HVGuard: Utilizing Multimodal Large Language Models for Hateful Video Detection

Yiheng Jing[1*], Mingming Zhang[1*], Yong Zhuang[1†], Jiacheng Guo[1], Juan Wang[1†], Xiaoyang Xu[1], Wenzhe Yi[1], Keyan Guo[2], Hongxin Hu[2]
[1]Key Laboratory of Aerospace Information Security and Trusted Computing,
Ministry of Education School of Cyber Science and Engineering, Wuhan University
[2]University at Buffalo

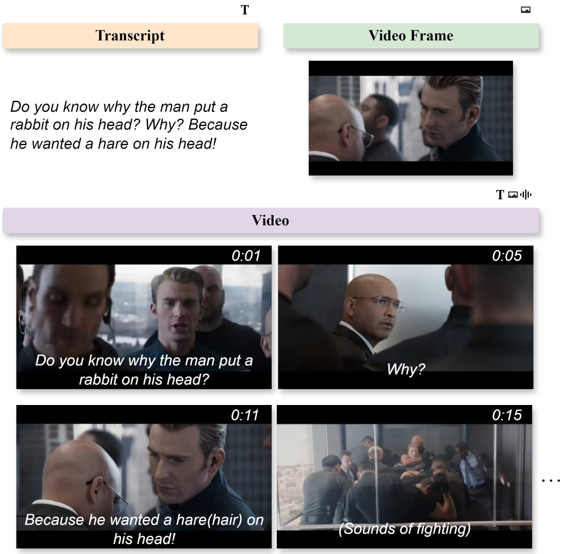
Abstract

The rapid growth of video platforms has transformed information dissemination and led to an explosion of multimedia content. However, this widespread reach also introduces risks, as some users exploit these platforms to spread hate speech, which is often concealed through complex rhetoric, making hateful video detection a critical challenge. Existing detection methods rely heavily on unimodal analysis or simple feature fusion, struggling to capture cross-modal interactions and reason through implicit hate in sarcasm and metaphor. To address these limitations, we propose HVGUARD, the first reasoning-based hateful video detection framework with multimodal large language models (MLLMs). Our approach integrates Chain-of-Thought (CoT) reasoning to enhance multimodal interaction modeling and implicit hate interpretation. Additionally, we design a Mixture-of-Experts (MoE) network for efficient multimodal fusion and final decision-making. The framework is modular and extensible, allowing flexible integration of different MLLMs and encoders. Experimental results demonstrate that HVGUARD outperforms all existing advanced detection tools, achieving an improvement of 6.88% to 13.13% in accuracy and 9.21% to 34.37% in M-F1 on two public datasets covering both English and Chinese.

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

1 Introduction

Video platforms like YouTube (Google, 2005), Bilibili (Kuanyu, 2009), and TikTok (ByteDance, 2016) have transformed information dissemination and fueled multimedia growth. However, this also brings risks, as some users exploit these platforms to spread harmful content like hate speech (Ottoni et al., 2018). Hate speech refers to language that targets individuals or groups on the basis of characteristics such as race, religion, or gender (Hee et al., 2024b; Fortuna and Nunes, 2018), thereby fueling social tension and posing tangible risks to both individuals and communities. Thus, effectively detecting hate speech on video platforms (Alcântara et al., 2020; Das et al., 2023; Wu and Bhandary, 2020) has become an urgent challenge.



Do you know why the man put a rabbit on his head? Why? Because he wanted a hare on his head!

Do you know why the man put a rabbit on his head?

Why?

Because he wanted a hare(hair) on his head!

(Sounds of fighting)

Figure 1: A typical example of hateful video. The offender mocks a bald individual using a pun that plays on the phonetic similarity between "hair" and "hare".

The multimodal nature of video, which combines visual, auditory, and textual elements, enables subtle and indirect expressions of hate. As shown in Figure 1, implicit hate speech is often difficult to detect, as it relies heavily on cross-modal cues and contextual understanding. Content that appears harmless within a single modality may reveal its offensive intent only when multiple modalities are analyzed together. Current hateful video de-

[*]Equal contribution. Emails: {yihengjing, mingmingzhang, yong.zhuang, jiachengg, jwang, xiaoyangx, wenzhey}@whu.edu.cn; {keyanguo, hongxinh}@buffalo.edu
[†]Corresponding authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8993–9006
November 4-9, 2025 ©2025 Association for Computational Linguistics

tection methods typically encode modalities separately or concatenate their features superficially (Yu et al., 2022; Wu and Bhandary, 2020; Wang et al., 2024a; Das et al., 2023), limiting their ability to capture nuanced interactions. Additionally, rhetorical devices such as metaphors, irony, and sarcasm frequently occur in hateful videos (Xu et al., 2024; Ge et al., 2023), further complicating detection tasks. Given the rapid proliferation of culturally contextualized hateful videos online (Ottoni et al., 2018), effective detection requires integrating multimodal interactions with advanced reasoning capabilities and external knowledge, underscoring the practical significance of improving hateful content moderation.

Recent multimodal large language models (MLLMs) (Bai et al., 2023; Team et al., 2024; Liu et al., 2024; Wang et al., 2024b) show strong potential for hateful video detection, given their deep semantic comprehension and rich world knowledge (Tang et al., 2025). To fully exploit their capabilities, we incorporate Chain-of-Thought (CoT) reasoning, guiding MLLMs to systematically analyze interactions across visual, auditory, and textual modalities. In this work, we first explore how effectively MLLMs with CoT reasoning handle multimodal interactions and rhetorical devices, such as metaphors, in hateful videos. Motivated by these insights, we propose HVGUARD[1], the first reasoning-based hateful video detection framework. HVGUARD leverages MLLMs to generate multimodal rationales via CoT reasoning, explicitly modeling cross-modal and rhetorical elements. Furthermore, we introduce a Mixture-of-Experts (MoE) network (Jacobs et al., 1991) that integrates low-level multimodal features with high-level semantic rationales for robust detection. Extensive experiments demonstrate that HVGUARD achieves up to 0.86 accuracy, significantly outperforming existing state-of-the-art methods.

