Reasoning-Aware Multimodal Fusion for Hateful Video Detection

Anonymous authors
Paper under double-blind review

Abstract

Hate speech in online videos is posing an increasingly serious threat to digital platforms, especially as video content becomes increasingly multimodal and context-dependent. Existing methods often struggle to effectively fuse the complex semantic relationships between modalities and lack the ability to understand nuanced hateful content. To address these issues, we propose an innovative Reasoning-Aware Multimodal Fusion (RAMF) framework. To tackle the first challenge, we design Local-Global Context Fusion (LGCF) to capture both local salient cues and global temporal structures, and propose Semantic Cross Attention (SCA) to enable fine-grained multimodal semantic interaction. To tackle the second challenge, we introduce adversarial reasoning—a structured three-stage process where a vision-language model generates (i) objective descriptions, (ii) hate-assumed inferences, and (iii) non-hate-assumed inferences—providing complementary semantic perspectives that enrich the model's contextual understanding of nuanced hateful intent. Evaluations on two real-world hateful video datasets demonstrate that our method achieves robust generalisation performance, improving upon state-of-the-art methods by 3% and 7% in Macro-F1 and hate class recall, respectively. We will release the code after the anonymity period ends.

Disclaimer: This paper contains sensitive content that may be disturbing to some readers.

1 Introduction

Online videos have become a dominant communication medium, and their widespread reach has enabled hateful content to spread rapidly (Das et al., 2023). Such content exacerbates discrimination and social division, and can even incite offline violence (Townsend, 2025; Robertson, 2025). Current methods (Zhang et al., 2024; Koushik et al., 2025; Wang et al., 2024; Das et al., 2023; Yue et al., 2025) follow a standard of extracting and fusing features from video frames, audio, and transcribed text, but lack effective multimodal semantic interaction. As illustrated in Figure 1, existing hateful video detection faces two main challenges: (1) nuances of context understanding—hateful videos often convey harmful intent through nuanced contextual cues spanning time and modalities, such as the temporal resonance between specific visuals and statements (Yang et al., 2025; Wang et al., 2025b), or the implicit linkage between visual focus and auditory content (Rehman et al., 2025); and (2) fusion of multimodal semantic relations—current methods (Zhang et al., 2024; Koushik et al., 2025; Wang et al., 2024; Das et al., 2023; Yue et al., 2025) follow a standard of extracting and fusing features from video frames, audio, and transcribed text, but lack effective multimodal semantic interaction. Both challenges demand deep contextual reasoning capabilities that standard pipelines struggle to provide.

Recent approaches attempt to enhance contextual understanding through visual-language models (VLMs) generated reasoning (Lang et al., 2025; Hee et al., 2024). As shown in Figure 1, this standard fusion with reasoning methods incorporates direct reasoning or Chain-of-Thought (CoT) reasoning (Hee et al., 2025) as additional inputs. However, direct reasoning produces only surface-level descriptions lacking hateful semantic associations, while recent work reveals that CoT may generate misleading explanations that fail to reflect actual model reasoning (Barez et al., 2025), raising reliability concerns for sensitive detection tasks. Beyond

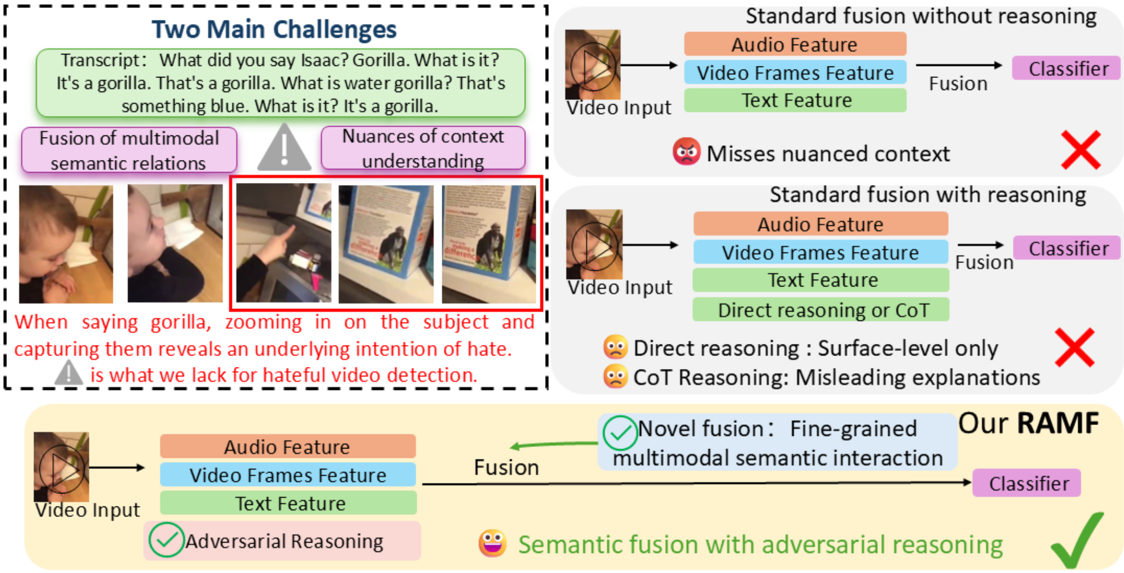


Figure 1: Left: Two main challenges—fusion of multimodal semantic relations and nuances of context understanding. Right: Standard paradigms vs. our RAMF.

reasoning quality, recent research (Wang et al., 2025b; Yang et al., 2025) of the hateful video dataset reveals that hate cues exhibit heterogeneous temporal distributions: they may erupt briefly within short segments or be dispersed across the entire video timeline. Concurrently, inference signals carry higher-order semantics at varying granularities, necessitating fine-grained integration with low-level multimodal features. These limitations reveal a fundamental gap: existing systems lack both reliable semantic reasoning and effective multimodal fusion mechanisms for hateful video detection.

To bridge this gap, we propose Reasoning-Aware Multimodal Fusion (RAMF), a unified framework that addresses both challenges. To tackle Challenge 1, we introduce adversarial reasoning—a structured three-stage process where a VLM generates (i) objective descriptions, (ii) hate-assumed inferences, and (iii) non-hate-assumed inferences—providing complementary semantic perspectives that enrich the model's contextual understanding while maintaining factual grounding. Unlike prior reasoning approaches (Lang et al., 2025; Hee et al., 2024; 2025), this adversarial design forces the model to explicitly consider both interpretations, providing complementary perspectives that enrich contextual understanding while maintaining factual grounding. To tackle Challenge 2, we design Local-Global Context Fusion (LGCF) to capture both local salient cues and global temporal structures, and propose Semantic Cross Attention (SCA) to enable fine-grained multimodal semantic interaction.

Our main contributions are: 1) Adversarial reasoning for nuanced and intent-aware semantic understanding. We propose a structured adversarial reasoning pipeline that generates objective descriptions, hate-assumed interpretations, and non-hate-assumed interpretations. This design provides complementary semantic views, enabling the understanding of nuanced hateful content in subtle contexts, improving robustness to nuanced scenarios. 2) LGCF and SCA for comprehensive multimodal fusion. We introduce LGCF to jointly capture local salient cues and global temporal patterns, and propose SCA to achieve fine-grained multimodal semantic interaction. This facilitates efficient fusion of heterogeneous modalities whilst integrating high-level reasoning signals. 3) Extensive experiments on HateMM and MultiHateClip demonstrate state-of-the-art performance, with improvements of 3% in Macro-F1 and 7% in hate class recall.

The source codes and data required to reproduce our results are available at `https://anonymous.4open.science/r/RAMF-FB34` and will be made public.

2 Related Work

2.1 Hateful Content Detection

For hate speech detection, early studies used manually designed text features. Chen et al. (2012) combined n-grams with grammatical rules, while Davidson et al. (2017) utilised sentiment analysis. Advances in deep learning have enabled automatic feature extraction, using convolutional neural networks (CNNs) (Le Cun et al., 1989) and recurrent neural networks (RNNs) (Elman, 1990) to identify hate patterns in text (Vashistha & Zubiaga, 2021; Menini et al., 2019; Corazza et al., 2020). Beyond text, hateful content also appears in multimodal forms (e.g., emojis and videos), necessitating the integration of multimodal methods. Das et al. (2023) developed the first hateful video dataset, HateMM, and proposed a standard paradigm for extracting multimodal information from video frames, audio, and transcribed text to detect hateful videos. Although recent models (Zhang et al., 2024; Lang et al., 2025; Céspedes-Sarrias et al., 2025; Rehman et al., 2025) incorporate multimodal information, they largely adhere to a standard fusion without reasoning or direct reasoning and fall short in capturing the nuanced, context-dependent nature of hateful content. For instance, while MoRE (Lang et al., 2025) leverages the BLIP visual language model to generate video frame descriptions, these captions remain surface-level and lack deeper semantic understanding of hate-related context.

a On the other hand, the successful application of VLMs in multimodal tasks (Tang et al., 2025) has sparked interest in using them for hateful content moderation (Hee et al., 2024; Hee & Lee, 2025; Hee et al., 2025; Rizwan et al., 2025). For example, IntMeme (Hee & Lee, 2025) fine-tunes VLM models using carefully designed prompts and contrastive learning objectives. InstructMemeCL (Hee et al., 2025) uses VLM to generate human-style explanations for memes and then uses CoT to generate more transparent VLM decision results. These works demonstrate the potential of VLM for hateful content, but there has been no exploration of complex contextual semantic understanding in hateful videos. Given this research gap, we have redesigned a structured VLM generation system to generate semantic explanations, thereby aiding comprehension of contextually hateful content within nuanced contexts.

