Biting Off More Than You Can Detect: Retrieval-Augmented Multimodal Experts for Short Video Hate Detection

Jian Lang
jian_lang@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Rongpei Hong
rongpei.hong@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Jin Xu
jin.xu@mu.ie
Maynooth University
Maynooth, County Kildare, Ireland

Yili Li
lylxzr@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Xovee Xu
xovee.xu@gmail.com
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Fan Zhou*
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Abstract

Short Video Hate Detection (SVHD) is increasingly vital as hateful content — such as racial and gender-based discrimination — spreads rapidly across platforms like TikTok, YouTube Shorts, and Instagram Reels. Existing approaches face significant challenges: hate expressions continuously evolve, hateful signals are dispersed across multiple modalities (audio, text, and vision), and the contribution of each modality varies across different hate content. To address these issues, we introduce **MoRE** (**M**ixture **o**f **R**etrieval-augmented multimodal **E**xperts), a novel framework designed to enhance SVHD. MoRE employs specialized multimodal experts for each modality, leveraging their unique strengths to identify hateful content effectively. To ensure model's adaptability to rapidly evolving hate content, MoRE leverages contextual knowledge extracted from relevant instances retrieved by a powerful joint multimodal video retriever for each target short video. Moreover, a dynamic sample-sensitive integration network adaptively adjusts the importance of each modality on a per-sample basis, optimizing the detection process by prioritizing the most informative modalities for each instance. Our MoRE adopts an end-to-end training strategy that jointly optimizes both expert networks and the overall framework, resulting in nearly a twofold improvement in training efficiency, which in turn enhances its applicability to real-world scenarios. Extensive experiments on three benchmarks demonstrate that MoRE surpasses state-of-the-art baselines, achieving an average improvement of 6.91% in macro-F1 score across all datasets.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '25, Sydney, NSW, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714560

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Computer vision**.

Keywords

Short video hate detection, retrieval augmentation, mixture of multimodal experts.

ACM Reference Format:
Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 2025. Biting Off More Than You Can Detect: Retrieval-Augmented Multimodal Experts for Short Video Hate Detection. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3696410.3714560

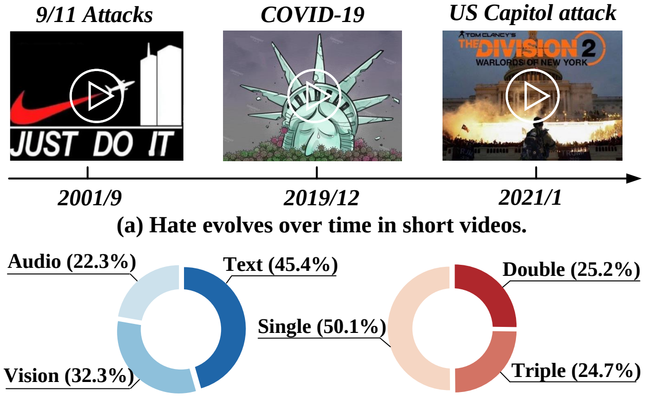
1 Introduction

Media consumption trends have increasingly shifted toward short videos, particularly on platforms like TikTok, YouTube Shorts, and Instagram Reels [12, 69, 75]. As a dynamic and immersive communication medium, short video can significantly boost user engagement and capture a larger share of daily screen time [2, 6, 13, 32, 74]. These videos seamlessly integrate diverse media modalities – such as audio, text, and vision – to convey information, exerting more substantial effects on mental health and social cohesion than content confined to a single modality.

However, this multimodal integration also enables the subtle and covert dissemination of hateful content[1], embedding harmful messages across various media forms. Hateful content in short videos often targets attributes like race, gender, or religion [10, 25, 45, 58, 66, 67] and can manifest through multiple modalities. Moreover, the prevalence of hateful content varies across modalities, with each contributing uniquely to its overall impact. The continual evolution of hateful content – driven by shifting social issues and advancements in tools for AI generated content (e.g., OpenAI's Sora [76]) – underscores the pressing need for highly effective methods to tackle the task of Short Video Hate Detection (SVHD).

Hateful content detection has been extensively studied in literature [1, 7–9, 15, 33, 41, 46, 49]. The majority of these works focus on

[1] **Disclaimer**: *This paper contains discussions of violence and discriminatory content that may be disturbing to some readers.*



(a) Hate evolves over time in short videos.

(b) Distribution of hateful content across different modalities.