The key contributions of this paper are as follows:

- **First Exploration of MLLMs and CoT in Hateful Video Understanding.** This is the first work to explore the potential of MLLMs and CoT reasoning for hateful video understanding, demonstrating their effectiveness in managing multimodal interactions and complex rhetorical devices, such as metaphors.
- **Novel Reasoning-Based Hateful Video Detection Framework.** We propose the first reasoning-based hateful video detection framework, integrating MLLMs with CoT reasoning to enhance multimodal interaction modeling and implicit hate interpretation. Additionally, we introduce a MoE network to efficiently fuse multimodal representations and MLLM-generated rationales, optimizing the decision-making process.
- **Extensive Evaluation of HVGUARD.** Experimental results show that HVGUARD achieves up to 0.86 accuracy, outperforming all existing detection tools with accuracy gains of 6.88% to 13.13% and M-F1 improvements of 9.21% to 34.37%. Extensive experiments on two public datasets, covering both English and Chinese, further validate its effectiveness in binary and ternary classification settings against five state-of-the-art baselines, including advanced MLLMs and existing detection tools.

2 Related Work

2.1 Hate Speech Detection

Modern hate speech detection systems can be categorized into unimodal and multimodal approaches based on data types. Unimodal detection is further divided into three primary modalities:

Text-based detection primarily addresses binary classification tasks, with advanced frameworks extending to ternary classification (hate speech, offensive speech, and normal speech). Foundational work by (Davidson et al., 2017) and (Founta et al., 2018) established robust text classification baselines, while recent studies have enhanced detection by analyzing contextual discourse (Yu et al., 2022) and decoding black humor nuances (Hee et al., 2024a).

Image-based detection focuses on visual hate expression, particularly in meme culture. Researchers have developed specialized datasets (Gasparini et al., 2022; Bhandari et al., 2023) and advanced methods like Pro-Cap (Cao et al., 2023) for implicit hate detection, with architectures such as MR.HARM (Lin et al., 2023) addressing multimodal hate meme analysis.

Audio-based detection employs CNNs to process spectral features, where studies like (Medina et al., 2022) and (Yousefi and Emmanouilidou, 2021) have advanced feature extraction techniques for improved acoustic hate speech identification.

[1] https://github.com/yihengjingWHU/HVGuard

Multimodal detection synergistically combines text, visual, and auditory cues, proving particularly effective for video analysis. Contemporary works (Das et al., 2023; Wang et al., 2024a) demonstrate superior performance through cross-modal fusion, though most approaches simply concatenate modality features. Our work advances this paradigm by modeling deep inter-modal interactions to capture the complex semantics of hate speech videos.

2.2 Multimodal Large Language Models (MLLMs)

The emergence of large language models (LLMs) has led to significant advances in natural language processing, enabling models like Gemini (Team et al., 2024) to handle multimodal inputs, such as images and text. While LLMs excel at reasoning and world knowledge, they lack the ability to "see" images, making them less effective at understanding multimodal data. Conversely, large visual models (VLMs) excel in image recognition but are limited in reasoning and world knowledge (Kirillov et al., 2023; Shen et al., 2024). The combination of LLMs and VLMs in MLLMs allows for more robust multimodal understanding, making them highly effective in tasks like image reasoning and video understanding (Wu et al., 2023). In our research, we leverage MLLMs to analyze and understand the complex interaction patterns in hate speech videos, providing valuable insights for reasoning models.

This integrated approach allows for more nuanced detection by simultaneously considering verbal content, visual context, and auditory cues, while explicitly modeling their synergistic relationships - a critical advancement for understanding sophisticated hate speech in multimedia environments.To further demonstrate the importance of multimodal information and MLLM rationale in hateful video understanding, we conducted preliminary study in Appendix A.

3 Method

3.1 Task Definition

The goal of hateful video detection is to extract features from videos and classify them based on these features. The video dataset is represented as $\mathcal{V} = \{v_1, \ldots, v_i, \ldots, v_{|\mathcal{V}|}\}$, where $|\mathcal{V}|$ is the number of videos. The task can be expressed as:

$$\arg \max_{c \in \{1,2,\ldots,|C|\}} P(c|v_i) \tag{1}$$

where $c \in \{1, 2, \ldots, |C|\}$ represents the classification categories. Our work focuses on utilizing rationale generated by MLLM and multimodal information from the video itself for detection. Therefore, this task can be re-expressed as:

$$\arg \max_{c \in \{1,2,\ldots,|C|\}} P(c|v_i^T, v_i^A, v_i^F, v_i^M) \tag{2}$$

where v_i^T represents the text information in the video (such as title, subtitles, or transcript), v_i^A represents the audio information of the video, v_i^F represents the frame information of the video, and v_i^M represents MLLM-derived rationales.

3.2 Overview

The overview of our framework, HVGuard, is shown in Figure 2. Based on preliminary study, we design this novel framework for hateful video detection, leveraging MLLM-derived rationales to address challenges in multimodal interaction and the interpretation of metaphors and rhetorical devices. This framework extracts text, audio, and video frames from the input video, providing a comprehensive semantic representation of the video. A CoT-based reasoning approach is then applied, progressively reasoning through the individual modalities and their interactions, to generate rationale from MLLM. In the final stage, these embeddings are ultimately integrated using a MoE network to yield the final classification results.

3.3 Multimodal Extraction Module

Considering that hateful videos encompass multiple modalities, we first extract features from the three main modalities: text, audio, and video frames.

For the audio signal v_i^A, we process it as a combination of semantic and emotional information. Specifically, we use FunASR (Gao et al., 2023), an open-source audio processing tool, to transcribe the spoken content into text v_i^{trans} and extract the corresponding emotional cues v_i^{emo}.

For the visual modality, we uniformly sample 32 frames per video at fixed intervals, following prior works such as ViViT (Arnab et al., 2021), VideoChat (Li et al., 2023), and Video-LLaVA (Lin et al., 2024), which demonstrate strong performance with this setting. Although using more frames may slightly improve accuracy, it significantly increases computational cost with limited gains. In our case, 32 frames offer an effective trade-off between efficiency and performance, as

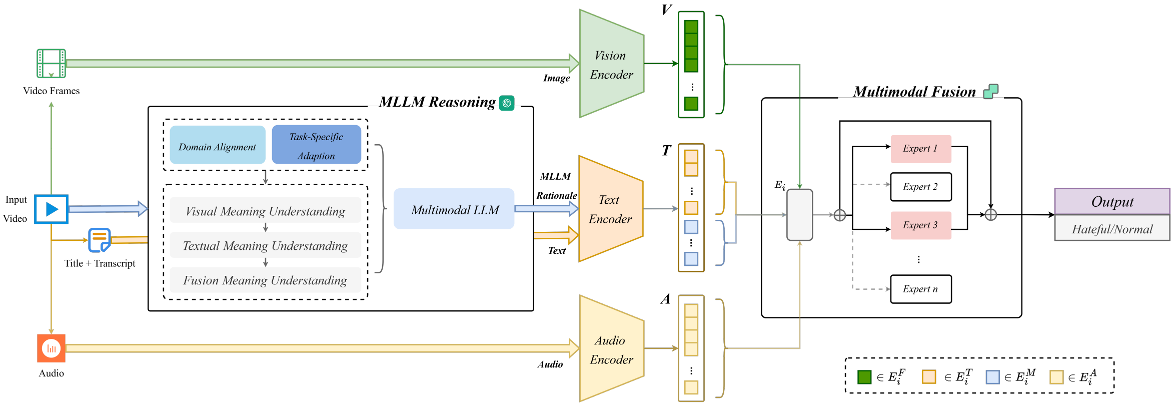


Figure 2: Overview of the proposed framework, consisting of three modules: Multimodal Extraction, MLLM Reasoning, and Multimodal Fusion. The original video is processed by Multimodal Extraction to extract information from different modalities. The MLLM then interprets this information and generates rationales, after which a Mixture-of-Experts (MoE) network is employed to fuse and process the multimodal data.