2.2 Multimodal Fusion in Hateful Video Detection

Das et al. (2023) developed the first hateful video dataset and established a baseline using a simple fusion method. Zhang et al. (2024) improved performance by employing complex cross-attention fusion techniques. Koushik et al. (2025) point out that existing fusion strategies have significant limitations in capturing complex multimodal semantic relationships and providing adaptable unified architectures. The latest work by Lang et al. (2025) enhanced hateful video detection performance through retrieval, expert fusion, and BLIP-based video description generation. However, the expert form of fixed modalities in it limits deep modal interaction and lacks multimodal semantic fusion.

Specifically for modelling and fusion methods, existing hateful video detection methods typically employ sequence models (e.g., long short-term memory networks (LSTM)) (Hochreiter & Schmidhuber, 1997) combined with attention mechanisms (Vaswani et al., 2017) to achieve modality interaction (Bai et al., 2018; Zhang et al., 2024; Das et al., 2023; Vashistha & Zubiaga, 2021; Mandal et al., 2024; Koushik et al., 2025). However, LSTM lacks effective local modelling capabilities, while traditional attention mechanisms (Vaswani et al., 2017), equipped with independent attention heads, lack direct interaction between heads. Recently, Multi-Token Attention (MTA) (Golovneva et al., 2025) addressed the head interaction issue by introducing key query convolutions and intra-group head mixing convolutions, thereby enabling information sharing among tokens. However, it mainly focused on contextual localisation and still lacked mechanisms to capture multimodal semantic structural dependencies. To address the above issues, we designed an improved attention module to achieve multi-semantic fusion of VLM-generated reasoning and modalities in traditional paradigms.

2.3 Reasoning

The application of VLMs to hate speech detection has made some progress (Hee et al., 2025; Hee & Lee, 2025; Wang et al., 2025a; Rizwan et al., 2025; Lang et al., 2025). However, the field of hateful video detection has still not achieved significant breakthroughs. Currently, VLM-based methods typically generate a single narrative for a given piece of content, use VLMs to determine whether memes contain hate speech (Hee & Lee, 2025), or employ CoT to provide inferential explanations (Hee et al., 2025). However, recent research (Barez et al., 2025) points out that CoT is not explainable, indicating that CoT cannot guarantee the model's fidelity to the reasoning process, thereby lacking explainability. The infidelity of chain reasoning poses a trust crisis for hateful video detection systems. VLMs may make judgments based on implicit biases or incorrect context, yet generate seemingly reasonable but erroneous explanations to mask their true decision-making mechanisms (Barez et al., 2025), failing to meet legal requirements for algorithmic transparency and fundamentally threatening the credibility and controllability of hate content detection. Given these limitations, we design a structured adversarial reasoning pipeline that provides complementary semantic perspectives without relying on potentially unfaithful reasoning chains.

3 Methodology

3.1 Problem Formulation

In the standard pipeline proposed by HateMM (Das et al., 2023), multimodal input videos consist of video frames V, audio segments A, and transcription texts T. Each modality $m \in \{T, A, V\}$ is independently encoded as an embedding sequence $X^m \in R^{L \times D^m}$, where L is the sequence length and D^m is the feature dimension. These modality-specific embeddings are fused into a unified semantic representation $Y_1 \in R^{D^Y}$ to predict whether a video contains hateful content ($y = 1$) or not ($y = 0$). However, the pipeline lacks

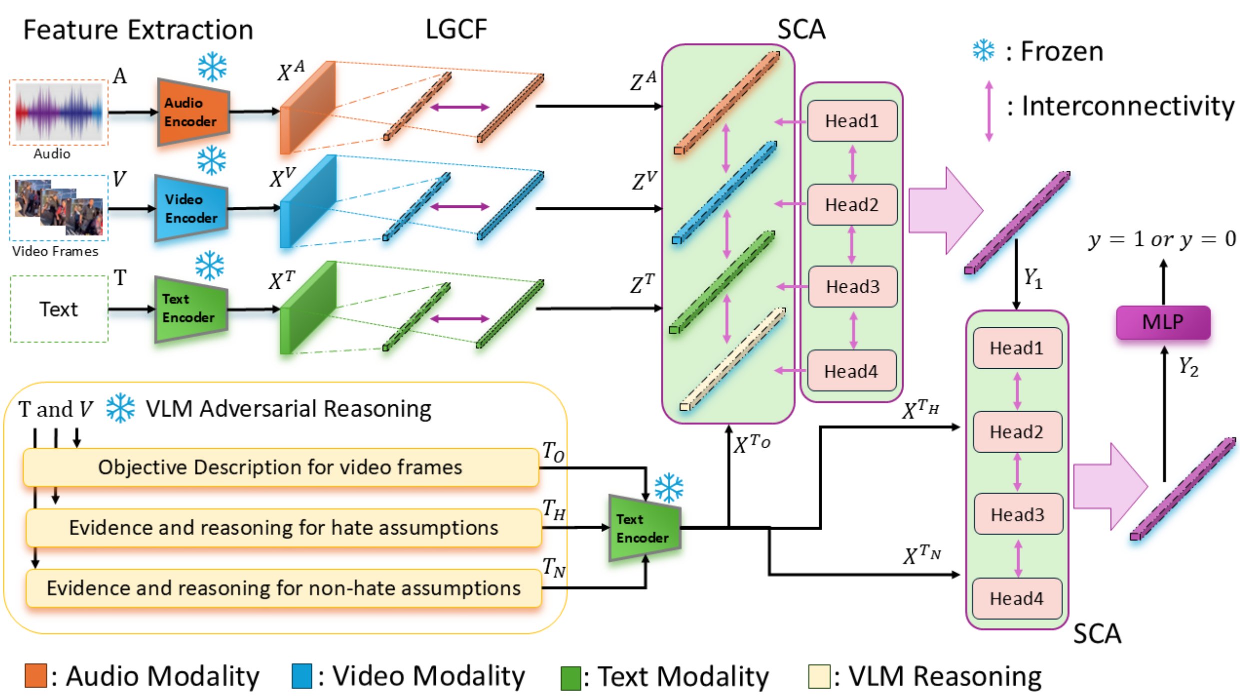


Figure 2: The overall architecture of the proposed framework, including the Local-Global Context Fusion (LGCF) module, the Semantic Cross Attention (SCA) mechanism.

additional semantic understanding knowledge, making it difficult to capture nuanced and context-dependent hateful clues. To address these issues, we introduce adversarial reasoning: objective descriptions (T_O),

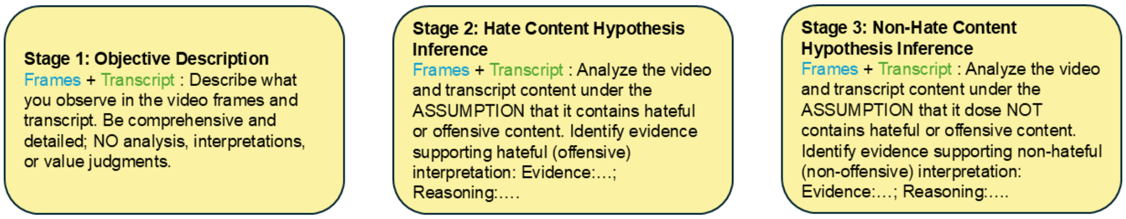


Figure 3: Prompts used in the three stages of our framework.

hate reasoning (T_H), and non-hate reasoning (T_N), generated through adversarial reasoning using a VLM. We adopt a two-stage fusion strategy to integrate six modalities. The original modalities with objective description $\{T, A, V, T_O\}$ are fused via the proposed LGCF and SCA to obtain a representation Y_1. Then, Y_1 is further fused with the adversarial reasoning texts $\{T_H, T_N\}$ through another SCA layer to produce the final semantic representation Y_2, which is used for classification. An overview of the proposed framework is shown in Figure 2.

3.2 Vision Language Model Adversarial Reasoning

To address the limitations of existing VLM-based reasoning approaches, we propose a structured three-stage adversarial reasoning pipeline that enhances contextual semantic understanding without relying on VLM's subjective reasoning explanations. Unlike single-narrative approaches, our adversarial reasoning explicitly guides the VLM through a space of contrasting assumptions, generating complementary evidence for both hateful and non-hateful interpretations within the same video. This design enhances contextual understanding through three mechanisms: 1) the structured prompting constrains the VLM to produce objective descriptions before interpretive reasoning, reducing hallucination and bias; 2) the adversarial format provides self-correction—even if one reasoning path is flawed, the complementary perspective can compensate; and 3) explicit instructions requiring visual evidence references strengthen factual grounding, enhancing the reliability of model outputs. The robustness to variations in VLM quality and the impact of reasoning quality are validated in our ablation study.

The three stages include (1) Objective description of content: The model generates an objective description of the visual elements observed in the video and the accompanying text, without involving interpretative judgments, to establish an unbiased representation (see Figure 3), denoted as T_O.

