More Than Sum of Its Parts: Deciphering Intent Shifts in Multimodal Hate Speech Detection

Content warning: This article contains examples of hateful content.

Runze Sun, Yu Zheng*, Zexuan Xiong, Zhongjin Qu, Lei Chen, Jie Zhou, Jiwen Lu
Department of Automation, Tsinghua University
{sunrz22,xiong-zx22,qzj23}@mails.tsinghua.edu.cn
{yu-zheng,leichenthu,jzhou,lujiwen}@tsinghua.edu.cn

Abstract

Combating hate speech on social media is critical for securing cyberspace, yet relies heavily on the efficacy of automated detection systems. As content formats evolve, hate speech is transitioning from solely plain text to complex multimodal expressions, making implicit attacks harder to spot. Current systems, however, often falter on these subtle cases, as they struggle with multimodal content where the emergent meaning transcends the aggregation of individual modalities. To bridge this gap, we move beyond binary classification to characterize semantic intent shifts where modalities interact to construct implicit hate from benign cues or neutralize toxicity through semantic inversion. Guided by this fine-grained formulation, we curate the **Hate via Vision-Language Interplay (H-VLI)** benchmark where the true intent hinges on the intricate interplay of modalities rather than overt visual or textual slurs. To effectively decipher these complex cues, we further propose the **A**symmetric **R**easoning via **C**ourtroom **A**gent **DE**bate **(ARCADE)** framework. By simulating a judicial process where agents actively argue for accusation and defense, ARCADE forces the model to scrutinize deep semantic cues before reaching a verdict. Extensive experiments demonstrate that ARCADE significantly outperforms state-of-the-art baselines on H-VLI, particularly for challenging implicit cases, while maintaining competitive performance on established benchmarks. Our code and data are available at: https://github.com/Sayur1n/H-VLI

1 Introduction

Hate speech, defined as attacks targeting individuals or groups based on protected characteristics (e.g., race, gender, religion), poses a severe threat to the safety of online communities (Gagliardone et al., 2015; Matsuda, 2018; Kiela et al., 2020; Cortese, 2005; Tsesis, 2002). Given the sheer volume and contagion of social media content, there is an urgent need for automated detection systems to safeguard both online and offline communities.

(a) *Metaphoric religious hate*

(b) *Pun-based racial hate*

Figure 1: Examples of implicit multimodal hate speech.

Unimodal hate speech detection has made considerable progress. Existing approaches now excel at identifying both explicit (Schmidt and Wiegand, 2017; Caselli et al., 2020; Davidson et al., 2017; Waseem and Hovy, 2016) and implicit (Ghosh et al., 2023; Ahn et al., 2024; Jafari et al., 2024; Zeng et al., 2025) hate speech within the text modality. However, as online content evolves into multimodal formats like memes and image–text pairs, hate speech has increasingly emerged from cross-modal interactions, shifting the battlefield toward Multimodal Hate Speech Detection (MMHSD) (Gomez et al., 2020; Kiela et al., 2020; Wang et al., 2025; Kapil and Ekbal, 2025).

Unlike explicit hate speech characterized by aggressive slurs or violent imagery, modern hate speech increasingly relies on implicit expressions. In these cases, the textual and visual modalities may appear benign individually, yet they mutually construct a hateful intent through complex semantic interactions, such as irony, metaphor, or cultural allusion.

Consider the examples in Figure 1. In both cases, the standalone text and images appear benign. However, their combination generates implicit hatefulness: Figure 1a uses a hostile metaphor invoking Islamophobic stereotypes, while Figure 1b relies on a visual-textual pun (the word "race") to convey racial hostility. Correctly deciphering these

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 19494–19514

instances requires models to move beyond superficial fusion and perform deep reasoning to capture how modalities amplify, contradict, or recontextualize each other.

However, existing research struggles with these complexities on two fronts. First, coarse and binary benchmarks lack interaction taxonomies, leading models to overfit surface cues. Second, standard direct fusion mechanisms fail to capture subtle semantic conflicts, leaving models vulnerable to "ambiguity traps" where they misinterpret benign satire or miss implicit attacks.

To bridge these gaps, our core contributions are highlighted as follows:

- **The SMI Paradigm:** We introduce a fine-grained **Stratified Multimodal Interaction (SMI)** paradigm. To characterize intent shifts, it operationalizes semantic interactions by assigning separate labels to the standalone text, the standalone image, and the combined pair. The visual-textual interaction is then explicitly classified based on the combination of these three labels.
- **The H-VLI Benchmark:** Guided by SMI paradigm, we curate **Hate via Vision-Language Interplay (H-VLI)**. This high-quality benchmark is constructed via a hybrid pipeline of consensus filtering, generative injection, and human-in-the-loop annotation, ensuring a remarkably high density of challenging implicit hate samples.
- **The ARCADE Framework:** We propose **A**symmetric **R**easoning via **C**ourtroom **A**gent **DE**bate (**ARCADE**), which simulates a judicial process using asymmetric agents: a Prosecutor (presuming guilt) and a Defender (presuming innocence). This adversarial dialectic forces the model to deeply scrutinize semantic interplay and cultural contexts before a Judge renders the final verdict, significantly enhancing detection accuracy and interpretability.

2 Related Work

Hate Speech Detection: Hate speech detection aims to identify malicious content targeting specific social groups (Schmidt and Wiegand, 2017; Davidson et al., 2017). While early unimodal studies focus on *explicit* surface linguistic cues (Nobata et al., 2016; Davidson et al., 2017; Caselli et al., 2020), recent works address *implicit* hate through contextual and semantic reasoning (Ghosh et al., 2023; Ahn et al., 2024; Wei et al., 2025; Zeng et al., 2025). However, these text-only methods remain insufficient for the increasingly multimodal nature of online content. To address this, Multimodal Hate Speech Detection (MMHSD) approaches, as a variant of multimodal forensics (Wang et al., 2018; Zhang et al., 2025; Zheng et al., 2025), typically integrate representations via feature fusion, cross-modal attention, or contrastive learning (Gomez et al., 2020; Dwivedy and Roy, 2023; Saddozai et al., 2025; Kapil and Ekbal, 2025). Although effective for explicit cases, these models often struggle when hateful meaning is conveyed indirectly through shared socio-cultural knowledge. More recently, methods based on Multimodal Large Language Models (MLLMs) have explored prompt-based, few-shot or fine-tuning paradigms (Wang et al., 2025; Rizwan et al., 2025; Li et al., 2025), yet they generally lack mechanisms to explicitly guide reasoning or systematically elicit the necessary socio-cultural context for complex implicit scenarios.

Multimodal Hate Speech Detection Benchmarks: Existing MMHSD datasets are generally categorized into text–image pair formats like MMHS150K (Gomez et al., 2020), and meme-style datasets where pieces of text are embedded within images (Kiela et al., 2020; Fersini et al., 2022a). While largely advancing the field, they feature relatively explicit cross-modal cues and utilize coarse-grained labeling schemes. Consequently, existing benchmarks largely focus on explicit hate, with limited coverage of context-dependent neutralization and implicitly constructed hate arising from subtle metaphors and stereotypes—limiting progress toward robust real-world detection.

Multi-Agent Debate: Multi-agent debate (MAD) enhances LLM reasoning by assigning agents distinct roles for structured argumentation (Liang et al., 2024), and has been applied to misinformation detection and factual verification (Han et al., 2025; Ma et al., 2025). Asymmetric debate architectures further improve decision robustness (Park et al., 2024; Kumar et al., 2025; Jin et al., 2025). However, these text-centric methods overlook cross-modal semantic interplay, missing subtle cues essential for implicit content understanding. MV-Debate (Lu et al., 2025) extends MAD to the multimodal domain via view-specific vision-language agents, yet such consensus-driven approaches may still struggle with deceptive implicit hate, where modalities actively interact to obfuscate malicious intent through tropes or metaphors,

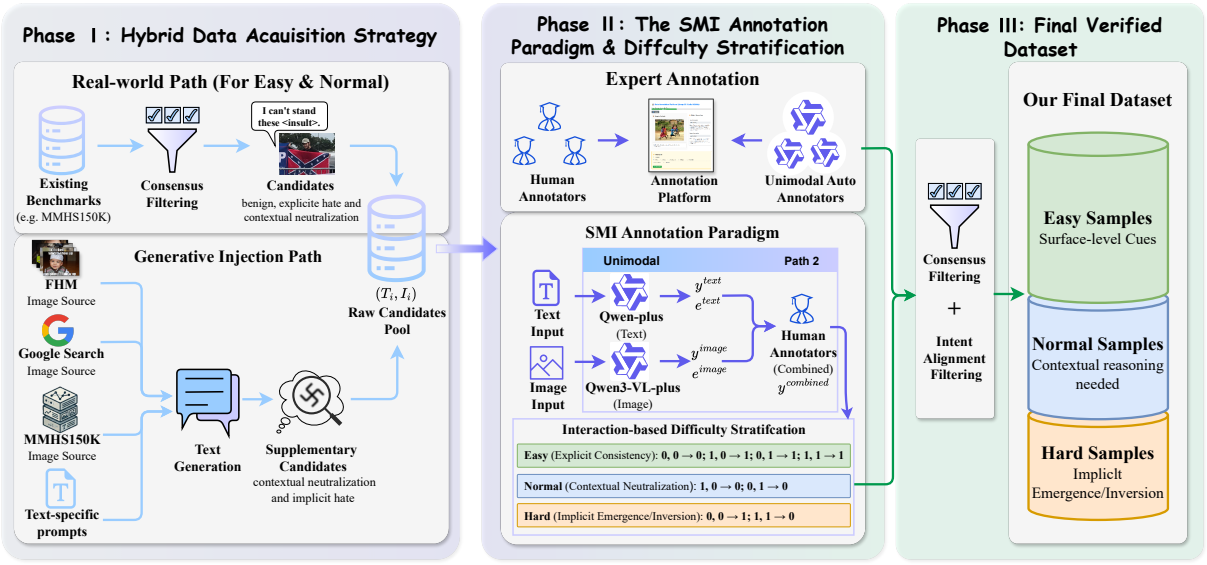


Figure 2: The construction pipeline of our **H-VLI** dataset. This process employs a hybrid strategy that combines consensus filtering of real-world samples with a generative injection path to ensure diversity. It utilizes the SMI paradigm to systematically categorize samples into eight distinct interaction patterns, which are further stratified into three difficulty levels (Easy, Normal, and Hard) based on the semantic interplay between modalities.

necessitating a shift from collaborative alignment to adversarial scrutiny.