also validated by our baseline Multihateclip (Wang et al., 2024a).

$$\begin{aligned} v_i^A, v_i^F &= extract(v_i), \\ v_i^{trans}, v_i^{emo} &= trans(v_i^A) \end{aligned} \tag{3}$$

where v_i^A represents the original audio signal, v_i^F represents the video frames, and v_i^{title} represents the video title.

Next, we construct the textual content v_i^T using the video title and transcript:

$$v_i^T = \{v_i^{title}, v_i^{trans}\} \tag{4}$$

3.4 MLLM Reasoning Module

To address the challenges in hateful video detection, such as metaphors, cultural contexts, and the complexity of multimodal interactions, it is necessary to leverage MLLMs to extract deep semantic information from the video. Based on preliminary study (Appendix A), we find that hateful video detection is a complex process, requiring the extraction of key cues from multiple modalities, including text, visuals, and audio. Inspired by the works of (Xu et al., 2024; Vishwamitra et al., 2024), we employ carefully designed CoT prompts to decompose this complex task, thereby enabling the understanding of multimodal hateful content within the video. Specifically, our CoT prompt is as follows:

Adaption Prompt. In the field of hateful content detection, domain alignment, role description and task-specific adaptation is critical, as it equips MLLMs with essential cultural context and contextual comprehension. This focuses the model's capabilities on addressing the specific challenges of understanding both nuanced and overt hateful content, thereby improves its performance and reliability (Csurka, 2017; Qi et al., 2024). We employ the prompt:

> *This is a video that may contain harmful content, such as hate speech, explicit violence, discrimination, or other forms of harmful behavior. You are a content moderation specialist. Your task is to identify any instances of hate speech, violent imagery, discriminatory actions, or any other content that could be considered harmful, abusive, or offensive. Ensure the answer's accuracy while keeping it concise and avoiding over-explanation.*

Visual Meaning Understanding. To guide the model to analyze the video progressively, starting with the visual information while ignoring the subtitles in the video frames. The focus is placed on analyzing the characters and scenes in the frames. To achieve this, we employ the following prompt:

> *Describe the video content based on {video frames}, ignoring subtitles in the frames. Pay attention to any special characters or scenes.*

Given the video frames v_i^F and this prompt X_{prompt}^F, the output computation is as follows:

$$res1 = MLLM(v_i^F, X_{prompt}^F) \tag{5}$$

Textual Meaning Understanding. We guide the model to focus on textual information by analyzing the video titles and transcripts, paying special attention to the presence of rhetorical devices such as puns and homophonic wordplay used as promotional strategies. Based on this, we employ the following prompt:

> *The video title is {video title}. The text in the video is {video transcript}. Please analyze the meaning of this text. Note that there may be homophonic memes and puns; distinguish and explain them.*

Given the textual input v_i^T and the prompt X_{prompt}^T , the output computation is as follows:

$$res2 = MLLM(v_i^T, X_{prompt}^T) \tag{6}$$

Fusion Meaning Understanding. Given the complex relationships between semantics across different modalities, it is essential to comprehensively consider the meaning conveyed by the video after multimodal fusion. As illustrated by figure 1, some videos may contain no obvious offensive content in their text or visuals individually, yet their combination can give rise to new meanings. Therefore, we aim for the model to synthesize the results from the first two steps and further integrate the video's raw information, including video frames, text, and extracted emotions of spoken content. This approach seeks to uncover deeper cross-modal interactions and analyze potential new metaphors. We employ the following prompt:

> *Please combine the {video title}, {video transcript}, {video frames}, {voice emotion}, {response1}, {response2} and analyze both the visual, textual and audio elements of the video to detect and flag any hateful content. No need to describe the content of the video, only answer implicit meanings and whether this video expresses hateful content further.*

The MLLM rationale is as follows:

$$v_i^M = MLLM(v_i^T, v_i^F, v_i^{emo}, res1, res2) \tag{7}$$

3.5 Multimodal Fusion Module

After obtaining rationale generated by MLLM reasoning module, we designed a multimodal fusion module to fuse information from the aforementioned modalities. We employ modality-specific encoders for each type of modality to obtain their respective embedding representations:

$$\begin{aligned} E_i^T &= f_T(v_i^T), \\ E_i^A &= f_A(v_i^A), \\ E_i^F &= f_F(v_i^F) \end{aligned} \tag{8}$$

where f_T, f_A, and f_F represent the text, audio, and vision modality encoders, while E_i^T, E_i^A, and E_i^F represent corresponding embeddings. To reduce the inference burden, we designed an embedding cache, allowing the above process to be executed only once on the dataset.

The rationale v_i^M generated by the MLLM is presented in textual form. We treat it as additional textual input and feed it into the text modality encoder to obtain embeddings:

$$E_i^M = f_T(v_i^M) \tag{9}$$

To fuse the embeddings from different modalities, we designed a mixture of experts network. First, all embeddings are concatenated into a single long vector as the representation embedding E_i for the entire video:

$$E_i = concat(E_i^T, E_i^A, E_i^F, E_i^M) \tag{10}$$

Next, we constructed n identical expert networks and one gating network, where n is the number of experts. These experts and the gating network share the same input E_i. Each expert network extracts high-level information specific to certain feature. The output of the k-th expert is denoted as O_k and is computed as follows:

$$O_k = f_k(E_i; \theta_k), \quad k \in \{1, 2, \dots, n\} \tag{11}$$

where f_k represents the mapping function of the k-th expert network, and θ_k denotes its parameters.

