SAGE: Synergistic Adaptive Gating of Experts for Hateful Video Detection

Jie Huang[1,2,3*], **Xin Liao**[1,2,3*], **Junjie Wang**[1,2,3†], **Mingyang Li**[1,2,3], **Wenshuo Wang**[1,2,3], **Ziyou Jiang**[1,2,3], **Shoubin Li**[1,2,3] and **Qing Wang**[1,2,3†]
[1]State Key Laboratory of Complex System Modeling and Simulation Technology, Beijing, China [2]Science and Technology on Integrated Information System Laboratory
Institute of Software Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences
{huangjie, junjie, mingyang2017, wangwenshuo2024, ziyou2019, shoubin, wq}@iscas.ac.cn, wd2234016@stu.ahu.edu.cn

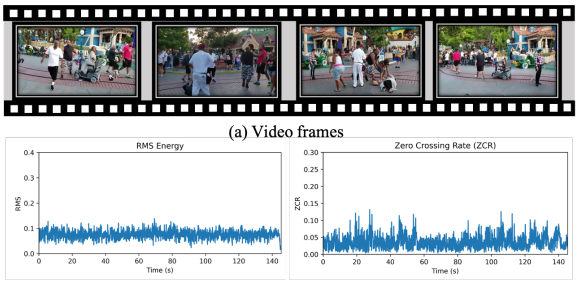
Abstract

With the rise of short-video platforms, hate speech has evolved from static text and memes into more covert and aggressive hateful video formats, profoundly impacting social dynamics and public sentiment. Existing detection methods typically rely on multimodal feature fusion, which blurs the distinct boundaries of modality-specific information. This leads to the feature dilution problem, where dominant benign modalities often overwhelm sparse, localized hateful cues. To address this, we propose SAGE (**S**ynergistic **A**daptive **G**ating of **E**xperts), a novel framework that shifts the paradigm from blind feature mixing to decision-level arbitration. Mimicking human cognitive processes, SAGE instantiates disentangled experts to rigorously preserve modality-specific semantics, facilitates global expert deliberation for context-aware refinement, and convenes an instance-level tribunal to dynamically arbitrate the final verdict based on evidentiary salience. Extensive experiments on HateMM and MultiHateClip benchmarks demonstrate that SAGE significantly outperforms state-of-the-art methods, achieving accuracy gains of 6.37% to 21.23% and macro-F1 score gains of 6.77% to 28.01%.

Disclaimer: This paper contains hateful content, which has the potential to be offensive and may disturb readers.

1 Introduction

With the rapid development of social media, the landscape of information dissemination has undergone fundamental changes. Information carriers have evolved from plain text to images and videos, offering enhanced immersiveness and expressiveness. This change has profoundly altered the way users perceive and interpret the world. However, this trend has also provided richer and more insidious channels for the propagation of hate speech. Hateful content targets social groups based on race, religion, gender, and other attributes, propagating prejudice and discrimination through explicit derogatory expressions or implicit narrative structures, visual metaphors, and audio cues (Silva et al., 2016; Chhabra and Vishwakarma, 2023). With the rise of short-video platforms, the cross-modal dissemination of hateful videos has intensified the challenges of content moderation, making automated hateful video detection a necessary solution.



(a) Video frames

(b) Audio representation

... girls' trying to get a boy's cash then you even see a spook with a big mustache... He don't do nothing but sleep and doze, he don't wake until his eyes are tired of being closed....

(c) Textual Content

Figure 1: An example of a video labeled "Hateful" where hateful content is primarily expressed through the **text modality**. (a) Video frames show people initially enjoying themselves before a fight breaks out. (b) Audio features (RMS and ZCR) indicate a neutral atmosphere without obvious hateful sentiment. (c) Text content contains implicit hateful remarks that reveal the underlying hostility.

Existing hateful video detection approaches generally fall into three categories: unimodal, multimodal, and large language model (LLM)-based methods. Unimodal approaches (Yousefi and Emmanouilidou, 2021; Mei et al., 2024; Tonneau et al., 2024) focus on isolated channels, inevitably missing cross-modal cues. LLM-based approaches (Jing et al., 2025; Wang et al., 2025;

*Equal contribution.
†Corresponding author.

Rehman et al., 2025; Lang et al., 2025; Koushik et al., 2025; Ma et al., 2025a; Yang et al., 2025) have demonstrated strong capabilities in semantic understanding and knowledge reasoning, but their deployment is often constrained by high computational costs and inference latency, limiting their applicability in real-time moderation scenarios. The most similar paradigm with this work is multimodal approaches (Mariconti et al., 2019; Wang et al., 2024a; Koushik et al., 2025) which jointly process text, audio, and visual features, and utilize feature concatenation or fusion techniques for mixing features into a unified, modality-agnostic representation (Das et al., 2023; Wang et al., 2024b; Lin et al., 2024). Although mechanisms like cross-modal attention can learn inter-modal relationships, this process inevitably blurs the distinct decision boundaries inherent to each modality. Consequently, the fused representation suffers from feature dilution, where subtle, modality-specific hateful cues become indistinguishable from the dominant benign context, ultimately leading to detection failure. For instance, as illustrated in Figure 1, while the visual content depicts people engaging in recreational activities followed by a conflict, and the audio features (RMS and ZCR) exhibit neutral patterns without explicit hateful expressions, the hateful nature is primarily conveyed through implicit textual remarks. In such scenarios, existing multimodal methods tend to indiscriminately fuse these conflicting cross-modal signals, allowing the neutral visual and audio cues to dilute the hateful textual content, thereby leading to misclassification.

This limitation stands in contrast to how human analyze potentially hateful content. It usually follows a profile-centric perspective: different profiles of the content, such as imagery, spoken words, on-screen text, or tone of voice, are examined both independently and in relation to each other. One might first identify a suspicious audio cue (audio profile), then cross-reference it with the visual context to confirm intent, and finally base the judgment primarily on the audio evidence while discounting the misleading visual scenery.

Driven by this observation, we argue that an effective framework must address two core challenges corresponding to this cognitive process: (1) how to enable deep, expert-level understanding of each modality to preserve unique semantics, while simultaneously allowing experts to exchange necessary context? (2) how to synthesize these independent analyses at the decision level, prioritizing the most salient expert for a given video while suppressing misleading or neutral ones?

To address this, we propose SAGE[1] (Synergistic Adaptive Gating of Experts), a novel hateful video detection framework that shifts the paradigm from blind feature mixing to cue-guided expert orchestration. SAGE first cultivates specialized "experts" for each modality, each tasked with deepening its own profile-specific representation and identifying potential sensitive cues (addressing challenge 1). These experts do not work in isolation; a global cross-modal interaction module allows them to exchange context, effectively letting auxiliary profiles refine or reinforce the primary cue-bearing profile. Most importantly, instead of merging all features indiscriminately, an instance-level adaptive gating mechanism dynamically assesses the contribution weight of each expert based on the salience and confidence of the cues it identifies (addressing challenge 2). This enables SAGE to emulate human-like reasoning: it prioritizes profiles with clear hateful indicators, uses supporting profiles to strengthen the decision, and filters out irrelevant or neutral information.

We conducted extensive experiments on two benchmark datasets, HateMM and MultiHateClip. Results demonstrate that SAGE achieves significant improvements over unimodal, multimodal, and LLM-based approaches, with accuracy gains ranging from 6.37% to 21.23% and macro-F1 score gains from 6.77% to 28.01%. Particularly on the HateMM dataset, SAGE attains 87.10% accuracy and 86.28% macro-F1 score.

The main contributions of this paper are summarized as follows:

- We identify the feature dilution problem in existing concatenation-based multimodal methods, revealing how benign modalities can mask localized hateful cues during feature fusion.
- We propose SAGE, a novel framework that employs synergistic experts to preserve modality-specific semantics and an adaptive gating mechanism to dynamically prioritize salient evidence at the decision level.
- We achieve state-of-the-art performance on HateMM and MultiHateClip datasets, validating that our expert-based arbitration strat-

[1]The SAGE implementation is available at https://github.com/XinLiao04/SAGE

egy effectively mitigates feature dilution compared to traditional fusion baselines.

2 Related work

Static Hateful Content Detection Early research primarily focused on unimodal text classification (Warner and Hirschberg, 2012; Tonneau et al., 2024), evolving from traditional machine learning to deep neural networks for analyzing tweets and comments. With the shift towards visual media, attention turned to hateful meme detection (Cao et al., 2023; Mei et al., 2024; Ma et al., 2025b), which requires joint reasoning over visual imagery and OCR-extracted text. While these approaches addressed the interplay between image and text, they remain limited to static content and fail to capture the temporal dynamics and acoustic cues inherent in video data.