(2) Hate-Assumed Inference: Assuming the content contains hate speech, the model explores discriminatory expressions and offensive content targeting specific groups, and provides contextual evidence and reasons (see Figure 3), denoted as T_H.

(3) Non-Hate-Assumed Inference: Assuming the content does not contain hate speech, the model explores reasonable alternative interpretations, such as artistic expression, satirical context, and personal conflicts, and provides corresponding contextual evidence and reasoning (see Figure 3), denoted as T_N.

3.3 Encoder Module

The encoder module comprises modality-specific preprocessing and feature extraction. Text, T, is obtained by transcribing speech with OpenAI's Whisper model (Radford et al., 2023), then tokenised and processed using Bidirectional Encoder Representations from Transformers (Bert or multilingual Bert (mBert)) (Devlin et al., 2019) and the HateXplain (HXP) model (Mathew et al., 2021) to extract 768-dimensional embeddings, with zero-padding or truncated to 100 fixed sequence length, denoted as X^T. The VLM reasoning text, T_O, T_H, T_N also processed with Bert or HXP to get X^{T_O}, X^{T_H} and X^{T_N}, then pass individually Multi-Layer Perceptron (MLP) with $\{512, 256\}$ to obtain unified dimensionality.

Audio, A, is resampled to 16kHz to extract 40-dimensional Mel Frequency Cepstral Coefficients (MFCC) (Muda et al., 2010), and to 48kHz to extract 512-dimensional semantic embeddings using the pretrained

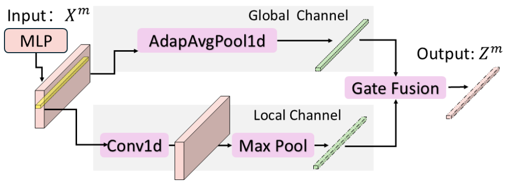


Figure 4: Structure of the LGCF, fusing local and global contextual information.

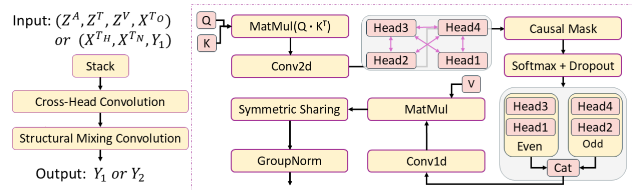


Figure 5: Structure of the SCA.

Contrastive Language-Audio Pretraining (CLAP) model (Elizalde et al., 2023), with zero-padding if needed or downsampling with 100 fixed stride, denoted as X^A.

Video frames, V, are processed by Vision Transformer (ViT) (Dosovitskiy et al., 2020) or Video Vision Transformer (Vivit) (Arnab et al., 2021) to extract 768-dimensional visual embeddings, and by the pretrained Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021) to extract 512-dimensional semantic embeddings, with black frames padded if needed, denoted as X^V.

3.4 Local-Global Context Fusion

Based on observations of the hateful video, hateful content can be either concentrated in a short period of time or spread throughout long videos (Yang et al., 2025). This highlights the importance of local modelling and global understanding for traditional three modalities. However, existing LSTM-based methods have weak local modelling capabilities (Das et al., 2023; Zhang et al., 2024). Inspired by Bai et al. (2018), who studied the advantages of convolution for sequence modelling, the proposed LGCF module addresses issues by adaptively combining local salient features and global context within the sequence, while preserving discriminative cues and overall temporal structure, as illustrated in Figure 4.

Given extracted embeddings X^m, we first apply modality-specific MLP with $\{512/128, 256\}$ to obtain the unified representation X^m_{MLP} (128 with MFCC). Subsequently, X^m_{MLP} is fed into two parallel channels to extract local and global temporal features. In the Local Temporal Channel (LTC), a one-dimensional convolution with kernel size 3 and padding 1 is applied across the time dimension to extract local context. The convolution kernel size is a non-critical parameter, as analysed in Figure 6. Then, maximum pooling is performed on the time axis to capture the maximum activation value of each feature:

$$v_{\text{local}} = \text{MaxPool1D}(\text{Conv1D}(X^m_{\text{MLP}})) \tag{1}$$

In the Global Temporal Channel (GTC), a global average pooling over the original sequence is computed:

$$v_{\text{global}} = \text{AdapAvgPool1D}(X^m_{\text{MLP})} \tag{2}$$

The local and global vectors are concatenated and passed through a learned gate to adaptively combine them:

$$\begin{aligned} Z^m &= g \odot v_{\text{local}} + (1-g) \odot v_{\text{global}}, \\ g &= \sigma(W[v_{\text{local}} \oplus v_{\text{global}}] + b) \end{aligned} \tag{3}$$

where σ is the sigmoid function, W is the weight matrix, b is the bias term, $\oplus$ denotes vector concatenation, and $\odot$ indicates element-wise multiplication. This design is crucial for detecting sparsity and implicit hate speech, ultimately compressing X^m for each modality into a compact and information-rich representation Z^m.

3.5 Semantic Cross Attention

Inspired by cross-head interactions in MTA (Golovneva et al., 2025), we propose the SCA mechanism. SCA introduces Cross-Head Convolution (CHC) and Structural Mixing Convolution (SMC) to facilitate comprehensive and fine-grained multimodal semantic fusion, as illustrated in Figure 5. To enable communication between heads and model the Key Query space structure, we apply 2D convolutions to the attention logits. Specifically, we treat the attention tensor as a 3D array of shape $[H, N, D^Z]$, where H is the number of attention heads and N is the sequence length. Each $[N \times D^Z]$ slice corresponds to an attention map for a single head. By treating the heads as convolution channels, we perform shared 2D convolutions. Unlike MTA (Golovneva et al., 2025), which assigns independent convolution kernels to each head, we apply a single shared convolution to all heads, achieving effective fusion of local information without additional parameter increases or modifications to the underlying operators, while reducing inference time, as analysed in Table 5.

The features from different modalities are stacked to form $Z^s \in R^{B \times 3/4 \times D}$, which is then fed into the SCA module (Z^s will be omitted in the following text). CHC are employed to effectively capture high-order correlations across modalities,

$$A = \mathrm{softmax}\left(\mathrm{Conv2D}_{\mathrm{heads}}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right) + M\right) \tag{4}$$

where Q_h, K_h and d_h are the query, key matrices and dimensionality of each attention head, respectively. The causal mask M is added to prevent access to future information. A 3×3 symmetric convolution operation with padding 1×1, denoted by $\mathrm{Conv2D}_{\mathrm{heads}}$. The 3×3 convolution kernel size used here is to accommodate the length of the input sequence Z^s, which is mostly 3.

$$\begin{aligned} A' &= \mathrm{HeadMix}(A) \\ &= \mathrm{Conv1D}_{\mathrm{groups}=N/2}(A_{\mathrm{even}} \oplus A_{\mathrm{odd}}) \end{aligned} \tag{5}$$

For SMC, the attention weights A are split into even- and odd-indexed heads, denoted as A_{even} and A_{odd}. The odd-even grouping topology achieves interleaving and mixing between heads of different distances, improving the robustness and generalisation ability of the model, as analysed in Table 2. Then, concatenate to the sequence dimension, represented as $\oplus$. Group convolution $\mathrm{Conv1D}_{\mathrm{groups}=N/2}$ is used to process mixed views, similar to MTA (Golovneva et al., 2025). The default setting for the convolution kernel size is 2 with a stride of 2. The performance results for different sizes are basically stable, as analysed in Figure 6. Finally, the output of the module is as follows:

$$Y = \mathrm{GN}\left(\mathrm{OutProj}\left(A'V\right)\right) \tag{6}$$

where V_h is the value matrix for each attention head. $A'V$ denotes the application of shared attention weights to the value matrix. The result is reshaped back to the original layout, producing the mixed attention weights A'. This is to extend the attention distribution after multimodal interaction to the entire attention space, considering structural alignment and semantic consistency. More analysis can be seen from the ablation study. Then followed by an output projection layer, OutProj, and group normalisation, GN.

For the SCA of the first layer, $Z^s \in \{X^T, X^A, X^V, X^{T_O}\}$, and get Y_1. For second layer SCA, $Z^s \in \{Y_1, X^{T_H}, X^{T_N}\}$, and get Y_2. For classification, Y_2 or Y_1 passes through an average pooling layer and enters the MLP classification layer with $\{128, 64, 2\}$ to obtain the final classification result. For RAMF, use Y_2. For a traditional paradigm-based Multimodal Fusion (MF) model, use Y_1.

4 Experiments

A total of 1,083 HateMM videos and 959/964 videos from the Chinese/English MHC subsets are used. These numbers differ from the 1,000 videos per subset reported by Wang et al. (2024) because we re-collected and re-partitioned the data to construct a more rigorous five-fold cross-validation setup with mutually exclusive test sets. During this process, some original videos were found to be unavailable due to removals from Bilibili and YouTube, resulting in slight deviations in dataset size. We adopt a 70%/10%/20% split for training, validation, and testing within each fold and perform 5-fold cross-validation. Unlike prior work, which fixed a single data split and only varied random seeds across runs (Lang et al., 2025; Wang et al., 2024), our re-partitioning ensures that test sets across folds are strictly non-overlapping, enabling a more generalisable and realistic evaluation. In the binary classification task for MHC, hate and offensive labels are merged into a single hate label, following consistent practice in MHC. Our models are trained using the Adam optimiser with a learning rate of 10^{-4} and cross-entropy loss; more detailed configurations are provided in the appendix C.

