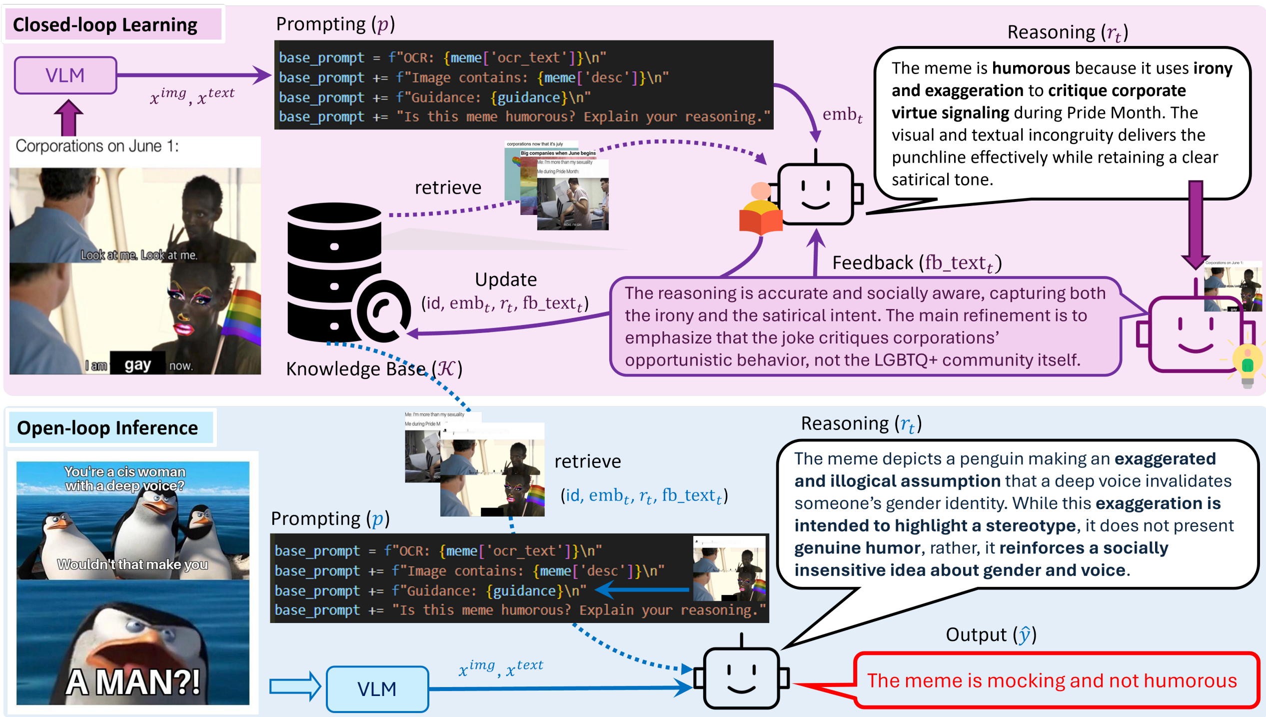


Figure 2: **FLoReNce framework.** In the *closed-loop learning* (top), the VLM reasons over memes, receives judge feedback, and stores judged embeddings in the Knowledge Base ($\mathcal{K}$). In the *open-loop inference* (bottom), $\mathcal{K}$ provides retrieval-based guidance for adaptive prompting on unseen memes, distinguishing humorous from non-humorous cases.

VLM $\mathcal{R}_\theta$, which, given a guidance prompt p, outputs a humor score and a textual rationale; it also exposes a hidden representation used for retrieval. We write this as

$$\begin{aligned} (\hat{y},\, r) &= \mathcal{R}_\theta\big(x^{\text{img}}, x^{\text{text}}, p\big), \\ \text{emb} &= \Phi_\theta\big(x^{\text{img}}, x^{\text{text}}, p\big) \in R^d, \end{aligned} \tag{1}$$

where $\hat{y} \in [0, 1]$ is the predicted humor probability, r is the generated reasoning text, and emb is the mean-pooled hidden embedding.

A numeric *control vector* c is mapped to guidance text via a prompt-mapper

$$p = \Psi(c), \tag{2}$$

which turns the control signal into succinct instructions (e.g., "be conservative," "check sarcasm," "verify setup→twist") used by $\mathcal{R}_\theta$. Intuitively, Ψ is the interface that converts *continuous* feedback into *discrete, interpretable* prompting that steers the VLM's reasoning.