Multimodal Hateful Video Detection The rise of short-video platforms has shifted the research focus to dynamic, multimodal detection. Early attempts relied on unimodal approaches, independently analyzing frames for hateful symbols (Nardelli and Comminiello, 2024), audio for aggressive tone (Yousefi and Emmanouilidou, 2021), or subtitles for toxic language (Lang et al., 2025). However, unimodal methods inherently miss cross-modal context. Consequently, multimodal fusion became the dominant paradigm (Koushik et al., 2025; Yang et al., 2025). Representative methods (Das et al., 2023; Wang et al., 2024b; Céspedes-Sarrias et al., 2025; Hossain et al., 2025) typically employ cross-modal attention or feature concatenation to merge modalities into a unified representation. Crucially, these "fuse-then-classify" strategies suffer from a significant limitation: they tend to mix features indiscriminately. In scenarios characterized by semantic dissonance, where hate is localized in one modality (e.g., audio) but contradicted by benign signals in others, static fusion strategies lead to feature dilution. The dominant benign features overwhelm the sparse hateful cues, causing detection failures. Unlike these entangled approaches, our SAGE framework adopts a disentangled, expert-based arbitration strategy to preserve modality-specific evidence.

Recent advancements in multimodal large language models (MLLMs), such as GPT-4V (Achiam et al., 2023) and Gemini (Team et al., 2024), have demonstrated impressive capabilities in semantic reasoning and zero-shot video understanding (Jing et al., 2025; Ma et al., 2025a). However, their deployment in real-world content moderation faces two hurdles: (1) Hallucination and Instability: MLLMs can exhibit response inconsistencies across modalities (Gong et al., 2025); (2) Computational Prohibitiveness: The massive parameter size and inference latency make them unsuitable for real-time, high-volume video moderation tasks. In contrast, SAGE offers a lightweight, specialized solution that achieves superior performance with significantly lower computational overhead.

3 Methodology

In this section, we elaborate on the architecture of SAGE, a framework designed to resolve the feature dilution problem where benign modalities obscure localized hateful cues. SAGE abandons traditional static fusion in favor of a dynamic, profile-centric paradigm that mimics human cognitive arbitration. It conceptualizes multimodal understanding as a dual process of collaboration and competition among modality-aware experts. As illustrated in Figure 2, the framework unfolds in three distinct modules: Profile Representation and Expert Initialization, Global Expert Deliberation, and Instance-Level Expert Tribunal. First, SAGE instantiates decoupled experts to rigorously encode the core semantic representations of each modality, thereby preserving modality-specific information and preventing feature entanglement. Subsequently, the Global Expert Deliberation mechanism facilitates cross-modal contextualization, reinforcing latent hateful cues while maintaining the stability of each expert's semantic profile. Finally, at the instance level, SAGE adaptively evaluates and arbitrates the contribution of each expert, amplifying the most salient evidence while suppressing irrelevant noise to render the final verdict.

3.1 Profile Representation and Expert Initialization

The foundation of SAGE lies in cultivating specialized agents capable of deep, modality-specific reasoning. Rather than treating multimodal inputs as mere signal streams, we view them as distinct semantic profiles (linguistic, acoustic, and visual) that require independent analysis before interaction. We instantiate three decoupled experts to extract raw evidence and align them into a unified semantic space.

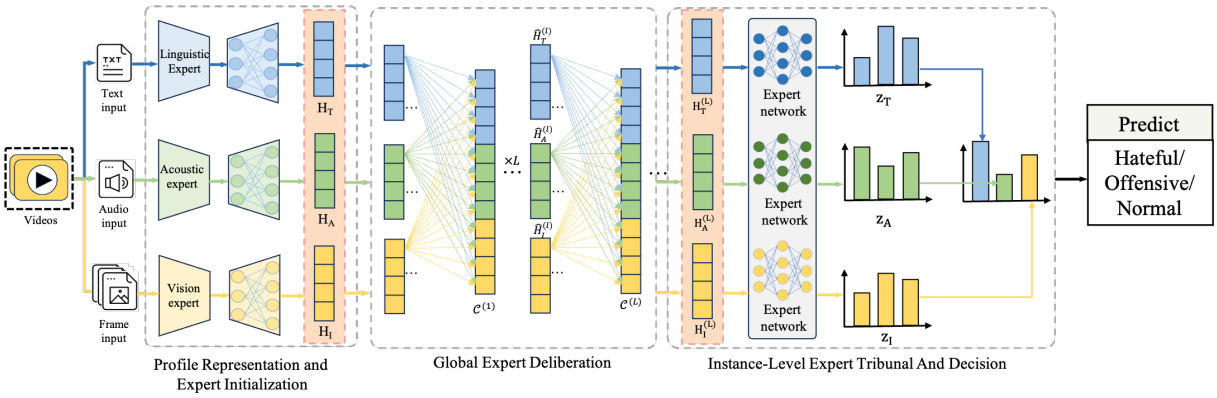


Figure 2: Overview of SAGE.

3.1.1 Unimodal Expert Instantiation

To construct the linguistic profile, the *Linguistic Expert* aggregates the video title, description, and transcripts obtained via automatic speech recognition. This raw textual sequence is processed by a pre-trained language encoder to capture explicit hate speech and implicit narrative structures, resulting in a sequence of token-level features denoted as $\mathbf{X}_T \in \mathbb{R}^{L_T \times d_T}$. Simultaneously, the *Acoustic Expert* focuses on paralinguistic cues such as tone, pitch, and volume, which are critical for detecting aggressive intent. By processing sampled audio frames through an acoustic feature extractor, we obtain the audio representation sequence $\mathbf{X}_A \in \mathbb{R}^{L_A \times d_A}$. Complementing these, the *Visual Expert* is tasked with modeling human appearances, actions, and scene contexts. It takes uniformly sampled video frames as input and utilizes a spatiotemporal visual encoder to extract high-level visual embeddings, yielding the feature sequence $\mathbf{X}_I \in \mathbb{R}^{L_I \times d_I}$. Here, $L_{\{T,A,I\}}$ and $d_{\{T,A,I\}}$ denote the sequence lengths and the original feature dimensions of the corresponding modalities, respectively.

3.1.2 Semantic Space Alignment

Since the raw features $\mathbf{X}_T$, $\mathbf{X}_A$, and $\mathbf{X}_I$ originate from disparate latent spaces with varying dimensions, they cannot directly engage in cross-modal deliberation. To address this, we apply modality-specific linear projection layers $\phi_m(\cdot)$ to map each expert's representation into a shared semantic manifold with a unified dimension d. Formally, for each modality $m \in \{T, A, I\}$, the initialized expert state $\mathbf{H}_m^{(0)}$ is computed as:

$$\mathbf{H}_m^{(0)} = \phi_m(\mathbf{X}_m) = \mathbf{X}_m \mathbf{W}_m + \mathbf{b}_m \quad (1)$$

where $\mathbf{W}_m \in \mathbb{R}^{d_m \times d}$ and $\mathbf{b}_m \in \mathbb{R}^d$ are learnable parameters. At this stage, the experts remain distinct, grounded in their respective modality profiles, yet are structurally aligned to facilitate the subsequent deliberation stage.

3.2 Global Expert Deliberation

While decoupled experts excel at identifying low-level cues within their respective domains, they inherently lack the cross-modal reasoning capability required to decipher complex hateful narratives (e.g., visual imagery that becomes offensive only when paired with specific audio context). To bridge this gap, we introduce the Global Expert Deliberation module. Designed to facilitate iterative semantic refinement, this module enables experts to dynamically exchange information while strictly preserving their structural independence. Its core objective is twofold: (1) to distill salient evidence within each modality to prevent noise propagation, and (2) to query a global semantic horizon for contextual verification, thereby resolving ambiguities without compromising the expert's distinct identity.

3.2.1 Intra-Modal Self-Refinement

Prior to engaging in external exchange, each expert must perform internal introspection to consolidate its own stance. We employ Multi-Head Self-Attention (MHSA) not only to capture long-range dependencies but also to filter irrelevant noise. By assigning higher attention weights to semantically rich segments, the expert distills the most informative cues from its raw profile.

Let $\mathbf{H}_m^{(l-1)} \in \mathbb{R}^{L_m \times d}$ denote the input representation of expert $m \in \{T, A, I\}$ at layer l. The refined state is computed as:

$$\hat{\mathbf{H}}_m^{(l)} = \mathrm{MHSA}\big(\mathrm{LN}(\mathbf{H}_m^{(l-1)})\big) + \mathbf{H}_m^{(l-1)} \quad (2)$$

This step serves as a pre-filter, ensuring that

only high-confidence, modality-specific signals are propagated to the subsequent interaction stage.

3.2.2 Global Semantic Horizon Construction

To enable experts to see beyond their isolated profiles, we construct a *Global Semantic Horizon* $\mathcal{C}^{(l)}$ at each layer of the iterative deliberation process. This horizon provides a panoramic view of the video's multimodal content, synthesized by aggregating distilled knowledge from all modality experts:

$$\mathcal{C}^{(l)} = \mathrm{Concat}\big(\hat{\mathbf{H}}_T^{(l)},\ \hat{\mathbf{H}}_A^{(l)},\ \hat{\mathbf{H}}_I^{(l)}\big) \in \mathbb{R}^{L_{\text{total}} \times d} \tag{3}$$

where $L_{\text{total}} = L_T + L_A + L_I$ denotes the total sequence length across all modalities.

