Tackling Model Bias via Game-theoretic Multi-agent Collaboration Framework for Hateful Meme Classification

Yiwei Wei[1] Zhengliang Guo[2] Shaozu Yuan[3] Chengyin Hu[2] *
Zhiyang Jia[2] Jiujiang Guo[5]* Meng Chen[4] Peiying Wang[3] Longbiao Wang[1,6]*
[1] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2] China University of Petroleum-Beijing at Karamay, China
[3] Meituan, China
[4] Oracle AI, Australia
[5] School of Computer Science and Technology, Tianjin University, China
[6] Huiyan Technology (Tianjin) Co., Ltd, China

Abstract

*Hateful meme classification aims to identify memes containing hateful content and has become increasingly important in the era of social media dominance. Large multimodal models (LMMs) have significantly enhanced the understanding of multimodal content, advancing this field. However, cognitive biases in LMMs can impede effective collaboration among models. To tackle this issue, we introduce **GECO**, a **G**ame-theoretic multi-ag**E**nt **C**ollaboration framew**O**rk that organizes multiple LMMs into interacting agents and employs game-theoretic principles to guide them toward an optimal cooperative equilibrium. GECO further integrates a mixed bonus scheme with both individual accuracy and cross-model agreement, which together drive the system toward a consistent cooperative solution. In addition, we implement efficient policy learning and introduce a penalty coefficient to optimize the framework effectively and ensure training stability. Extensive experiments on five public datasets demonstrate that our framework achieves new state-of-the-art performance. We release our code.*[†]

1. Introduction

The spread of harmful memes on social media exacerbates social polarization, disrupts public discourse, and leads to significant physical and psychological harm [13, 32]. Consequently, the importance of Hateful Meme Classification [37], which aims to detect hateful emotions within multimodal content, has become increasingly pronounced. Earlier studies [46, 51] addressed this task by leveraging pre-trained vision-language models [21, 26] and fine-tuning lightweight, task-specific classifiers on top of these backbones. Due to the advanced reasoning capabilities of large multimodal models (LMMs), LMM-based methods have been widely adopted for hateful meme classification. Some efforts focus on enhancing a single LMM's performance by improving few-shot capabilities through the integration of various LoRA-based modules [15], or by leveraging retrieval techniques to boost performance [29].

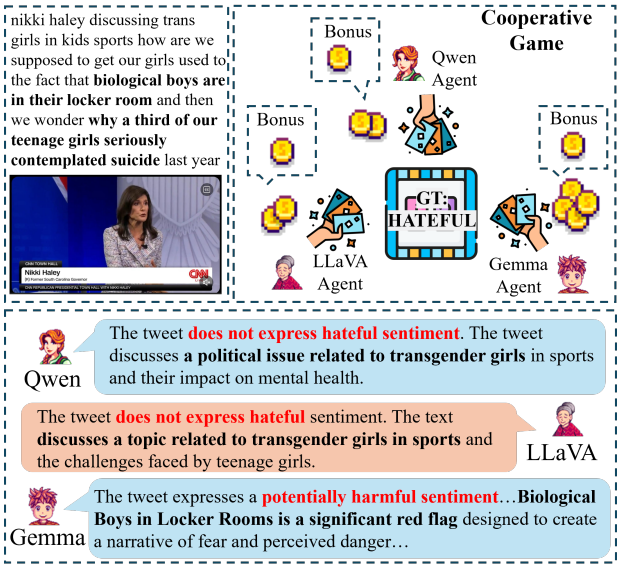


Figure 1. **Illustration of model bias in Hateful Meme Classification.** For the hateful meme ("nikki haley … last"), three LMMs (Qwen, Gemma, LLaVA) present varied interpretations. Within the game, only Gemma, having provided the correct interpretation, received the highest bonus, thereby guiding the entire cooperative framework towards the accurate answer.

*Corresponding authors. Email: cyhu@cupk.edu.cn, jiujiangguo@tju.edu.cn, longbiao_wang@tju.edu.cn

[†]Code is available at https://github.com/NagisaG/GECO.

Despite the advancements, single models often exhibit biases due to limitations in training data and paradigms. Consequently, multi-agent ensemble strategies [4, 14, 23, 27] have been developed to improve robustness and interpretability. Those ensemble methods can be generally categorized into **voting-based** [38, 50] and **debate-based** [5, 17] paradigms. Specifically, Mod-Hate [4] employs the majority rule to aggregate multiple LoRA-based models, stabilizing decisions, while ExplainHM [23] leverages LMMs to generate explanations from diverse viewpoints, using a judge model to evaluate these perspectives for better meme detection.

Though cooperation can moderately reduce single-model bias, it is inadequate to address the bias among models. Voting-based methods may amplify errors when a majority of agents share similar biases, and debate-based methods rely heavily on the fairness of the judging model, which can introduce biases, leading debates to biased conclusions. In Figure 1, the meme depicts Nikki Haley expressing concern about biological boys entering locker rooms, subtly conveying a sentiment of hatred. For this meme, the LMMs—LLaVA [25], Qwen [49], and Gemma [44]—provided differing interpretations. Both Qwen and LLaVA misinterpreted the content, while only Gemma successfully identified the latent message of fear and danger, pointing to an emotion of hatred. In this scenario, ensemble strategies like voting are unable to provide the correct prediction, while the judging model in debate strategies is easily misled by the majority of biased interpretations.