Closed-Loop Learning

Agent reasoning. Given m and a current prompt $p_t = \Psi(c_t)$, the agent produces $(\hat{y}_t, r_t)$ and an embedding emb$_t$ as above. This step collects the agent's *current belief* (score) and *argument* (rationale) under the present guidance.

Judge feedback and control. The Judge $\mathcal{J}_\phi$, with access to ground truth, critiques the agent by outputting a scalar error e_t, a textual critique fb_text$_t$, and a low-dimensional feedback vector f_t:

$$\begin{aligned} (e_t,\, \text{fb\_text}_t,\, f_t) &= \mathcal{J}_\phi\big(m, \hat{y}_t, r_t\big), \\ e_t = y - \hat{y}_t, &\quad f_t \in R^3. \end{aligned} \tag{3}$$

Here e_t measures *how far* the agent is from the target label, while f_t summarizes the critique semantics along three interpretable axes (e.g., irony/sarcasm, narrative structure, layout cues). The Controller $\mathcal{C}$ converts error history into a stabilizing action using PID dynamics:

$$u_t \;=\; K_P\, e_t \;+\; K_I \sum_{\tau=1}^{t} e_\tau \;+\; K_D\big(e_t - e_{t-1}\big). \tag{4}$$

We aggregate numeric signals into a control vector

$$c_t \;=\; \big[u_t,\; f_t^\top,\; k_t^\top\big]^\top \in R^{1+3+3}, \tag{5}$$

where $k_t \in R^3$ is an optional compact KB signal (zero at the beginning of training). The new prompt is then $p_t = \Psi(c_t)$, which nudges the agent toward the aspects highlighted by the critique (e.g., "check for sarcasm" when $f_{t,1}$ is large).

Experience-based KB update. After critique, we store *feedback-informed* experience into a non-parametric KB:

$$\mathcal{K} \;\leftarrow\; \mathcal{K} \,\cup\, \big\{(id,\, \text{emb}_t,\, r_t,\, \text{fb\_text}_t)\big\}. \tag{6}$$

Unlike retrieval systems that index raw training pairs, $\mathcal{K}$ preserves the *reasoning trace* and the *judge critique* alongside the embedding. This makes later retrieval *experience-aware*: the system recalls not only "what it saw," but also "how it was corrected."

Open-Loop Inference

Query embedding and retrieval. At test time there is *no judge*: the system adapts by leveraging $\mathcal{K}$ to shape prompts. We first form a neutral/base prompt $p_0 = \Psi(\mathbf{0})$ and compute a query embedding

$$q \;=\; \Phi_\theta\big(x^{\text{img}}, x^{\text{text}}, p_0\big). \tag{7}$$

We retrieve top-K neighbors by cosine similarity:

$$\text{sim}(q, \text{emb}_j) \;=\; \frac{q^\top \text{emb}_j}{\|q\|\,\|\text{emb}_j\|}, \qquad j \in \mathcal{N}_K(q). \tag{8}$$

This step *grounds* the current meme in past, judge-corrected experiences that were semantically similar.

Compact KB signal and control. We summarize retrieved entries into a compact memory signal

$$k \;=\; \frac{1}{K} \sum_{j \in \mathcal{N}_K(q)} g\big(\text{emb}_j\big) \;\in\; R^3, \tag{9}$$

where $g : R^d \to R^3$ is a fixed projection used by the controller. With no judge present, we set $f = \mathbf{0}$ and assemble

$$c \;=\; \big[u,\; f^\top,\; k^\top\big]^\top \;=\; \big[u,\; \mathbf{0}^\top,\; k^\top\big]^\top, \tag{10}$$

where u is a policy-driven scalar derived from k (e.g., more conservative when k signals risk of false positives). The prompt is $p = \Psi(c)$, which *adapts* the agent's attention toward failure modes seen in similar past cases (e.g., stereotype inversion, sarcasm cues).