Following prior unified protocol (Das et al., 2023; Wang et al., 2024), we sample an average of 100 frames per video in HateMM and 32 frames in MHC. For VLM reasoning, we employ the Qwen 2.5-VL-32B (Bai et al., 2025), using 16 sampled frames per video as visual input. This configuration is selected to accommodate hardware constraints. For baseline models, we strictly adhere to the settings specified in their original papers. We maintain the same evaluation metrics used in the HateMM and MHC, including macro-F1 score, accuracy, F1 score for hate class, precision for hate class, and recall for hate class. The best model for each fold is selected based on the macro F1 score on the validation set and evaluated on the test set.

4.1 Baselines

We evaluate the effectiveness of the proposed model by comparing it with recent unimodal and multimodal approaches on the HateMM and MHC datasets, which are two real-world video datasets in the field. For unimodal baselines, we adopt CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2020) (Vivit (Arnab et al., 2021) in MHC), CLAP (Elizalde et al., 2023), MFCC (Muda et al., 2010), HXP (Mathew et al., 2021) and BERT (Devlin et al., 2019) (mBert in MHC–Chinese). We apply average pooling with an MLP to each unimodal input, following HateMM (Das et al., 2023) and MHC (Wang et al., 2024). We additionally include LLMs and VLMs, i.e., GPT4–o (OpenAI, 2024), LLama (3.1–405B and 4–17B) (Patterson et al., 2022), and Qwen–Max (Qwen, 2024) for comparison by evaluating their zero-shot performance on hateful video classification. LLMs and VLMs are guided via prompts to analyse textual and video-text inputs; further details are given in the appendix A.5. For multimodal models, HateMM (Das et al., 2023) and MHC (Wang et al., 2024) are used as baseline models, along with CMFusion (Zhang et al., 2024) and the state-of-the-art MoRE (Lang et al., 2025) model.

4.2 Quantitative Results

Table 1 shows the performance comparison between our proposed model and existing methods on hateful video detection. Our model achieves the best performance on all datasets and different feature combinations, demonstrating excellent robustness and generalisation ability. The MF row in Table 1 corresponds to a configuration without VLM inference, designed to demonstrate the advantages of the fusion module while still achieving significant improvements over previous fusion methods, validating the effectiveness of our proposed SCA and LGCF in multimodal semantic fusion. RAMF further improves performance, demonstrating that our novel adversarial reasoning framework effectively enhances the model's robustness against nuanced and context-dependent hate content, particularly by simultaneously improving macro-F1 and recall, which are crucial for hate video detection. More ablation analysis demonstrating superiority over CoT methods.

4.3 Ablation Study

For the ablation study presented in Table 2, we analyse the individual contributions of each proposed component. In the RAMF ablation, we compare different strategies: RAMF^1 represents the proposed framework using Qwen 2.5-VL-32B, while RAMF^2 substitutes LLaMA 4-17B as the reasoning generator. The marginal performance difference between RAMF^1 and RAMF^2 demonstrates that our framework is robust to variations in VLM quality. RAMF^3 fuses the objective description in the second SCA layer instead of the first. The performance drop observed when using the CoT reasoning method (MF-CoT) instead of adversarial reasoning highlights the effectiveness of the latter (detailed CoT implementation is provided in the appendix A.3). Further, removing the second-layer SCA (w/o hierarchical fusion) and instead processing all information in a single SCA results in performance degradation. Similarly, excluding either the objective descriptions (w/o ObjDesc) or the adversarial reasoning (w/o Assumption) leads to reduced performance. Notably, eliminating the adversarial reasoning capability leads to a decline in MF1 by over 2%, demonstrating its significant role.

Table 1: Video classification performance (%)(five-fold average). MF1(F1): Macro-F1 and F1 for hate class; Acc: Accuracy; P(H): Precision for hate class; R(H): Recall for hate class. The MF row corresponds to a configuration without VLM inference. Results with detailed standard deviations are provided in the appendix C.

Model	HateMM				MHC(Chinese)				MHC(English)			
	MF1(F1)	Acc	P(H)	R(H)	MF1(F1)	Acc	P(H)	R(H)	MF1(F1)	Acc	P(H)	R(H)
Unimodal												
BERT^{T1}	78.6 (74.5)	79.5	74.3	74.9	58.8 (45.8)	63.1	44.8	49.2	62.6 (53.3)	65.1	49.4	58.6
HXP^{T2}	80.6 (77.4)	81.2	74.7	80.8	–	–	–	–	64.0 (53.4)	67.8	54.0	56.8
MFCC^{A1}	66.9 (62.4)	67.7	58.9	67.1	54.4 (38.3)	60.3	38.8	39.2	47.8 (24.8)	58.4	31.2	21.0
CLAP^{A2}	72.2 (64.9)	74.2	71.1	59.7	59.1 (42.2)	66.2	48.2	38.0	57.9 (41.2)	64.8	47.4	36.6
ViT^{V1}	71.2 (64.5)	72.8	67.1	62.6	61.9 (50.6)	65.5	48.2	54.3	56.6 (42.6)	61.2	43.8	41.9
CLIP^{V2}	73.4 (67.8)	74.6	69.5	66.4	58.3 (37.5)	68.6	54.6	28.6	63.6 (52.3)	67.3	52.2	52.7
LLM												
GPT-4o	78.0 (76.6)	78.2	66.8	89.8	54.0 (26.6)	70.4	71.8	16.4	63.7 (45.8)	72.4	70.1	34.2
Qwen-Max	66.8 (69.1)	67.0	55.2	92.3	61.6 (45.4)	68.4	53.5	40.3	70.2 (61.2)	73.0	60.6	62.3
LLaMA 3.1	72.1 (68.2)	72.7	64.5	73.9	–	–	–	–	69.1 (56.6)	74.3	67.2	49.1
VLM												
Qwen-VL	70.3 (70.9)	70.3	58.2	90.8	60.8 (41.6)	70.3	59.8	32.4	71.3 (59.9)	75.7	69.4	53.3
LLaMA 4	69.8 (70.7)	69.8	57.8	91.2	–	–	–	–	69.4 (57.2)	74.4	66.9	50.2
Multimodal												
T1·A1·V1												
HateMM	79.3 (76.5)	79.8	73.9	79.5	64.0 (51.8)	68.3	51.8	53.1	62.3 (48.2)	67.6	54.0	44.0
CMFusion	79.1 (74.4)	80.2	77.1	72.2	61.4 (44.9)	68.6	53.5	39.3	60.8 (45.8)	66.8	52.3	40.6
MoRE	81.0 (76.6)	82.1	80.0	73.5	60.2 (40.8)	69.6	57.1	31.7	62.8 (49.6)	67.5	51.6	47.8
MF	82.3 (79.0)	82.9	78.0	80.2	65.9 (53.9)	70.3	55.1	53.2	63.5 (51.5)	67.8	53.3	50.9
RAMF	**83.7 (80.9)**	**84.3**	**78.6**	**83.7**	**69.3 (60.2)**	**72.4**	**58.4**	**63.8**	**64.1 (51.9)**	**68.5**	**53.2**	**52.2**
T2·A2·V2												
HateMM	82.4 (79.6)	83.0	79.4	80.2	63.3 (48.9)	69.0	54.3	47.3	68.4 (57.8)	72.3	59.1	57.1
CMFusion	81.9 (77.9)	82.8	80.2	75.8	60.6 (45.3)	66.7	49.0	42.4	67.7(56.6)	71.7	60.6	55.1
MoRE	82.1 (78.3)	82.9	79.7	77.0	62.5 (46.8)	69.1	54.1	41.2	67.4 (54.5)	72.5	61.1	49.3
MF	83.1 (80.2)	83.7	78.3	82.6	66.2 (59.2)	70.3	55.2	54.9	69.6 (59.1)	73.4	62.1	56.8
RAMF	**85.1 (82.5)**	**85.6**	**79.8**	**85.5**	**70.9 (61.3)**	**74.5**	**61.3**	**62.2**	**71.7 (63.8)**	**74.0**	**62.1**	**67.4**

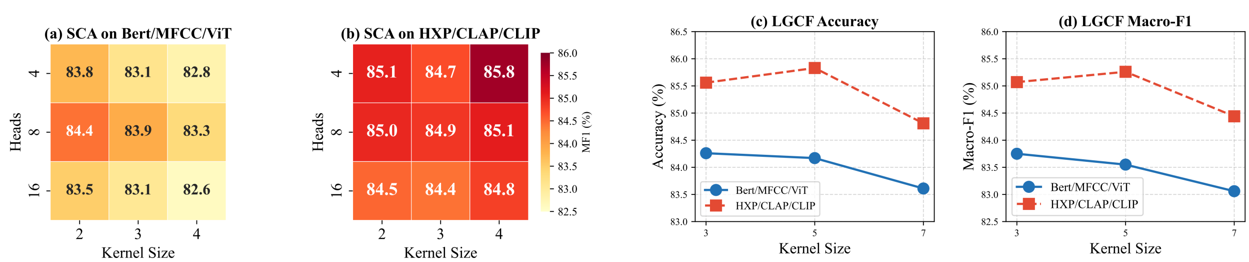


Figure 6: Hyperparameter analysis across two feature configurations on the HateMM dataset. Acc denotes accuracy, and MF1 denotes macro-F1.