To fill this gap, we utilize foundational concepts from game theory [39], traditionally applied in strategic gameplay to guide each participant toward an optimal state, known as the Nash equilibrium [33]. However, applying game-theoretic formulations to mitigate model bias remains challenging, as independently optimizing each agent's prediction does not guarantee consensus, potentially resulting in unstable or inconsistent collective decisions and leaving inter-model bias unresolved. In light of this, we propose GECO, the **G**ame-theoretic multi-ag**E**nt **C**ollaboration framew**O**rk, to address the model bias. This framework includes three types of agents: the reasoning agent, the learnable agent, and the master agent, with the reasoning agent instantiated as three LMMs serving as reasoning modules, while the learnable and master agents adapt their policies accordingly. To coordinate these agents effectively, we design a mixed bonus scheme that incorporates both individual correctness and collective agreement into the optimization objective. This scheme surpasses traditional game-theoretic formulations by explicitly promoting consensus on the correct label within the collaborative framework. Moreover, efficient policy learning is implemented to update agents, reducing noise from multiple sampling and minimizing extra computation. We also employ a penalty coefficient to stabilize the training process. This cooperative optimization steers the system toward a cohesive cooperative solution, mitigating individual agent bias and yielding robust overall performance. Experimental results across five benchmark datasets demonstrate that our model achieves state-of-the-art performance, proving the superiority of our approach.

Overall, our contributions are three-fold:

- This is the first work to introduce game-theoretic thinking to multiple agents for hateful meme detection, paving the way for applying game-theoretic concepts to classification tasks.
- We propose GECO: various LMMs are optimized to reach mutual agreement on the correct label via a designed mixed bonus scheme, mitigating the model bias, and enhancing overall decision accuracy. We also design an efficient policy learning and penalty coefficient to ensure training stability.
- Through extensive experiments on widely-used benchmarks, we demonstrate that the proposed GECO significantly outperforms existing state-of-the-art methods.

2. Related Work

2.1. Hateful Meme Classification

Hateful meme detection is critical for curbing harmful content, yet hinges on parsing subtle image–text semantics [34]. Benchmark releases have catalyzed progress while underscoring the task's difficulty [9, 19, 34, 40]. Early systems used dual-stream encoders with late fusion or fine-tuned multimodal backbones, but implicit, culture-bound cues limit attention-based and shallow-fusion methods [18, 19, 24, 31, 43, 46]. Recent work leverages LMMs to generate rationales for distillation or to augment decisions with auxiliary modules [4, 14, 23]. The inherent biases of LMMs can sometimes lead to contradictions within the system's detection of hateful sentiments [8]. When this occurs among the majority of models within the system, strategies such as voting become ineffective.

2.2. Game Theory

Game theory [10, 41] is a mathematical framework that studies how to find optimal decisions among multiple players. The aim of game theory is to achieve a Nash equilibrium(stable state) in which no player can achieve a better payoff by changing their strategy. Game-theoretic formulations are typically categorized into cooperative and non-cooperative games. Cooperative game theory [2, 6, 48] is specifically designed to address the division of predetermined values among members, providing standard mechanisms for handling binding agreements. In contrast, non-cooperative game theory is widely applied in strategic gaming scenarios, ranging from Poker [3] and Star-

Craft [47] to Go [42] and Diplomacy [1, 7]. In classification tasks involving multiple agents, the combination of different agents can be viewed as a form of non-cooperative game. However, directly applying game-theoretic formulations does not guarantee consensus, often leading to unstable collective decisions and persistent inter-model bias. To address this, we model multimodal classification as a multi-agent game and introduce GECO, which integrates specialized agents with a consensus-driven objective to enable bias-aware collaborative prediction among multiple LMMs.

3. Method

Our approach, GECO, is illustrated in Figure 2. We first define all agents and the actions required to complete the game. Next, we introduce our proposed optimization strategy to alleviate the bias of different models and achieve consistency in performance.

3.1. Game Agents Definition

3.1.1. Agents Introduction

In this framework, there are three types of agents: the reasoning agent, the learnable agent, and the master agent.

Reasoning agent. To achieve complementary multimodal reasoning interpretation, we select three mainstream LMMs, including Qwen2-VL (v_Q) [49], LLaVA-1.5 (v_L) [25], and Gemma3 (v_G) [44] as *reasoning agents*, leveraging their capabilities in multimodal understanding.

Before the game, all reasoning agents are adapted through lightweight LoRA fine-tuning [15]. Once the game begins, these agents are kept frozen, preserving their inherent multimodal reasoning capabilities and providing stable features for hateful meme detection. For a labeled sample ξ_k, each LMM is guided to produce a *single target token* that converts the label $s(y_k)$ into text, optimizing the LMM using a standard language modeling loss:

$$\mathcal{L}_i^{\text{LMM}} = -\sum_{k=1}^{N} \log p_{\theta_i}\big(\hat{y}_k^{\text{LMM}} = s(y_k) \,\big|\, \xi_k\big), \tag{1}$$

where p_{θ_i} denotes the token-level conditional distribution parameterized by the LMM with LoRA [15] adapters. Each reasoning agent $i \in \{v_Q, v_L, v_G\}$ generates a hidden representation $f_i^k = h_i(\xi_k)$, where $h_i(\cdot)$ denotes the backbone; and $\{f_i^k\}$ corresponds to the last-token hidden state. Subsequently, we project $\{f_i^k\}$ into the decision space $\mathcal{D}$ to generate the reasoning agent representation:

$$z_i^k = \phi_i(f_i^k) \in \mathbb{R}^{\mathcal{D}}. \tag{2}$$

Learnable Agent. Since the three fine-tuned reasoning agents are non-learnable during the subsequent strategic game, we introduce a *learnable agent* v_C to enhance the understanding of strategies. To reduce computational burden, we choose a CLIP-based [36] model as the learnable agent due to its low parameter count and strong multimodal understanding capabilities.