3 Our Dataset H-VLI

To advance the detection of implicit multimodal content (where hateful intent is implicitly constructed from benign signals or semantically neutralized by context), it is essential to move beyond binary labels and decipher the specific semantic interactions between modalities. However, existing benchmarks typically rely on coarse-grained categories and often suffer from substantial label noise, as reflected by low inter-annotator agreement (e.g., MMHS150K in Table 2). To bridge this gap, we first formulate the problem through a fine-grained lens, introduce the **Stratified Multimodal Interaction (SMI)** paradigm, and finally construct **H-VLI**, a high-quality benchmark systematically populated based on these interaction patterns.

3.1 Problem Formulation

Many existing MMHSD studies formulate the task simply as a *binary classification* problem (Gomez et al., 2020; Kiela et al., 2020). Given a text–image pair (T_i, I_i), the goal is to predict a label $\hat{y}_i \in \{0, 1\}$ by maximizing:

$$\hat{y}_i = \arg\max_{y \in \{0,1\}} p(y \mid T_i, I_i), \tag{1}$$

However, merely distinguishing whether content is hateful is insufficient. This binary formulation fails to identify specific target groups and lacks natural language explanations, making it difficult to articulate the underlying reasoning, especially for implicit cases.

Difficulty	Pattern $(y^{Text}, y^{Image} \rightarrow y^{Combined})$	Interaction Mechanism
Easy	000, 011, 101, 111	Semantic Consistency
Normal	100, 010	Contextual Neutralization
Hard	001, 110	Implicit Emergence / Inversion

Table 1: Difficulty Stratification of Multimodal Interaction Patterns (0: NotHate, 1: Hate).

To address this, we formulate MMHSD as a six-class classification task with explanatory supervision. Our goal is to learn a model $\mathcal{F}$ that predicts a hate speech category $\hat{y}_i \in \mathcal{Y}$ together with a natural language explanation e_i (the category set $\mathcal{Y}$ is defined in Appendix A.):

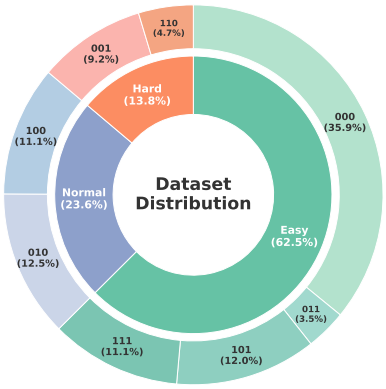
$$(\hat{y}_i, e_i) = \arg\max_{y \in \mathcal{Y},\, e} p(y, e \mid T_i, I_i). \tag{2}$$

This formulation requires the model to jointly analyze textual and visual information to determine the specific hate speech category.

3.2 Annotation and Difficulty Stratification

To capture the complexity of multimodal hate, particularly when modalities conflict, we introduce the Stratified Multimodal Interaction (SMI) paradigm. For each sample, we annotate a **five-tuple**, explicitly labeling unimodal sentiments alongside the final multimodal annotation:

$$\mathcal{A}_i = (y_i^{\text{text}}, e_i^{\text{text}}, y_i^{\text{image}}, e_i^{\text{image}}, y_i^{\text{combined}}) \tag{3}$$



Dataset	Source	# Samples	# Cls	SMI	Diff.	Agreem. (κ)
MMHS150K (Gomez et al., 2020)	Img+Txt	149,823	6	✗	✗	0.15†
FHM (Kiela et al., 2020)	Meme	10,000	2	✗	✗	0.68*
MultiOFF (Suryawanshi et al., 2020)	Meme	743	2	✗	✗	0.4 – 0.5†
Harm-C (Pramanick et al., 2021a)	Meme	3,544	3/4	✗	✗	0.68 / 0.78*
MAMI (Fersini et al., 2022b)	Meme	11,000	2/5	✗	✗	0.58 / 0.34†
H-VLI (Raw)	Img+Txt	8,438	6	✓	✓	0.59†
H-VLI (Final)		5,569				**0.94**†

Note: †Fleiss' κ; *Cohen's κ. SMI: Stratified Multimodal Interaction paradigm. Diff: Difficulty Grading.

Table 2: Comparison of our dataset (H-VLI) with existing multimodal hate speech datasets. **H-VLI (Raw)** represents data before consistency filtering.

where $y_i^{\text{text/image}}$,$e_i^{\text{text/image}}$ denote the unimodal labels and explanations respectively. y_i^{combined} represents the final multimodal ground-truth label.

Taxonomy of Multimodal Interaction: Under the SMI paradigm, the interplay between unimodal signals ($y^{\text{text}}, y^{\text{image}}$) and the combined outcome (y^{combined}) yields eight distinct interaction patterns (i.e., all 2^3 combinations of $(y^{\text{text}}, y^{\text{image}}, y^{\text{combined}}) \in \{0,1\}^3$). While these eight categories allow for fine-grained classification to ensure systematic and comprehensive sample collection during dataset construction, we group them into three difficulty levels for clearer and more practical evaluation, as summarized in Table 1:

- **Easy (Explicit Consistency):** The final verdict aligns with explicit unimodal polarity (e.g., $y^{combined} = y^{text} \vee y^{image}$). Hate is explicitly present in at least one modality or absent in both, requiring no complex cross-modal reasoning.
- **Normal (Contextual Correction):** A unimodal toxic signal is neutralized by the other modality (e.g., counter-speech), requiring the model to perform semantic correction.
- **Hard (Implicit Interactions):** The most challenging scenario where hatefulness emerges from the intersection of benign modalities (Emergence), or toxic elements are recontextualized into benign satire (Inversion).

3.3 Dataset Construction

To populate the SMI paradigm and address the challenge of obtaining diverse implicit multimodal samples in real-world distributions, we employ a hybrid pipeline (Figure 2) that combines rigorous filtering of existing benchmarks with targeted synthetic generation.

Sourcing Real-world Samples: We leverage MMHS150K (Gomez et al., 2020) as a foundation for Easy and Normal samples. To mitigate crowd-sourcing noise, we strictly filter for samples with unanimous inter-annotator agreement and yield 5,232 high-quality candidates, which effectively represent benign, explicit hate and contextual neutralization patterns.

Generative Data Injection: To populate the Normal and Hard subsets, we employ a generative injection strategy. Using Qwen3-VL-Plus (Bai et al., 2025) and Gemini-2.5-Pro (Anil et al., 2023), we craft synthetic captions that induce specific semantic interactions, leveraging strategies such as contextual inversion and victim-perspective narration to simulate real-world ambiguity (see Appendix B for detailed strategies). This yields 7,506 candidates covering complex reasoning scenarios often missing in organic data.

Human-in-the-loop Annotation: To ensure label reliability, we implement a model-assisted expert review process. To facilitate this, we developed a specialized annotation interface tailored to the SMI paradigm (displayed as "Annotation Platform" in Figure 2). To reduce cognitive load and enforce structured reasoning, the platform visualizes unimodal priors ($y^{text}, e^{text}, y^{image}, e^{image}$) pregenerated by Qwen-Plus (Bai et al., 2023a) and Qwen3-VL-Plus (Bai et al., 2025). This provides annotators with initial rationales alongside the multimodal input.

Step	Filter	Discarded	Remaining
–	Initial Candidates	–	8,802
1	Quality Control	364	8,438
2	Consensus Filtering	2,593	5,845
3	Intent Alignment	276	5,569

Table 3: Post-annotation filtering statistics.

Annotators then assign the final multimodal label (y) based on these cues (see Appendix F for interface details). To guarantee quality, the dataset undergoes strict post-annotation filtering, including an Intent Alignment check to ensure human

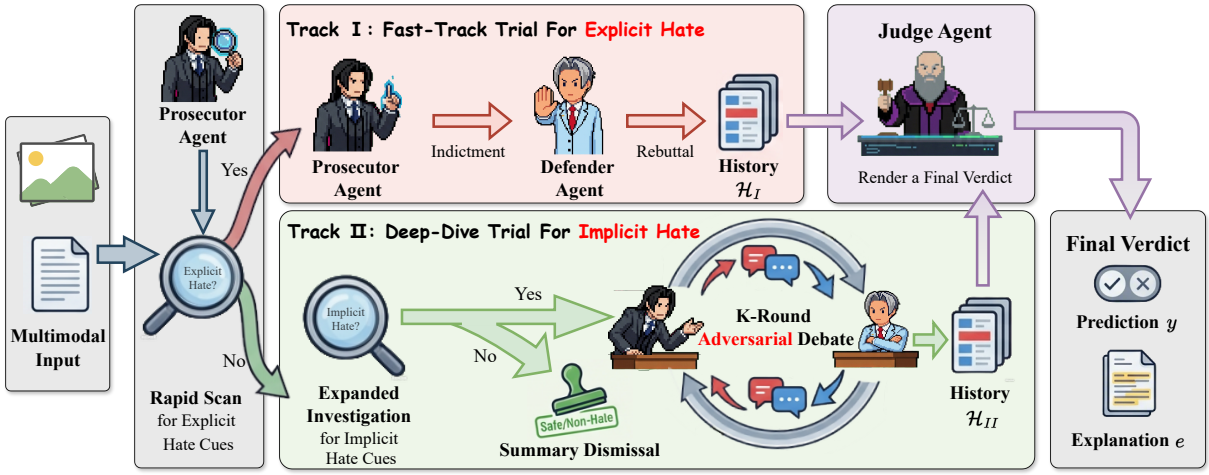


Figure 3: **The architecture of the ARCADE framework.** The model simulates an asymmetric courtroom debate involving Prosecutor and Defender agents. It features a Gated Dual-Track mechanism that routes samples to either a Fast-Track for explicit hate or a Deep-Dive trial for implicit reasoning, allowing the Judge agent to synthesize the debate history and render a final verdict.

consensus matches the generative objective. The three-step filtering process refines 8,802 candidates into 5,569 high-quality samples (Table 3). Consequently, H-VLI achieves superior inter-annotator agreement ($\kappa = 0.94$, Table 2) compared to existing benchmarks. Detailed dataset statistics are presented alongside.