This shared semantic horizon serves as a global knowledge repository against which each expert can contextualize and verify its modality-specific evidence.

3.2.3 Profile-Anchored Context Inquiry

This mechanism implements the philosophy of "active query" rather than "passive fusion". We employ Multi-Head Cross-Attention (MHCA) to allow each expert to selectively retrieve complementary evidence from the Global Semantic Horizon.

Crucially, a specific expert m acts as the **Querier** ($\mathbf{Q}_m$), actively retrieving evidence from the global horizon, which serves as the source of **Keys** ($\mathbf{K}_{\text{ctx}}$) and **Values** ($\mathbf{V}_{\text{ctx}}$). This profile-anchored interaction is formulated as:

$$\tilde{\mathbf{H}}_m^{(l)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}_{\text{ctx}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{ctx}} \tag{4}$$

By anchoring the Query to the expert's own profile, the resulting representation $\tilde{\mathbf{H}}_m^{(l)}$ remains structurally aligned with the original modality (e.g., the visual expert remains vision-centric). However, its semantic content is now enriched: ambiguous visual signals can be clarified by textual context, and subtle audio cues can be reinforced by visual evidence.

Finally, the context-aware representation is refined through a Feed-Forward Network (FFN) with residual connections:

$$\mathbf{H}_m^{(l)} = \mathrm{FFN}\Big(\mathrm{LN}\Big(\tilde{\mathbf{H}}_m^{(l)} + \hat{\mathbf{H}}_m^{(l)}\Big)\Big) + \tilde{\mathbf{H}}_m^{(l)} + \hat{\mathbf{H}}_m^{(l)} \tag{5}$$

Through L layers of such deliberation, the experts evolve from isolated detectors into context-aware reasoners, preparing them for the final instance-level tribunal.

3.3 Instance-Level Expert Tribunal And Decision

Traditional multimodal fusion often relies on static concatenation, which leads to indiscriminate feature blending. Consequently, a strong hateful signal in one profile (e.g., audio) can be statistically overwhelmed by benign signals in others. To counter this, SAGE functions as an Instance-Level Tribunal, treating the three deliberated experts as independent witnesses and dynamically arbitrating the final verdict based on evidentiary salience.

3.3.1 Independent Expert Diagnosis

Following the global deliberation stage, each expert possesses a refined understanding that integrates its core profile semantics with cross-modal context. Each expert $m \in \{T, A, I\}$ first independently diagnoses the content. Specifically, we aggregate the sequence representations $H_m^{(L)}$ using an attention pooling layer to obtain a focused profile-level representation $h_m \in \mathbb{R}^d$. This representation is then fed into a modality-specific classifier to produce the expert's prediction logits $\mathbf{z}_m$:

$$\mathbf{z}_m = \mathbf{W}_m^{(\text{cls})}\mathbf{h}_m + \mathbf{b}_m^{(\text{cls})} \tag{6}$$

where $\mathbf{W}_m^{(\text{cls})} \in \mathbb{R}^{K \times d}$ and $\mathbf{b}_m^{(\text{cls})} \in \mathbb{R}^K$ denote the learnable weight matrix and bias vector for expert m, respectively, and K represents the number of target classes.

This process enables each expert to cast an independent "vote" based on its specialized perspective.

3.3.2 Tribunal Arbitration Mechanism

Simultaneously, the Tribunal Network serves as the semantic judge, arbitrating the competition among experts to assess the trustworthiness of each expert for the current video instance. Unlike static fusion strategies, this mechanism operates dynamically by surveying the global evidence landscape and identifying which modality conveys the most salient information.

Specifically, we construct a tribunal state by concatenating the pooled representations from all experts and compute the evidential arbitration weights:

$$h_{\text{tribunal}} = \mathrm{Concat}(h_T, h_A, h_I) \tag{7}$$

$$\boldsymbol{\alpha} = \mathrm{Softmax}\big(W_{\text{gate}}\, h_{\text{tribunal}} + b_{\text{gate}}\big) \tag{8}$$

where $\boldsymbol{\alpha} = [\alpha_T, \alpha_A, \alpha_I]$ denotes the instance-specific arbitration weights assigned to the linguistic, acoustic, and visual experts, respectively.

To further suppress noisy or uninformative modalities and encourage decision sparsity (e.g., completely ignoring a silent audio track), we apply a *Top-K* selection strategy. Only the k most confident experts are retained, and their corresponding weights are re-normalized to obtain the final arbitration weights $\boldsymbol{\alpha}'$.

The final prediction $\hat{y}$ is derived through the tribunal's consensus by aggregating the experts' logits under the arbitrated weights:

$$\hat{y} = \sum_{m \in \{T,A,I\}} \alpha'_m \cdot \mathbf{z}_m \tag{9}$$

Through this mechanism, SAGE realizes dynamic dominance. For instance, if the acoustic expert detects high-confidence racial slurs, the tribunal can assign $\alpha_A \approx 1$, thereby allowing the audio evidence to override benign visual context. This ensures that localized hateful cues are explicitly surfaced rather than statistically diluted by non-informative modalities.

4 Experiment Design

To comprehensively evaluate the performance of SAGE, we propose the following three research questions (RQs):

- **RQ1: How does SAGE perform on the hate video detection task?**
- **RQ2: What is the contribution of each component within SAGE?**
- **RQ3: How computationally efficient is SAGE?**

Datasets Our study utilizes two widely used video datasets, HateMM (Das et al., 2023) and MultiHateClip (MHClip) (Wang et al., 2024b), each demonstrating a rich variety of hateful scenes and contexts. Please go to Appendix C for more details.

Baselines To comprehensively evaluate the effectiveness of our model on the Hateful Video Detection task, we select three categories of models as baselines for comparison. Detailed information about the baseline models is provided in Appendix D.

Metric To comprehensively evaluate our model and enable fair comparison with existing approaches, we adopt four evaluation metrics following previous work (Wang et al., 2024b; Lang et al., 2025): Accuracy (ACC), Macro-Precision (M-P), Macro-Recall (M-R), and Macro-F1 (M-F1) score. Additionally, to assess computational efficiency, we report the average per-sample inference time (Inf. Time) as an efficiency metric.

Implementation Details All experiments are conducted on a server running Ubuntu 24.04 LTS with dual Intel Xeon Gold 6133 CPUs and four NVIDIA RTX A6000 GPUs (48GB memory each). Training details of the proposed SAGE framework, including hyperparameters and optimization configurations, are provided in Appendix E. Detailed processing procedures for multi-modal features of video data are provided in Appendix G.

5 Experimental Results and Analysis

5.1 Overall Performance (RQ1)

Table 1 reports the binary classification results on the MHClip and HateMM datasets, while more fine-grained three-class classification results are provided in Appendix H.1. Overall, SAGE significantly outperforms existing baselines. Compared with the multimodal baseline models, SAGE achieves accuracy improvements ranging from 6.37% to 10.89% and M-F1 gains of 6.77% to 16.21%. Based on these experimental results, we draw the following conclusions:

(1) Multimodal Synergy vs. Unimodal Blindness. Unimodal methods generally exhibit performance ceilings due to inherent information incompleteness. While text-based methods perform relatively well (e.g., 76.04% accuracy on HateMM), they fail to capture semantic dissonance, such as hateful audio overlaid on benign visuals. This confirms the necessity of multimodal synergy. However, simply having multiple modalities is insufficient. As seen in the results, standard multimodal baselines (e.g., simple concatenation) improve over unimodal ones but still lag significantly behind SAGE.

(2) Validating the resolution of feature dilution. While traditional multimodal fusion methods (e.g., HateMM, MHClip baselines) generally achieve over 80% accuracy, they struggle to break through the performance bottleneck. This empirical gap validates our hypothesis of feature dilution: static fusion allows benign modalities to overwhelm localized hateful cues. In contrast, SAGE's superior performance demonstrates that our instance-level expert tribunal successfully identifies and amplifies the most salient evidence (e.g., a short audio

Table 1: Experimental results of all baselines and our proposed SAGE on HateMM and MHClip datasets for binary classification. T: Text, A: Audio, I: Image, V: Video. The "AVG. Improve" row is calculated as the average difference between SAGE and all baselines of the current type. The best results are in **bold** and the second-best are underscored.