In the MF ablation study, removing any single module leads to a performance drop, validating the overall design and necessity of each component. More granular ablation of the SCA reveals that eliminating core mechanisms such as CHC or SMC reduces performance, proving the effectiveness of these mechanisms for semantic communication and integration. Comparisons with standard attention (StdAttn) (Vaswani et al., 2017) fusion mechanisms, cross attention (CrossAtten) (Tsai et al., 2019) fusion mechanisms, MTA (Golovneva et al., 2025) and replacing structured odd-even grouping with simple concatenation (Concat), further affirm the superiority of the enhanced attention architecture. Ablation results from the LGCF module confirm the necessity of both the gating mechanism and the dual channel structure. Additionally, replacing the LGCF with an LSTM architecture results in a performance decline, indicating that the proposed

Table 2: Ablation study on the HateMM dataset using BERT, MFCC, and ViT. MF1 denotes macro-F1, and Acc denotes accuracy. Values in parentheses indicate the relative decrease in macro-F1 compared to the RAMF[1] or MF baseline.

RAMF Ablation		
RAMF[1]	84.26	83.75
RAMF[2]	84.35	83.62
RAMF[3]	83.61	82.87 (↓0.88)
MF-CoT	83.24	82.61 (↓1.14)
w/o Hier. Fusion	82.78	81.95 (↓1.80)
w/o ObjDesc	83.80	83.14 (↓0.61)
w/o Assumption	82.41	81.69 (↓2.06)

MF Ablation		
MF	82.96	82.32
w/o MLP	80.83	80.08 (↓2.24)
w/o LGCF	80.74	79.87 (↓2.45)
w/o SCA	80.28	79.41 (↓2.91)

SCA Module Ablation		
w/o CHC	80.83	80.12 (↓2.20)
w/o SMC	82.13	81.23 (↓1.09)
Concat	81.94	81.11 (↓1.21)
MTA	81.94	81.18 (↓1.14)
StdAttn	79.07	77.89 (↓4.43)
CrossAttn	78.89	78.28 (↓1.84)

LGCF Module Ablation		
w/o Gate Fusion	80.56	79.44 (↓2.88)
w/o GTC	79.63	78.87 (↓3.45)
w/o LTC	77.87	77.21 (↓5.11)
LSTM	77.96	76.82 (↓5.50)

(a) HateMM (b) MoRE (c) MF (d) RAMF

Figure 7: UMAP (McInnes et al., 2020) space visualisations across different models (HXP/CLAP/CLIP). Class 0: Non-hate. Class 1: Hate.

framework effectively captures local and global spatio-temporal features without relying on conventional sequential processing constraints.

4.4 Hyperparameter Analysis

We analyse the sensitivity of RAMF to key hyperparameters, including the number of attention heads and convolutional kernel sizes in both SCA and LGCF modules (Figure 6). The results demonstrate that RAMF is relatively insensitive to moderate changes in these settings. The performance across different configurations remains stable, indicating the robustness of the proposed architecture.

4.5 Qualitative Analysis

Figure 7 visualises the feature space. Compared with prior methods, the boundaries between hateful and non-hateful samples in the baseline feature space are blurred, whereas the distribution of RAMF embeddings is more compact and better separated.

To demonstrate the advantage of our adversarial reasoning design, Figure 8 presents a qualitative comparison between CoT reasoning and RAMF. As shown on the left, CoT produces long but weakly grounded

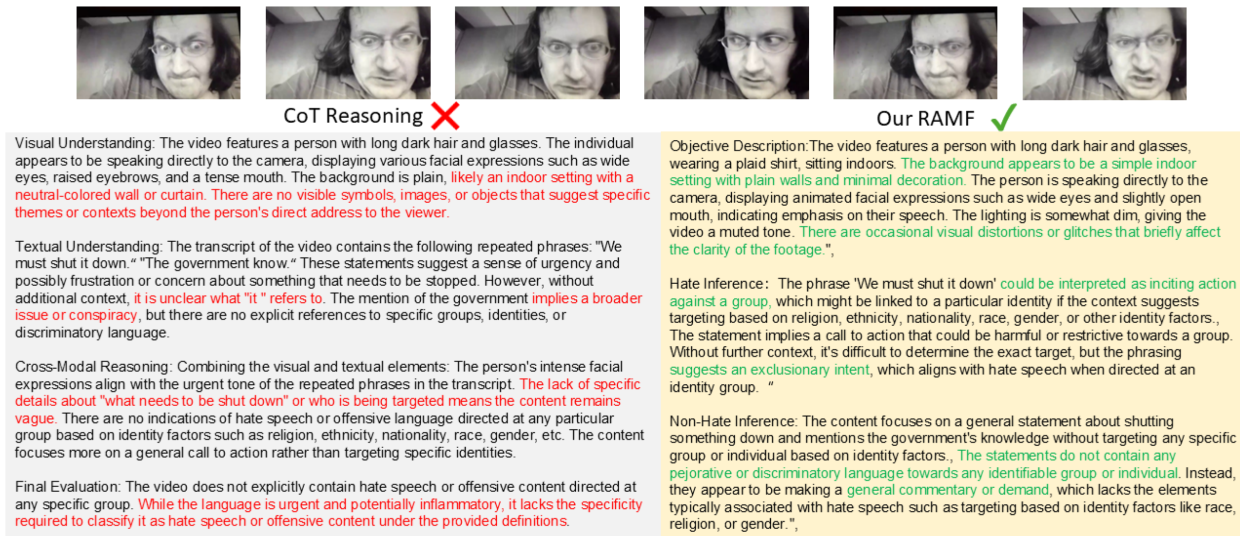


Figure 8: Comparison of reasoning strategies

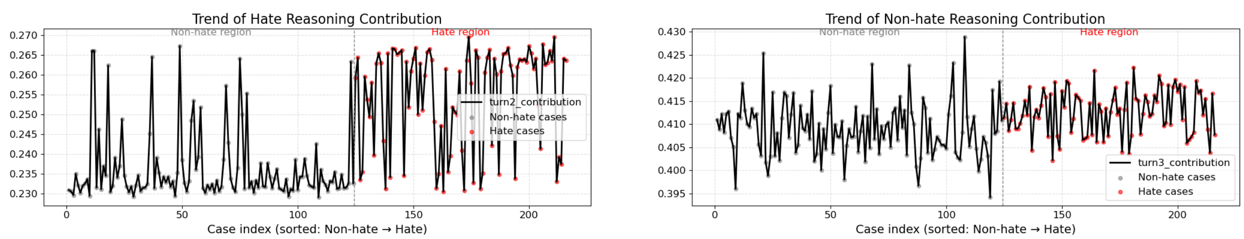


Figure 9: Trends of hate and non-hate reasoning contributions across the HateMM dataset. The left panel shows the variation in hate reasoning contribution, while the right panel illustrates the non-hate reasoning contribution. The dataset is sorted from non-hate to hate instances, with the right-side region corresponding to hate cases.

explanations that rely heavily on speculative observations and the absence of explicit hateful cues. It fails to resolve ambiguous textual references (e.g., "it"), and its cross-modal analysis remains superficial, offering no mechanism to complete missing contextual information. In contrast, RAMF generates structured, adversarial reasoning that systematically explores both hateful and non-hateful interpretations. The objective description provides a neutral, hallucination-free account of the visual scene, while the hate-assumed and non-hate–assumed inferences offer complementary semantic hypotheses. This adversarial setup forces the model to surface potential identity-targeting implications (when present) and, equally importantly, to articulate legitimate non-hateful explanations when the evidence supports neutrality. As reflected in the figure, RAMF not only resolves ambiguous textual cues but also supplies balanced, evidence-grounded reasoning, enabling more reliable intent interpretation and reducing false positives driven by surface-level correlations.

To further analyse RAMF's behaviour across different types of instances, we plot the trend of reasoning contributions in Figure 9. The upper panel presents the hate reasoning contribution, showing a clear increase when entering the hate-case region (right side of the plot). In contrast, the lower panel displays the non-hate reasoning contribution, which remains relatively stable across non-hate cases and slightly decreases in the hate region. This trend demonstrates that RAMF adaptively shifts the focus of its adversarial reasoning depending on whether the input contains hateful intent, aligning well with the expected behaviour of a robust hate-speech detection model.

Figure 10 demonstrates RAMF's ability to interpret nuanced multimodal cues and accurately infer intent across diverse scenarios. Whether hateful signals appear subtly in text, are embedded within neutral or

	Case 1	Case 2	Case 3
			
Visual	Static frames showing a vintage Columbia record label; no explicit hateful visual cues.	A speaker on a dim stage giving a talk; no aggressive or harmful actions depicted.	Black-and-white surveillance-style video showing a confrontation outside a house; tense scene with multiple individuals interacting.
Text	The song title contains a highly offensive racial slur, representing a *subtle but strong hate cue* that requires precise text understanding.	Mentions historically sensitive groups (e.g., Jewish communities), but used in *informational or analytical context*, not in a derogatory way.	Overlay text includes hateful and racially derogatory comments added in post-production, providing strong hate cues.
Audio	No meaningful speech content; audio is non-informative for classification.	Academic lecture tone; calm delivery with no hateful intent conveyed through prosody.	Urgent shouting and commands suggesting distress, aligning with the heated scenario.
Ground True	Hate	Non-Hate	Hate
Baseline	✗	✗	✗
RAMF	✓	✓	✓

Figure 10: Representative cases comparing the baseline HateMM model and our proposed RAMF.

analytical discussions, or are mixed with noisy or ambiguous visuals, RAMF consistently aligns linguistic, visual, and contextual information to recover the correct meaning. These examples highlight RAMF's strength in capturing fine-grained semantics and context-dependent cues, enabling robust and intent-aware understanding of complex video content.