4 Methodology

We propose the **A**symmetric **R**easoning via **C**ourtroom **A**gent **DE**bate (**ARCADE**) framework to enhance multimodal hate speech detection through adversarial dialectics. As illustrated in Figure 3, ARCADE simulates a judicial process with three specialized agents (Prosecutor, Defender, and Judge) to address the reasoning challenges in deciphering multimodal intent shifts.

4.1 Asymmetric Agent Design

Unlike symmetric frameworks where agents share generic personas, ARCADE establishes distinct cognitive priors to simulate an adversarial trial.

The Prosecutor (Risk Discovery): Adopting a "presumption of guilt", the Prosecutor agent (A_{pros}) maintains high sensitivity to potential risks. It actively hypothesizes malice and explicitly maps visual symbols to textual metaphors to uncover latent hate, rather than performing neutral classification.

The Defender (Contextual Safety): The Defender agent (A_{def}) operates under a "presumption of innocence". It serves as a dynamic safety alignment which scrutinizes evidence for benign motivations (e.g., satire, self-deprecation, or educational documentation) and aims to invalidate the Prosecutor's claims through logical contextualization.

The Judge (Final Arbiter): The Judge (A_{judge}) evaluates the validity of the debate history $\mathcal{H}$ against the raw input to render the final verdict $\hat{y}$ and explanation e, without participating in argument generation.

4.2 Gated Dual-Track Litigation Process

Given a multimodal sample $S_i = (T_i, I_i)$, the Prosecutor performs a rapid scan for explicit hate symbols. A gating function $\Phi(S_i)$ determines the procedural path:

$$\Phi(S_i) = \begin{cases} 1, & \text{if } A_{pros} \text{ detects explicit hate cues} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Based on $\Phi(S_i)$, the sample is routed to either a Fast-Track or a Deep-Dive trial as follows:

Track I: Fast-Track Trial (Explicit Hate): When $\Phi(S_i) = 1$, the primary challenge is context verification (e.g., distinguishing hate from quotations). The trial is streamlined into a single turn: The Prosecutor submits an indictment E_{exp}, and the Defender immediately provides a contextual rebuttal. The history is recorded as $\mathcal{H}_I = \{E_{exp}, \text{Rebuttal}(S_i, E_{exp})\}$.

Track II: Deep-Dive Trial (Implicit Reasoning): When $\Phi(S_i) = 0$, the sample lacks overt hate symbols but requires rigorous socio-cultural reasoning. The Prosecutor first conducts an expanded investigation. If no evidence is found, the process terminates via **Summary Dismissal**. Conversely,

if potential implicit cues are identified, ARCADE initiates an **Adversarial Debate**. Let u_k denote the utterance at turn k (with $u_0^{pros} = u_0^{def} = \varnothing$). The state transition proceeds as:

$$u_k^{pros} = A_{pros}(S_i, u_{k-1}^{pros}, u_{k-1}^{def}) \quad (5)$$

$$u_k^{def} = A_{def}(S_i, u_k^{pros}, u_{k-1}^{def}) \quad (6)$$

This iterative process forces agents to deepen their reasoning, exposing logical links between neutral images and biased texts. The resulting history is recorded as $\mathcal{H}_{II} = \{u_1^{pros}, u_1^{def}, \ldots, u_K^{def}\}$.

Algorithm 1: ARCADE Inference Process

Input: Multimodal Sample S, Max Turns K
Output: Prediction y, Explanation e

Initialize: Agents $A_{pros}, A_{def}, A_{judge}$
$g \leftarrow \Phi(S)$ via Eq. (4)
if $g = 1$ **then**

Track I: Fast-Track

$E_{exp} \leftarrow \text{INDICT}(A_{pros}, S)$
$u_{reb} \leftarrow \text{REBUT}(A_{def}, S, E_{exp})$
$\mathcal{H} \leftarrow \{E_{exp}, u_{reb}\}$

else

Track II: Deep-Dive

$u_1^{pros} \leftarrow \text{DETECTIMPLICIT}(A_{pros}, S)$
if $u_1^{pros} = \emptyset$ **then**
// Summary Dismissal
return (Non-Hate, *"No implicit risks"*)
end if
$\mathcal{H} \leftarrow \{u_1^{pros}\}$
$u_1^{def} \leftarrow A_{def}(S, u_1^{pros}, \emptyset)$ via Eq. (6)
$\mathcal{H} \leftarrow \mathcal{H} \cup \{u_1^{def}\}$
for $k = 2$ to K **do**
$u_k^{pros} \leftarrow A_{pros}(S, u_{k-1}^{pros}, u_{k-1}^{def})$ via Eq. (5)
$u_k^{def} \leftarrow A_{def}(S, u_k^{pros}, u_{k-1}^{def})$ via Eq. (6)
$\mathcal{H} \leftarrow \mathcal{H} \cup \{u_k^{pros}, u_k^{def}\}$
end for

end if

$(\hat{y}, e) \leftarrow A_{judge}(S, \mathcal{H})$ via Eq. (7)
return (y, e)

Verdict: Finally, the Judge aggregates the debate history $\mathcal{H}$ to get the prediction. By conditioning on the adversarial exchange, the Judge distinguishes between genuine hateful intent and benign usage:

$$(\hat{y}_i, e_i) = \arg\max_{y \in Y, e} p(y, e \mid S_i, \mathcal{H}) \quad (7)$$

The complete procedure including the summary dismissal logic is outlined in Algorithm 1.

5 Experiment

In this section, we evaluate existing methods and our proposed ARCADE on our constructed H-VLI dataset and FHM (Kiela et al., 2020) dataset.

5.1 Tasks and Evaluation Metrics

To comprehensively assess the compared models, we formulate the evaluation as two distinct tasks:
Task 1: Fine-grained Hate Categorization: Moving beyond simple binary detection, this task challenges models to discern specific hate types (e.g., racist, sexist) alongside non-hateful content. We report *Accuracy* for difficulty subsets (Easy, Normal, Hard) and *Accuracy*, *Macro-F1*, and *Weighted-F1* for the overall dataset.
Task 2: Binary Hate Detection: Following previous works (Gomez et al., 2020; Kapil and Ekbal, 2025), we aggregate the five subtypes into a single "hateful" class against "non-hateful" to evaluate fundamental detection capabilities. We report *Accuracy* across difficulty subsets, with *Accuracy*, *Recall*, and *F1-score* as overall metrics.

5.2 Datasets and Configurations

Split	Total	Hate	Non-Hate
Train	4,391	1,533(34.9%)	2,858(65.1%)
Test	1,178	462(39.2%)	716(60.8%)
Overall	5,569	1,995(35.8%)	3,574(64.2%)

Table 4: Distribution of hate vs. non-hate samples in the H-VLI benchmark.

Our Dataset (H-VLI): We partitioned the H-VLI dataset using a strict disjoint split based on image sources and semantic topics to prevent information leakage and ensure the test set contains truly unseen samples. This process yielded a training set of 4,391 samples and a test set of 1,178 samples (as detailed in Table 4). In our evaluation, training-based methods utilize this train-test split, whereas training-free methods are directly evaluated on the test set.
FHM Dataset: To assess the generalization capability, we utilized the *Facebook Hateful Memes* (FHM) dataset (Kiela et al., 2020) which contains 10,000 multimodal memes designed to facilitate the detection of hateful content. Following prior protocols (Rizwan et al., 2025; Wang et al., 2025), we evaluate on the 500-sample dev split (as the test set labels are withheld). Since FHM only provides binary labels, it is used exclusively for Task 2.