Type	Model	Modality				Inf. Time (ms)	HateMM				MHClip-YouTube				MHClip-BiliBili			
		T	A	I	V		ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
Uni-Modal	mBert (Devlin et al., 2019)	✓				6.93	0.7604	0.7216	0.7883	0.7137	0.6905	0.4982	0.7188	0.5446	0.6975	0.5826	0.6259	0.5820
	MFCC (Davis and Mermelstein, 1980)		✓			99.57	0.7281	0.6746	0.7604	0.6730	0.6750	0.5037	0.5357	0.5191	0.6111	0.4976	0.5036	0.5029
	ViViT (Arnab et al., 2021)			✓		369.94	0.7558	0.7216	0.7704	0.7138	0.6813	0.5639	0.5818	0.5624	0.6914	0.4444	0.5981	0.5111
	LB (Zhu et al., 2023)				✓	542.54	0.7604	0.7398	0.7558	0.7336	0.7063	0.4984	0.5897	0.5280	0.7037	0.6201	0.6409	0.6141
	AVG. Improve	-	-	-	-	-	11.98%↑	14.73%↑	10.23%↑	14.86%↑	15.03%↑	28.01%↑	25.01%↑	11.41%↑	21.23%↑	16.29%↑	16.29%↑	19.04%↑
Multi-Modal	HateMM (Das et al., 2023)	✓	✓	✓		490.01	0.8065	0.7985	0.7975	0.7998	0.7188	0.6281	0.6451	0.6211	0.7160	0.6635	0.6656	0.6618
	Mo-Hate (Tomar et al., 2023)	✓	✓	✓		553.08	0.8157	0.8049	0.8098	0.8014	0.7312	0.6781	0.6752	0.6817	0.7284	0.6408	0.6776	0.6320
	MHClip (Wang et al., 2024b)	✓	✓	✓		578.21	0.8065	0.7910	0.8062	0.7838	0.7250	0.6270	0.6523	0.6190	0.7037	0.4667	0.7263	0.5255
	HCC1 (Koushik et al., 2025)	✓	✓	✓		567.21	0.7834	0.7702	0.7756	0.7667	0.7125	0.4876	0.6118	0.5259	0.7160	0.5522	0.6829	0.5677
	MoRE (Lang et al., 2025)	✓	✓	✓		618.76	0.8110	0.8004	0.8061	0.7966	0.7400	0.6662	0.7219	0.6563	0.7500	0.6994	0.7153	0.6911
	MM-HSD (Céspedes-Sarrias et al., 2025)	✓	✓	✓		662.33	0.8203	0.8054	0.8230	0.7972	0.7438	0.7172	0.7113	0.7489	0.6914	0.6495	0.6462	0.6550
	AVG. Improve	-	-	-	-	-	6.37%↑	6.77%↑	6.80%↑	6.62%↑	10.89%↑	16.21%↑	13.59%↑	14.65%↑	7.25%↑	13.63%↑	6.94%↑	12.08%↑
MLLM	Qwen-VL 7B (Bai et al., 2023)	✓		✓		1982.98	0.6452	0.6286	0.6290	0.6286	0.6469	0.6363	0.6542	0.6798	0.6579	0.6368	0.6381	0.6566
	LLaMA-3.2V 11B (Grattafiori et al., 2024)	✓		✓		1166.29	0.6843	0.6417	0.6421	0.6413	0.7149	0.6559	0.6533	0.6591	0.6593	0.5479	0.5678	0.5499
	Keye-VL 8B (Team et al., 2025)				✓	7099.30	0.7872	0.7293	0.8035	0.7125	0.6947	0.5881	0.6231	0.5859	0.6706	0.5629	0.5934	0.5649
	GPT-4 (Achiam et al., 2023)	✓				1526.19	0.7558	0.7550	0.7683	0.7777	0.7937	0.7127	0.7666	0.6932	0.7593	0.7371	0.7305	**0.7593**
	GPT-4V (Achiam et al., 2023)	✓		✓		7654.09	0.7327	0.7323	0.7870	0.7746	0.7518	0.7300	0.7267	0.7739	0.6914	0.6866	0.7214	0.7546
	AVG. Improve	-	-	-	-	-	14.99%↑	16.54%↑	14.51%↑	15.02%↑	11.71%↑	13.16%↑	12.07%↑	11.03%↑	10.24%↑	11.41%↑	10.48%↑	8.59%↑
Ours	SAGE	✓	✓	✓		567.98	**0.8710**	**0.8628**	**0.8711**	**0.8572**	**0.8375**	**0.7962**	**0.8055**	**0.7887**	**0.7901**	**0.7484**	**0.7551**	0.7430

slur) while suppressing noise, thereby resolving the dilution issue that plagues static fusion models.

(3) Specialized Experts vs. Generalist MLLMs. Despite the impressive generalized reasoning capabilities of MLLMs, they fall short in this domain-specific task. As indicated in the results, LLM-based methods (relying on prompt engineering) generally underperform compared to our specialized framework. Specifically, on the HateMM dataset, SAGE achieves a improvement of 8.48 percentage points in accuracy and 13.35 points in m-F1 over the best-performing MLLM baseline. This performance gap indicates that, under the current zero-shot prompting setting, general-purpose MLLMs struggle to capture the subtle, high-frequency acoustic and visual cues characteristic of hate speech. In contrast, SAGE's specialized experts are explicitly designed to extract these fine-grained discriminative features, leading to superior detection accuracy.

(4) Analysis of Fine-grained Detection. The three-class classification task (Appendix H.1) reveals the challenge of distinguishing "offensive" from "hateful" content. While all models experience performance drops in this more fine-grained setting, SAGE still achieves the best performance among all methods. This indicates that our disentangled expert deliberation mechanism enables more granular feature discrimination, effectively separating implicit hate from general offensive expressions better than entangled fusion methods.

5.2 Ablation Study (RQ2)

Table 2 summarizes the contribution of each component within SAGE. Overall, removing any module consistently degrades performance, validating the framework's holistic design. Regarding expert profiles, the sharpest drop occurs in the w/o Linguistic Expert variant, confirming text as the primary source of explicit hate.

Regarding architecture, the w/o Semantic Alignment variant suffers the most severe degradation, confirming that projecting heterogeneous features into a unified space is a prerequisite for effective interaction. Removing Global Expert Deliberation (GED) reduces experts to isolated islands, hindering the cross-modal contextualization needed for complex narratives. The significant drop in the w/o Tribunal Network variant empirically validates our core hypothesis: dynamic arbitration is essential to prevent feature dilution, where dominant benign signals otherwise obscure localized hateful cues.

Furthermore, we replace SAGE's tribunal network with a feature-level gating mechanism. Specifically, this variant performs weighted fusion of multimodal features via gating weights prior to classification, rather than relying on an expert-based arbitration module for explicit decision making. Experimental results show that this variant achieves only 82.95% accuracy and 82.47% macro-F1 on the HateMM dataset, representing a decrease of 4.15% and 3.81%, respectively, compared to the original SAGE. This comparison further demonstrates that, compared to soft selection at the feature fusion stage, dynamic arbitration at the decision level is more effective in capturing fine-grained hate-related signals while suppressing irrelevant

Table 2: Ablation study evaluating the core components of SAGE on HateMM and MHClip datasets.

Model	Modality			HateMM				MHClip-YouTube				MHClip-BiliBili			
	T	A	I	Acc	M-F1	M-P	M-R	Acc	M-F1	M-P	M-R	Acc	M-F1	M-P	M-R
w/o Linguistic expert		✓	✓	0.7834	0.7702	0.7756	0.7667	0.7688	0.6344	0.7670	0.6238	0.6975	0.5488	0.6233	0.5598
w/o Acoustic expert	✓		✓	0.8571	0.8510	0.8504	0.8517	0.8000	0.7527	0.7563	0.7494	0.7840	0.7455	0.7469	0.7441
w/o Visual expert	✓	✓		0.8295	0.8182	0.8269	0.8128	0.8063	0.7680	0.7635	0.7733	0.7716	0.7367	0.7333	0.7407
w/o Alignment	✓	✓	✓	0.8203	0.8084	0.8169	0.8032	0.7750	0.7217	0.7249	0.7189	0.7222	0.6167	0.6704	0.6109
w/o GED	✓	✓	✓	0.8387	0.8351	0.8322	0.8444	0.7812	0.7237	0.7330	0.7168	0.7469	0.6824	0.7020	0.6730
w/o Tribunal network	✓	✓	✓	0.8433	0.8334	0.8411	0.8283	0.7937	0.7530	0.7488	0.7580	0.7654	0.7361	0.7296	**0.7473**
SAGE	✓	✓	✓	**0.8710**	**0.8628**	**0.8711**	**0.8572**	**0.8375**	**0.7962**	**0.8055**	**0.7887**	**0.7901**	**0.7484**	**0.7551**	0.7430

noise.

5.3 Efficiency Analysis (RQ3)

Figure 3 benchmarks SAGE against competitive methods. Despite a larger parameter count necessitated by specialized experts, SAGE incurs a negligible inference overhead of only 0.1 seconds while delivering an accuracy gain of over 6%. In practical deployment, the end-to-end latency consists of two main components: data preprocessing and model inference. For a typical video, the complete processing pipeline takes approximately 5.97 seconds, with the following breakdown: ASR transcription accounts for the largest proportion of time (3.8 seconds), followed by video frame sampling (16 frames, 1.3 seconds), audio extraction (0.2 seconds), model inference (0.5 seconds), and feature embedding of the three modalities—text, audio, and vision (0.02 seconds, 0.04 seconds, and 0.11 seconds, respectively).