5 Conclusion

In this work, to tackle the challenges of nuanced context understanding and multimodal semantic fusion in hateful video detection, we propose a novel Reasoning-Aware Multimodal Fusion framework. This framework consists of two core components: (1) Adversarial Reasoning, which generates complementary hate/non-hate perspectives through a structured three-stage VLM process, providing contextually grounded semantic information that avoids the limitations of direct reasoning and CoT approaches. (2) A novel fusion mechanism comprising Local-Global Context Fusion that captures both local salient cues and global temporal structures, and Semantic Cross-Attention that enables fine-grained multimodal semantic interaction. Extensive experiments on two benchmarks demonstrate the strong detection ability and generalizability of RAMF, providing a promising solution to context-aware hateful video detection.

Broader Impact Statement

This work aims to improve the detection of hateful content in multimodal videos, contributing to safer online platforms and more effective content moderation. By enhancing context understanding and multimodal reasoning, the proposed framework may help reduce the spread of harmful speech and protect vulnerable communities. However, automated detection systems may still produce false positives or inherit biases from underlying models, potentially affecting fair moderation. We emphasise that such systems should be deployed with human oversight, bias evaluation, and transparent governance to ensure responsible use.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. URL https://arxiv.org/abs/2103.15691.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. URL https://arxiv.org/abs/1803.01271.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v2, 2025.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 71–80, 2012. doi: 10.1109/SocialCom-PASSAT.2012.55.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2), March 2020. ISSN 1533-5399. doi: 10.1145/3377323. URL https://doi.org/10.1145/3377323.

Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. Mm-hsd: Multi-modal hate speech detection in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 2546–2555. ACM, October 2025. doi: 10.1145/3746027.3754558. URL http://dx.doi.org/10.1145/3746027.3754558.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 03 2017. doi: 10.1609/icwsm.v11i1.14955.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations, 2023. URL https://arxiv.org/abs/2309.05767.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(90)90002-E. URL https://www.sciencedirect.com/science/article/pii/036402139090002E.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Multi-token attention, 2025. URL https://arxiv.org/abs/2504.00927.

Ming Shan Hee and Roy Ka-Wei Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions, 2025. URL https://arxiv.org/abs/2502.11073.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. Recent advances in online hate speech moderation: Multimodality and the role of large models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4407–4419, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.254. URL https://aclanthology.org/2024.findings-emnlp.254/.

Ming Shan Hee, Zihan Gao, Yinglong Wang, Xiangxiang Chu, Roy Ka-Wei Lee, and Zengchang Qin. Contrastive instruction fine-tuning large multimodal model for hateful meme classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):760–773, Jun. 2025. doi: 10.1609/icwsm.v19i1.35844. URL https://ojs.aaai.org/index.php/ICWSM/article/view/35844.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Girish A. Koushik, Diptesh Kanojia, and Helen Treharne. 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*, WWW '25, pp. 2014–2023, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3718382. URL https://doi.org/10.1145/3701716.3718382.

Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 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*, WWW '25, pp. 2763–2774, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714560. URL https://doi.org/10.1145/3696410.3714560.

Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'89, pp. 396–404, Cambridge, MA, USA, 1989. MIT Press.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. Attentive fusion: A transformer-based approach to multimodal hate speech detection, 2024. URL https://arxiv.org/abs/2401.10653.

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

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. A system to monitor cyberbullying based on message classification and social network analysis. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the Third Workshop on Abusive Language Online*, pp. 105–110, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3511. URL https://aclanthology.org/W19-3511/.

Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink, 2022. URL https://arxiv.org/abs/2204.05149.

Qwen. Qwen2.5 technical report, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Dr. Nagendra Kumar. 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 (Volume 1: Long Papers)*, pp. 17209–17221. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.842. URL http://dx.doi.org/10.18653/v1/2025.acl-long.842.

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, and Animesh Mukherjee. Exploring the limits of zero shot vision language models for hate meme detection: The vulnerabilities and their interpretations. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1669–1689, Jun. 2025. doi: 10.1609/icwsm.v19i1.35894. URL https://ojs.aaai.org/index.php/ICWSM/article/view/35894.

Adi Robertson. Facebook ranks worst for online harassment, according to a global activist survey. *The Verge*, 2025. URL https://www.theverge.com/news/713976/online-harassment-meta-social-media-environmental-activists. Accessed: 2025-07-31.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2025. URL https://arxiv.org/abs/2312.17432.

Mark Townsend. Uk a 'powder keg' of social tensions a year on from summer riots, report warns. *The Guardian*, 2025. URL https://www.theguardian.com/uk-news/2025/jul/15/social-tensions-british-people-polarisation-research. Accessed: 2025-07-31.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Neeraj Vashistha and Arkaitz Zubiaga. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1), 2021. ISSN 2078-2489. doi: 10.3390/info12010005. URL https://www.mdpi.com/2078-2489/12/1/5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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

Han Wang, Rui Yang Tan, and Roy Ka-Wei Lee. Cross-modal transfer from memes to videos: Addressing data scarcity in hateful video detection. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 5255–5263. ACM, April 2025a. doi: 10.1145/3696410.3714534. URL http://dx.doi.org/10.1145/3696410.3714534.

Han Wang, Zhuoran Wang, and Roy Ka-Wei Lee. Hateclipseg: A segment-level annotated dataset for fine-grained hate video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 13304–13310. ACM, October 2025b. doi: 10.1145/3746027.3758289. URL http://dx.doi.org/10.1145/3746027.3758289.

Shuonan Yang, Tailin Chen, Rahul Singh, Jiangbei Yue, Jianbo Jiao, and Zeyu Fu. Revealing temporal label noise in multimodal hateful video classification. In *Proceedings of the 4th International Workshop on Multimodal Human Understanding for the Web and Social Media*, MUWS '25, pp. 26–34, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718380. doi: 10.1145/3728481.3762164. URL https://doi.org/10.1145/3728481.3762164.

Jiangbei Yue, Shuonan Yang, Tailin Chen, Jianbo Jiao, and Zeyu Fu. Multimodal hate detection using dual-stream graph neural networks. *ArXiv*, abs/2509.13515, 2025. URL https://api.semanticscholar.org/CorpusID:281333169.

Yinghui Zhang, Tailin Chen, Yuchen Zhang, and Zeyu Fu. Enhanced Multimodal Hate Video Detection via Channel-wise and Modality-wise Fusion . In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 183–190, Los Alamitos, CA, USA, December 2024. IEEE Computer Society. doi: 10.1109/ICDMW65004.2024.00030. URL https://doi.ieeecomputersociety.org/10.1109/ICDMW65004.2024.00030.

A VLM Inference Experiment

A.1 Large Language Models (Text-only)

Each English or Chinese transcript is sent to a Large Language Model (LLM) with the fixed prompt below: *“Please determine whether the following English/Chinese text contains hateful content. If it contains hateful content, please return 1; if it does not contain hateful content, please return 0. Just return 1 or 0, no other words. Here is the following text:* {text}"

A.2 Vision–Language Models (Image + Text)

For multimodal inference, we uniformly extracted five frames from each video. Although the Qwen-VL-Max (Qwen, 2024) model supports more than five images, the Llama4 (Patterson et al., 2022) recommends a maximum of five images. To maintain consistency, we have standardised the input to five images. All five frames, in chronological order, are provided together with the transcript through the following prompt: *“Please analyse both the video frames and the following text to determine if they contain hateful content. If contain hateful content, please return 1; if not contain hateful content, please return 0. Just return 1 or 0, no other words. Here is the text:* {text} *"*

The placeholder *{text}* is replaced by the raw transcript. The model therefore performs binary classification in a strict zero-shot setting and must output exactly “1” or “0”.

A.3 Implementation Details of Chain-of-Thought

Chain-of-Thought (CoT) prompting has become a widely used approach to improve reasoning quality in complex tasks. In this work, we design a structured CoT prompt to guide the model through a stepwise analysis of potentially harmful video content, as shown in Figure 11. The prompt explicitly breaks down the task into four stages—visual understanding, textual analysis, multimodal reasoning, and final evaluation—encouraging the model to consider different aspects of the content systematically before making a judgment.

This is a video that may contain harmful content such as hate speech, explicit violence, discrimination, or other offensive behavior. You are a content moderation expert. Analyze this video using a reasoning process. The video is represented by: Visual frames and Transcript text (could be none).

HATE SPEECH DEFINITION: Any communication that attacks or uses pejorative/discriminatory language targeting a person or group based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. OFFENSIVE CONTENT: Contains hurtful, derogatory, or socially inappropriate content, even if it does not constitute direct hate speech. IMPORTANT: Simple conflicts, arguments, insults, or violent behaviour between individuals may not constitute hate speech.