Model	Training	Task 1: Fine-grained Hate Categorization						Task 2: Binary Hate Detection					
		Easy	Normal	Hard	Overall			Easy	Normal	Hard	Overall		
		Acc	Acc	Acc	Acc	Mac-F1	W-F1	Acc	Acc	Acc	Acc	Recall	F1
BERT (Devlin et al., 2019)	✓	65.83	72.66	29.59	62.31	57.40	63.03	73.33	85.47	39.64	71.48	68.83	65.43
ViT-b-16 (Dosovitskiy, 2021)	✓	41.94	68.86	37.28	47.88	14.52	43.32	48.89	61.94	55.62	53.06	40.69	40.47
BERT+ViT	✓	68.33	77.51	33.73	65.62	61.25	66.19	75.83	84.78	38.46	72.67	71.00	67.08
MMBT (Kiela et al., 2019)	✓	68.61	80.28	33.14	66.38	62.17	66.99	74.44	88.58	36.69	72.50	68.18	66.04
Momenta (Pramanick et al., 2021b)	✓	60.69	76.47	30.18	60.19	51.86	60.35	70.42	76.47	33.14	66.55	65.37	60.52
PromptHate (Cao et al., 2022)	✓	68.33	73.36	36.69	65.03	60.55	65.47	76.67	69.20	37.87	69.27	71.86	64.72
Qwen-VL-Plus (Bai et al., 2023b)	✗	85.97	10.03	44.24	61.41	59.75	63.16	**96.94**	23.34	14.55	67.32	80.35	65.77
Qwen-VL-Plus w/ Ours	✗	83.50	61.46	28.48	70.29	63.89	70.23	83.66	73.17	29.88	73.52	61.01	64.19
Qwen-VL-Max (Bai et al., 2023b)	✗	89.44	9.34	55.15	64.91	66.60	65.31	95.00	17.77	45.45	69.11	90.83	69.68
Qwen-VL-Max w/ Ours	✗	86.85	65.97	39.39	75.00	70.27	75.42	89.69	68.64	38.79	77.35	77.46	72.76
GPT-4.1-mini (OpenAI et al., 2024)	✗	91.11	28.72	66.27	72.24	72.26	72.92	91.67	50.52	63.91	77.59	84.85	74.81
GPT-4.1-mini w/ Ours	✗	84.98	82.99	58.18	80.72	74.80	80.41	91.52	78.47	66.06	84.73	83.37	80.98
GPT-4o (OpenAI et al., 2024)	✗	91.94	36.33	64.50	74.36	73.19	74.94	94.72	39.45	61.54	76.40	94.37	75.83
GPT-4o w/ Ours	✗	87.76	73.26	75.76	82.51	79.53	82.93	89.57	77.08	78.79	84.98	86.43	81.78
Qwen3-VL-Plus (Bai et al., 2025)	✗	**92.22**	34.26	40.00	70.61	69.42	71.16	94.29	47.04	31.52	73.85	80.92	70.69
Qwen3-VL-Plus w/ Ours	✗	87.48	77.43	51.52	79.95	75.04	79.95	92.76	67.83	51.52	80.84	82.89	77.14
Gemini-2.5-Flash (Anil et al., 2023)	✗	90.28	55.02	69.23	78.61	75.33	79.26	93.19	53.29	69.23	79.97	92.42	78.35
Gemini-2.5-Flash w/ Ours	✗	83.17	74.31	74.55	79.78	75.19	80.24	85.95	77.78	75.15	82.42	87.31	79.48
GPT-5-mini (OpenAI, 2025b)	✗	91.25	29.76	49.11	70.12	68.62	70.89	94.86	31.83	12.43	67.57	77.92	65.34
GPT-5-mini w/ Ours	✗	91.66	70.83	63.64	82.59	80.22	83.08	94.44	72.57	61.82	84.47	**94.75**	82.63
Gemini-2.5-Pro (Anil et al., 2023)	✗	88.75	52.94	74.56	77.93	77.32	78.55	91.11	61.25	78.11	81.92	93.07	80.15
Gemini-2.5-Pro w/ Ours	✗	83.03	73.26	**78.79**	80.03	76.03	80.49	87.48	73.96	**79.39**	83.02	86.21	79.84
GPT-5 (OpenAI, 2025a)	✗	91.51	55.56	69.23	80.48	74.94	81.39	93.59	50.67	24.79	74.32	68.25	64.56
GPT-5 w/ Ours	✗	90.88	**88.64**	65.54	**86.92**	**81.73**	**86.99**	92.58	**87.23**	69.13	**88.06**	85.96	**84.20**

Table 5: Results on our H-VLI dataset. **Bold** indicates the best, underline indicates the second best. (Gray rows indicate trained models; bold/underline highlighting only applies to training-free models. Rows highlighted in color indicate models enhanced with our framework.)

Implementation Details: We fine-tuned all supervised baselines on our dataset splits on a single NVIDIA RTX 4090 GPU. MLLMs were evaluated via official APIs, where "Baseline" refers to zero-shot single-turn prompting. For ARCADE, we employ **Qwen3-VL-Plus** as the fixed Auxiliary Model to evaluate different Judge backbones, setting reasoning rounds to $K = 3$ for H-VLI and $K = 2$ for FHM. Unless otherwise specified, all reported results are obtained from a single experimental run. See Appendix D for further details.

5.3 Main Results

Table 5 comprehensively compares the traditional supervised models and various Multimodal Large Language Models (MLLMs) equipped with our proposed framework. The results demonstrate the effectiveness of our method across both fine-grained categorization and binary detection tasks. Notably, ARCADE helps mitigate the over-sensitivity issue common in standard MLLMs, where models refuse to analyze sensitive content due to safety triggers (see Appendix D.4).

Supervised vs. Baseline MLLMs: Traditional supervised models struggle with complex reasoning. While MMBT performs decently on Normal samples (80.28%) in Task 1, its performance collapses on the Hard subset (33.14%). This highlights a lack of sociocultural knowledge and deep multimodal understanding. In contrast, benefiting from large-scale pre-training, modern MLLMs like GPT-5 and Gemini-2.5-Pro show superior zero-shot single-turn prompting baselines.

Effectiveness of Our Framework: Integrating our ARCADE yields consistent improvements across almost all backbones by bridging the reasoning gap. For GPT-5-mini, our method boosts Macro-F1 by over 11% compared to the base model. Similarly, on the binary Task 2, ARCADE pushes GPT-5-mini to 84.47% Accuracy, demonstrating that our dual-track mechanism effectively unlocks the latent potential of smaller models.

Robustness on Normal and Hard Samples: The key advantage of our framework is most pronounced in complex scenarios. As shown in Table 5, standard MLLMs often fail on complex cases involving implicit hate and semantic inversion. For example, Qwen-VL-Plus scores only 14.55% on the Hard subset (Task 2). Our method doubles this performance to 29.88%. Similarly, we boost GPT-4o's accuracy on Hard samples from 61.54% to 78.79%. This indicates that the "Prosecutor-Defender" debate successfully elicits the deep reasoning chains necessary to disentangle nuanced hateful content. For a concrete demonstration of these reasoning dynamics, we provide a detailed qualitative analysis in Appendix E.

Model	Training	Acc	Recall	F1
Text BERT (Devlin et al., 2019)	✓	57.12	-	-
Image Region (Kiela et al., 2020)	✓	52.34	-	-
CLIP-BERT (Pramanick et al., 2021b)	✓	58.28	-	-
DisMultiHate (Pramanick et al., 2021a)	✓	62.42	-	-
ViLBERT CC (Sharma et al., 2018)	✓	64.70	-	-
MMBT-Region (Kiela et al., 2019)	✓	65.06	-	-
PromptHate (Cao et al., 2022)	✓	67.82	-	-
BLIP (Li et al., 2022)	✓	69.20	-	-
ALBEF (Li et al., 2021)	✓	70.58	-	-
Pro-CapPromptHate (Cao et al., 2023)	✓	72.28	-	-
LLaVA (Vicuna 13B) (Mei et al., 2024)	✓	77.30	-	-
IDEFICS 9B (Laurençon et al., 2023)	✗	49.80	-	-
LLAVA-1.5 7B (Liu et al., 2024)	✗	60.00	-	-
INSTRUCTBLIP VICUNA 7B (Dai et al., 2023)	✗	53.06	-	-
Spark-VL (Wang et al., 2025)	✗	73.20†	-	-
Qwen-VL-Max (Wang et al., 2025)	✗	72.80†	-	-
GPT-4 (Wang et al., 2025)	✗	**78.60**†	-	-
Qwen-VL-Max (Bai et al., 2023b)	✗	70.39	64.88	68.71
Qwen-VL-Max w/ Ours	✗	73.70	77.50	74.70
Qwen3-VL-Plus (Bai et al., 2025)	✗	69.77	54.96	64.56
Qwen3-VL-Plus w/ Ours	✗	73.49	83.40	75.99
GPT-4o (OpenAI, 2024)	✗	73.80	77.60	74.76
GPT-4o w/ Ours	✗	75.31	81.40	76.80
GPT-5-mini (OpenAI, 2025b)	✗	73.00	58.00	68.53
GPT-5-mini w/ Ours	✗	74.69	**83.47**	**76.81**

Table 6: Results on the FHM Dataset for Binary Hate Detection. **Bold** and underline indicate the best and the second best results. Gray rows indicate training-based models. Rows highlighted in blue indicate models enhanced with our framework. Models marked with † utilize additional knowledge via Retrieval-Augmented Generation (RAG) in a few-shot setting.

5.4 Generalization on Public Benchmarks

To evaluate the generalization capability of ARCADE, we compare its performance on the FHM dataset against state-of-the-art methods (Table 6). Notably, recent RAG-based approaches (Wang et al., 2025) achieve the upper bound (78.60% Accuracy) by retrieving external socio-cultural knowledge and few-shot exemplars. In contrast, ARCADE operates without knowledge base or external retrieval yet delivers competitive performance solely through internal reasoning, enabling GPT-4o to reach 75.31% Accuracy and 76.80% F1.

Furthermore, ARCADE effectively bridges the capability gap for smaller models. For example, Qwen-VL-Max enhanced by ARCADE (73.70%) rivals the performance of the much stronger GPT-4o baseline (73.80%). Note that ARCADE maintains a balanced trade-off between precision and recall across all backbones, exemplified by GPT-5-mini, which achieves the highest Recall (83.47%) and F1-score (76.81%), confirming robust generalization beyond dataset artifacts.

To further validate this generalization capability on unseen data, we conducted an additional evaluation on the FHM test set (comprising 1,000 samples: 490 hate, 510 non-hate). As presented in Table 7, the integration of ARCADE consistently improves accuracy, recall, and F1 scores on this independent test set. Specifically, the framework boosts the F1 scores of Qwen3-VL-Plus and GPT-5-mini to 74.45% and 77.91%, respectively, thereby firmly substantiating our claims regarding robust generalization.

Model	Training	Acc	Recall	F1
Qwen3-VL-Plus	✗	69.46	67.92	68.56
Qwen3-VL-Plus w/ Ours	✗	73.82	77.87	74.45
GPT-5-mini	✗	75.80	66.33	72.87
GPT-5-mini w/ Ours	✗	**77.12**	**82.29**	**77.91**

Table 7: Results on the FHM test set for Binary Hate Detection.

Method	On H-VLI		On FHM	
	Acc	Mac-F1	Acc	F1
Baseline	70.61	69.42	69.77	64.56
Multiround†	74.47	71.43	72.23	74.47
Ours ($K = 1$)	78.03	73.80	73.28	73.98
Ours ($K = 2$)	77.71	73.07	-	-
Ours ($K = 3$)	-	-	71.93	74.67
Ours ($K = 4$)	77.80	73.93	-	-
Qwen → Qwen	62.00	54.15	-	-
Qwen → Qwen3	73.81	67.40	-	-
Qwen3 → Qwen	70.29	63.89	-	-
Ours (Proposed)‡	**79.95**	**75.04**	**73.49**	**75.99**

Table 8: Results of ablation studies. **Bold** indicates the best, underlined indicates the second best. Note that for our default settings: †Multiround uses $K = 3$ for Our Dataset and $K = 2$ for FHM. ‡Ours uses $K = 3$ for Our Dataset and $K = 2$ for FHM. The notation 'Model A → Model B' signifies using Model A as the Auxiliary (Prosecutor/Defender) and Model B as the Judge.