Simultaneously, the gating network $g(E_i; \phi)$ dynamically generates weights w_k to adjust the contribution of each expert's output. To prevent weight polarization, dropout is applied to the gating network's output weights. The gating network computes these weights as:

$$\begin{aligned} w_k = \text{Dropout}\left(\frac{\exp(g_k(E_i; \phi))}{\sum_{j=1}^{n} \exp(g_j(E_i; \phi))} \right), \\ k \in \{1, 2, \dots, n\} \end{aligned} \tag{12}$$

where $g_k(E_i; \phi)$ is the unnormalized weight produced by the gating network, and ϕ represents the parameters of the gating network.

The final fused output O_{fusion} is obtained by combining the weighted outputs of all experts:

$$O_{fusion} = \sum_{k=1}^{n} w_k \cdot O_k \tag{13}$$

3.6 Final Decision

During training, we optimize the parameters of the expert and gating networks by minimizing a loss function. Assuming the ground truth labels are y and the final decision outputs are $\hat{y}$, we use a cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{14}$$

where m denotes the number of samples.

4 Experiments

4.1 Dataset

In our study, we employ two high-quality, up-to-date public datasets for hateful video detection:

HateMM(Das et al., 2023). The HateMM dataset consists of 1,083 videos sourced from BitChute, a platform with lenient content moderation, resulting in a higher prevalence of hateful content. Videos are labeled as either Hate or Non-Hate.

MultiHateClip(Wang et al., 2024a). The MultiHateClip dataset is a multilingual benchmark dataset for hateful video detection, including 2,000 videos from YouTube and Bilibili, with 1,000 videos in English and 1,000 in Chinese. Each video is classified as Hateful, Offensive, or Normal.

Dataset	Language	Total	H	O	N
HateMM	English	1,066	427	0	639
Multihateclip	English	891	72	218	601
	Chinese	897	112	180	605

Table 1: Overview of datasets. H:hateful, O:offensive, N:normal

To enhance data reliability, we filtered out corrupted and blurry videos. Additionally, to ensure high-quality textual information, we re-annotated the video transcripts using the speech transcription tool FunASR (Gao et al., 2023), improving the accuracy of multimodal analysis. The dataset we use is summarized in Table 1.

4.2 Experiment Settings

We randomly split all datasets into training, testing, and validation sets with a 7:2:1 ratio. For the ternary classification task on the MultiHateClip dataset, the labels used are Hateful, Offensive, and Normal. For binary classification on both the MultiHateClip and HateMM datasets, we combine Hateful and Offensive into a single category, keeping the Normal label unchanged.

All models are trained with a learning rate of 1e-4, a batch size of 32, and early stopping after 100 epochs. Experiments are conducted on three Tesla V100-32G GPUs. Model performance is primarily evaluated using macro-averaged F1 score (M-F1) and accuracy (acc). We employ GPT-4o(Achiam et al., 2023), XLM(Conneau et al., 2020), Vit(Dosovitskiy, 2020), and Wav2Vec(Baevski et al., 2020) as the fundamental MLLM and modality encoders.

4.3 Baseline Models

We evaluate HVGuard with five baselines, including three advanced MLLMs and two state-of-the-art methods in hateful video detection: (1) **GPT-4o** (Achiam et al., 2023): An advanced MLLM by OpenAI, with high-level reasoning capabilities. (2) **Gemini-1.5-pro** (Team et al., 2024): A sophisticated multimodal model by Google DeepMind, capable of handling diverse reasoning tasks and understanding multiple modalities, including audio, images, videos, and text. (3) **Qwen-VL-7B** (Bai et al., 2023): An open-source vision-language model by Alibaba Cloud, excelling in tasks like image captioning, question answering, and visual localization. (4) **HateMM** (Das et al., 2023): A multimodal hateful video detection model that combines text, audio, and visual pretrained models through a trainable fusion layer to make final predictions. (5) **MultiHateClip** (Wang et al., 2024a): A model that processes each modality's features through independent fully connected layers, concatenates them, and performs final classification to determine whether the video contains hate speech.

For the MLLMs used, we employ a generalized prompt to detect hateful videos: "*Analyze whether the video contains hateful content.*" To ensure test consistency, we reproduced all the baselines and conducted a unified evaluation.

4.4 Evaluation Results

To evaluate the effectiveness of HVGuard, we report results in Table 2. The Multihateclip dataset, containing both English and Chinese videos, is used to test cross-lingual generalization. We consider both binary and ternary classification to reflect

Dataset	Number of categories	Model	Acc	M-F1	F1(H)	R(H)	P(H)	F1(O)	R(O)	P(O)
Multihateclip(English)	3	GPT-4o	0.7326	0.3280	0.2957	0.2361	0.3953	0.4923	0.4486	0.5455
		Gemini-1.5-pro	0.6319	0.4458	0.2143	0.2000	0.2308	0.3409	0.3488	0.3333
		Qwen-VL	0.5618	0.4060	0.2051	**0.6154**	0.1231	0.2258	0.1556	0.4118
		HateMM	0.6966	0.4894	0.1333	0.1667	0.1111	0.5217	0.5516	0.5345
		Multihateclip	0.7079	0.4946	0.1667	0.1667	0.1667	0.4928	0.5780	0.4750
		HVGuard	**0.8090**	**0.6646**	**0.4556**	0.4722	**0.5000**	**0.6488**	**0.6270**	**0.6994**
	2	GPT-4o	0.7989	0.5019	/	/	/	0.6455	0.5699	0.7443
		Gemini-1.5-pro	0.7198	0.6020	/	/	/	0.3855	0.2759	0.6400
		Qwen-VL	0.6573	0.6549	/	/	/	0.6258	**0.9273**	0.4722
		HateMM	0.7191	0.6646	/	/	/	0.5421	0.4722	0.6548
		Multihateclip	0.7416	0.6806	/	/	/	0.5544	0.4861	0.7269
		HVGuard	**0.8539**	**0.7714**	/	/	/	**0.6308**	0.5819	**0.7619**
Multihateclip(Chinese)	3	GPT-4o	0.6444	0.4460	0.2326	0.1852	0.3125	0.2941	0.3448	0.2564
		Gemini-1.5-pro	0.6648	0.4393	0.2069	0.1500	0.3333	0.2985	0.2703	0.3333
		Qwen-VL	0.5719	0.4472	0.3333	**0.6875**	0.2200	0.2491	0.1889	0.3656
		HateMM	0.6889	0.4163	0.0741	0.0476	0.1667	0.3667	0.3889	0.4722
		Multihateclip	0.7111	0.4573	0.1667	0.1111	0.3333	0.3778	0.3889	0.4167
		HVGuard	**0.8045**	**0.5643**	**0.3563**	0.2917	**0.5278**	**0.4417**	**0.4190**	**0.6139**
	2	GPT-4o	0.7389	0.6900	/	/	/	0.5766	0.5714	0.5818
		Gemini-1.5-pro	0.7443	0.6188	/	/	/	0.4000	0.2632	0.8333
		Qwen-VL	0.6704	0.6684	/	/	/	0.6424	**0.9298**	0.4907
		HateMM	0.7444	0.6908	/	/	/	0.5694	0.5694	0.5826
		Multihateclip	0.7778	0.6904	/	/	/	0.5299	0.4028	0.7833
		HVGuard	**0.8603**	**0.8219**	/	/	/	**0.7408**	0.6905	**0.8274**
HateMM	2	GPT-4o	0.7308	0.7306	0.7238	0.8806	0.6144	/	/	/
		Gemini-1.5-pro	0.7874	0.7872	0.7933	0.8554	0.7396	/	/	/
		Qwen-VL	0.7089	0.7089	0.7075	**0.8824**	0.5906	/	/	/
		HateMM	0.7500	0.7454	0.7430	0.7259	0.7614	/	/	/
		Multihateclip	0.7614	0.7594	0.7611	0.7537	0.7690	/	/	/
		HVGuard	**0.8563**	**0.8597**	**0.8479**	0.8228	**0.8809**	/	/	/