Final reasoning and inference. The agent produces the final judgment and rationale

$$(\hat{y},\, r) = \mathcal{R}_\theta\big(x^{\text{img}}, x^{\text{text}}, p\big), \tag{11}$$

completing a memory-driven loop without supervision. Intuitively, the KB plays the role of a non-parametric prior over reasoning behavior: it nudges the agent toward historically successful interpretations for similar memes. Because $\mathcal{R}_\theta$ is frozen, this achieves inference-time adaptation not by changing weights, but by modulating prompts with control signals learned from prior feedback.

Experiments Setup

Evaluation Datasets

We evaluate on PrideMM (Shah et al. 2024), a multimodal dataset of 5,063 text-embedded images (memes, posters, infographics) related to the LGBTQ+ movement, collected from Facebook, Twitter/X, and Reddit during 2020–2024. OCR text is extracted (with standard cleaning) and paired with the image.

Unless otherwise specified, we follow the split protocol used in the PrideMM paper's experiments, i.e., a predefined 85/5/10 train/validation/test split. In our closed-loop learning stage, we build a feedback-informed KB on the training split without updating model weights: the Reasoning Agent is frozen, and the Judge accesses labels to issue critiques. We report final results on the held-out test split. At inference time, the Judge is disabled; the system performs retrieval from the KB and adaptive prompting without labels.

Baselines

We benchmark FLoReNce against a comprehensive set of text-only, vision-only, multimodal, and agentic/prompted baselines that collectively span the evolution of humor understanding in multimodal memes.

Classical and Multimodal Fusion. Early multimodal approaches treat humor as a feature-level fusion problem. ResNet50+MLP (visual-only) and T5+MLP (text-only) provide unimodal lower bounds. MOMENTA (Pramanick et al. 2021) combines BERT-based textual features and ResNet-based visual features through multimodal transformers to detect image–text incongruity. MemeCLIP (Shah et al. 2024) leverages CLIP embeddings to align vision and text in a shared contrastive space and then trains a classifier for humor or hate recognition. These models encode multimodal correlation but lack explicit reasoning or feedback; their predictions are static once trained.

Prompt-based and Agentic Reasoning. Recent work explores LLM/VLMs as reasoning agents. PromptHate (Cao et al. 2022) reformulates meme classification as a textual entailment problem using prompt templates over RoBERTa. LoReHM (Huang et al. 2024) fine-tunes the LLaVA-34B VLM with low-rank adapters to reason about humor and harmfulness through visually grounded instructions. COLA (Lan et al. 2024) introduces collaborative multi-agent stance reasoning using GPT-3.5-Turbo, extending chain-of-thought and debate prompting to generate stance-aware textual judgments. MiND (Liu et al. 2025b) employs Qwen2.5-VL-32B with iterative self-reflection for meme interpretation, representing a state-of-the-art agentic reasoning baseline on PrideMM.

These methods exploit in-context or few-shot prompting and LLM-as-agent reasoning, yet they remain open-loop: no explicit control feedback regulates their reasoning trajectory once a prompt is issued.

Evaluation Metrics

Predictive performance. We report three thresholded metrics: Accuracy, Macro-F1 (unweighted mean of per-class F1), and Matthews Correlation Coefficient (MCC), a robust, single-number summary under class imbalance. Unless stated otherwise, we binarize $\hat{y} \in [0, 1]$ at 0.5 (optionally calibrated on `val`).

Reasoning quality. Let $y \in \{0, 1\}$ be the ground truth and $\hat{y} \in \{0, 1\}$ the predicted label. Denote N_1 and N_0 the number of positive and negative examples, and TP, TN the counts of correct predictions in each class. We define a basic correctness-aligned reasoning score

$$\text{RQ} \;=\; \frac{1}{2}\left(\frac{\text{TP}}{N_1} \;+\; \frac{\text{TN}}{N_0}\right)$$

Implementation Details

Reasoning Agent $\mathcal{R}_\theta$: Qwen2.5-VL-32B-Instruct (frozen). **Judge** $\mathcal{J}_\phi$: same backbone in supervision mode; produces $(e_t, \text{fb\_text}_t, f_t)$ with $e_t = y - \hat{y}_t$ and $f_t \in R^3$ from MiniLM `all-MiniLM-L6-v2` (first 3 components).

Table 1: Results on PrideMM. Predictive Performance reports *Accuracy*, *Macro-F1*, *MCC*, and *RQ*.