5.5 Ablation Study

Table 8 summarizes the ablation study. Unless otherwise specified, we standardize **Qwen3-VL-Plus** for both Auxiliary (Prosecutor/Defender) and Judge roles to isolate component effectiveness.

Impact of Rounds in Track II (K): We first investigated the impact of the number of reasoning rounds (K) in Track II (Deep-Dive Trial). Results indicate that performance improves with K up to a saturation point ($K = 3$ for H-VLI, $K = 2$ for FHM) and deviation from this optimum degrades results. This indicates that a balanced number of rounds is crucial: insufficient rounds lead to inadequate discussion, while excessive iterations risk introducing noise or over-interpretation.

Effectiveness of Dual-Track Mechanism: Removing the initial gate mechanism and the Fast-Track (forcing deep debates on all samples, denoted as "Multiround") drops accuracy from 79.95% to 74.47% on H-VLI and from 73.49% to 72.23% on FHM. This confirms that the Gatekeeper effectively prevents over-interpretation of explicit sam-

ples. Notably, even "Multiround" still outperforms the single-turn prompting Baseline, validating the debate format's intrinsic utility.

Auxiliary vs. Judge Roles: We examined the impact of model capabilities on the Auxiliary and Judge roles using Qwen-VL-Plus (Qwen, Weak) and Qwen3-VL-Plus (Qwen3, Strong) on H-VLI. Apart from the best performance achieved by the all strong setup (79.95% Accuracy), **Weak Aux** $\rightarrow$ **Strong Judge** significantly outperforms the reverse configuration (73.81% vs. 70.29%). This indicates the **Judge is the bottleneck**: a capable arbiter is critical to synthesize evidence, whereas a weak Judge fails to render correct verdicts even when presented with high-quality arguments.

6 Further Discussion

6.1 Impact of Model-Generated Hints on Annotator Bias

During the human-in-the-loop annotation phase, there is a potential risk that providing model-generated hints could introduce anchoring bias, inadvertently influencing human judgment. To empirically investigate this impact, we conducted an ablation study. Specifically, we tasked our original annotators with labeling a fresh, unseen subset of 500 samples strictly without any model-generated unimodal labels or explanations, while keeping all other procedures identical.

Setting	Pre-Filter	Post-Filter
With hints	$\kappa = 0.59$	$\kappa = 0.94$
Without hints	$\kappa = 0.57$	$\kappa = 0.94$

Table 9: Inter-annotator agreement comparison.

As shown in Table 9, the initial agreement without hints remained comparable to the setting with hints ($\kappa = 0.57$ vs. 0.59). More importantly, after applying our post-annotation filtering, both settings converged to an identical high consistency ($\kappa = 0.94$). This demonstrates that the hints had a negligible impact on reinforcing model priors. They effectively reduced cognitive load, while our rigorous post-filtering stage served as the primary driver of dataset reliability.

6.2 Detailed Routing Statistics of the Dual-Track Mechanism

Our gating mechanism is primarily designed to reliably distinguish whether explicit hateful cues are present, thereby ensuring that samples are routed to the most appropriate reasoning track. To demonstrate its reliability, Table 10 provides a detailed sample routing breakdown by the gating function $\Phi(S)$ on the H-VLI (Task I) using Qwen3-VL-Plus. Note that out of the 1,178 test samples, 6 were rejected due to API constraints, leaving 1,172 samples successfully processed by the gating function.

Type	Total	Track I	Track II	Dismiss
000	361	31(8.6%)	156(43.2%)	174(48.2%)
001	99	60(60.6%)	39(39.4%)	0(0.0%)
010	169	149(88.2%)	20(11.8%)	0(0.0%)
011	40	39(97.5%)	1(2.5%)	0(0.0%)
100	119	98(82.4%)	21(17.6%)	0(0.0%)
101	126	117(92.9%)	9(7.1%)	0(0.0%)
110	66	66(100.0%)	0(0.0%)	0(0.0%)
111	192	188(97.9%)	4(2.1%)	0(0.0%)

Table 10: Sample routing statistics of $\Phi(S)$ on H-VLI using Qwen3-VL-Plus.

The results indicate that $\Phi(S)$ reliably routes over 90% of samples with explicit hate labels (i.e., those containing a "1" in either of the first two modalities) into Track I. Conversely, the vast majority of "000" samples (where both modalities are individually benign) are appropriately directed to Track II (Full Debate) or Summary Dismiss. The minor fraction of "000" samples entering Track I primarily consists of text containing seemingly hateful keywords utilized in harmless contexts (e.g., self-deprecation). This confirms the high reliability and dynamic adaptability of $\Phi(S)$ within our framework.

7 Conclusion

This paper tackles the detection of multimodal hate speech, positing that the hateful or non-hateful intent emerges from cross-modal interplay rather than simple summation. Moving beyond binary classification, we established the Stratified Multimodal Interaction (SMI) paradigm and the H-VLI benchmark to systematically decipher the intent shifts between modalities, ranging from contextual neutralization to implicit emergence. To reason through such ambiguity, we proposed ARCADE, which leverages a dual-track asymmetric courtroom debate to scrutinize deep semantic cues. Our approach demonstrates superior performances and interpretability compared to existing methods. We hope this work paves the way for more interpretable and reasoning-driven content moderation systems.

8 Limitations

Despite the advancements presented, we acknowledge several limitations in our work.

First, despite enriching data density, synthetic samples in our benchmark may not fully capture the chaotic linguistic noise of organic social media. Second, the multi-agent debate prioritizes reasoning depth over real-time efficiency, involving higher computational costs that future model distillation could mitigate. Third, a performance trade-off exists: while our framework significantly boosts detection on normal and hard samples, it may yield a marginal regression on easy ones, though this effect diminishes with stronger backbones. Finally, our current framework is English-centric; extending it to multilingual and diverse cultural contexts remains a vital direction for future research.

9 Ethical Considerations

Note: Any hateful content examples cited or generated within this paper are strictly for analytical purposes and do not represent the authors' personal views.

Motivation and Risk Balance. We aim to comprehensively advance the field of multimodal hate speech detection by establishing a robust framework (ARCADE) and a challenging benchmark (H-VLI). Existing datasets are increasingly saturated, often lacking the high-difficulty samples necessary to pressure-test modern MLLMs. Consequently, we employed "Generative Injection" to construct complex cases that effectively simulate the nuanced reasoning required in real-world moderation. We acknowledge the dual-use risks inherent in synthesizing hateful content. However, the vulnerability of current safety systems to these sophisticated attacks is an objective reality. Our work exposes these blind spots to facilitate the development of more resilient defenses, rather than to exacerbate the threat.

Safety Protocols. To mitigate potential harms, we strictly enforce the following protocols: (1) The dataset and generation prompts will be accessible solely to credentialed academic researchers under a restrictive license, strictly prohibiting redistribution. (2) A rigorous human-in-the-loop review process was implemented to filter out content that is illegal or violates fundamental safety guidelines beyond the scope of research purposes. (3) All data were collected from publicly accessible sources, and no additional personal information beyond what was originally visible was gathered or inferred.

Licensing and Terms of Use. All source data used in this work are obtained from publicly accessible resources and are processed strictly in accordance with their original licenses, platform terms of service, and intended research purposes. The constructed dataset (H-VLI) is explicitly intended for non-deployment, research-only use and will be released under a restrictive license that prohibits redistribution and commercial use, consistent with the original access conditions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62441616, Grant 62336004, Grant 62125603, Grant 62306031, Grant 62506198, in part by the China Postdoctoral Science Foundation under Grant 2024M761674.

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10444–10455, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and 1331 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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 *ACM MM*, pages 5244–5252.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Anthony Cortese. 2005. *Opposing hate speech*. Bloomsbury Publishing.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36:49250–49267.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *AAAI*, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Alexey Dosovitskiy. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Vishwajeet Dwivedy and Pradeep Kumar Roy. 2023. Deep feature fusion for hate speech detection: a transfer learning approach. *Multimedia Tools Appl.*, 82(23).

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

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

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *EMNLP*, pages 6159–6173.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, pages 1470–1478.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *EMNLP*, pages 15125–15140.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. In *SIGIR*, pages 117–122.

Weiqiang Jin, Dafu Su, Tao Tao, Xiujun Wang, Ningwei Wang, and Biao Zhao. 2025. Courtroom-fnd: a multi-role fake news detection method based on argument switching-based courtroom debate. *Journal of King Saud University Computer and Information Sciences*, 37(3):33.

Zhenchao Jin. 2022. Imagedl: Search and download images from specific websites. `https://github.com/CharlesPikachu/imagedl/`.

Prashant Kapil and Asif Ekbal. 2025. A transformer based multi task learning approach to multimodal hate speech detection. *Nat. Lang. Process. J.*, 11.

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

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33.

Sandeep Kumar, Abhijit A Nargund, and Vivek Sridhar. 2025. CourtEval: A courtroom-based multi-agent evaluation framework. In *Findings of ACL*, pages 25875–25887.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset

of interleaved image-text documents. *NeurIPS*, 36:71683–71702.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. 2025. Skyra: Ai-generated video detection via grounded artifact reasoning. *arXiv preprint arXiv:2512.15693*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, pages 17889–17904.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Rui Lu, Jinhe Bi, Yunpu Ma, Feng Xiao, Yuntao Du, and Yijun Tian. 2025. Mv-debate: Multi-view agent debate with dynamic reflection gating for multimodal harmful content detection in social media. *arXiv preprint arXiv:2508.05557*.