Fig. 1: Illustration of motivation. (a): As new social events emerge, the expressions of hateful content undergo constant evolution. (b): Multimodal distribution of hateful content in the MHClip-B dataset [63]. The blue donut chart illustrates the distribution of hateful content across different modalities – audio, text, and vision. The red donut chart depicts the proportions of short videos that contain hateful content in one, two, or all three modalities.

text-based analyses [1, 15, 33, 49] within microblogging platforms such as Twitter and Facebook. With the increasing integration of images in social media posts and the advancements in image processing technologies, researchers have expanded their efforts to identify hateful elements in text-image posts and memes [7–9, 41, 46], utilizing pre-trained models and incorporating task-specific classification layers. However, despite the rapid rise in the popularity of short videos, research on hate detection in short videos remains very limited [14, 63]. Short videos encompass multiple modalities, which can subtly and covertly facilitate the dissemination of hateful content. In addition, the prevalence of hateful content varies across these modalities in short videos, which necessitates a dynamic and adaptable detection framework that can effectively identify hateful content across diverse modalities. Moreover, as hateful content is subject to continuous evolution, developing an effective and robust framework for SVHD entails addressing several significant challenges, which are summarized as follows:

Challenge 1: Adapting to the Evolution of Hateful Content. Hateful content continuously evolves in response to societal shifts, becoming more subtle and increasingly difficult to detect. Fig. 1(a) illustrates an evolution example through three short videos. Initially, hate expressions employed imagery related to the 9/11 attacks to overtly criticize terrorism in the USA. Subsequently, during the COVID-19 pandemic, more nuanced and veiled content emerged, satirizing the response of American society. More recently, a combination of video game imagery and photos from the US Capitol attack has been utilized to critique American politics. This progression underscores the adaptive nature of hateful content over time. Consequently, it is imperative to develop detection frameworks that remain current and can generalize across increasingly evolving forms of hate in short videos.

Challenge 2: Harnessing Multiple Modalities for Hateful Content Analysis. Short videos encompass multimodal information such as audio, text, and visual content. Effectively utilizing data from different modalities for hate detection poses a significant challenge. The left side of Fig. 1(b) shows the modality-wise distribution of hateful content in the MHClip-B [63] dataset, highlighting that each modality contributes essential information for detecting hateful content, which can manifest in various forms. For instance, hate speech may be embedded in textual overlays, discriminatory lyrics may be presented in background music, and offensive gestures may appear in visual streams. Therefore, it is critical to develop a multimodal framework that can effectively integrate all modalities to detect various types of hateful content in short videos.

Challenge 3: Managing Modality-Specific Influences in Hate Detection. Not all modalities in short videos contribute equally to hate detection; each modality plays a distinct role. As shown in the right of Fig. 1(b), 75.3% short videos in the MHClip-B dataset contain hateful content presented in only one or two modalities. This distribution suggests that indiscriminately integrating all modalities could be counterproductive. The detection model may overemphasize noisy or redundant information, misleading the learning process and degrading detection performance. Thus, focusing on the most informative modalities and content is crucial for accurate detection. A more adaptive and selective multimodal fusion approach is needed to dynamically adjust each modality's contribution at the sample level, ensuring more precise hate detection.

To address these challenges, we propose a novel **M**ixture **o**f **R**etrieval-augmented multimodal **E**xperts (**MoRE**) framework. It introduces contextual knowledge-augmented multimodal experts designed to well adapt the dynamic and evolving hateful content and effectively harnesses data dispersed across multiple modalities in short videos for detection (i.e., Challenges 1 & 2). First, a basic expert is developed to focus on individual modalities, including audio, text, and vision. To adapt to the evolving nature of hateful content – mimicking human learning processes [28, 29] – our model retrieves relevant information to deepen its understanding on specific topics. The basic expert is subsequently augmented with contextual knowledge retrieved via a powerful joint multimodal video retriever, which integrates audio, textual, and visual modalities for fine-grained video-to-video retrieval. By leveraging contextual knowledge from the retrieved videos, the experts remain aware of the evolving expressions of hate, thereby enhancing their capability to generalize to emerging forms of hateful content. These contextual knowledge-augmented multimodal experts not only improves the adaptability of MoRE to new hate expressions but also ensures more accurate and comprehensive detection of hateful content across multiple modalities.

To address the varying significance of each modality in hate detection for different short videos (Challenge 3), MoRE incorporates a novel sample-sensitive integration network. This network includes a modality-mixture soft router which identifies the specific contributions of each modality's features to hate detection in each video, prioritizing those with the most significant impact for each video sample. Consequently, the network accurately determines the contributions of different modalities at the sample level, enhancing detection performance and providing interpretability regarding the roles of various modalities in hate detection for each short video.