Please reason in four steps, then summarize your final judgment.
Step 1: Visual Understanding. Describe the visual content in the frames. Focus on characters, scenes, and potential symbolic or hateful imagery.
Step 2: Textual Understanding. Analyze the transcript of the video. Pay attention to metaphors, puns, homophones, or rhetorical devices that may express hate or implicit bias.
Step 3: Multimodal Reasoning. Integrate insights from the visuals and the transcript. Identify any \text{implicit meanings}, contradictions, or cross-modal cues that enhance or modify the hateful nature of the content.
Step 4: Final Evaluation. Does this video express harmful or hateful content? Explain in 1-2 sentences why. Keep the answer concise, structured (following these four steps), and professional.

Figure 11: Chain-of-Thought prompt.

A.4 Implementation Details of Adversarial Reasoning

Figure 12 presents the full prompt template used to implement the adversarial reasoning procedure. The prompt is explicitly structured into three sequential stages: (1) objective visual description, which restricts the model to observable entities, actions, and textual content without interpretation; (2) hate-assumed inference, where the model is instructed to identify potential hateful or offensive signals under the assumption that such content is present; and (3) non-hate-assumed inference, which enforces an alternative interpretation assuming the absence of hate speech. Each stage defines clear task boundaries and output constraints, enabling consistent and reproducible generation of reasoning text that can be directly encoded and fused with multimodal features.

Step 1: Objective Description
You are a content analysis expert. Provide a detailed, objective description of this video content.
Task: Describe what you observe in the video frames and transcript:
1. objective_visual_description: Describe what you see - objects, people, actions, settings, text, symbols, colors, expressions, gestures, etc.
Rules:
Be comprehensive and detailed;
NO analysis, interpretations, or value judgments;
Simply report what you observe;

Step 2: Hate-assumed Inferences
You are a hateful content detection specialist. Analyse the video and transcript content under the ASSUMPTION that it contains hateful or offensive content.
HATE SPEECH DEFINITION: Any communication that attacks or uses pejorative/discriminatory language targeting a person or group based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
OFFENSIVE CONTENT: Contains hurtful, derogatory, or socially inappropriate content, even if it does not constitute direct hate speech.
IMPORTANT: Simple conflicts, arguments, insults, or violent behaviour between individuals may not constitute hate speech.
Task: Identify evidence supporting hateful(offensive) interpretation:
1. evidence: Describe specific evidence supporting the hateful nature of the content 2. reasoning: Explain why it is hateful content.
Focus on:
- Language targeting religion, ethnicity, race, gender, nationality, etc.
- Group-based dehumanizing language or symbols
- Identity-based discriminatory attacks (not personal insults)
- Incitement against identity groups

Step 3: Non-hate-assumed Inferences
You are a content analyst. Analyze this content under the ASSUMPTION that it does NOT constitute hate speech and offensive content.
HATE SPEECH DEFINITION: Any communication that attacks or uses pejorative/discriminatory language targeting a person or group based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
OFFENSIVE CONTENT: Contains hurtful, derogatory, or socially inappropriate content, even if it does not constitute direct hate speech.
IMPORTANT: Simple conflicts, arguments, insults, or violent behaviour between individuals may not constitute hate speech.
Task: Identify evidence supporting non-hate(non-offensive) interpretation:
1. evidence: Describe specific evidence supporting the non-hateful nature of the content
2. reasoning: Explain why this content does not hateful
Consider:
- Is this a personal dispute rather than group targeting?
- Are insults directed at individuals rather than identity groups?
- Is there artistic, satirical, or educational context?
- Does the content lack group-based discriminatory language?

Figure 12: Adversarial reasoning prompt.

A.5 Model Versions

Table 3: Model used in zero-shot evaluation.

Category	Model identifier
LLM	GPT-4o-2024-05-13
	Qwen-Max-2025-01-25
	Llama-3.1-405B-Instruct
VLM	Qwen-VL-Max-2025-01-25
	Llama-4-Maverick-17B-128E-Instruct

Table 3 lists the models evaluated in this study, covering both large language models (LLMs) and vision-language models (VLMs). The LLMs include GPT-4o (OpenAI, 2024), Qwen-Max(Qwen, 2024), and Llama-3.1 (Patterson et al., 2022), while the VLMs include Qwen-VL-max (Qwen, 2024) and Llama-4-Maverick (Patterson et al., 2022). GPT-4o inferences were conducted via the official OpenAI API, and all Qwen and Llama variants were accessed through Alibaba Cloud's generative-AI service.

B End-to-End Implementation of the RAMF Framework

Algorithm 1 summarises the end-to-end training and inference procedure of the proposed RAMF framework. For each input video instance, the pipeline first invokes a vision–language model to generate three types of reasoning texts—objective description, hate-assumed inference, and non-hate-assumed inference—following the predefined adversarial prompting strategy. Multimodal features are then extracted independently from text, audio, video, and reasoning outputs using modality-specific encoders and unified through lightweight MLP projections. The fused representations are processed by the Local–Global Context Fusion module to capture complementary temporal cues, followed by a hierarchical Semantic Cross Attention mechanism that integrates both low-level modalities and high-level reasoning signals. Finally, the aggregated representation is passed to a classifier to produce the hate/non-hate prediction.

Algorithm 1 Training of RAMF for hateful video detection.

Input: The hateful video dataset $\mathcal{S} = \{S_1, \cdots, S_N\}$.
Output: Predicted category $\hat{y}$ (Hate or Non-hate).

for each instance S_i in $\mathcal{S}$ **do**
  /* *VLM Adversarial Reasoning* */
  Generate objective description T_O using VLM with video frames and transcript of S_i.
  Generate hate-assumed inference T_H under the assumption that the content contains hate speech.
  Generate non-hate-assumed inference T_N under the assumption that content does not contain hate speech.
  /* *Feature Extraction* */
  Extract features X^m, $m \in \{a, v, t\}$ from audio, video, and text modalities of S_i using MFCC/CLAP, ViT/CLIP, and BERT/HXP, respectively.
  Encode reasoning texts T_O, T_H, T_N using text encoder to obtain X^{T_O}, X^{T_H}, X^{T_N}.
  /* *Local-Global Context Fusion (LGCF)* */
  for $m \in \{t, a, v, T_O\}$ **do**
    Apply MLP to X^m to obtain unified representation X^m_{MLP}.
    Extract local features: $v_{\text{local}} = \text{MaxPool1D}(\text{Conv1D}(X^m_{\text{MLP}}))$.
    Extract global features: $v_{\text{global}} = \text{AdapAvgPool1D}(X^m_{\text{MLP}})$.
    Compute gating weight $g = \sigma(W[v_{\text{local}} \oplus v_{\text{global}}] + b)$.
    Fuse features: $Z^m = g \odot v_{\text{local}} + (1 - g) \odot v_{\text{global}}$.
  end for
  /* *First-Layer Semantic Cross Attention (SCA)* */
  Stack the modality features: $Z_s^{(1)} = [Z^t, Z^a, Z^v, Z^{T_O}]$.
  Apply SCA with Cross-Head Convolution (CHC) and Structural Mixing Convolution (SMC) to obtain Y_1.
  /* *Second-Layer Semantic Cross Attention* */
  Encode adversarial reasoning into feature space and stack: $Z_s^{(2)} = [Y_1, X^{T_H}, X^{T_N}]$.
  Apply SCA to $Z_s^{(2)}$ to obtain final representation Y_2.
  /* *Classification* */
  Apply average pooling to Y_2: $Y_{\text{pool}} = \text{AvgPool}(Y_2)$.
  Feed Y_{pool} into MLP classifier to obtain prediction $\hat{y}_i$.
end for

C Additional Experimental Details

We re-partition the five-fold dataset such that the test sets across all five folds are mutually exclusive—each test set contains a distinct subset of the data. While training and validation splits may partially overlap across folds, this design ensures that the model is evaluated on entirely different test data in each fold. This setup enables a more generalisable and comprehensive evaluation, as it avoids repeated testing on the same examples and better reflects performance under diverse data conditions.

For the HateMM dataset, the MF model was trained for 60 epochs with a batch size of 64. For the MultiHateClip (MHC) dataset, MF was trained for 20 epochs with a batch size of 32. For both datasets, the RAMF model was trained for 20 epochs using a batch size of 16.

Model training and testing were conducted on a laptop equipped with an Intel(R) Core(TM) i9-14900HX processor, 96 GB of system RAM, and an NVIDIA GeForce RTX 4090 Laptop GPU with 16 GB of VRAM. A fixed random seed of 2021 was used across all experiments, following the configuration reported in the original HateMM implementation. For CoT and AR experiments, use L40 46GB GPU memory experiments.

Importantly, some results reported in our experiments deviate from those in the original publications. This discrepancy primarily arises from our re-partitioning of the five-fold datasets, and this modification enables a more generalisable and comprehensive evaluation. In contrast, previous experimental work only fixed the data set division and changed the random seed five times (Lang et al., 2025; Wang et al., 2024).

Table 4: Performance comparison of all models across three datasets (HXP/CLAP/CLIP). MHC(C): MultiHateClip Chinese dataset. MHC(E): MultiHateClip English dataset.