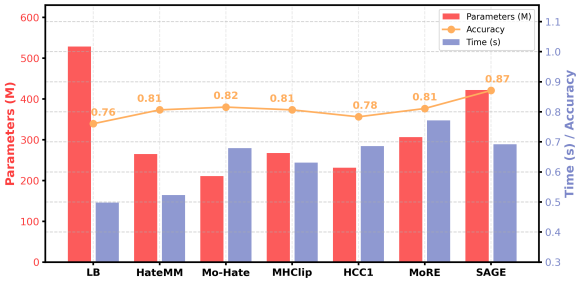


Figure 3: Comparison of the performance of multimodal models in terms of parameters (in **red**), inference time (in **purple**) and ACC performance (in **yellow**) on HateMM dataset.

These results demonstrate that SAGE achieves a strong balance between model complexity and computational efficiency. It offers high accuracy and low latency, meeting the dual requirements of timeliness and precision in content moderation systems, and demonstrating practical value for large-scale deployment.

5.4 Robustness Analysis

We analyze the robustness of SAGE along two dimensions: the effect of ASR transcription quality and the impact of missing modalities.

Robustness to ASR Quality. Table 3 reports SAGE's performance across three datasets under varying ASR transcription quality. Higher transcription quality consistently yields improved performance. More importantly, performance degradation remains moderate across all datasets when transcription quality decreases, indicating that SAGE does not over-rely on textual signals and is robust to ASR errors.

Table 3: Robustness analysis under different transcription qualities.

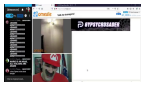
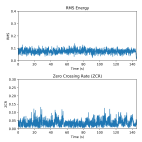
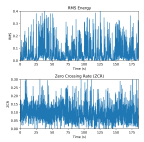
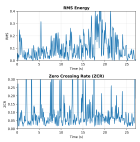
Trans Quality	Dataset	Acc	M-F1	M-P	M-R
Original Trans	HateMM	0.8525	0.8432	0.8511	0.8379
Better Trans	HateMM	0.8710	0.8628	0.8711	0.8572
Original Trans	MHClip-YouTube	0.8000	0.7673	0.7582	0.7818
Better Trans	MHClip-YouTube	0.8375	0.7962	0.8055	0.7887
Original Trans	MHClip-BiliBili	0.7716	0.7393	0.7343	0.7463
Better Trans	MHClip-BiliBili	0.7901	0.7484	0.7551	0.7430

Robustness to Missing Modalities. Table 4 reports performance on the HateMM dataset when each modality is individually removed. The absence of audio or visual modalities leads to only a moderate performance drop, whereas removing the text modality results in a more pronounced decline, which is consistent with the observation that hateful intent is predominantly conveyed through language. This limitation is not unique to SAGE, but reflects a common challenge shared by existing multimodal models and human moderators when critical information is unavailable.

Table 4: Robustness Analysis under Modality Missing Settings.

Modality Removed	Acc	M-F1	M-P	M-R
Text	0.7972	0.7663	0.8337	0.7542
Audio	0.8664	0.8568	0.8696	0.8494
Vision	0.8187	0.8020	0.8145	0.7954

Table 5: Visualization of the Instance-Level Expert Tribunal. We report the independent predictions from Linguistic (T), Acoustic (A), and Visual (I) experts, alongside the dynamic arbitration weights assigned by the tribunal. GT: Ground Truth.

Hatefulness	(A) **GT: Hateful**	(B) **GT: Hateful**	(C) **GT: Harmful**
Vision			
Audio			
Text	...girls' trying to get a boy's cash then...He don't do nothing but sleep...	Why you're not? because I can't. what do you think of the niggers....	straight women hate men more than anybody because lesbians can't hate men that much...
Expert Diagnosis	0.83 / 0.41 / 0.96	0.38 / 0.57 / 0.81	0.49 / 0.54 / 0.58
Tribunal Weights	0.21 / 0.00 / 0.66	0.00 / 0.01 / 0.99	0.00 / 0.37 / 0.62
Output	**Hateful** (0.81) ✓	**Hateful** (0.81) ✓	**Harmful(0.56)** ✓

5.5 Case Study: Visualizing the Expert Tribunal

Table 5 visualizes the Instance-Level Expert Tribunal to demonstrate how SAGE arbitrates evidence across modalities. We reference Zero Crossing Rate (ZCR) and RMS energy to validate acoustic noise (Das et al., 2023; Wang et al., 2024b).

(A) Synergistic Evidence Aggregation. The video conveys hate through both racial slurs (Text) and violent imagery (Visual), while the audio remains calm. SAGE demonstrates synergistic reasoning by assigning high weights to both Linguistic and Visual experts, effectively aggregating complementary evidence while silencing the non-informative Acoustic expert. **(B) Robustness against Acoustic Noise.** This scenario involves high-energy acoustic noise (high ZCR/RMS) that mimics aggression. Despite this interference, the Tribunal correctly identifies the audio as noise rather than evidence. It assigns a decisive weight (0.99) to the Linguistic expert (containing explicit hate), proving its robustness in filtering out misleading high-energy signals. **(C) Collaborative Evidence Reinforcement.** In this scenario, hate is distributed across modalities with moderate intensity. Both the acoustic and linguistic experts detect potential hate with comparable confidence, while the visual modality remains ambiguous. The Tribunal executes collaborative reinforcement: instead of relying on a single dominant view, it assigns significant weights to both text (0.62) and audio (0.37) to cross-validate the decision, while silencing the non-contributing visual.

6 Conclusion

In this work, we address the critical challenge of feature dilution in multimodal hateful video detection. We propose SAGE, a novel framework that shifts the paradigm from static feature mixing to dynamic evidence arbitration. By conceptualizing the detection process as a committee of disentangled experts, SAGE enables synergistic deliberation to capture cross-modal context without compromising semantic independence, followed by an instance-level tribunal that dynamically weighs evidentiary salience. Extensive experiments on the HateMM and MHClip benchmarks demonstrate that SAGE establishes new state-of-the-art performance on both binary and fine-grained classification tasks. Our analysis confirms that SAGE effectively resolves semantic dissonance, successfully identifying hate speech even when hidden in a single modality. Furthermore, SAGE offers a highly efficient alternative to resource-intensive LLMs, providing a robust and cost-effective solution for real-time content moderation. We hope the design principles of "disentangle-then-arbitrate" offer valuable insights for future multimodal research.

Limitations

Despite achieving state-of-the-art performance, SAGE has limitations that warrant further exploration. First, regarding expert initialization, we employ three independently pre-trained encoders (RoBERTa, MFCC, VideoMAE) to ensure disentangled profile representation. While this preserves modality specificity, the inherent semantic heterogeneity between these disparate feature spaces poses challenges for perfect alignment. Future work could explore initializing experts with jointly pre-trained backbones or employing nonlinear manifold mapping to bridge the semantic gap more effectively without sacrificing expert independence. Second, regarding the comparison paradigm, due to the scarcity of domain-specific video data and computational constraints, our evaluation of MLLM was limited to zero-shot inference. While this setting partially highlights the efficiency advantages of SAGE, it does not fully exploit the potential of MLLMs, as more sophisticated prompting strategies or fine-tuning may yield stronger performance. Future research could inves-

tigate knowledge distillation techniques to transfer the complex reasoning capabilities of MLLMs into the lightweight SAGE framework, combining the best of both paradigms (deep reasoning and inference efficiency).

Ethical Considerations

Our research aims to mitigate the dissemination of hateful videos and reduce their societal harm. However, the deployment of automated moderation tools warrants strict ethical scrutiny. First, regarding data privacy, we use established public benchmarks (HateMM, MHClip) solely for academic research purposes and strictly adhere to their data usage licenses. Second, regarding algorithmic fairness, we acknowledge that multimodal models risk learning spurious correlations between toxicity and specific demographics (e.g., dialects or skin tones). While SAGE's expert-based arbitration enhances interpretability, we advocate using the model as a triage tool to assist rather than replace human moderators, minimizing the risk of suppressing protected speech or marginalized voices. Finally, to address the potential psychological impact of exposure to hateful audio-visual content, we provide regular psychological counseling for researchers involved in qualitative analysis.

Acknowledgement

We sincerely appreciate all the reviewers for their constructive suggestions. This work was supported by the National Key Research and Development Program of China (No. 2024YFF0618800), National Natural Science Foundation of China Grant (No.62402484).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. 2025. Mm-hsd: Multi-modal hate speech detection in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2546–2555.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Md Mithun Hossain, Md Shakil Hossain, Sudipto Chaki, and Muhammad Firoz Mridha. 2025. Co-attendwg: Co-attentive dimension-wise gating and expert fusion for multi-modal offensive content detection. *IEEE Transactions on Artificial Intelligence*.

Yiheng Jing, Mingming Zhang, Yong Zhuang, Jiacheng Guo, Juan Wang, Xiaoyang Xu, Wenzhe Yi, Keyan Guo, and Hongxin Hu. 2025. Hvguard: Utilizing multimodal large language models for hateful video detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Girish A Koushik, Diptesh Kanojia, and Helen Treharne. 2025. Towards a robust framework for multimodal hate detection: A study on video vs. image-based content. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2014–2023.

Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 2025. Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection. In *Proceedings of the ACM on Web Conference 2025*, pages 2763–2774.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Jin Ma, Mohammed Shujaa Aldeen, Feng Luo, and Long Cheng. 2025a. Few-shot detection of hate videos using multi-modal large language models. In *Proceedings of the 1st ACM Workshop on Deepfake, Deception, and Disinformation Security*, pages 32–35.

Yihan Ma, Xinyue Shen, Yiting Qu, Ning Yu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025b. From meme to threat: On the hateful meme understanding and induced hateful content generation in open-source vision language models. In *USENIX Security Symposium (USENIX Security). USENIX*.

Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 5333–5347.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940.

Pietro Nardelli and Danilo Comminiello. 2024. Josenet: A joint stream embedding network for violence detection in surveillance videos. *arXiv preprint arXiv:2405.02961*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR.

Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Nagendra Kumar. 2025. Implihatevid: A benchmark dataset and two-stage contrastive learning framework for implicit hate speech detection in videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 17209–17221.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 687–690.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*.

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. 2023. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933.

Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9020–9040.

Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. 2024a. What do they "meme"? a metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294:111778.

Han Wang, Rui Yang Tan, and Roy Ka-Wei Lee. 2025. Cross-modal transfer from memes to videos: Addressing data scarcity in hateful video detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5255–5263.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024b. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26.

Shuonan Yang, Tailin Chen, Jiangbei Yue, Guangliang Cheng, Jianbo Jiao, and Zeyu Fu. 2025. Reasoning-aware multimodal fusion for hateful video detection. *arXiv preprint arXiv:2512.02743*.

Midia Yousefi and Dimitra Emmanouilidou. 2021. Audio-based toxic language classification using self-attentive convolutional neural network. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 11–15.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

A Problem Definition

We formulate hateful video detection as a multimodal multi-class classification task. Let $\mathcal{D} = \{v_i\}_{i=1}^{N}$ denotes a dataset containing N videos. For each video instance v_i, we extract three distinct modalities: textual features T_i (derived from titles, descriptions, and ASR transcripts), acoustic features A_i (sampled audio frames), and visual features I_i (sampled video frames). The objective is to learn a mapping function F that predicts the semantic label $\hat{y}_i$ from a pre-defined label set $\mathcal{Y} = \{c_1, c_2, \ldots, c_K\}$, where K denotes the number of classes.

Formally, given an input tuple (T_i, A_i, I_i), the model estimates the posterior probability distribution over the class space $\mathcal{Y}$. The predicted label $\hat{y}_i$ is obtained by selecting the class with the maximum likelihood:

$$\hat{y}_i = \arg\max_{c \in \mathcal{Y}} P(c \mid T_i, A_i, I_i; \Theta) \tag{10}$$

where Θ denotes the set of learnable parameters of the SAGE framework.

B End-to-End Optimization Objective

We introduce a comprehensive end-to-end training objective to optimize the SAGE framework. To ensure a robust and unbiased tribunal, our loss function is designed to achieve three goals concurrently: accurate final arbitration, competent individual experts, and fair expert utilization.

Let $\mathcal{D} = \{(v^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote a training batch, where $y^{(i)}$ is the ground-truth label. The total objective $\mathcal{L}$ is defined as a weighted sum of the arbitration loss, expert supervision loss, and a fairness regularization term:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{L}_{\text{arb}}^{(i)} + \lambda_{\text{exp}} \mathcal{L}_{\text{exp}}^{(i)} + \lambda_{\text{fair}} \mathcal{L}_{\text{fair}}^{(i)} \right), \tag{11}$$

where λ_{exp} and λ_{fair} are hyperparameters that control the trade-off between different objectives.

The primary objective is to minimize the error of the final arbitrated decision $\hat{y}^{(i)}$. We employ the standard Cross-Entropy (CE) loss:

$$\mathcal{L}_{\text{arb}}^{(i)} = \mathcal{L}_{\text{CE}}(\hat{y}^{(i)}, y^{(i)}). \tag{12}$$

Optimizing this term guides the tribunal to dynamically weigh multimodal evidence for correct classification.

To prevent experts from free-riding (i.e., relying solely on cross-modal context without learning distinctive profile features), we enforce explicit supervision on each expert's independent prediction $y_m^{(i)}$:

$$\mathcal{L}_{\text{exp}}^{(i)} = \sum_{m \in \{T,A,I\}} \mathcal{L}_{\text{CE}}(y_m^{(i)}, y^{(i)}). \tag{13}$$

This constraint ensures that every expert remains individually discriminative, providing high-quality testimony for the tribunal.

A common issue in dynamic gating mechanisms is modal collapse, where the model over-relies on a dominant modality (e.g., vision) while suppressing others. To prevent such bias and encourage diverse expert participation, we introduce a fairness regularization term (also known as a load-balancing loss (Shazeer et al., 2017)):

$$\mathcal{L}_{\text{fair}}^{(i)} = \sum_{m \in \{T,A,I\}} g_m^{(i)} \cdot f_m^{(i)}. \tag{14}$$

Here, $g_m^{(i)}$ denotes the tribunal weight assigned to expert m, and $f_m^{(i)}$ represents the utilization frequency (load) of that expert within the batch.

Minimizing this term penalizes extreme imbalances in expert usage, encouraging the tribunal to consult all experts fairly rather than developing a static bias toward a single modality.

C Dataset

In our experiments, we evaluated the performance of our SAGE model on two available datasets, HateMM (Das et al., 2023) and MultiHateClip (Wang et al., 2024b), and conducted comprehensive comparisons against a range of baseline models. The statistical information of the dataset is shown in Table 6 .

Table 6: Summary of dataset in our experiments. H: Hateful, O: Offensive, N: Normal.

Dataset	Language	H	O	N	Total
HateMM	English	428	-	652	1080
MHClip	Chinese	100	168	571	839
	English	67	190	573	830

HateMM The HateMM dataset comprises 1083 videos from BitChute, manually labeled as *Hate* or *Non-Hate*, and serves as a foundational dataset for researches on hateful video detection tasks. Additionally, the dataset includes temporal segments identifying when hateful content occurs (hate snippet), along with annotations specifying the target of hatred (target).

MHClip The multilingual MultiHateClip dataset provides fine-grained annotations for hateful video analysis by introducing an *Offensive* label. It contains a total of 2,000 videos, 1,000 from YouTube (denoted as MHClip-YouTube) and 1,000 from Bilibili (denoted as MHClip-Bilibili). Each video is annotated with *Hateful*, *Offensive* or *Normal*. In addition to specifying the hate snippet and the target, the dataset also identifies the particular modality in which the hateful content appears.

Task Specification For binary classification task, we use both the HateMM and MHClip dataset to comprehensively test the model's performance. For the MHClip dataset, we merge the *Offensive* and *Hateful* labels into a single category *Harmful* while retaining the *Normal* label unchanged. To validate the effectiveness of SAGE method on fine-grained labels, we additionally performed a three-class classification experiment exclusively on MHClip (Bilibili and YouTube) dataset and the label used are *Hateful*, *Offensive* and *Normal*.

D Baselines

During our experiments, we selected a total of 14 widely adopted hateful content detection methods, covering unimodal, multimodal, and LLM-based models, to evaluate their performance in hateful video detection.

Unimodal hateful video detection model

- mBert (Devlin et al., 2019): BERT has been proven to exhibit high efficiency and superior performance in hateful speech detection (Mozafari et al., 2019). For the textual content of videos, after feature extraction via BERT, we take the [CLS] token as the global representation of the text and subsequently feed the textual feature into an MLP with two FC layers for classification.
- MFCC (Davis and Mermelstein, 1980): MFCC has been widely used to characterize audio signals and serves as an effective input for both traditional and deep learning models (Yousefi and Emmanouilidou, 2021). For each audio, we extract a 128-dimensional MFCC feature as its audio representation, and feed the feature into an MLP with two FC layers for classification.
- ViViT (Arnab et al., 2021): ViViT performs global modeling on video frames, effectively capturing long-term temporal and spatial dependencies to enhance classification performance. We take the [CLS] token as the global feature representation and feed the feature into an MLP with two FC layers for classification.
- LanguageBind (Zhu et al., 2023): LanguageBind is a language-centric multimodal framework that aligns multiple modalities into a unified linguistic semantic space. Leveraging its strong cross-modal alignment capability and its ability to take video as direct input, we feed the extracted feature into a MLP with two FC layers for classification.

Multimodal hateful video detection model

- HateMM (Das et al., 2023): The HateMM model performs single-modal feature modeling through dense layers and LSTMs, and subsequently concatenates the feature of the three modalities for subsequent classification.
- Mo-Hate (Tomar et al., 2023): Mo-Hate is a multimodal fusion approach based on a modified BART architecture. It leverages an attention

mechanism to systematically fuse textual, visual, and acoustic information for hateful content detection, while accounting for semantic, contextual, and temporal dependencies.

- MHClip (Wang et al., 2024b): The MHClip model performs single-modal feature modeling via multiple FC layers, and subsequently concatenates the feature of the three modalities for subsequent classification.