		Predictive Performance			Reasoning Performance
Model	Backbone	Accuracy	Macro-F1	MCC	RQ (%)
Visual Only	ResNet50 + MLP	66.08	61.67	0.33	-
Text Only	T5 + MLP	67.85	66.10	0.36	-
MemeCLIP (Shah et al. 2024)	CLIP	**78.30**	76.99	0.57	-
MOMENTA (Pramanick et al. 2021)	CLIP	73.57	69.92	0.47	-
PromptHate (Cao et al. 2022)	RoBERTa	73.77	73.46	0.49	-
LoReHM (Huang et al. 2024)	LLaVA-34B	70.09	64.07	0.39	64.8
COLA (Lan et al. 2024)	GPT-3.5-Turbo	53.25	59.34	0.07	58.5
MiND (Liu et al. 2025b)	Qwen2.5-VL-32B	54.45	50.43	0.05	52.6
FLoReNce (K=1)	Qwen2.5-VL-32B	73.40	77.08	0.48	74.0
FLoReNce (K=3)		73.73	**77.36**	0.48	74.3
FLoReNce (K=5)		73.80	77.33	0.48	74.4
FLoReNce (K=10)		73.60	77.33	0.47	74.2

Controller $\mathcal{C}$: PID with (K_P, K_I, K_D); states $(\sum e_t, e_{t-1})$ persist across samples during closed-loop learning.

Knowledge Base $\mathcal{K}$: stores $(\mathrm{id}, \mathrm{emb}, r, \mathrm{fb\_text})$, where `emb` is the mean of the last hidden layer from Qwen-VL. Retrieval uses cosine similarity on CPU tensors; the KB is JSONL (reproducible, low memory). The prompt mapper Ψ converts control into guidance.

Closed-Loop Learning For each train meme m_t, the agent produces $(\hat{y}_t, r_t)$ and emb_t; the Judge returns (e_t, f_t); the Controller computes u_t; and $\mathcal{K}$ is augmented with $(\mathrm{id}_t, \mathrm{emb}_t, r_t, \mathrm{fb\_text}_t)$. PID states $(\sum e_t, e_{t-1})$ are preserved across steps. No test items are added to $\mathcal{K}$, and no model weights are updated.

Open-Loop Inference For each test meme, we compute a query embedding q_t, retrieve top-K neighbors from $\mathcal{K}$, and summarize them into $k_t \in R^3$. With Judge disabled ($f_t = 0$), we form $c_t = [u_t,\ 0^\top,\ k_t^\top]$ and generate a controlled prompt $\Psi(c_t)$ for final $(\hat{y}_t, r_t)$. Unless specified, we set $K =$ `TOP_K` from the config.

Hyperparameters and Reproducibility Unless stated, we set $(K_P, K_I, K_D) = (1.0, 0.5, 0.1)$. We cap generation at 128 tokens (rationales). All experiments run on NVIDIA L40S (48GB).

Results and Discussions

The results in Table 1 show that FLoReNce attains predictive performance on PrideMM that is comparable to, and in several cases slightly better than, strong multimodal baselines. Classical unimodal systems (Visual Only, Text Only) remain in the mid-60% range for accuracy, confirming that humor in LGBTQ+ memes is genuinely multimodal. Established multimodal approaches such as MemeCLIP reach higher accuracy (78.30%) and macro-F1 (76.99%), indicating that better image-text alignment helps detect incongruity. Prompt-based or agentic baselines (PromptHate, LoReHM, MiND) mostly fall in the low-to-mid 70% accuracy range, with macro-F1 between 64% and 73%. Our framework, FLoReNce with retrieval and control at $K = 3$, reaches 73.73% accuracy and 77.36% macro-F1, which is notable because the macro-F1 gain is larger than the gain in accuracy. This suggests that the feedback-informed KB and control signal are especially helpful for the harder class, improving class-balanced performance rather than only the dominant class.

Across the FLoReNce configurations (K=1, 3, 5, 10), the performance is remarkably stable: accuracy remains around 73–74%, macro-F1 stays around 77%, and the reasoning-quality score hovers near 74%. Since RQ is defined in a balanced-accuracy style over correctly reasoned instances, this pattern implies that once the KB has been populated in the closed-loop phase, the open-loop inference can exploit even a small number of retrieved neighbours (K=1 or 3) to produce consistent, feedback-aligned reasoning. Increasing K beyond 3 does not deteriorate performance, which indicates that the retrieved experiences are semantically coherent and that the prompt-mapping function Ψ can absorb slightly richer control signals. Overall, these findings support the central claim of FLoReNce: integrating control-style feedback with a non-parametric, feedback-built memory improves not only raw prediction but also the stability and consistency of humor reasoning.