Xiaochen Ma, Guozheng Rao, Lina Xu, Xin Wang, Zaiming Fan, and Zhe Zhang. 2025. Guided and knowledgeable multi-agent debate for fact verification. *Expert Systems with Applications*, page 130103.

Mari J Matsuda. 2018. *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Routledge.

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

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*, pages 145–153.

OpenAI. 2024. Gpt-4o system card. https://openai.com/research/gpt-4o-system-card.

OpenAI. 2025a. Gpt-5. https://platform.openai.com/docs/models/gpt-5. Accessed: 2026-01-04.

OpenAI. 2025b. Gpt-5-mini. https://platform.openai.com/docs/models/gpt-5-mini. Accessed: 2026-01-04.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-agent-based debate simulation for generalized hate speech detection. In *EMNLP*, pages 20963–20987.

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

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

Naquee Rizwan and 1 others. 2025. Exploring the limits of zero shot vision language models for hate meme detection. In *ICWSM*, volume 19.

Furqan Khan Saddozai, Sahar K Badri, Daniyal Alghazzawi, Asad Khattak, and Muhammad Zubair Asghar. 2025. Multimodal hate speech detection: a novel deep learning framework for multilingual text and images. *PeerJ Computer Science*, 11:e2801.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

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 (TRAC-2020)*. Association for Computational Linguistics.

Alexander Tsesis. 2002. *Destructive messages: How hate speech paves the way for harmful social movements*, volume 27. NYU Press.

Xiuxian Wang, Lanjun Wang, Yuting Su, Hongshuo Tian, Guoqing Jin, and An-An Liu. 2025. Few-shot in-context learning for implicit semantic multimodal content detection and interpretation. *IEEE TCSVT*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL SRW*.

Lu Wei, Liangzhi Li, Tong Xiang, Liu Xiao, and Noa Garcia. 2025. Cracking the code: Enhancing implicit hate speech detection through coding classification. In *TrustNLP*.

Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. 2025. Sheep's skin, wolf's deeds: Are LLMs ready for metaphorical implicit hate speech? In *ACL*, pages 16657–16677.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. 2025. D3qe: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16292–16301.

Yu Zheng, Boyang Gong, Fanye Kong, Yueqi Duan, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. 2025. Learning counterfactually decoupled attention for open-world model attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 122–132.

Appendix Overview

This supplementary material provides extended technical details, implementation specifics, and qualitative analyses to support the main manuscript. The contents are organized as follows:

- **Data Construction (§A – §C):** Detailed category definitions, generative injection strategies for different difficulty levels, prompt templates for data synthesis, and the multi-stage human-expert filtering pipeline.
- **Experimental Setup (§D.1 – §D.3):** Comprehensive list of evaluated MLLM backbones, hyperparameter configurations, and the full prompt templates for our multi-agent framework (ARCADE).
- **Extended Analysis (§D.4 – §E):** Quantitative analysis of MLLM over-refusal behavior and qualitative case studies (including failure modes) to provide deeper insights into model reasoning.
- **Human Annotation (§F):** Details regarding the custom annotation platform, annotator recruitment, and ethical considerations.

A Hate Categorization

As shown in Table 11, we adopt the hate category definitions proposed by Gomez et al. (2020).

Label	Hate Category
0	NotHate
1	Racist
2	Sexist
3	Homophobic
4	Religious Hate
5	OtherHate (e.g., disability, age)

Table 11: Hate category definitions

B Generative Injection Details

B.1 Strategies

The purpose of generating synthetic samples is to compensate for the intrinsic limitations of the samples drawn from the MMHS150K dataset. MMHS150K is collected from Twitter using a *textual hate keyword–based retrieval* strategy. While effective for harvesting explicit hate content, this collection paradigm introduces several critical biases.

First, almost all samples in the dataset contain surface-level hate keywords in the text modality. As a result, the dataset severely under-represents samples whose textual content appears non-hateful on the surface (i.e., $y_i^{\text{text}} = 0$). Second, samples of *Normal* and *Hard* difficulty levels are extremely scarce, leading to insufficient coverage of medium- and high-difficulty cases, especially implicit hate samples. Third, the annotation quality of MMHS150K is moderate, with relatively low inter-annotator agreement, as shown in Table 2.

To address these issues, we collect real-world images and adopt targeted generative strategies to construct diverse multimodal samples. Specifically, we leverage Multimodal Large Language Models (MLLMs) to generate image-aligned augmented texts, thereby forming samples of different hate types and difficulty levels.

Image Sources. Our image pool is constructed from three sources. (1) From MMHS150K, we apply Qwen3-VL-plus to perform image-only annotations and select 1,500 benign images and 1,232 harmful images. (2) From the FHM dataset, we sample 500 benign and 500 harmful images. (3) Using keywords related to various protected groups, we collect 6,875 *culturally charged but visually neutral* images by querying Google Images through a image downloading tool (Jin, 2022).

Construction Strategies.

Easy Samples. For *Easy* samples, which do not involve semantic inversion across modalities, we design simple prompts that instruct the MLLM to generate either benign descriptions, neutral opinions, or explicitly hateful statements conditioned on the image, such that the combined multimodal semantics meet the target label. To avoid triggering the built-in safety mechanisms of MLLMs, we require the model to replace the attacked entity with a placeholder token `<insult>` when generating hateful text, and to additionally output the corresponding target group (`target_group`). We then use a manually curated hate lexicon and randomly sample a hate expression associated with `target_group` to replace the placeholder, producing the final hateful text.

Semantic Reversal Samples. For samples involving semantic reversal across modalities (i.e., types 010, 100, and 110), we analyze common linguistic and pragmatic patterns and summarize four representative construction methods: *(i) opposition-based*, *(ii) contextual inversion*, *(iii) meta-commentary* (e.g., educational or critical quotation), and *(iv) victim-perspective narration*. For each category, we design dedicated prompts to guide the MLLM to generate samples where one modality contains hateful content in isolation, but the combined multimodal semantics are non-hateful. For texts that include hateful expressions, we employ the same placeholder-based strategy described above to ensure controllability and safety during generation.

Implicit Multimodal Hate Samples (001). For implicit multimodal hate samples of type 001, we collect stereotype-based and metaphor-based texts targeting different protected groups. These texts are then paraphrased by Gemini-2.5-Pro (Anil et al., 2023) under a meaning-preserving constraint, producing an augmented set of *surface-benign trigger texts*. Finally, these texts are randomly paired with images associated with the corresponding protected groups, yielding the desired 001-type samples.

Overall, the proposed generative injection strategies substantially enrich the coverage of benign-text samples, medium- and high-difficulty cases, and implicit multimodal hate samples, thereby mitigating the biases inherent in keyword-based dataset construction.

B.2 Prompts

Here is an example prompt used to generate texts for 010-type samples in a *contextual inversion* method (Figure 4).

B.3 Representative Samples

To demonstrate the diversity and complexity of the H-VLI benchmark, we present representative samples across different SMI types in Figure 5. These examples highlight the necessity of fine-grained multimodal reasoning.

C Filtering Strategies

To ensure the reliability of our benchmark and mitigate the noise inherent in both crowd-sourced annotations and model-generated content, we implement a rigorous multi-stage filtering pipeline.

Filtering on Real-world Samples The original MMHS150K (Gomez et al., 2020) dataset contains a significant amount of noise due to subjective crowd-sourcing. To extract a high-quality subset, we apply a strict **Consensus Filtering** strategy. We

Prompt Template for 010-type samples

```
You are generating synthetic multimodal
samples where the image is HATEFUL (labels
1-5) but the text is NOT hateful. First
observe the image carefully and determine
the hate type.
[Hate category definitions]
[TARGET_OPTIONS]
Requirements:
1. The tweet text must NOT contain any
   hate speech. It should give a
   harmless or alternative explanation
   for the hateful image content, showing
   a context where the image is not
   actually promoting hate.
2. The text should make sense relative to
   the hateful elements implied by the
   image.
3. Choose text_label = 0.
4. target_group is the target of hate in
   the image chosen from TARGET_OPTIONS.
5. final_label = 0, since the combination
   is NOT hateful.
6. The generated tweet_text must resemble
   a natural social-media post...
[Output JSON instructions]
```

Figure 4: The prompt template used for generating contextual inversion samples.

Text: All it takes is a slight chromosomal shift to make us human instead of just **another animal**.

Type 001 (Implicit Hate)
Label: **Sexist**

Text: Me at black people who let their white and non black friends say **nigga** <url>

Type 110 (Implicit Inversion)
Label: **NotHate**

Text: Oh, look, another piece of media relying on tired, classist stereotypes. It's profoundly unoriginal. Can we please move past this lazy attempt at social commentary? #BoringTropes

Type 010 (Contextual Neut.)
Label: **NotHate**

Text: It's just one big, ugly **dyke** holding everything back.

Type 100 (Contextual Neut.)
Label: **NotHate**

Figure 5: Showcase of representative samples from the H-VLI dataset. **Type 001** represents implicit hate where benign components construct malice. **Type 110** shows benign intent despite toxic modalities. **Type 010/100** demonstrate contextual neutralization where text or image corrects the other's toxicity.

retain only those samples where all three original annotators reached a unanimous agreement (i.e., perfect consensus) on the hate speech label. This process effectively removes ambiguous or controversial samples, providing a solid foundation of clear-cut real-world cases.

Filtering on Annotated Samples To ensure the reliability of our benchmark, we employ a comprehensive three-step filtering mechanism applied to the **entire pool of annotated candidates**:

- **Step 1: Quality Control and Expert Adjudication.** We first exclude samples flagged by annotators as "Low Quality" (e.g., blurred images, nonsensical text). For samples marked as "Not Sure" (indicating ambiguity), rather than immediate exclusion, they undergo a secondary review by another expert. The expert's adjudication replaces the uncertain label, allowing the sample to re-enter the filtering pipeline for valid assessment.