Given the labeled sample ξ_k, the CLIP text encoder maps T_k to token embeddings $S_k = \{s_1, \ldots, s_n, s_{\text{cls}}\}$, and the CLIP image encoder maps I_k to patch embeddings $V_k = \{v_{\text{cls}}, v_1, \ldots, v_m\}$. We concatenate the two sequences and feed the result into a K-layer Transformer encoder [45] to model cross-modal interactions. The resulting aggregate vectors are denoted by $\tilde{v}_{\text{cls}}$ and $\tilde{s}_{\text{cls}}$. We then perform a lightweight fusion in the feature space:

$$\begin{aligned} (p_s,\, p_v) &= \text{softmax}\big(W_f([\tilde{s}_{\text{cls}}; \tilde{v}_{\text{cls}}])\big), \\ f_{v_C}^k &= p_s\, \tilde{s}_{\text{cls}} + p_v\, \tilde{v}_{\text{cls}}\,, \end{aligned} \tag{3}$$

where W_f is a learnable mapping. We project $f_{v_C}^k$ into the space $\mathcal{D}$ to generate the learnable agent representation:

$$z_{v_C}^k = \phi_{v_C}(f_{v_C}^k) \in \mathbb{R}^{\mathcal{D}}. \tag{4}$$

Master agent. The *master agent* v_F aggregates the agents' embeddings and produces the final policy used for prediction. Acting as the ultimate decision-maker within the system, it derives its payoff not only from its own actions but also from those of other agents, thereby enhancing the robustness of decision-making.

The projected features z_i^k are concatenated to generate the master agent representation:

$$z_{v_F}^k = [\, z_{v_L}^k;\, z_{v_Q}^k;\, z_{v_C}^k;\, z_{v_G}^k \,] \in \mathbb{R}^{4\mathcal{D}}. \tag{5}$$

3.1.2. Action Execution

Each agent $v_i \in \mathcal{V} = \{v_L, v_Q, v_C, v_G, v_F\}$ provides a *policy* π_i over its binary action space $\mathcal{A}_i = \{0, 1\}$, constructing the joint action set $\mathcal{A}^5 = \{a_{v_L}, a_{v_Q}, a_{v_C}, a_{v_G}, a_{v_F}\}$. The policy is defined as a probability distribution $\pi_i(\cdot \mid \xi_k) \in \Delta(\mathcal{A}_i)$. To map the representation z_i^k to a concrete action preference, we use an agent-specific policy head ψ_i to produce logits ℓ_i^k, followed by a temperature-scaled softmax:

$$\begin{aligned} \ell_i^k &= \psi_i(z_i^k) \in \mathbb{R}^{|\mathcal{A}_i|}, \\ \pi_i(a_i \mid \xi_k) &= \frac{\exp\big((\ell_i^k)_{a_i}/\tau_i\big)}{\sum_{a' \in \mathcal{A}_i} \exp\big((\ell_i^k)_{a'}/\tau_i\big)}\,. \end{aligned} \tag{6}$$

We define the mixed bonus vector x_i of agent v_i as $x_i(a) \equiv \pi_i(a \mid \xi_k)$ for all $a \in \mathcal{A}_i$, i.e.,

$$x_i = (x_i(a_1), x_i(a_2), \ldots, x_i(a_{|\mathcal{A}_i|})) \in \mathcal{X}_i. \tag{7}$$

where $\mathcal{X}_i$ is the set of agent strategies i. Let $x = \{x_i \mid i \in \mathcal{V}\}$ denote the strategy profile, the predicted label $\hat{y}_k$ is:

$$\hat{y}_k = \arg\max_{a \in \mathcal{A}_{v_F}} x_{v_F}(a). \tag{8}$$

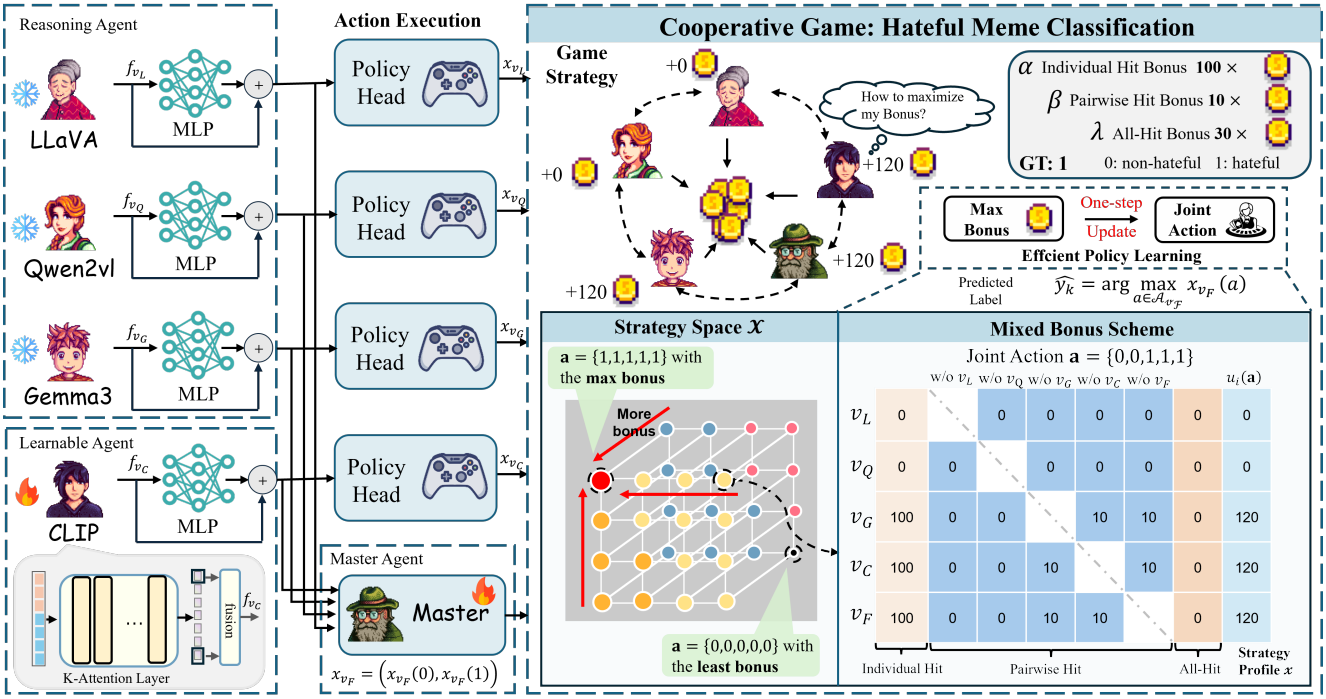


Figure 2. Overview of the proposed GECO. Reasoning agents (LLaVA, Qwen2-VL, Gemma3) and a learnable CLIP-based agent interact under a regularized normal-form game, coordinated by a master agent. We utilize a mixed bonus scheme to optimize the entire framework, employing efficient policy learning to ensure stable equilibrium convergence.