Table 2: Results of different methods on the task of hateful video detection. H:hateful, O:offensive, Acc:accuracy, M-F1:macroF1, R:recall, P:precision.

different moderation needs: binary classification supports rapid filtering, while ternary classification enables finer-grained control by introducing an "Offensive" category.

Overall, HVGuard outperformed all other baselines, with an improvement of 6.88% to 13.13% in accuracy and 9.21% to 34.37% in M-F1 compared to existing SOTA detection tools. We then explored further conclusions through the following analysis.

HVGuard achieved SOTA performance on both English and Chinese hateful video datasets, demonstrating its multilingual adaptability. Additionally, it outperformed other baselines in both ternary and binary classification tasks.

We also achieved superior performance on most metrics for crucial labels of "Hateful" and "Offensive," demonstrating the HVGuard ability for hateful video detection. Notably, Qwen-VL achieved the highest recall rate for "Hate" category, but performed poorly in accuracy and M-F1. This suggests that Qwen-VL tends to classify videos as "Hate", leading to the misclassification of some normal videos. In practical applications, an excessively high false positive rate may negatively impact normal information flow within online communities.

To more clearly demonstrate the effectiveness of the proposed framework, we present a case study in Appendix B. Moreover, our framework achieves a very low false positive rate, additional analysis further indicates that the few remaining misclassifications are often associated with sensitive terms or identity-related topics (see Appendix C).

4.5 Effectiveness of Components in HVGuard

Model	Ternary		Binary	
	Acc	M-F1	Acc	M-F1
w/o Vision encoder	0.7865	0.4760	0.8202	0.7397
w/o Text encoder	0.7753	0.5633	0.8258	0.7090
w/o Audio encoder	0.7697	0.5807	0.8258	0.7413
w/o Modal features	0.7584	0.4816	0.8146	0.7126
w/o CoT	0.7416	0.4715	0.7921	0.5512
MoE→MLP	0.7809	0.5936	0.8371	0.7466
MoE→Cross attention	0.8034	0.6525	0.8427	**0.8037**
HVGuard	**0.8090**	**0.6646**	**0.8539**	0.7714

Table 3: Ablation study for the components in HVGuard.

Table 3 summarizes the results of the ablation study on the MultiHateClip(English) dataset using HVGuard. Removing the visual, text, or audio components individually resulted in performance declines, indicating that each modality plays a crucial role in hate detection. Furthermore, ablation of all modal features, relying solely on MLLM

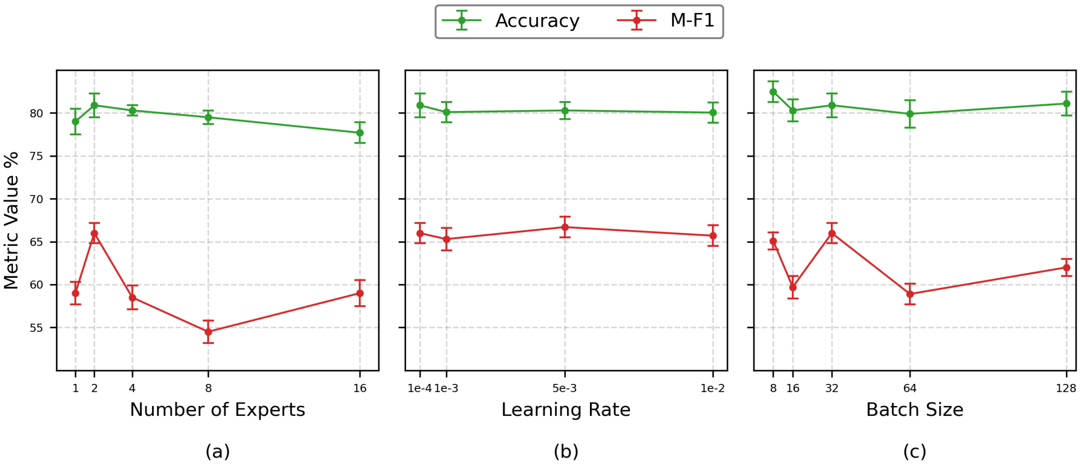


Figure 3: (a) Number of experts hyper-parameter study. (b) Learning rate hyper-parameter study. (c) Batch size hyper-parameter study.

rationale—led to a noticeable decrease in performance. These findings underscore the importance of integrating comprehensive multimodal information for accurate detection.

Moreover, removing the CoT guidance for the MLLM and relying solely on generalized prompt templates resulted in a significant performance drop. This demonstrates that the CoT approach generates more informative supplementary features, enabling the multimodal fusion module to make more accurate predictions.

Furthermore, replacing the MoE in the model with a standard MLP or a cross attention layer also led to a performance decline. This indicates that MoE is crucial for the multimodal tasks in this context. MoE leverages information from different modalities, along with the rationale provided by the MLLM, to enhance hateful video detection. More details can be found in Appendix D.

In addition, we conducted comprehensive experiments on different combinations of MLLMs, Text encoders, Vision encoders, and Audio encoders, demonstrating the deployment flexibility of HVGuard. Details are shown in Appendix E.

4.6 Hyper-parameter Study

To investigate the effects of the hyper-parameters in HVGuard, we show the impact of hyper-parameters on the performance trend.