Case Study

Figure 3 illustrates three representative examples from PrideMM showing how FLoReNce leverages feedback and retrieval for adaptive humor reasoning. In the first two cases (top rows), the system successfully classifies the memes as *humorous*. For the "*Corporations on June 1*" meme, the retrieved example from the KB contains prior judge feedback emphasizing irony, contrast, and corporate opportunism, which guides the Reasoning Agent to identify the satirical intent and deliver a socially aware explanation. Similarly, in the "*Doctor / Gender*" meme, retrieval provides feedback cues such as subversive punchline, shock-based humor, and social taboo, helping the agent contextualize the blunt dialogue as intentional irony rather than literal bias. These examples demonstrate how feedback-informed retrieval enables FLoReNce to internalize semantic corrections, transforming judge supervision into reusable reasoning guidance. The third case ("*Christians vs. Gay People*") highlights a

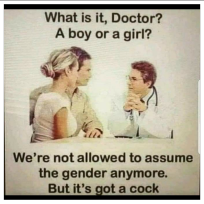
Meme	Rationale	Retrieved Meme	Feedback
 Label: Humor; Prediction: Humor	This meme is indeed **humorous**, but its humor stems from a combination of **satire, irony**, and **social commentary**. The meme starts by highlighting the precise transition between May 31st at 11:59 PM and June 1st at 12:00 AM. This specific timing sets up the context for the joke, which revolves around the beginning of June — Pride Month. By pinpointing the exact moment when June begins, the meme emphasizes the sudden influx of companies that jump on the Pride bandwagon, often seen as **opportunistic or insincere**. The Brain with a Rainbow Flag: a clever visual metaphor suggesting corporate exploitation of LGBTQ+ symbolism for profit rather than genuine support. The logo "Every Single Company in June" reinforces this **ironic contrast** and delivers the **punchline**.		The agent's prediction is **correct.** The meme is indeed **humorous**, and the reasoning provided aligns well with the structure and intent of the meme. The agent correctly identified the key elements contributing to the **humor**, including the **contrast** between the initial supportive statement and the subsequent rejection depicted in the image. The agent understood the following: – Textual Contrast: The top text portrays corporations as strong allies of the LGBT community, using **warm language** and **colorful emojis** to emphasize their commitment.
 Label: Humor; Prediction: Humor	This meme can be considered **humorous** due to tension between traditional binary gender assumptions and modern sensitivity around gender identity. The phrase "We're not allowed to assume the gender anymore" reflects contemporary discussions about gender fluidity and inclusivity, while the doctor's blunt response — "But it's got a cock" — delivers a **direct, subversive punchline** that **violates expectation** and creates **shock-based humor**. The humor therefore arises from **contrast**, **irony**, and **social taboo** surrounding **gender norms** and **political correctness**.		Image: The cartoon face shows an **exaggerated expression of shock or horror**, which sets the tone for the meme. Text: "No, They're young. They don't know what's best for them." This statement implies a **judgment** about the appropriateness of decisions made for young individuals, particularly in a cultural or religious context. Supporting Image: Three children wearing hijabs — suggesting a **contrast** between **innocence** and **adult authority** and indicating **social commentary**.
 **Label**: No Humor ; **Prediction**: Humor	Top Panel: A woman yelling at a cat, representing the conflict between Christians and gay people who did nothing wrong. Bottom Panel: A person pointing at a TV screen that delivers a moral statement about disapproval versus shaming. The meme critiques the treatment of LGBTQ+ individuals by religious groups, using **absurdity** and **irony** to expose **hypocrisy**. The humor arises from the juxtaposition of serious social criticism with a comedic visual format, producing **dark or satirical humor** rather than lighthearted amusement.		Top Panel: "When abusers in the church have not been charged for their crimes." – Image: a person sleeping peacefully with the text "i sleep." Bottom Panel: "When Trans people exist." – Image: the same person with glowing eyes and an intense expression, captioned "real shit." The clear **contrast** between the two panels highlights **hypocrisy** and **social double standards**, forming the basis of **satirical humor**.