- HCC1 (Koushik et al., 2025): The HCC1 model utilizes HateXplain (Mathew et al., 2021), CLAP (Elizalde et al., 2023), and CLIP (Radford et al., 2021) for textual, audio, and visual feature extraction, respectively, and subsequently employs a simple fusion strategy for classification.

- MoRE (Lang et al., 2025): MoRE utilizes a joint multimodal video retriever to retrieve contextual knowledge, deploys context-enhanced multimodal experts tailored to the evolution of hateful content, and integrates a sample-aware integration network for dynamic modality weighting. This integrated architecture equips MoRE with the ability to adapt to the evolving nature of hateful content across multiple modalities.

LLM-based hateful video detection model

- Qwen-VL (Bai et al., 2023): A multimodal LLM capable of processing both textual and visual inputs. It leverages large-scale pretraining on text-image pairs to generate rich semantic representations, which can be directly used for hateful content detection in videos.

- LLaMA-3.2 vision (Grattafiori et al., 2024): A multimodal extension of LLaMA-3.2 that can process both texts and images. By integrating visual features from video frames with textual information, it is capable of reasoning over multimodal content and detecting hateful videos.

- Keye-VL (Team et al., 2025): A vision-language model specifically designed for short-video understanding. It combines temporal modeling of video frames with visual feature extraction and textual context integration. By jointly reasoning over visual, textual, and temporal cues, Keye-VL can effectively detect subtle and context-dependent hateful content.

- GPT-4 (Achiam et al., 2023): A text-only large language model endowed with strong logical and causal reasoning capabilities, comprehensive cross-domain knowledge, and robust long-range contextual understanding of sequential text.

- GPT-4V: A multimodal extension of GPT-4, capable of processing both text and video frames. By combining textual and visual cues, GPT-4V can analyze the full video content to detect hateful elements, benefiting from its powerful reasoning and cross-modal understanding.

E Implementation Details

We randomly split the HateMM dataset into training, validation, and test sets with a ratio of 7:1:2. For the MHClip dataset, we follow the original data splitting strategy (Wang et al., 2024b) to ensure fair comparison with existing baselines. We adopt the AdamW optimizer with a learning rate of 1e-4, weight decay of 5e-5, and batch size of 16. A linear warmup followed by cosine decay learning rate scheduler is employed to stabilize training and accelerate convergence. To assess statistical significance, we conduct a one-tailed paired t-test based on results obtained from multiple random seeds, comparing SAGE with each baseline model on the test set. The results show highly significant differences ($p < 0.001$), indicating that the model's performance is stable and reliable.

F Prompt Design

To leverage large language and vision-language models for hateful video detection, we design task-specific prompts that guide the models to analyze both visual frames and textual content. This prompt-based adaptation allows the models to transfer their general reasoning and multimodal understanding capabilities to our specific task without task-specific fine-tuning.

For models that take modal information extracted from video as input, the designed prompts are shown in Table 7.

Since Keye-VL directly takes raw video as input while the other adopts extracted video information, we have specifically designed two distinct prompts to align the input format as shown in Table 8.

G Data Processing

Modality Extraction: For each video, we uniformly sample 16 frames to represent the visual modality, ensuring balanced temporal coverage while maintaining computational efficiency. For

Table 7: Example prompts used for Qwen-VL, LLaMA-3.2, GPT-4 and GPT-4V in the hateful video detection task. The prompt includes the textual features (in **blue**), visual features (in **yellow**), the instruction question (in **green**), and the generated outputs example (in **pink**).

Inputs: Title, Description, Transcription, Video frames.

Binary Classification Prompt:

System: You are a video content moderation expert. Given the textual and visual information of the video, your task is to pinpoint any instances of hate speech, discrimination, or other types of content that could be regarded as hateful.

User: Video with title:{**Title**}, description:{**Description**}, transcript:{**Transcript**}; Visual content are shown in images: {**Video frames**}. Combine the video information above, determine whether the video contains hateful content. **If it is hateful, output 'Hateful'; if it is normal, output 'Normal'**.

Output: *Hateful / Normal.*

Three-class Classification Prompt:

System: You are a video content moderation expert. Given the textual and visual information of the video, your task is to pinpoint any instances of hate speech, discrimination, or other types of content that could be regarded as hateful.

User: Video with title:{**Title**}, description:{**Description**}, transcript:{**Transcript**}; Visual content are shown in images: {**Video frames**}. Combine the video information above, determine whether the video contains harmful content. **If it is hateful, output 'Hateful'; if it is offensive, output 'Offensive'; if it is normal, output 'Normal'**.

Output: *Hateful / Offensive / Normal.*

Table 8: Example prompts used for Keye-VL in the hateful video detection task. The prompt includes the video (in **purple**), the instruction question (in **green**), and the generated outputs example (in **pink**).

Inputs: Video.

Binary Classification Prompt:

System: You are a video content moderation expert. Given a raw video, your task is to pinpoint any instances of hate speech, discrimination, or other types of content that could be regarded as hateful.

User: You receive a {**Video**} and need to determine whether the video contains hateful content. **If it is hateful, output 'Hateful'; if it is normal, output 'Normal'**.

Output: *Hateful / Normal.*

Three-class Classification Prompt:

System: You are a video content moderation expert. Given a raw video, your task is to pinpoint any instances of hate speech, discrimination, or other types of content that could be regarded as hateful.

User: You receive a {**Video**} and need to determine whether the video contains hateful content. **If it is hateful, output 'Hateful'; if it is offensive, output 'Offensive'; if it is normal, output 'Normal'**.

Output: *Hateful / Offensive / Normal.*

the audio modality, we extract acoustic feature using the MFCC implementation from the torchaudio library, delivering a lightweight but effective representation of speech characteristics. Additionally, to obtain high-quality transcribed text from the audio, we adopt OpenAI Whisper Turbo, a large-scale automatic speech recognition (ASR) model. This choice achieves a balance between transcription accuracy and computational efficiency, providing reliable multilingual transcriptions to support textual analysis.

Encoder Configuration: For textual feature extraction, we adopt RoBERTa (cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual), with the maximum text length fixed to 128, to obtain $\mathbf{X}_T$ with $d_T = 768$. For the audio modality, the feature dimension is set to 128. We use an FFT window length of 400, a hop length of 160, and 128 Mel filter banks to extract Mel-frequency features, producing audio representations $\mathbf{X}_A$ with $d_A = 128$. For the visual modality, the sampled video frames are first augmented using random flipping, scaling, and cropping to enhance robustness. Feature extraction is then performed using VideoMAE (MCG-NJU/videomae-base), yielding visual representations $\mathbf{X}_I$ with $d_I = 768$.

H Additional Experiments

H.1 Three-class Classification Task

To further evaluate the detection performance of SAGE on hateful videos, we additionally conducted experiments on the three-class classification task using the MHClip dataset with fine-grained annotations. The performance of all models on MHClip-YouTube and MHClip-Bilibili is presented in the Table 9 and Table 10.

The experimental results demonstrate that when extending from binary classification to fine-grained three-class classification task, the performance of all models undergoes a significant decline. This is primarily due to the fact that fine-grained classification requires models to possess more robust information integration and more refined discriminative capabilities. Nevertheless, the relative performance hierarchy of the models in the three-class task remains consistent with the basic conclusions we observed in the binary classification experiments. Notably, some baseline models have almost completely lost their ability to detect *Hateful* content, often misclassifying it as *Offensive*, indicating fun-

Table 9: Experimental results of all baselines and our proposed SAGE on MHClip-YouTube dataset for ternary classification. T: Text, A: Audio, I: Image, V: Video. H: Hateful, O: Offensive. The best results are in **bold** and the second-best are underscored.

Type	Model	Modality				MHClip-YouTube							
		T	A	I	V	**Acc**	**M-F1**	**F1(H)**	**R(H)**	**P(H)**	**F1(O)**	**R(O)**	**P(O)**
Uni-Modal	mBert	✓				0.7063	0.3478	0.0000	0.0000	0.0000	0.2222	0.1471	0.4545
	MFCC		✓			0.7000	0.2916	0.0000	0.0000	0.0000	0.0526	0.0294	0.2500
	ViViT			✓		0.6875	0.4152	0.1111	0.0833	0.1667	0.3214	0.2647	0.4091
	LB				✓	0.7063	0.2935	0.0000	0.0000	0.0000	0.0541	0.0294	0.3333
Multi-Modal	HateMM	✓	✓	✓		0.7063	0.4079	0.0000	0.0000	0.0000	0.3939	0.3824	0.4062
	Mo-Hate	✓	✓	✓		0.7188	0.5420	0.3333	**0.5000**	0.2500	0.4528	0.3529	**0.6316**
	MHClip	✓	✓	✓		0.7125	0.4509	0.1111	0.0833	0.1667	0.4138	0.3529	0.5000
	HCC1	✓	✓	✓		0.7125	0.3531	0.0000	0.0000	0.0000	0.2273	0.1471	0.5000
	MoRE	✓	✓	✓		0.7000	0.4925	0.2069	0.1200	0.7500	0.4155	0.4604	0.5161
MLLM	Qwen-VL 7B	✓		✓		0.5319	0.4189	0.2222	0.2727	0.1875	0.3516	0.5517	0.2581
	LLaMA-3.2V 11B	✓		✓		0.6950	0.4754	0.2353	0.1818	0.3333	0.3492	0.3793	0.3235
	Keye-VL 8B				✓	0.7086	0.4578	0.3158	0.2727	0.3750	0.2326	0.1562	0.4545
	GPT-4	✓				0.6809	0.5018	0.2500	0.2727	0.2308	0.4167	0.5172	0.3488
	GPT-4V	✓		✓		0.7021	0.5581	0.3636	0.3636	0.3636	0.4737	**0.6207**	0.3830
Ours	SAGE	✓	✓	✓		**0.7500**	**0.6127**	**0.4706**	0.3333	**0.8000**	**0.5278**	0.5588	0.5000

damental limitations in their capacity to distinguish the boundary between *Hateful* and *Offensive*. In contrast, the SAGE model still maintains robust overall performance and outperforms all baseline models in terms of Accuracy and Macro-F1 score.