Figure 3 illustrates the impact of varying numbers of experts, learning rate and batch size on the performance through a line chart, showing that the model achieves optimal performance when the number of experts is 8, and the learning rate and batch size have little to no impact on the performance. Despite experimenting with different values for these hyperparameters, the model's performance remained relatively stable across the variations, indicating that the performance is primarily influenced by the number of experts rather than the learning rate or batch size.

5 Conclusion

In this work, we propose a hateful video detection framework named HVGuard, which is the first reasoning-based hateful video detection framework with MLLMs. This framework carefully designs a CoT reasoning strategy to fully leverage the reasoning ability of MLLMs and introduces a MoE network for the efficient utilization of rationale and multimodal features. Experiments demonstrate that the proposed framework achieves SOTA performance on two public datasets, containing both English and Chinese videos. In the future, we aim to improve the framework by incorporating larger, more diverse, and multilingual datasets to enhance its performance and adaptability across different contexts and languages. This expansion will help address the complexities of detecting hateful content in a broader range of scenarios.

Limitations

We only evaluated HVGuard on the Chinese and English datasets and did not evaluate other languages. This limits our further exploration of the language generalizability of the framework.

Moreover, we believe that fine-grained detection of hateful videos is of great importance. Although we have considered both binary and ternary classification scenarios, more refined categorization may be more beneficial for the application of such research in real-world contexts.

Ethical Considerations

Our work presents HVGUARD, a framework for hateful video detection, with the goal of enhancing online safety by mitigating the spread of hate speech. While HVGUARD demonstrates effectiveness, automated moderation inevitably involves trade-offs, including the possibility of false positives that may affect benign content and false negatives that may overlook nuanced hate expressions. The focus on English and Chinese datasets also limits generalizability across cultures and languages. We view HVGUARD as a research contribution that highlights the potential and challenges of reasoning-based multimodal moderation, and we encourage future work to expand to more diverse datasets, conduct fairness-oriented evaluation, and explore human–AI collaboration in practical applications.

Acknowledgements

This work was supported by the Key R&D Projects in Hubei Province (No. 2023BAB165) and the Smart Grid–National Science and Technology Major Project (No. 2024ZD0803000). The experiments were conducted on the research platform provided by the School of Cyber Science and Engineering, Wuhan University.

References

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

Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: dataset and baseline results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4309–4319.

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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

ByteDance. 2016. tiktok. https://www.tiktok.com.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440. Association for Computational Linguistics.

Gabriela Csurka. 2017. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, pages 1–35.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Google. 2005. Youtube. `https://www.youtube.com`.

Ming Shan Hee, Rui Cao, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024a. Understanding (dark) humour with internet meme analysis. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1276–1279.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Lee. 2024b. Recent advances in online hate speech moderation: Multimodality and the role of large models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kuanyu. 2009. Bilibili. `https://www.bilibili.com`.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Robin Matthew Medina, Judith Nkechinyere Njoku, and Dong-Seong Kim. 2022. Audio-based hate speech detection for the metaverse using cnn. In *KICS*.

Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr, and Virgílio Almeida. 2018. Analyzing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM conference on web science*, pages 323–332.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13062.

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*.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

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

Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. 2024. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 788–806. IEEE.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. 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.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 585–590. IEEE.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

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. IEEE.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930.

A Preliminary Study

With the advancement of artificial intelligence, MLLMs have become the focal point of the latest developments. The complementarity of LLMs and VLMs has given rise to MLLMs, such as Gemini 1.5(Team et al., 2024) and GPT-4 series (Achiam et al., 2023). They can receive, reason, and output multi-modal information, showing impressive capabilities in various multi-modal tasks, including image reasoning and video understanding (Wu et al., 2023; Fu et al., 2025), thus opening up new ways to solve complex and novel challenges in the multi-modal field.

Model	Hate	Offensive
GPT-4o	0.9513	0.8909
Gemini-1.5-pro	0.9120	0.8001

Table 4: Semantic similarity between MLLM and human definitions

The prerequisite for using MLLM to assist in the understanding of hateful content in videos is that the MLLM's comprehension of terms such as Hate, Offensive, and Normal aligns with human understanding of these terms.

We adopted MultiHateCLIP (Wang et al., 2024a) (one of the baselines) for its definitions of these terms:

- Hateful: Videos that incite discrimination or demean individuals or groups based on attributes such as race, ethnicity, nationality, religion, disability, age, veteran status, sexual orientation, gender identity, etc.

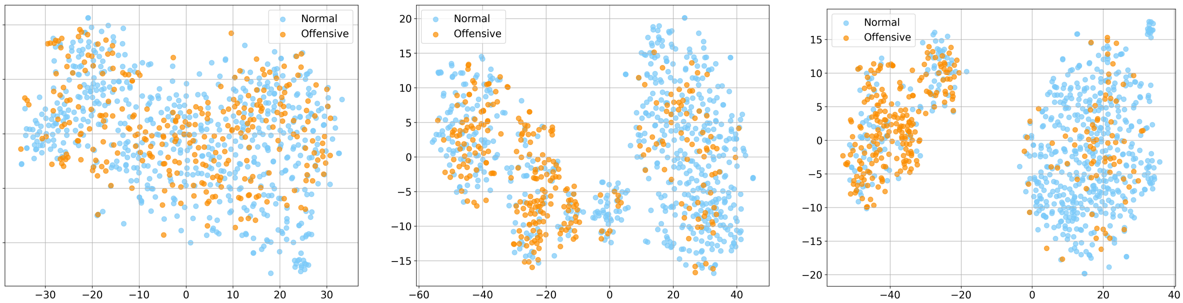
- Offensive: Videos that may cause discomfort or distress but do not meet the criteria for hateful content as defined above.

- Normal: Content free from hatefulness or offensiveness.

We analyzed the MLLM's comprehension to verify its alignment with MultiHateCLIP's definitions. Taking GPT-4o's responses as an example:

- Hateful: Videos that promote, incite, or glorify hatred, discrimination, or violence against individuals or groups based on attributes such as race, ethnicity, nationality, religion, gender, sexual orientation, disability, or other protected characteristics. This includes slurs, dehumanization, and calls for harm against a group.

- Offensive: Videos that contain derogatory, profane, or inappropriate language but do not necessarily incite violence or target protected



(a) Embeddings from Bert.

(b) Embedding from GPT-4o.

(c) Embedding from GPT-4o with CoT reasoning.

Figure 4: Visualization of features used by different methods. (a) Embedding of video titles, transcripts. (b) Embedding of MLLM rationale. (c) Embedding after incorporating the CoT prompts.

groups. This includes insults, strong language, or rude remarks that may be considered inappropriate but do not meet the threshold of hate speech.