- **Step 2: Consensus Filtering.** We evaluate the consistency among the three annotators and categorize samples into three levels: (1) *Perfect*: All three annotators agree on the exact fine-grained label (0–5); (2) *Strong*: All annotators agree that the content is hateful (labels 1–5), achieving unanimous binary consensus despite disagreement on the specific hate category; (3) *Weak*: Cases where annotators disagree, but a clear majority vote (e.g., 2 vs. 1) can still determine the final label. We retain samples from these three categories to ensure label validity while maximizing data diversity, discarding only those where no majority consensus can be reached.

- **Step 3: Intent Alignment.** To ensure that the content faithfully reflects the intended semantic construction, we perform an *Intent Alignment* check. We compare the majority vote of human annotators with the original generative target (Machine Label) or source metadata. Only samples where the human consensus

strictly matches the intended semantic target are retained. This step effectively eliminates cases involving generation failure, hallucination, or label misalignment.

Through this rigorous pipeline, we ensure that the final dataset consists of high-quality, human-verified samples with clear semantic intent. Consequently, the Inter-Annotator Agreement (Fleiss' κ) of our dataset significantly improved from 0.59 (Raw) to 0.94 (Final), as detailed in Table 2.

D Experiment Details

D.1 Evaluated Multimodal Large Language Models

We evaluate our framework using a set of representative state-of-the-art multimodal large language models (MLLMs) drawn from three widely used model families: Qwen, GPT, and Gemini. The selected models span different capability tiers and design choices, enabling a comprehensive assessment of our method across heterogeneous MLLM backbones.

Qwen Models (Bai et al., 2023b, 2025). From the Qwen family, we use *Qwen-VL-Plus*, *Qwen-VL-Max*, and *Qwen3-VL-Plus*. These models combine vision encoders with large language models and are designed for general-purpose vision–language understanding and reasoning. They serve as strong open or semi-open MLLM baselines and allow us to evaluate the effectiveness of our framework on advanced non-proprietary backbones.

GPT Models (OpenAI, 2024) From the GPT family, we evaluate *GPT-4.1-mini*, *GPT-4o*, *GPT-5-mini*, and *GPT-5*. These models represent different scales and capability levels within OpenAI's multimodal model lineup and exhibit strong generalization and reasoning abilities across vision–language tasks.

Gemini Models (Anil et al., 2023). From the Gemini family, we include *Gemini-2.5-Flash* and *Gemini-2.5-Pro*, which differ in model capacity and inference efficiency. Their inclusion enables evaluation of our framework across both lightweight and high-capacity proprietary MLLMs.

Evaluation Setting. All MLLMs are accessed via their official APIs and evaluated in a training-free, zero-shot setting. For each backbone, we report results for both the original model behavior (*Baseline*) and the enhanced reasoning enabled by our framework (ARCADE), without any additional fine-tuning. This ensures a fair comparison and isolates the performance gains introduced by structured multi-agent reasoning.

D.2 Implementation Details

Hyperparameter Configuration. We carefully calibrated the sampling temperature for each agent to align with their functional roles. The **Gatekeeper** is set to a temperature of 0 to ensure deterministic and consistent detection of explicit surface-level cues. For the auxiliary agents (**Prosecutor** and **Defender**), we utilize a higher temperature of 0.8. This promotes generation diversity, encouraging the agents to explore a broader range of implicit interpretations and creative rebuttals rather than converging on generic arguments. Conversely, the **Judge** agent operates at a low temperature of 0.1 to prioritize stability and precision, ensuring the final verdict is strictly grounded in the debate history with minimal randomness.

Error Recovery. To ensure evaluation accuracy, we re-evaluated samples that triggered non-safety-related failures, such as JSON formatting errors or incomplete outputs. These cases were re-processed to distinguish technical execution errors from genuine model refusals, ensuring that the reported metrics reflect the model's actual reasoning capabilities.

D.3 Agent Prompt Templates

We provide the detailed prompt templates used for each agent. For brevity, we first define the core task specifications shared across all prompts, followed by the general templates for the Prosecutor, Defender, and Judge. Note that templates for different stages (e.g., Track I and Track II) exhibit only minor instructional variations tailored to guide the model's specific focus.

Global Task Definitions

[Core Definition]
Hate speech includes abusive, insulting, demeaning, or threatening expressions targeting protected groups... Targets based on political affiliation, ideology, or policy positions do not constitute hate speech.

[Hate Category Definitions]
(0-NotHate, 1-Racist, 2-Sexist, 3-Homophobe, 4-Religion and 5-OtherHate.)

Safety Classifier Prompt (Baseline)

```
You are an expert AI safety classifier...
Your task is to analyze the provided tweet
(text and image) and determine if it
expresses hate speech...
[Core Definition]
[Hate Category Definitions]
[Output JSON instructions]
```

Prosecutor Agent Prompt

```
You are the Prosecutor in a simulated
courtroom... identify at most 3 potential
hate speech cues. Be maximalist,
hyper-vigilant, and strict.
[Core Definition]
Task Requirements:
• Direct: Explicit slurs, symbols, or
  clear threats.
• Socio-cultural: Implicit hate relying on
  stereotypes, history, or dog whistles.
• Metaphor: Dehumanization equating
  protected groups to animals/objects
  (Tenor vs. Vehicle).
[Hate Category Definitions]
[Output JSON instructions]
```

Defender Agent Prompt

```
You are the Defense Attorney... critically
examine the Prosecutor's cues and
determine whether they can be reasonably
refuted based on concrete evidence.
[Core Definition]
Defense Principles:
• Rebuttals must be grounded strictly in
  explicit evidence.
• Acknowledge plausible cues if concrete
  counter-evidence is unavailable.
• Focus on whether hatred is directed at
  oneself, individuals, or non-human
  entities.
[Output JSON instructions]
```

Judge Agent Prompt

```
You are the Judge... deliver a final
verdict based on The Post, Prosecutor's
Arguments, and Defense's Rebuttal.
[Core Definition]
Final Decision Rules:
• Reject hate if context is
  self-referential or non-protected.
• If one or more cues remain credible
  after Defense, assign hate label (1-5).
• Evaluate for Direct, Socio-cultural, and
  Metaphorical cues.
[Hate Category Definitions]
[Output JSON instructions]
```

Model	Fine-grained Task		Binary Task	
	Ref.#	Ref.%	Ref.#	Ref.%
Qwen-VL-Plus (Bai et al., 2023b)	4	0.34	6	0.51
Qwen-VL-Plus w/ Ours	10	0.85	11	0.93
Qwen-VL-Max (Bai et al., 2023b)	4	0.34	6	0.51
Qwen-VL-Max w/ Ours	10	0.85	8	0.68
GPT-4.1-mini (OpenAI et al., 2024)	0	0.00	0	0.00
GPT-4.1-mini w/ Ours	6	0.51	6	0.51
GPT-4o (OpenAI et al., 2024)	0	0.00	0	0.00
GPT-4o w/ Ours	6	0.51	6	0.51
Qwen3-VL-Plus (Bai et al., 2025)	4	0.34	8	0.68
Qwen3-VL-Plus w/ Ours	6	0.51	9	0.76
Gemini-2.5-Flash (Anil et al., 2023)	0	0	0	0.00
Gemini-2.5-Flash w/ Ours	6	0.51	6	0.51
GPT-5-mini (OpenAI, 2025b)	0	0.00	0	0.00
GPT-5-mini w/ Ours	6	0.51	6	0.51
Gemini-2.5-Pro (Anil et al., 2023)	0	0.00	0	0.00
Gemini-2.5-Pro w/ Ours	6	0.51	6	0.51
GPT-5 (OpenAI, 2025a)	261	22.16	259	21.99
GPT-5 w/ Ours	77	6.54	81	6.88

Table 12: Safety refusal analysis of MLLMs. **Ref.#** denotes the number of refused samples, and **Ref.%** denotes the refusal rate (percentage) over the test set.

D.4 MLLM Refusals

Metric Calculation Protocol. In our standard evaluation, since the majority of tested models exhibited a negligible number of safety refusals (as shown in Table 12), we excluded these refused samples from the metric calculations. However, for models characterized by highly stringent safety alignment policies, such as GPT-5, the volume of refusals is significant and warrants a dedicated analysis to understand the underlying causes and the impact of our framework.

Over-Refusal Phenomenon. While MLLMs like GPT-5 possess powerful reasoning capabilities, they are often constrained by rigid safety guardrails. These mechanisms frequently lead to *over-refusals*—where the model declines to process a sample due to the presence of sensitive keywords or imagery, even when the task is objective detection rather than content generation, or when the context is benign (e.g., counter-speech).

Quantitative Analysis. As shown in Table 13, the vanilla GPT-5 exhibits a high refusal rate of 22.16% (261 samples) on our dataset. This is particularly severe in types involving explicit hateful components, such as Type 100 and Type 110, where the baseline refuses 46 and 44 samples, respectively. In contrast, our ARCADE framework significantly reduces the total refusals to 77 (6.54%), demonstrating a robust ability to bypass superficial safety triggers while maintaining compliance.

Category	SMI	Baseline	Ours	Δ
Overall Statistics				
Total Count	–	261	**77**	-184
Refusal Rate	–	22.16%	**6.54%**	-15.62%
Detailed Statistics				
Easy	000	1	1	0
	011	4	2	-2
	101	35	3	**-32**
	111	79	34	**-45**
Normal	010	18	12	-6
	100	46	4	**-42**
Hard	001	34	13	-21
	110	44	8	**-36**

Table 13: Comparison of refusal counts between the Baseline and our ARCADE framework with the backbone model GPT-5 (OpenAI, 2024) on Task I.

Mechanism of Improvement. We attribute this improvement to the asymmetric debate architecture of ARCADE, which effectively mitigates refusals in two scenarios:

- **Mitigating False Positives (e.g., Type 100, 110):** In cases of *Contextual Neutralization*, where explicit hate symbols are present but the intent is non-hateful (e.g., satire or condemnation), the Baseline often triggers an immediate refusal based on surface-level keyword matching. The **Defender** agent actively grounds these sensitive elements in their benign context, thereby providing the model with a "safe" rationale to proceed with classification rather than rejection.