Figure 3: Examples of retrieved memes with feedbacks. Correct prediction in green. Incorrect prediction in red.

failure: although the meme was labeled non-humorous, the system predicted humorous because retrieved feedback on similar religious-context memes overemphasized absurdity and irony, causing misalignment between form (satirical tone) and intent (mocking content). This case underlines the challenge of distinguishing satire that critiques power from mockery that targets marginalized groups, a boundary that FLoReNce aims to learn more robustly through future refinement of judge feedback and retrieval filtering.

Ablation Studies

To understand the contribution of each component in FLoReNce, we progressively removed the KB, the control path, and the judge-derived semantic feedback f_t, while keeping the retrieval size fixed at $K = 3$. The plain VLM only achieve mid-range performance on PrideMM, while adding the KB alone yields a considerable gain, showing that retrieving feedback-informed experiences helps even without extra control. Adding only the controller also improves over the base model, though slightly less than KB-only, indicating that control is more effective when it can condition on meaningful memory signals. When we keep PID and KB but drop the semantic feedback vector f_t, performance decreases compared to the full model, which confirms that judge critiques carry information that cannot be recovered from embeddings alone. The best results are obtained when all three ingredients are present (PID + KB + f_t), supporting our claim that humorous meme understanding benefits from a closed-loop design that fuses numeric control, semantic feedback, and non-parametric memory.

Table 2: Component ablation on PrideMM (K=3).

Variant	Acc	Macro-F1	MCC
Base VLM (no KB, no control)	64.20	58.10	0.22
+ KB only (no control)	68.30	63.90	0.35
+ Control only (no KB)	72.00	69.40	0.44
– f_t (PID+KB, drop feedback vec)	73.00	70.20	0.46
– PID (KB signal only)	72.60	70.00	0.45
Full FLoReNce (PID+KB+f_t)	**73.73**	**77.36**	**0.48**

Conclusion

In this paper, we delved into multimodal humor understanding and proposed FLoReNce, which treats the task as a regulated, experience-aware process rather than a one-shot classification problem. We introduced a feedback-loop formulation that transforms judge critiques into control signals and stores them as non-parametric experience, enabling iterative refinement of both predictions and rationales. On PrideMM, this yields improvements in accuracy and in the stability and consistency of generated reasonin even under minimal retrieval, showing that feedback-informed prompting can be a practical alternative to full model fine-tuning for subjective, nuance-heavy phenomena. Collectively, these findings position FLoReNce as a general recipe for controllable, critique-driven reasoning in multimodal settings, and open avenues for richer memory design.

References

Băroiu, A.-C.; and Trăușan-Matu, 2022. Automatic sarcasm detection: Systematic literature review. *Information*, 13(8): 399.

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 321–332.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Guo, Y.; Huang, J.; Dong, Y.; and Xu, M. 2020. Guoym at SemEval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1120–1125.

Gupta, A.; Kataria, H.; Mishra, S.; Badal, T.; and Mishra, V. 2020. BennettNLP at SemEval-2020 task 8: multimodal sentiment classification using hybrid hierarchical classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1085–1093.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, J.; Lin, H.; Liu, Z.; Luo, Z.; Chen, G.; and Ma, J. 2024. Towards low-resource harmful meme detection with LMM agents. *arXiv preprint arXiv:2411.05383*.

Kalloniatis, A.; and Adamidis, P. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2): 43.

Kasu, S. K. R.; Rehman, M. Z. U.; Dar, S. S.; Junghare, R. B.; Namboodiri, D. S.; and Kumar, N. 2025. D-humor: Dark humor understanding via multimodal open-ended reasoning. *arXiv preprint arXiv:2509.06771*.

Kumari, G.; Adak, C.; and Ekbal, A. 2024. Mu2STS: A Mu ltitask Mu ltimodal S arcasm-Humor-Differential T eacher-S tudent Model for Sarcastic Meme Detection. In *European Conference on Information Retrieval*, 19–37.