3.2. Game Strategy

3.2.1. Mixed Bonus Scheme

Unlike typical applications in game theory that focus on individual agent performance, our approach prioritizes overall consensus within the collaborative framework to overcome inherent model bias and enhance complementarity. Hence, we design a mixed bonus scheme that encourages both individual accuracy and mutual agreement on the correct label. Given a joint action $\mathbf{a} = \{a_{v_L}, a_{v_Q}, a_{v_C}, a_{v_G}, a_{v_F}\} \in \mathcal{A}^5$, $\mathcal{A}^5$ is the joint action set, and the label y_k, the mixed bonus scheme for agent i can be defined as:

$$\begin{aligned} u_i(a_i, a_{-i}) = \alpha \cdot \mathbb{I}(a_i = y_k) \\ + \lambda \sum_{j \in \mathcal{V} \setminus \{i\}} \mathbb{I}(a_i = y_k)\, \mathbb{I}(a_j = y_k) \\ + \beta \cdot \mathbb{I}\big(\forall j \in \mathcal{V},\, a_j = y_k\big), \end{aligned} \tag{9}$$

where $\mathbb{I}(\cdot)$ is the indicator function and a_{-i} collects the actions of all agents except i. The scheme consists of three terms: an individual hit bonus α, a pairwise hit bonus λ, and an all-hit bonus β.

The objective is to optimize all agents' expected utility. Let the joint action space be $\times_{j \in \mathcal{V}} \mathcal{A}_j$, the expected utility of agent i under strategy profile $x = \{x_j\}_{j \in \mathcal{V}}$ is

$$U_i(x_i, x_{-i}) = \sum_{\mathbf{a} \in \times_{j \in \mathcal{V}} \mathcal{A}_j} u_i(\mathbf{a}) \cdot \prod_{j \in \mathcal{V}} x_j(a_j). \tag{10}$$

where $\mathbf{a}$ is the joint action.

3.2.2. Efficient Policy Learning

Our goal is for each agent to learn the overall optimal strategy in the game through sampling. Unlike traditional game theory definitions [11], since each agent must choose from a binary action space—and incorrect classifications do not contribute to the overall expectation—we restrict sampling to the correct actions for each agent. On one hand, this approach enhances computational efficiency, and on the other, due to the reduced action space, we implement single-step updates to avoid the noise associated with multiple sampling [12, 30]. We define the expected conditional utility $U_i(a_i, x_{-i})$ as the expected bonus for agent i when taking action a_i while other agents adhere to actions x_{-i}:

$$U_i(a_i, x_{-i}) = \sum_{a_{-i} \in \times_{j \in \mathcal{V} \setminus \{i\}} \mathcal{A}_j} u_i(a_i, a_{-i}) \cdot \prod_{j \in \mathcal{V} \setminus \{i\}} x_j(a_j). \tag{11}$$

This computation is efficiently implemented using vectorized tensor operations, producing a vector of expected bonuses for all actions. To ensure that training can achieve

the Nash equilibrium, we introduced regularization based on theoretical insights [30]. Thus, we define the regularized advantage vector $\mathbf{F}_i^x$ and its scalar components $F_i^x(a_i)$:

$$\begin{aligned} \mathbf{F}_i^x &= -\nabla_{x_i} u_i(x) + \eta \log(\mathbf{x}_i), \\ F_i^x(a_i) &= -U_i(a_i, x_{-i}) + \eta \log x_i(a_i), \end{aligned} \tag{12}$$

where η is the balance parameter. For improved training stability, the regularized advantage is centered by subtracting a baseline value B_i, computed under a baseline strategy $\hat{x}_i$ (e.g., uniform distribution):

$$B_i = \langle \mathbf{F}_i^x, \hat{x}_i \rangle = \sum_{a_i} \hat{x}_i(a_i) F_i^x(a_i). \tag{13}$$

The centered advantage vector is then given by:

$$\mathbf{A}_i = \mathbf{F}_i^x - B_i \mathbf{1}, \tag{14}$$

where $\mathbf{1}$ denotes an all-ones vector.

The Regularized Nash Advantage loss $\mathcal{L}_{\text{RNA}}(x)$ for agent i is defined as the inner product between its current strategy x_i and the stop-gradient version of the centered advantage vector:

$$\begin{aligned} \mathcal{L}_{\text{RNA}}(x) &= \sum_{i \in \mathcal{V}} \left\langle \text{sg}(\mathbf{F}_i^x - \langle \mathbf{F}_i^x, \hat{x}_i \rangle \, \mathbf{1}) \, , \, x_i \right\rangle \\ &= \sum_{i \in \mathcal{V}} \left\langle \text{sg} \, \mathbf{A}_i, x_i \right\rangle . \end{aligned} \tag{15}$$

The stop-gradient operator (sg) treats $\mathbf{A}_i$ as a fixed constant during backpropagation, simplifying gradient computation and improving numerical stability. By minimizing this loss, the strategy x_i is directed towards actions with a positive advantage, guiding the system to maximize overall benefit.