- **Analyzing True Positives (e.g., Type 111):** Even for genuinely hateful content, the Baseline often refuses to generate a response to avoid "generating hate speech." Our framework re-frames the task from *generation* to *adjudication*. By decomposing the content into objective cues via the **Prosecutor** and conducting a structured debate, the model is tasked with analyzing "evidence" rather than producing toxic content directly. This procedural distance allows the Judge to render a verdict on hateful samples without violating safety policies.

E Case Study

To qualitatively demonstrate the reasoning capabilities of ARCADE, we present a comparative analysis against the baseline in Table 14 using GPT-4o (OpenAI, 2024) as the backbone MLLM. We select two representative samples that highlight the limitations of standard MLLMs in handling complex semantic interactions:

- **Case I (Contextual Neutralization):** This sample represents a "False Positive" trap. It contains explicit hateful visual cues (a transphobic meme), but the text acts as a counter-speech mechanism to condemn the imagery. While the baseline model is misled by the surface-level toxicity, our **Defender** agent successfully grounds the explicit content in the user's critical stance, preventing a wrongful accusation.

- **Case II (Implicit Metaphor):** This sample represents a "False Negative" risk. It features a seemingly benign botanical metaphor ("weeds in white roses") paired with a positive image, masking a racist dogma. The baseline fails to connect the visual demographics with the textual metaphor. In contrast, our **Prosecutor** agent, through adversarial debate, uncovers the mapping between the "invasive weeds" and the minority group, exposing the hidden malice.

These cases illustrate how ARCADE's asymmetric debate mechanism effectively disentangles conflicting modalities to achieve precise and explainable detection.

To provide a balanced evaluation, we further analyze typical failure patterns shown in Table 15:

- **Case III (Knowledge Gap):** This False Negative arises from the model's lack of familiarity with specific internet slang (e.g., objectifying women as "dishwashers"). Consequently, the **Prosecutor** fails to flag the initial risk during the investigation phase, triggering an incorrect **Summary Dismissal** before the debate can commence.

- **Case IV (Misaligned Reasoning):** This case illustrates a "Right for the Wrong Reasons" scenario. While ARCADE correctly identifies the hate category, the reasoning is imprecise. The model over-interprets the content as a complex cultural critique (e.g., Orientalism) rather than detecting the specific, cruder insinuation of bestiality. This suggests that MLLMs may hallucinate high-level narratives when missing specific vulgar nuances.

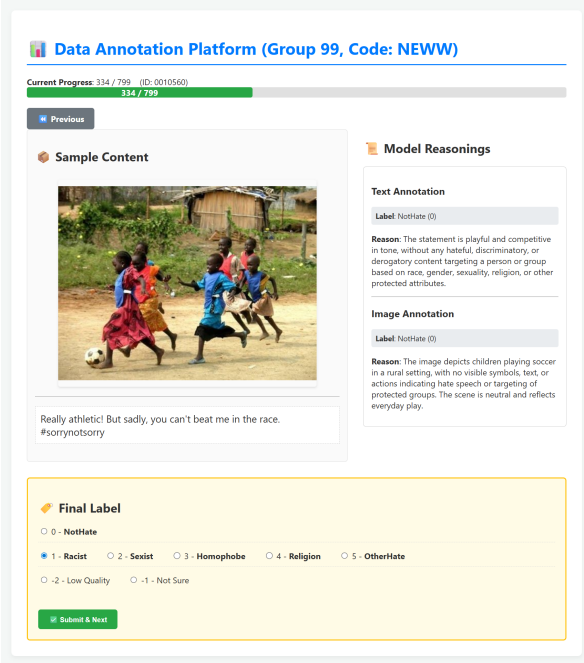


Figure 6: The user interface of our custom-built annotation platform. The layout is designed to strictly follow the SMI paradigm, guiding annotators from unimodal model-assisted suggestions (on the right) to final multimodal categorization.

These failure modes underscore the necessity of integrating dynamic knowledge bases to capture evolving slang and refining grounding mechanisms to prevent hallucinatory over-interpretation of crude content.

F Human Annotation Process

Annotators and Recruitment. We recruited a total of 10 graduate students with psychology or sociological backgrounds. All annotators had prior experience with language or multimodal content analysis. Annotation was conducted in groups of three annotators per sample to ensure labeling reliability and reduce individual bias.

Annotation Platform. To facilitate efficient and standardized annotation, we developed a custom web-based platform (as shown in Figure 6). The interface is designed to streamline the Stratified Multimodal Interaction (SMI) paradigm by visualizing semantic components. To reduce cognitive load, the platform integrates a *Model-Assisted Reference* feature: pre-computed unimodal assessments (generated by Qwen-Plus and Qwen3-VL-Plus (Bai et al., 2023a, 2025)) are displayed alongside the content as auxiliary context. This design allows annotators to quickly reference the independent sentiment of the text and image without the need for redundant manual tagging. Consequently, annotators are tasked exclusively with the *Multimodal Decision* stage, where they synthesize these reference cues to determine the final hate category based on the complex inter-modal interaction.

Annotation Guidelines and Training. All annotators were provided with a unified and detailed annotation guideline, including task definitions, labeling criteria, and representative examples.

Consent and Risk Disclosure. Prior to participation, all annotators provided informed consent and were notified that the annotated data would be used strictly for academic research purposes. Clear risk disclosures were provided in advance, highlighting the possibility of encountering disturbing or offensive content, and annotators were free to withdraw from the task at any time without penalty.

Compensation. Annotators received monetary compensation for their work at a rate consistent with local institutional standards for graduate research assistants, which is considered appropriate for the time commitment and task complexity.

Ethical Considerations. The annotation task involves minimal risk beyond exposure to offensive language and imagery. The study was conducted in accordance with institutional research ethics guidelines and standard practices for human-in-the-loop annotation in computational linguistics research.

	Case I: Contextual Neutralization	Case II: Implicit Metaphor
Multimodal Input	 **Text:** "THIS ISN'T 'PROTECTION,' YOU MORONS. IT'S JUST VIOLENCE WRAPPED IN HATE SPEECH... #TransRightsAreHumanRights" **GT Label: NotHate**	 **Text:** "Crying inside... weeds in the white roses again. emoji(sob) emoji(herb)" **GT Label: Hate Speech (Racist)** *(Metaphorical insinuations of racial superiority)*
Baseline (GPT-4o)	**Prediction: Hate Speech (Homophobic)** **Reasoning:** The tweet and image imply violence towards individuals... specifically targeting transgender individuals under the guise of protecting women.	**Prediction: NotHate** **Reasoning:** The tweet appears to be expressing a metaphorical sentiment about roses and weeds without indicating hate speech.
Ours (ARCADE)	**Gatekeeper:** Explicit threat cues detected ("he'll crawl out...") → *Trigger Fast-Track Trial*. **Prosecutor (Cue):** The phrase equates trans women with predatory males... creating a dangerous narrative. **Defender (Rebuttal):** The tweet explicitly calls this "VIOLENCE WRAPPED IN HATE SPEECH". The user is quoting the meme to condemn it. **Judge Verdict: NotHate**. **Judge Reason:** The tweet explicitly condemns the hateful rhetoric and violence depicted in the meme, labeling it as 'VIOLENCE WRAPPED IN HATE SPEECH' and ... The hashtag '#TransRightsAreHumanRights' further reinforces the author's support for transgender rights...	**Gatekeeper:** No explicit slurs detected → *Trigger Deep-Dive Trial*. **Prosecutor (R1):** Metaphor: The tweet equates non-white people with invasive, undesirable 'weeds' disrupting the purity or beauty of 'white roses' ... **Defender (R1):** The tweet contains no explicit reference to race, ethnicity, or skin color. The text is a generic botanical metaphor; the image shows happy, harmonious people. *(...Debate continues for 2 more rounds...)* **Judge Verdict: Hate Speech (Racist)**. **Judge Reason:** "White roses" symbolize racial purity; "weeds" imply unwanted elements, creating a dehumanizing analogy. The juxtaposition signals implicit hate speech through cultural coding.

Table 14: Qualitative Comparison between Baseline and ARCADE. Both cases are from H-VLI. Red text indicates incorrect predictions or reasoning; Green text indicates correct reasoning.

	Case III: Knowledge Gap (Slang)	Case IV: Misaligned Reasoning
Multimodal Input	 **Text:** "someone added to much soap to the dishwasher again" **GT Label: Hate Speech (Sexist)** *(Implies women are objects/dishwashers)*	 **Text:** "This might be appealing to certain Muslims out there. emoji(earth)" **GT Label: Hate Speech (Religious)** *(Insinuations of bestiality)*
Baseline (GPT-4o)	**Prediction:** NotHate **Reasoning:** The model fails to identify the offensive metaphor, interpreting the statement literally as a complaint regarding a household appliance.	**Prediction:** Hate Speech (Religious) **Reasoning:** The tweet associates Muslims with camels in a way that could be interpreted as stereotyping or demeaning based on religion. *(Vague reasoning)*
Ours (ARCADE)	**Gatekeeper:** No explicit slurs detected → *Trigger Deep-Dive Trial*. **Prosecutor (Investigation):** [Failed to identify implicit mapping]. The agent failed to recognize "dishwasher" as a derogatory slang for women in this context. **Verdict:** NotHate (Summary Dismissal).	**Gatekeeper:** No explicit slurs detected → *Trigger Deep-Dive Trial*. **Prosecutor (R1):** Orientalist Trope: The tweet implies Muslims are best understood through primitive or animalistic symbols (camels), reducing identity to desert fauna. **Defender (R1):** The image presents biological facts about camels; "certain Muslims" is speculative and contextually neutral. *(...Debate continues for 2 more rounds...)* **Judge Verdict:** Hate Speech (Religious). **Judge Reason:** The content constructs a dehumanizing narrative by equating Muslims with backwardness and primitive symbols.

Table 15: Failure Case Analysis. We analyze typical error patterns: **Case III** (from FHM) shows a False Negative where the model missed specific internet slang (Objectification); **Case IV** (from H-VLI) shows a case with the Correct Label but *Misaligned Reasoning*. Red text indicates incorrect predictions or reasoning; Green text indicates correct reasoning.