To further investigate MLLMs' comprehension of these terms, we analyze their semantic similarity to MultiHateCLIP's definitions in Table 4, demonstrating that MLLMs can effectively distinguish between hateful and offensive content.

To more clearly demonstrate how the reasoning capability of MLLMs aids in understanding of hateful content in videos, we conducted a visual analysis of embedding representations on the hateful video dataset Multihateclip (Wang et al., 2024a). Figure 4a visualizes the embeddings of pure textual information (video title and transcript) extracted using the pre-trained text encoder Bert (Devlin, 2018), which exhibit significant overlap with no discernible class separability. This indicates the insufficiency of traditional approaches with single modality. However, when analyzing videos with MLLMs (Figure 4b), a certain degree of class separability becomes observable. By further incorporating the CoT prompting strategy (detailed in Section 3.4), we guide the MLLM to clarify rhetorical devices such as metaphors and puns in the videos, ultimately achieving sharper classification boundaries (Figure 4c). Thus, MLLMs provide effective rationale for hateful video understanding, and the CoT prompting strategy further amplifies this capability.

B Case Study

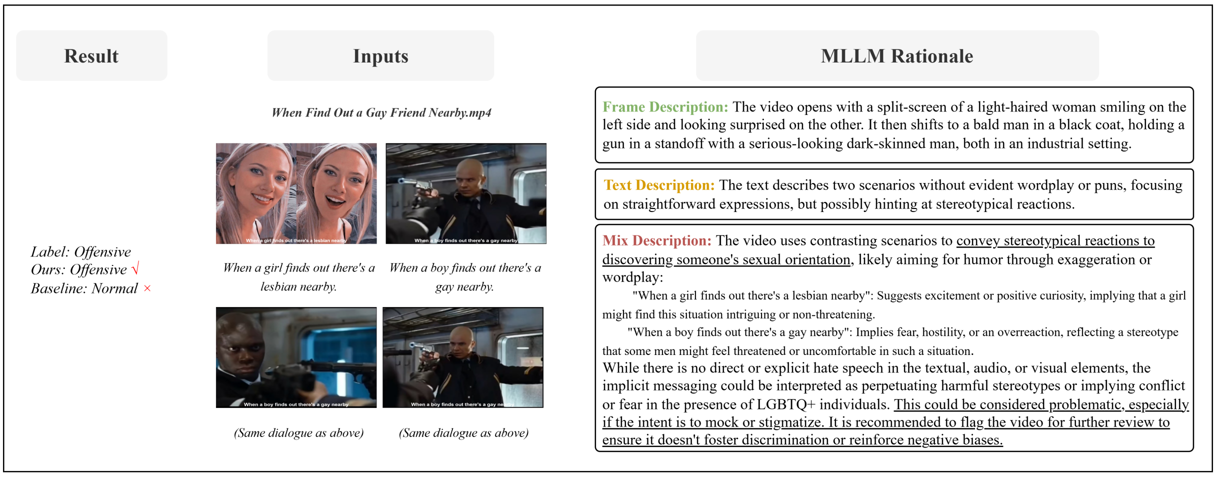
To provide a more comprehensive demonstration of HVGuard's effectiveness, we present a detailed case study in Figure 5. In this example, a video titled "*When Find Out a Gay Friend Nearby.mp4*" is processed, where understanding the reactions of different gender groups to homosexuality requires analyzing both visual and textual modalities. In HVGuard, MLLM leverages CoT prompts to guide reasoning from both video frames and transcripts, with the analysis from these modalities integrated to accurately interpret the video content. In contrast, baseline methods lacking MLLM reasoning fail to capture the complementary information between the visuals and the text, leading to incomplete analysis and misclassification.

C False Positive and Bias Analysis

Dataset	Number of categories	Model	F1(N)	R(N)	P(N)
Multihateclip (English)	3	HateMM	0.7899	0.8547	0.7434
		Multihateclip	0.7521	0.7186	0.8255
		HVGuard	**0.8895**	**0.9025**	**0.8787**
	2	HateMM	0.7532	0.8321	0.6937
		Multihateclip	0.7809	0.8765	0.7178
		HVGuard	**0.9120**	**0.9472**	**0.8815**
Multihateclip (Chinese)	3	HateMM	0.8082	0.9373	0.7276
		Multihateclip	0.8273	0.9620	0.7409
		HVGuard	**0.8948**	**0.9861**	**0.8218**
	2	HateMM	0.8123	0.8158	0.8116
		Multihateclip	0.8509	0.9485	0.7723
		HVGuard	**0.9031**	**0.9340**	**0.8786**
HateMM	2	HateMM	0.6941	0.6643	0.7393
		Multihateclip	0.7578	0.7668	0.7493
		HVGuard	**0.8715**	**0.8937**	**0.8567**

Table 5: Results of different methods on the task of hateful video detection. N:normal, R:recall, P:precision.

In real-world video platform scenarios, it is crucial not only to ensure accurate detection of hateful content but also to minimize the false positive rate on normal videos, so as to avoid negatively impacting the user experience of legitimate content creators. The results for the Normal category are shown in Table 5. As illustrated, HVGuard achieves strong performance in identifying normal videos, demonstrating its effectiveness in distinguishing between hateful and non-hateful content.


Result
Inputs
MLLM Rationale
Label: Offensive
Ours: Offensive √
Baseline: Normal ×
When Find Out a Gay Friend Nearby.mp4
When a girl finds out there's a lesbian nearby
When a boy finds out there's a gay nearby
When a girl finds out there's a lesbian nearby.
When a boy finds out there's a gay nearby.
When a boy finds out there's a gay nearby
When a boy finds out there's a gay nearby
(Same dialogue as above)
(Same dialogue as above)

Frame Description: The video opens with a split-screen of a light-haired woman smiling on the left side and looking surprised on the other. It then shifts to a bald man in a black coat, holding a gun in a standoff with a serious-looking dark-skinned man, both in an industrial setting.

Text Description: The text describes two scenarios without evident wordplay or puns, focusing on straightforward expressions, but possibly hinting at stereotypical reactions.

Mix Description: The video uses contrasting scenarios to convey stereotypical reactions to discovering someone's sexual orientation, likely aiming for humor through exaggeration or wordplay:

"When a girl finds out there's a lesbian nearby": Suggests excitement or positive curiosity, implying that a girl might find this situation intriguing or non-threatening.

"When a boy finds out there's a gay nearby": Implies fear, hostility, or an overreaction, reflecting a stereotype that some men might feel threatened or uncomfortable in such a situation.