3.2.3. Regularization and Final Objective

To stabilize updates and reduce oscillations, we regularize the current policy $p = \pi_{v_F}$ toward a slowly moving reference policy q, maintained by EMA [22]: $q_t \leftarrow \mu q_{t-1} + (1-\mu) p_t$. We define a symmetric KL-style regularizer as

$$J_\gamma(p, q) = (1-\gamma) \, D_{\text{KL}}(q \| p) + \gamma \, D_{\text{KL}}(p \| q). \tag{16}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{RNA}} + J_\gamma(p, q), \tag{17}$$

where $\mathcal{L}_{\text{RNA}}$ optimizes the game objective and J_γ enforces smooth policy evolution.

4. Experimental Setup

The main experiments were conducted on the five publicly available datasets: **Hateful Meme** [19], **PrideMM** [40], **MultiOff** [43], **HarMeme** [34], and **MAMI** [43].

4.1. Baselines

We evaluate our model against two categories of baselines: (1) **CLIP-based methods**, including CLIP [36], MOMENTA [35], Hate-CLIPper [20], MemeCLIP [40], and RGCL [28]; (2) **LMM-based methods**, including IntMeme [14], ExplainHM [23], Mod-Hate [4],LoReHM [16], M2KE [27], and RA-HMD [29], the previous state-of-the-art (SOTA) model.

4.2. Implementation Details

We adopt LLaVA-1.5-13B [25], Qwen2-VL-7B [49], and Gemma3-4B [44] as LMM agents, covering a spectrum from 4B to 13B. Unlike prior works [14, 23, 29], which typically employ 7B–13B backbones for multimodal reasoning, our selection intentionally includes both larger and smaller variants to balance computational cost and to assess the scalability of the cooperative framework across model sizes. All agents are projected into a unified decision space $\mathcal{D}$ with a dimensionality of 768. We employ the AdamW optimizer with an initial learning rate of 2×10^{-5} for non-CLIP parameters and 5×10^{-5} for CLIP-related modules. The regularization coefficient η for the regularized advantage term is set to 0.35. The individual, pairwise, and all-hit bonuses are set to $\alpha = 1.0$, $\lambda = 0.5$, and $\beta = 1.0$, respectively. The mixing coefficient γ is set to 0.5. Following previous work [29], we report Accuracy (Acc), F1-Score (F1), and Area Under the Curve (AUC) as evaluation metrics.

5. Results and Discussions

5.1. Main Results

We evaluate the proposed GECO by comparing it with CLIP-based and LMM-based methods on five publicly available datasets. The main results are shown in Table 1. Our analysis yields the following insights:

(1) GECO achieves consistent SOTA results across all benchmarks, confirming its effectiveness in enhancing cooperative decision-making among heterogeneous agents. On PrideMM, GECO surpasses the previous best, RA-HMD, by +4.74% in accuracy, while on MAMI, it attains 81.50% accuracy and 82.84% F1, outperforming all prior methods. Moreover, GECO demonstrates strong robustness under low-resource conditions. On the MultiOff dataset, which contains fewer than 500 samples and poses a significant challenge for single-model systems, GECO achieves 78.52% accuracy, exceeding RA-HMD by +7.41%. The performance gain becomes particularly evident in such limited supervision scenarios, where cooperative optimization allows agents to exchange complementary visual–textual cues and form more stable decision boundaries. These results collectively validate GECO's advantage in improving agreement and generalization across diverse data regimes.

Table 1. Comparison with baseline systems across five benchmark datasets. Best performance is highlighted in **bold**, while "–" indicates that the relevant paper lacks results for this dataset, and † denotes the debate-based ensemble methods.

Category	Method	PrideMM		HatefulMemes		MAMI		HarMeme		MultiOff	
		Acc	F1	Acc	AUC	Acc	AUC	Acc	AUC	Acc	F1
CLIP-Based	CLIP [36] (2021)	72.39	72.33	72.04	79.81	77.70	68.40	76.78	82.63	62.41	48.14
	MOMENTA [35] (2021)	72.23	71.78	61.34	69.17	72.10	81.68	80.48	86.32	–	–
	Hate-CLIPper [20] (2022)	75.53	74.08	76.10	85.48	74.80	87.20	84.80	89.72	62.40	54.80
	MemeCLIP [40] (2024)	76.06	76.09	–	–	–	–	84.72	83.74	–	–
	RGCL [28] (2024)	76.34	76.50	78.92	87.04	78.40	89.40	87.03	91.91	67.13	58.11
LMM-Based	ExplainHM† [23] (2024)	–	–	75.60	–	–	–	87.00	–	–	–
	Mod-Hate [4] (2024)	–	–	59.07	64.69	61.10	67.20	69.85	73.26	–	–
	LoReHM [16](2024)	–	–	65.60	–	75.40	–	73.73	–	–	–
	M2KE† [27] (2025)	–	–	75.76	75.62	75.85	–	–	–	–	–
	IntMeme [14] (2025)	–	–	71.52	81.50	72.30	81.89	81.92	89.35	–	–
	RA-HMD [29] (2025)	78.10	78.70	82.10	91.10	79.90	90.40	88.10	93.20	71.11	64.80
	GECO	**82.84**	**82.84**	**84.35**	**91.57**	**81.50**	**91.80**	**89.11**	**93.95**	**78.52**	**77.90**

(2) LMM-based methods surpass CLIP-based approaches due to their stronger reasoning ability, enabling better understanding of subtle semantics in hateful meme detection. However, the reliance on a single LMM perspective often leads to the entrenchment of inherent biases, ultimately compromising both the fairness and robustness of detection systems. In contrast, GECO significantly outperforms LMM-based models, demonstrating that such biases are not inevitable. Instead, they can be effectively mitigated by introducing heterogeneity and collaborative mechanisms, which not only address the limitations of individual models but also substantially enhance overall performance.