H.2 Hyper-Parameter Analysis

In this section, we will discuss the effects of two key hyperparameters on model performance: the number of iterations L in the global expert deliberation module and the value of K in the *Top-K* expert selection within instance-level expert tribunal and decision module. The experimental results on the HateMM dataset are shown in Table 11.

Since the two modules in SAGE are independent of each other, we evaluate their impact on model performance by varying one hyperparameter while keeping the other fixed. The analysis is divided into two parts: the effect of L and the effect of K.

Effect of L: With K fixed at 2, the results show that when $L = 2$, the model achieves optimal performance while maintaining a moderate parameter size. When $L = 1$, the global expert deliberation module contains only a single layer, resulting in insufficient cross-modal fusion and a subsequent drop in performance. In contrast, when $L = 3$, multiple global expert deliberation layers may induce modality interference and noise accumulation, which amplify noise across all modalities, impairing modal discriminative ability and diminishing generalization performance.

Effect of K: With L fixed at 2, the model attains optimal performance when $K = 2$. Setting $K = 1$ (i.e., only a single expert is selected) results in inadequate decision-level fusion and may lead to the loss of complementary cues from other neutral or potential hateful modalities. Although $K = 3$ theoretically enables the utilization of more modal information, in practice, noise accumulation and cross-modal interference may elevate misclassification rates and degrade overall performance.

Overall, a moderate number of global expert deliberation layer and an adaptive expert selection strategy achieve an effective balance between information utilization and noise robustness.

H.3 Error Analysis

Although SAGE demonstrates strong performance in binary classification, it still faces challenges in fine-grained three-class classification, particularly in distinguishing between *Offensive* and *Hateful* content. Errors mainly arise from unclear boundaries between *Offensive* and *Hateful*. In addition, the model sometimes fails to effectively correct biases in expert predictions.

Case A is essentially normal content objectively discussing a sensitive topic. However, the model incorrectly classifies it as offensive because the visual expert dominates the decision with a weight of 0.57, assigning the highest offensive probability (0.51). In contrast, the audio modality is discarded, and the textual expert's hateful cues are underuti-

Table 10: Experimental results of all baselines and our proposed SAGE on MHClip-BiliBili dataset for ternary classification. T: Text, A: Audio, I: Image, V: Video. H: Hateful, O: Offensive. The best results are in **bold** and the second-best are underscored.

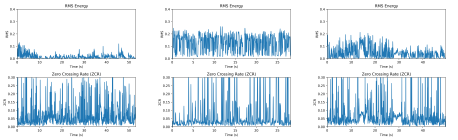
Type	Model	Modality				MHClip-BiliBili							
		T	A	I	V	**Acc**	**M-F1**	**F1(H)**	**R(H)**	**P(H)**	**F1(O)**	**R(O)**	**P(O)**
Uni-Modal	mBert	✓				0.6852	0.3982	0.0952	0.0556	0.3333	0.2917	0.2188	0.4375
	MFCC		✓			0.6605	0.2822	0.0000	0.0000	0.0000	0.0526	0.0312	0.1667
	ViViT			✓		0.6914	0.3416	0.1000	0.0556	**0.5000**	0.1053	0.0625	0.3333
	LB				✓	0.6852	0.5030	0.4444	0.3333	0.6667	0.2545	0.2188	0.3043
Multi-Modal	HateMM	✓	✓	✓		0.6975	0.4288	0.1000	0.0556	**0.5000**	0.3704	0.3125	0.4545
	Mo-Hate	✓	✓	✓		0.7160	0.4725	0.1730	0.1111	0.4000	0.3934	0.3750	0.4138
	MHClip	✓	✓	✓		0.7037	0.4057	0.1000	0.0556	**0.5000**	0.2857	0.2188	0.4118
	HCC1	✓	✓	✓		0.6667	0.4039	0.2400	0.1667	0.4286	0.1702	0.1250	0.2667
	MoRE	✓	✓	✓		0.7076	0.5620	0.4117	0.3589	0.4827	0.4354	0.4500	0.4218
MLLM	Qwen-VL 7B	✓		✓		0.5062	0.4260	0.2927	0.3333	0.2609	0.3333	0.5312	0.2429
	LLaMA-3.2V 11B	✓		✓		0.6111	0.4734	0.2667	0.2222	0.3333	0.3913	0.5625	0.3000
	Keye-VL 8B				✓	0.6975	0.3434	0.0952	0.0556	0.3333	0.1111	0.0625	**0.5000**
	GPT-4	✓				0.5988	0.4951	0.3077	0.2222	**0.5000**	0.4505	**0.7812**	0.3165
	GPT-4V	✓		✓		0.6481	**0.6019**	**0.5882**	**0.5556**	0.6250	0.4554	0.7188	0.3333
Ours	SAGE	✓	✓	✓		**0.7469**	0.5676	0.3077	0.2222	**0.5000**	**0.5278**	0.5938	0.4750

Table 11: The impact of the global expert diliberation layer N and the selected experts counts K in the expert gating mechanism on model performance on the HateMM dataset. Para: Parameters without encoders. The best results are highlighted in **bold**.

Fixed Variable	Variable	Value	Acc	M-F1	M-P	M-R	Para
Selected Experts Count K=2	Iteration Number N	N = 1	0.8433	0.8356	0.8371	0.8343	16M
		N = 2	**0.8710**	**0.8628**	**0.8711**	**0.8572**	28M
		N = 3	0.8571	0.8531	0.8497	0.8597	41M
Iteration Number N=2	Selected Experts Count K	K = 1	0.8479	0.8371	0.8492	0.8301	28M
		K = 2	**0.8710**	**0.8628**	**0.8711**	**0.8572**	28M
		K = 3	0.8618	0.8537	0.8593	0.8495	28M

lized. As a result, the model confuses objective references to sensitive vocabulary with genuine offensive intent. Case B corresponds to offensive content containing mixed and relatively weak hate signals, where the intent is closer to provoking public discussion rather than promoting targeted hatred. Nevertheless, the model misclassifies it as hateful because text contains hateful topic 'racism'. Equal weights amplify the textual expert's slight bias toward the hateful category and the visual expert's ambiguous judgment, while the audio modality, which could have provided useful discriminative cues, is discarded. Case C is intrinsically hateful with explicit racial hate speech. All three experts lean toward offensive, which results in the final prediction being offensive regardless of how the weights are allocated. Consequently, fused biased expert judgments made the model overlook the core hateful nature and misclassify it as Offensive.

Table 12: Examples of misclassifications by the SAGE on MHClip-YouTube dataset. Each expert's decision in the three-class setting is shown in the format $m : N/O/H$, where $m \in \{T, A, I\}$. Component: Modality-level harmfulness, and GT: Ground Truth.

Hatefulness	(A) **GT: Normal**	(B) **GT: Offensive**	(C) **GT: Hateful**
Vision			
Audio			
Text	accused of sexual misconduct...that cruelty and misogyny...and it was horrifying...,	do wonyong have princess syndrome? lots of fans are accusing her of racism and choleism ...	all white people are racist,period point blank end of story. you can't be white and not racist..
Component	Text, Audio	Text, Audio, Image	Text, Audio
Expert Diagnosis	**I**: 0.15 / 0.51 / 0.34 **A**: 0.24 / 0.36 / 0.40 **T**: 0.17 / 0.35 / 0.48	**I**: 0.25 / 0.38 / 0.37 **A**: 0.25 / 0.36 / 0.39 **T**: 0.29 / 0.32 / 0.39	**I**: 0.21 / 0.65 / 0.14 **A**: 0.16 / 0.43 / 0.41 **T**: 0.21 / 0.57 / 0.22
Tribunal Weights	0.57 / 0.00 / 0.24	0.37 / 0.00 / 0.37	0.00 / 0.37 / 0.32
Output	**Offensive** ×	**Hateful** ×	**Offensive** ×