While there is no direct or explicit hate speech in the textual, audio, or visual elements, the implicit messaging could be interpreted as perpetuating harmful stereotypes or implying conflict or fear in the presence of LGBTQ+ individuals. This could be considered problematic, especially if the intent is to mock or stigmatize. It is recommended to flag the video for further review to ensure it doesn't foster discrimination or reinforce negative biases.

Figure 5: Example of case study.

In addition to overall detection results, we further analyzed false positives and potential model biases. Across all datasets, approximately 74.12% of false positives involved sensitive terms such as profanity or emotionally charged expressions. For instance, in one MultihateClip (English) video from The Walking Dead, characters used strong language under stress, leading our model to incorrectly flag the clip as hateful despite the absence of offensive intent. This illustrates that contextually appropriate but emotive language can be misclassified as harmful content.

Dataset	Sensitive terms	LGBTQ+ topics
HateMM	78.29%	6.98%
MultihateClip (Chinese)	54.67%	4.00%
MultihateClip (English)	89.41%	22.35%

Table 6: Proportion of sensitive terms and LGBTQ+ topics in false positive samples across datasets.

We also examined the prevalence of sensitive topics in false positive samples across datasets, summarized in Table 6. While a substantial proportion of videos contain sensitive terms (e.g., 89.41% in MultihateClip-English), the proportion of videos explicitly involving LGBTQ+ topics is notably higher in English (22.35%) than in Chinese (4.00%). We observed that the model tends to overestimate offensiveness in the presence of such sensitive or identity-related themes, even when the content is neutral or supportive. This pattern suggests that pretraining biases and limited contextual reasoning contribute to false positives.

These findings highlight an important fairness concern: misclassification disproportionately affects creators addressing sensitive identities or social issues, potentially resulting in over-moderation. As part of future work, we plan to expand the analysis to finer-grained hate-related categories, conduct cross-cultural sensitivity testing, and explore bias mitigation strategies such as adaptive prompting, human-in-the-loop moderation, and culturally grounded evaluation.

D More Details on the Model Architecture

Considering that the features of this task are formed by concatenating multiple modalities, we employ a MoE network composed of multiple experts for processing. Different experts focus on different part of the features, enabling a profound understanding of multimodal features. Simultaneously, we utilize a gating network to modulate the weights of different experts, ensuring that each expert's contribution can be dynamically adjusted based on the properties of the input data.

The design of MoE. Each expert is implemented as a two-layer feedforward network with ReLU activation. The first linear layer projects the input (concatenated multimodal embeddings) into a hidden space, while the second layer produces the expert-specific output. Multiple experts focus on different parts of the features, thereby achieving better utilization of multimodal features.

The design of gating network. The gating network is implemented as a lightweight linear layer followed by softmax, which computes weights for combining the expert outputs. Taking the same input as the experts, it produces a distribution over the experts to indicate their relevance for the given input. To prevent polarization—avoiding over-reliance on or neglect of specific experts—a

Dataset	Categories	Model	Acc	F-F1
Multihateclip (English)	3	Multihateclip	0.7079	0.4946
		HVGuard(w/o gate)	0.8034	0.5605
		HVGuard	**0.8090**	**0.6646**
	2	Multihateclip	0.7416	0.6806
		HVGuard(w/o gate)	0.8315	**0.8045**
		HVGuard	**0.8539**	0.7714
Multihateclip (Chinese)	3	Multihateclip	0.7111	0.4573
		HVGuard(w/o gate)	0.7709	0.4402
		HVGuard	**0.8045**	**0.5643**
	2	Multihateclip	0.7778	0.6904
		HVGuard(w/o gate)	0.8315	0.8045
		HVGuard	**0.8603**	**0.8219**
Hatemm	2	Multihateclip	0.7614	0.7594
		HVGuard(w/o gate)	0.8218	0.8041
		HVGuard	**0.8563**	**0.8597**

Table 7: Results of removing gating network

dropout layer is applied to the gating weights, thereby enhancing generalization ability. During the forward pass, all experts process the input in parallel, and their outputs are combined through a weighted sum based on the gating weights, ensuring MoE's strong capability in handling complex inputs. We conducted ablation studies (Table 7) by removing the gating network entirely and simply averaging expert outputs, which demonstrated the necessity of learned gating weights.

For HVGUARD, we adopt a single gating mechanism for several reasons. First, since the model focuses on video classification—a multimodal but single-task learning scenario—the gate effectively balances expert contributions across different feature aspects while maintaining computational efficiency. This design aligns with the original MoE framework proposed by (Jacobs et al., 1991) and widely adopted in later work (e.g., (Shazeer et al., 2017)), where single gating has proven effective for resource-constrained multi-modal tasks. In contrast, multiple gating networks are typically reserved for multi-task learning, as seen in (Ma et al., 2018), where gates optimize for diverse objectives.

E Flexibility of framework component

Table 8 shows the impact of different combinations of MLLMs and Encoders. We conducted tests on the ternary classification scenario of Multihateclip(English). The combination of GPT-4o(Achiam et al., 2023), XLM(Conneau et al., 2020), Vit(Dosovitskiy, 2020), and Wav2Vec(Baevski et al., 2020) achieved the highest M-f1 value, while the combination of Qwen-VL(Bai et al., 2023), Bert(Devlin, 2018), ViViT(Arnab et al., 2021), and Wav2Vec achieved the highest accuracy. MFCC as an Audio Encoder significantly lowered the results, indicating that excellent modality encoders are necessary.

MLLM	**Text Encoder**	**Vision Encoder**	**Audio Encoder**	**Acc**	**M-F1**
GPT-4o	XLM	Vit	Wav2Vec	0.8090	**0.6646**
			MFCC	0.7809	0.4762
		ViViT	Wav2Vec	0.7921	0.5881
			MFCC	0.7865	0.5604
	Bert	Vit	Wav2Vec	**0.8202**	0.5562
			MFCC	0.7978	0.5590
		ViViT	Wav2Vec	0.8034	0.6175
			MFCC	0.8146	0.5384
Qwen-VL	XLM	Vit	Wav2Vec	0.7865	0.6276
			MFCC	0.7640	0.4759
		ViViT	Wav2Vec	0.7809	0.5744
			MFCC	0.7697	0.5637
	Bert	Vit	Wav2Vec	0.7921	0.5652
			MFCC	0.7753	0.5022
		ViViT	Wav2Vec	0.7978	0.5282
			MFCC	0.7809	0.4835

Table 8: Results of different model combinations on Multihateclip(English)

We found that different combinations have varying impacts on performance, with the capabilities of the MLLM being the most significant factor. However, even the least effective combination significantly outperformed the baseline, demonstrating the flexibility and generalizability of our proposed HVGUARD framework.