(3) GECO consistently outperforms debate-based methods such as ExplainHM and M2KE. Although debate methods introduce agent interaction, their reliance on fixed dialogue rules or biased judges limits stability and accuracy. In contrast, GECO's game-theoretic design enables adaptive cooperation among agents, yielding stronger consensus and higher overall performance across benchmarks.

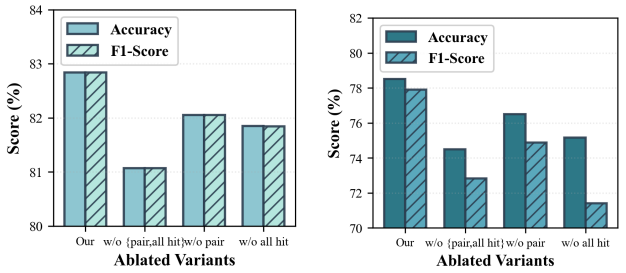
5.2. Ablation Study

To validate the proposed architecture and examine the individual and joint contributions of its heterogeneous agents, extensive ablation experiments were performed, as summarized in Table 2. In single-agent ablations, removing the classification agent v_F results in the most significant performance degradation, confirming the necessity of explicitly incorporating the classifier as a player within the cooperative game. The core reasoning agent v_L (LLaVA) ranks second in importance, reflecting its central role in vision–language understanding. The divergent agent v_C (CLIP) contributes essential cross-modal semantic alignment, while the complementary agents v_Q (Qwen) and v_G (Gemma) provide additional contextual reasoning and linguistic diversity. Beyond individual impacts, two-agent ablations reveal pronounced synergistic dependencies: removing any pair of agents leads to performance degradation substantially exceeding the additive effect of their independent removals. These results substantiate the effectiveness of the hierarchical reward design in fostering cooperative interactions. Overall, optimizing heterogeneous agents to a regularized Nash equilibrium via Nash Advantage Learning effectively mitigates single-model biases and leverages complementary knowledge, yielding superior fairness and robustness in hateful meme detection.

Table 2. Ablation study on PrideMM and MultiOff datasets.

Category	Mode	PrideMM		MultiOff	
		ACC	F1	ACC	F1
	Full Model	**82.84**	**82.84**	**78.52**	**77.90**
Single Agent	w/o v_L	81.66	81.62	74.50	71.94
	w/o v_C	82.05	82.05	75.17	73.24
	w/o v_Q	82.45	82.44	73.83	73.00
	w/o v_G	82.25	82.25	75.84	74.06
	w/o v_F	62.88	62.16	62.42	42.97
Multiple Agents	w/o $\{v_L, v_Q\}$	79.88	79.88	73.15	70.72
	w/o $\{v_L, v_C\}$	79.68	79.65	73.15	70.72
	w/o $\{v_L, v_G\}$	80.08	80.06	72.48	67.16
	w/o $\{v_Q, v_G\}$	80.47	80.45	76.51	74.88
	w/o $\{v_Q, v_C\}$	80.67	80.66	71.81	68.39
	w/o $\{v_C, v_G\}$	80.87	80.84	73.83	71.32

To deeply validate the specific contributions and synergistic effects of each component within our designed cooperative reward mechanism, we conducted a detailed ablation study, with results presented in Figure 3. Our full model, which simultaneously utilizes individual hit (α), pairwise hit (λ), and all hit (β) bonuses, significantly outperforms all ablated variants. Specifically, when all cooperative re-



(a) PrideMM Performance (b) MultiOff Performance

Figure 3. Ablation study of cooperative reward mechanism on PrideMM and MultiOff datasets.

Table 3. Comparison between supervised fine-tuning (SFT) and our game-theoretic variant (GECO) across different LMM backbones on the PrideMM and MultiOff datasets.

Model	Method	PrideMM		MultiOff	
		ACC	F1	ACC	F1
LLaVA	SFT	78.70	78.74	67.79	**58.62**
	GECO	**81.45**	**81.64**	**70.46**	57.10
Qwen	SFT	77.51	77.20	66.44	44.44
	GECO	**78.90**	**78.90**	**68.46**	**44.47**
Gemma	SFT	75.74	75.74	65.10	65.79
	GECO	**77.91**	**77.78**	**76.51**	**71.54**

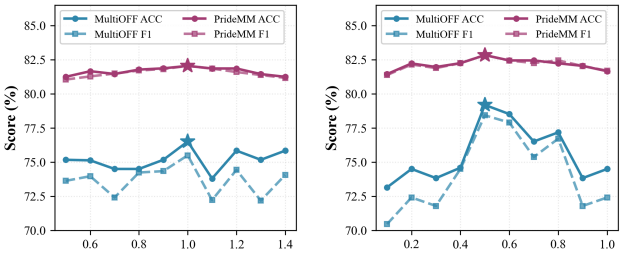
wards are removed, model performance drops to its lowest point, clearly indicating that a lack of cooperative incentives causes agents to focus solely on individual objectives, thus failing to coordinate effectively. Furthermore, removing only the pairwise hit bonus ("wo-pair") or only the all hit bonus ("wo-coop") also results in significant performance degradation. This study forcefully demonstrates that each reward component we designed is indispensable for promoting effective agent cooperation; it is this carefully orchestrated combination of individual incentives and multi-level cooperative incentives that successfully guides the heterogeneous agents to transcend their inherent biases and, through negotiation, converge on a collectively optimal and more robust classification decision.

5.3. Non-game-theoretic Strategies Comparison

Table 3 compares the performance of three LMMs optimized with our GECO against their directly supervised fine-tuned (SFT) counterparts. It is observed that GECO surpasses SFT on the vast majority of metrics across both PrideMM and MultiOff datasets, demonstrating the effectiveness of introducing strategic coordination among agents. By enabling cooperative optimization through a mixed bonus–penalty mechanism, GECO encourages agents to exploit complementary strengths while mitigating individual biases, leading to more consistent and robust single-model performance.