Dataset	Model	Accuracy	Macro F1	F1 (H)	Precision (H)	Recall (H)
HateMM	HateMM	0.8300 ± 0.0264	0.8243 ± 0.0248	0.7964 ± 0.0256	0.7941 ± 0.0412	0.8025 ± 0.0519
	CMFusion	0.8287 ± 0.0203	0.8195 ± 0.0210	0.7790 ± 0.0255	0.8029 ± 0.0263	0.7580 ± 0.0417
	MoRE	0.8290 ± 0.0245	0.8210 ± 0.0212	0.7830 ± 0.0224	0.7700 ± 0.0345	0.7970 ± 0.0567
	MF	0.8370 ± 0.0352	0.8325 ± 0.0333	0.8063 ± 0.0294	0.7752 ± 0.0601	0.8434 ± 0.0290
	RAMF	**0.8556 ± 0.0183**	**0.8507 ± 0.0199**	**0.8246 ± 0.0266**	**0.7978 ± 0.0190**	**0.8547 ± 0.0540**
MHC(C)	MHC	0.6903 ± 0.0450	0.6327 ± 0.0618	0.4898 ± 0.1052	0.5427 ± 0.0750	0.4727 ± 0.1569
	CMFusion	0.6674 ± 0.0332	0.6064 ± 0.0388	0.4533 ± 0.0635	0.4901 ± 0.0474	0.4243 ± 0.0777
	MoRE	0.6910 ± 0.0434	0.6250 ± 0.0451	0.4688 ± 0.0845	0.5410 ± 0.0712	0.4120 ± 0.1211
	MF	0.7029 ± 0.0411	0.6624 ± 0.0445	0.5469 ± 0.0635	0.5523 ± 0.0811	0.5491 ± 0.0848
	RAMF	**0.7446 ± 0.0403**	**0.7096 ± 0.0434**	**0.6126 ± 0.0677**	**0.6128 ± 0.0647**	**0.6224 ± 0.1151**
MHC(E)	MHC	0.7233 ± 0.0215	0.6843 ± 0.0473	0.5784 ± 0.0976	0.5911 ± 0.0590	0.5712 ± 0.1296
	CMFusion	0.7171 ± 0.0229	0.6770 ± 0.0270	0.5663 ± 0.0535	0.6059 ± 0.0663	0.5514 ± 0.1065
	MoRE	0.7250 ± 0.0312	0.6740 ± 0.0353	0.5450 ± 0.0645	0.6110 ± 0.0856	0.4930 ± 0.1345
	MF	0.7335 ± 0.0336	0.6960 ± 0.0435	0.5910 ± 0.0712	0.6209 ± 0.0788	0.5682 ± 0.0903
	RAMF	**0.7398 ± 0.0418**	**0.7166 ± 0.0430**	**0.6376 ± 0.0572**	**0.6213 ± 0.0934**	**0.6740 ± 0.1109**

During inference, the language model was configured with a maximum of 2048 new tokens, temperature set to 0.7, top-p sampling with a threshold of 0.9, and sampling enabled. The pad token ID was set to the end-of-sequence token from the tokenizer. These settings were applied consistently across reasoning tasks, including CoT and AR, to ensure coherent yet diverse outputs.

D Additional Results

Table 4 presents the performance comparison of all evaluated models across three datasets, reporting the mean and standard deviation over multiple runs. The results are computed under the same experimental settings as in the main experiments, using HXP, CLAP, and CLIP features. This table provides a more detailed view of performance variability and serves as a supplementary reference to the main results.

D.1 Efficiency Analysis

Table 5 presents the computational efficiency of RAMF. Compared to HateMM, RAMF increases parameters from 1.36M to 3.78M and FLOPs from 2.18G to 3.10G, with inference latency rising from 0.025ms to 0.20ms per sample. This moderate overhead is justified by the 5% macro-F1 improvement (79.3% → 84.3%). Replacing SCA with MTA slightly reduces parameters (3.56M) but causes a 1.05% macro-F1 drop; substituting with standard attention increases parameters (4.61M) while yielding inferior performance (82.46%). These results validate that SCA achieves an optimal efficiency-effectiveness balance.

Table 5: Efficiency and performance comparison in the HateMM dataset. "Params" and "FLOPs" refer to the number of trainable parameters and the computational cost per forward pass, respectively. "Time" denotes the average per-sample inference latency. '→' means 'replaced by'.

Model	Params	FLOPs	Time	MF1(%)
HateMM	1.36M	2.18G	0.025ms	79.3
RAMF	3.78M	3.10G	0.20ms	84.3
SCA→MTA	3.56M	2.74G	0.30ms	83.25
SCA→StdAttn	4.61M	2.80G	0.17ms	82.46

E Comprehensive Failure Case Analysis

Figure 13 and 14 present representative failure cases of the proposed RAMF framework. For each case, we provide a modality-wise breakdown of visual, textual, and audio cues, together with an analysis explaining why the model fails under these specific conditions. The cases illustrate remaining challenges for reasoning-aware multimodal fusion, particularly in scenarios involving ambiguous contextual signals, conflicting modality cues, or nuanced pragmatic language use. This analysis complements the quantitative results by highlighting the limitations of the proposed approach.

	Case 1
Ground True: Hate	
Visual	The video consists of a sequence of scenes from a video game environment, depicting characters gathering near a pickup truck, driving through urban and industrial areas, and engaging in combat-related actions such as running, aiming, and shooting. The visuals primarily convey action-oriented gameplay elements typical of crime or shooter-themed games. No explicit hateful symbols, gestures, or identity-targeting visual cues are present throughout the video.
Text	The transcribed text mainly describe narrative actions and situational dialogue related to movement, preparation, and confrontation (e.g., characters being ready to go, pointing weapons, or coordinating actions). The language does not contain explicit slurs or direct attacks against protected groups. However, certain phrases can implicitly allude to violence or aggression without clearly specifying an identity-based target, making the hateful intent highly context-dependent and difficult to infer from surface-level textual semantics alone.
Audio	The background audio features a lively and upbeat background music (BGM), creating a cheerful and energetic atmosphere that contrasts sharply with the violent actions depicted in the video. No explicit hateful speech or discriminatory verbal expressions are present in the audio channel. The positive emotional tone of the music acts as a strong neutralizing signal, potentially misleading the model into interpreting the content as entertainment-oriented rather than harmful.
Analysis	This case illustrates a failure scenario where the baseline model incorrectly classifies the video as non-hateful due to the absence of explicit hate cues in any single modality. The visual content appears as generic video game violence, the textual modality lacks direct identity-based attacks, and the audio modality introduces a strong positive emotional bias through cheerful background music. The multimodal inconsistency—violent actions paired with upbeat audio—dilutes the perceived severity of the content and obscures potential implicit hateful intent. Without deeper contextual reasoning to reconcile these conflicting signals, the baseline model overly relies on surface-level modality cues and fails to capture the nuanced, implicit nature of hate in this scenario. This highlights the limitation of standard multimodal fusion approaches when faced with subtle or masked hate expressions, especially in cases where audio or visual styles distract from underlying harmful semantics.

Figure 13: Representative failure case 1.

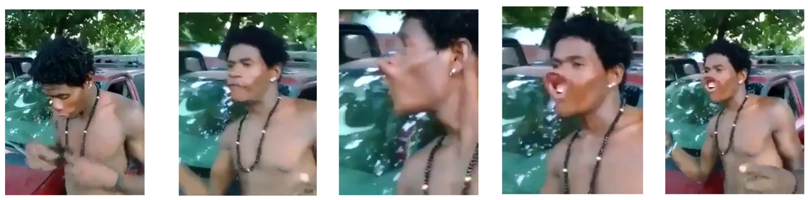
	Case 2
Ground True: Non hate	
Visual	The video shows a shirtless Black individual standing outdoors in what appears to be a parking lot, with cars and greenery visible in the background. The individual wears a beaded necklace and uses exaggerated facial expressions and playful gestures, including puckering lips and animated hand movements. Surrounding people are visible reacting with smiles and laughter.
Text	The transcript contains the phrase "you nigger", which is a highly offensive racial slur when interpreted in isolation. Due to its strong lexical association with hate speech, the presence of this term dominates the textual signal and triggers hate detection mechanisms.
Audio	The speaker's tone is energetic and playful, accompanied by laughter from people nearby. The laughter appears to be a natural response to the individual's exaggerated expressions and humorous behavior, rather than mockery or derision. No aggressive shouting, hostile intonation, or emotionally negative prosody is present.
Analysis	This case represents a false positive caused by over-reliance on surface-level textual cues without sufficient integration of cultural and contextual information across modalities. While the transcript includes a highly offensive racial slur, the visual and audio signals strongly indicate a playful, in-group interaction characterized by humor, exaggerated expression, and mutual amusement. The laughter and relaxed body language suggest social bonding rather than ridicule or hate. The baseline model fails to account for reclaimed or intra-group language usage and lacks pragmatic reasoning to reconcile the contradiction between explicit textual signals and non-hostile visual–audio context. This highlights a key limitation of current hate detection systems: lexical sensitivity without contextual grounding can lead to misclassification, particularly in cases involving culturally specific language use and expressive, non-adversarial social interactions.

Figure 14: Representative failure case 2.