Lan, X.; Gao, C.; Jin, D.; and Li, Y. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, 891–903.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lee, X. Y.; Akatsuka, S.; Vidyaratne, L.; Kumar, A.; Farahat, A.; and Gupta, C. 2025. Reliable Decision-Making for Multi-Agent LLM Systems. *arXiv preprint arXiv:2406.04092*.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2024a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. arXiv.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, Y.; Li, Y.; Zhang, S.; Liu, G.; Chen, Y.; Shang, R.; and Jiao, L. 2024b. An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency. *Knowledge-Based Systems*, 287: 111457.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Liu, Y.; Cao, J.; Li, Z.; He, R.; and Tan, T. 2025a. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

Liu, Z.; Fan, C.; Lou, H.; Wu, Y.; and Deng, K. 2025b. MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection. *arXiv preprint arXiv:2507.06908*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.

Milner, R. M. 2018. *The world made meme: Public conversations and participatory media*. mit Press.

Nguyen, N. T. V.; Childress, F. D.; and Yin, Y. 2025. Debate-Driven Multi-Agent LLMs for Phishing Email Detection. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, 1–5. IEEE.

Pham, C.; Liu, B.; Yang, Y.; Chen, Z.; Liu, T.; Yuan, J.; Plummer, B. A.; Wang, Z.; and Yang, H. 2023. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*.

Phan, K. N.; Lee, G.-S.; Yang, H.-J.; and Kim, S.-H. 2022. Little flower at memotion 2.0 2022: Ensemble of multimodal model using attention mechanism in memotion analysis. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Pramanick, S.; Akhtar, M. S.; and Chakraborty, T. 2021. Exercise? I thought you said'Extra Fries': Leveraging Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 513–524.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their

Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4439–4455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rahman, M.; Provath, M. A.-M.; Deb, K.; Dhar, P. K.; and Shimamura, T. 2025. CAMFusion: Context-Aware Multi-Modal Fusion Framework for Detecting Sarcasm and Humor Integrating Video and Textual Cues. *IEEE Access*.

Rehman, M. Z. U.; Raghuvanshi, D.; Jain, U.; Bansal, S.; and Kumar, N. 2025. A multimodal-multitask framework with cross-modal relation and hierarchical interactive attention for semantic comprehension. *Information Fusion*, 103628.

Schneider, J. 2025. Generative to agentic ai: Survey, conceptualization, and challenges. *arXiv preprint arXiv:2504.18875*.

Shah, S. B.; Shiwakoti, S.; Chaudhary, M.; and Wang, H. 2024. MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17320–17332. Miami, Florida, USA: Association for Computational Linguistics.

Shifman, L. 2013. *Memes in digital culture*. MIT press.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, A. K.; Kim, G.; Kim, J.; Park, H.-m.; Choi, B. J.; and De Neve, W. 2024a. RAMM: A Residual Attention Multimodal Model for Humor Detection. In *International Conference on Intelligent Human Computer Interaction*, 229–240. Springer.

Singh, G. V.; Firdaus, M.; Chauhan, D. S.; Ekbal, A.; and Bhattacharyya, P. 2024b. Well, now we know! unveiling sarcasm: Initiating and exploring multimodal conversations with reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18981–18989.

Smit, A.; Duckworth, P.; Grinsztajn, N.; Barrett, T. D.; and Pretorius, A. 2023. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*.

Vásquez, C.; and Aslan, E. 2021. “Cats be outside, how about meow”: multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171: 101–117.

Vlad, G.-A.; Zaharia, G.-E.; Cercel, D.-C.; Chiru, C.; and Trăușan-Matu, 2020. UPB at SemEval-2020 Task 8: Joint Textual and Visual Modeling in a Multi-Task Learning Architecture for Memotion Analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1208–1214.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xi, Z.; Yu, B.; and Wang, H. 2025. Multimodal sarcasm detection based on sentiment-clue inconsistency global detection fusion network. *Expert Systems with Applications*, 275: 127020.

Zhang, H.; Cui, Z.; Wang, X.; Zhang, Q.; Wang, Z.; Wu, D.; and Hu, S. 2025. If multi-agent debate is the answer, what is the question. *arXiv preprint arXiv:2502.08788*.

Zong, C.; Yan, Y.; Lu, W.; Shao, J.; Huang, E.; Chang, H.; and Zhuang, Y. 2024. Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. *arXiv preprint arXiv:2402.14320*.