Table 4. Comparison between GECO and non-game-theoretic ensemble strategies on the PrideMM and MultiOff datasets.

Category	Method	PrideMM		MultiOff	
		ACC	F1	ACC	F1
Non-Game	Voting	79.49	79.37	69.80	59.46
	Debate	79.88	79.87	73.15	70.72
Game	**GECO**	**82.84**	**82.84**	**78.52**	**77.90**



(a) All-hit bonus β (b) Pairwise hit bonus λ

Figure 4. Parameter Analysis of all hit bonus β and pairwise hit bonus λ on PrideMM and MultiOff datatsets.

We further compare GECO with multi-agent ensemble strategies, including Voting-based and Debate-based ensembles, as shown in Table 4. While ensemble methods improve over single SFT models, they still underperform compared to GECO. Voting-based ensembles treat all agents equally, overlooking the heterogeneity and varying reliability of different models. Debate-based ensembles introduce interaction among agents, yet their effectiveness heavily depends on the fairness and stability of the judging process, which may propagate bias rather than correct it. In contrast, GECO explicitly formulates agent cooperation as a game, driving adaptive alignment through strategic optimization. As a result, GECO achieves the highest accuracy and F1 across datasets, confirming its advantage in reducing model bias and improving decision reliability.

5.4. Parameter Analysis

To investigate the influence of the reward mechanism, we conduct a parameter analysis on both the pairwise hit bonus β and the all hit bonus λ, while holding action a constant at its default value of 1.0. The results on both the PrideMM and MultiOff datasets are presented in Figure 4. First, to isolate the effect of the all-hit bonus β, we set the pairwise hit bonus λ to 0. GECO exhibits stable performance as the β varies within the [0.9, 1.2] range, demonstrating our method's low sensitivity to this parameter. Subsequently, we fix the β to 1.0 to evaluate the impact of the pairwise hit bonus λ. We observe a significant degradation in both ACC and F1 when the λ is set to minimal (≤ 0.3) or excessive (≥ 0.9) values. This phenomenon suggests that an insuffi-

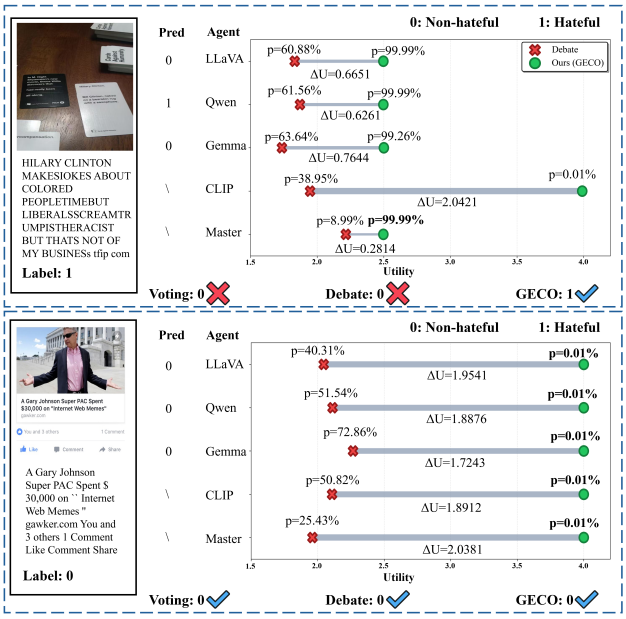


Figure 5. Case study illustrating the decision performance of different methods. ΔU represents the difference in conditional expected utility between our method and the ensemble method.

cient pairwise incentive fails to promote effective collaboration, while an overly strong incentive introduces detrimental bias and increases the risk of overfitting.

5.5. Case Study

We present two cases in Figure 5 to clarify the advantages of our GECO over voting and debate-based methods. In the first case, the agents exhibited disagreement on whether the content was hateful, leading the simple voting-based method to an incorrect classification. The debate-based model likewise failed, as the interaction among agents did not resolve the disagreement, and the biased judgment of the referee resulted in a wrong decision with extremely low confidence ($p = 8.99\%$). In contrast, GECO, through its cooperative game-theoretic framework, guides all agents to evolve toward a jointly optimal strategy with higher overall utility (U). The model achieves a superior U, successfully overcoming the disagreement and correctly classifying the sample with high confidence ($p = 99.99\%$). The effectiveness of this cooperative strategy is further demonstrated in the second case. Although all models achieved correct classifications, GECO still exhibits higher decision confidence, leveraging its superior U to reach near-deterministic certainty ($p = 0.01\%$). This stands in stark contrast to the Debate-based method, which remained uncertain ($p = 25.43\%$) due to inconsistent reasoning among agents.

To further illustrate how GECO leverages game-theoretic mechanisms to mitigate inherent cognitive biases, we present a visualized case analysis in Figure 6, which intuitively depicts the decision-making process of different

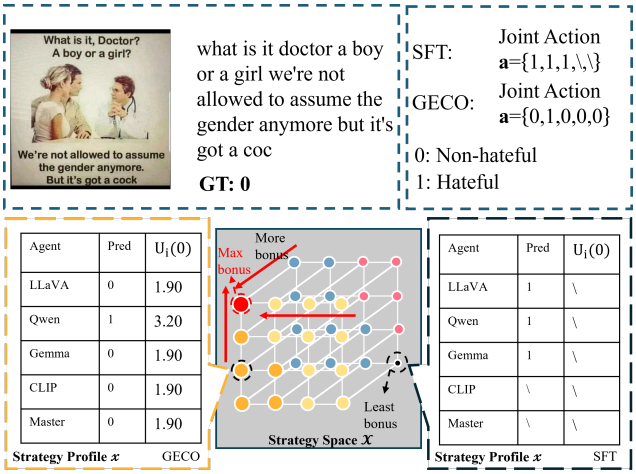


Figure 6. Visualized case analysis illustrating how GECO corrects biased collective predictions in the decision-making process.

agents. This case contains sensitive lexical cues that lead all SFT-based models to exhibit bias, erroneously classifying the sample as hateful. After applying GECO, the system successfully corrects this collective failure, steering the majority of agents toward the correct non-hateful prediction. Crucially, the utility analysis reveals the intrinsic driving dynamics of the game-theoretic mechanism: although Qwen initially retains an incorrect prediction, its potential utility for adopting the correct action (0) reaches 3.20—substantially higher than its current payoff. This disparity, driven by the strong incentive of the all-hit bonus, motivates the model to adjust toward the correct decision during subsequent optimization. Conversely, agents that have already reached the correct consensus reside in a stable Nash equilibrium ($U_i(0) = 1.90$), where any unilateral deviation would reduce their individual payoff. This analysis further demonstrates GECO's effectiveness in guiding the multi-agent system to evolve from a low-payoff, biased region toward a high-payoff region of correct consensus.

6. Conclusion and Future Work

In this paper, we introduced GECO, a game-theoretic multi-agent framework for mitigating bias in hateful meme detection. By coordinating heterogeneous LMMs through a cooperative objective, GECO improves agreement and reduces individual model bias, resulting in more robust and reliable predictions across diverse meme scenarios. Extensive experiments on multiple benchmarks demonstrate that GECO consistently surpasses state-of-the-art methods. Ablation analyses and detailed case studies further validate the necessity of orchestrating diverse agents and the effectiveness of the game-theoretic strategy. In future work, we will refine GECO's optimization and collaboration mechanisms to support more complex multimodal learning scenarios.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 6250010550), the Natural Science Foundation of Xinjiang Uygur Autonomous Region(No. 2024D01B99), and the Introduction Program for Young Doctors (Tianchi Talents) of Xinjiang Uygur Autonomous Region. We acknowledge Stardew Valley and ConcernedApe LLC for the original character artwork used in several figures of this paper. All rights to these materials remain with the copyright holder.

References

[1] Thomas Anthony, Tom Eccles, Andrea Tacchetti, János Kramár, Ian Gemp, Thomas Hudson, Nicolas Porcel, Marc Lanctot, Julien Pérolat, Richard Everett, et al. Learning to play no-press diplomacy with best response policy iteration. *Advances in Neural Information Processing Systems*, 33:17987–18003, 2020. 3

[2] Robert J Aumann. Markets with a continuum of traders. *Econometrica: Journal of the Econometric Society*, pages 39–50, 1964. 2

[3] Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. 2

[4] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, pages 4575–4584, 2024. 2, 5, 6

[5] Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2

[6] Gerard Debreu and Herbert Scarf. A limit theorem on the core of an economy. *International Economic Review*, 4(3): 235–246, 1963. 2

[7] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. 3

[8] Neil Fasching and Yphtach Lelkes. Model-dependent moderation: inconsistencies in hate speech detection across llm-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22271–22285, 2025. 2

[9] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022. 2

[10] Drew Fudenberg and Jean Tirole. Game theory mit press. *Cambridge, MA*, 86, 1991. 2

[11] Ian Gemp, Luke Marris, and Georgios Piliouras. Approximating nash equilibria in normal-form games via stochastic optimization. In *The Twelfth International Conference on Learning Representations*. 4

[12] Ian Gemp, Luke Marris, and Georgios Piliouras. Approximating nash equilibria in normal-form games via stochastic optimization. *arXiv preprint arXiv:2310.06689*, 2023. 4

[13] Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118, 2021. 1

[14] Ming Shan Hee and Roy Ka-Wei Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 774–785, 2025. 2, 5, 6

[15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 1, 3

[16] Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. Towards low-resource harmful meme detection with LMM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA, 2024. Association for Computational Linguistics. 5, 6

[17] Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. *arXiv preprint arXiv:2502.19130*, 2025. 2

[18] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 2

[19] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 2, 5

[20] Gokul Karthik Kumar and Karthik Nandakumar. Hateclipper: 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, 2022. 5, 6

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5265–5275, 2020. 1

[22] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 5

[23] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM web conference 2024*, pages 2359–2370, 2024. 2, 5, 6

[24] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and

Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 3, 5

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[27] Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. Is having rationales enough? rethinking knowledge enhancement for multimodal hateful meme detection. In *SIGIR*, pages 559–569, 2025. 2, 5, 6

[28] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In *ACL*, pages 5333–5347, 2024. 5, 6

[29] Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. pages 23817–23839, 2025. 1, 5, 6

[30] Linjian Meng, Wubing Chen, Wenbin Li, Tianpei Yang, Youzhi Zhang, and Yang Gao. Reducing variance of stochastic optimization for approximating nash equilibria in normal-form games. In *Forty-second International Conference on Machine Learning*, 2025. 4, 5

[31] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020. 2

[32] Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167, 2021. 1

[33] John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950. 2

[34] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, 2021. 2, 5

[35] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, 2021. 5, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 5, 6

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 1

[38] Karthik Raman. *Machine learning from human preferences and choices*. Cornell University, 2015. 2

[39] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010. 2

[40] Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. Memeclip: Leveraging clip representations for multimodal meme classification. In *EMNLP*, pages 17320–17332, 2024. 2, 5, 6

[41] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008. 2

[42] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359, 2017. 3

[43] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41, 2020. 2, 5

[44] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2, 3, 5

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[46] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020. 1, 2

[47] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019. 3

[48] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947. 2

[49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3, 5

[50] Xiutian Zhao, Ke Wang, and Wei Peng. An electoral approach to diversify llm-based multi-agent collective decision-making. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2727, 2024. 2

[51] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*, 2020. 1