Additionally, instead of a traditional two-stage training process [5, 68, 71], we introduce a unified and effective end-to-end training paradigm. This paradigm jointly optimizes both the experts and the overall framework, providing a scalable and applicable solution for SVHD. In summary, the key contributions of this work are as follows:

- **Contextual Knowledge-Augmented Multimodal Experts:** We design several multimodal experts to better adapt to the continuously evolving nature of hateful content in short videos and harness the multiple modalities in hate detection. By retrieving relevant instances through a powerful joint multimodal video retriever, the experts acquire contextual knowledge that deepens their understanding of specific topics, enabling them to keep pace with the evolving expressions of hate in short videos.
- **Sample-Sensitive Integration Network:** We propose a novel adaptive integration network that evaluates the varying contributions of different modalities within individual video samples to improve the performance of hate detection. This adaptive integration network dynamically adjusts the influence of each modality, prioritizing those with the most significant impact on detecting hateful content, thereby ensuring more precise and effective detection.
- **Unified End-to-End Training Paradigm:** We develop an effective end-to-end training paradigm that significantly enhances the model's scalability and applicability, making the model highly suitable for deployment in large-scale SVHD applications.

Extensive experiments on three real-world short video datasets demonstrate that MoRE outperforms state-of-the-art baselines. Notably, our model achieves an average improvement of 6.91% in macro-F1 score across all three datasets. Furthermore, our model surpasses three popular Large Vision-Language Models (LVLMs), highlighting its effectiveness and efficiency for SVHD, even when compared to large models trained on trillions of tokens and billions of parameters. The source codes and data required to reproduce our results are available at https://github.com/Jian-Lang/MoRE.

2 Related Work

Early studies primarily focused on identifying hate speech within text-based materials. Traditional machine learning approaches, such as Support Vector Machines and Naive Bayes classifiers [34, 65], have been commonly used for detection. With the rise of deep learning, more advanced methods have been developed for hate speech detection [1, 49]. Subsequently, multimodal hate detection, which analyzes both textual and visual information in posts and memes, has made significant progress [7–9, 41, 47]. For example, Pro-Cap [7] leverages pre-trained models and prompting techniques to generate image captions that identify hateful content. However, despite their effectiveness, these approaches are not directly applicable to hate detection in videos. Unlike text-image posts or memes, videos consist of multiple frames and incorporate various modalities, making it unclear which modality carries the hateful message, thereby highly increasing detection complexity.

Research on video-based hate detection remains limited. Recent advancements include the introduction of benchmark datasets such as HateMM [14] and MHClip [63]. Although baseline detection models were provided, they simply fused audio, text, and visual features equally for prediction. This simple design undermines their effectiveness in SVHD, as it overlooks the dynamic nature of hateful content and the varying significance of each modality in detecting hate across different short videos. In contrast, our proposed MoRE first retrieves the most relevant instances to construct the contextual knowledge-augmented multimodal experts that adapt to the evolving nature of hateful content. Then, a sample-sensitive integration network adaptively assigns weights to these experts at the sample level, further enhancing the prediction accuracy of MoRE in detecting hateful content in short videos. Additional research related to the multimodal retrieval and the Mixture of Experts (MoE), is reviewed in Appendix A.

3 Methodology

Problem Statement. Let $\mathcal{S} = \{S_1, \cdots, S_N\}$ denote the set of short videos on video platforms, where N is the number of short videos. Each short video S_i is characterized by its multimodal content, including audio, textual, and visual content, expressed as $S_i = \{s_i^a, s_i^t, s_i^v\}$. The objective of SVHD is to determine whether a given short video S_i is **hateful** or **non-hateful** by considering all its modal contents s_i^a, s_i^t, and s_i^v.

Feature Extraction. The extracted features are summarized as follows: the audio features $\mathbf{x}_i^a \in \mathbb{R}^{l \times d_a}$, the visual features $\mathbf{x}_i^v \in \mathbb{R}^{m \times d_v}$, and the textual features $\mathbf{x}_i^t \in \mathbb{R}^{n \times d_t}$, where l is the number of audio frames, m is the number of key frames sampled from the video, and n represents the number of word tokens. d_a, d_t, and d_v are the feature dimensions for each modality. The detailed feature extraction process is provided in Appendix B.

Fig. 2 provides an overview of our proposed MoRE framework and illustrates the relationship among its core components. The following sections will delve into each component of MoRE, providing detailed explanations of their roles and interactions.

3.1 Joint Multimodal Video Retriever

To provide relevant instances to make our framework better adapt to the complex and evolving nature of hateful content, we design a novel joint multimodal video retriever, which simultaneously incorporates audio, textual, and visual features to perform video-to-video retrieval, moving beyond the limitations of unimodal retrieval methods that rely on a single modality. By jointly considering all modalities, our strategy enables the retrieval of instances associated with the target video from multiple perspectives, leading to significantly improved retrieval precision.

3.1.1 Memory Bank Construction. To store high-quality semantic information as prior knowledge, we define the memory bank $\mathcal{B}$, which encodes audio, textual, and visual content using a collection of (audio, text, vision) triples. Detailed description of the construction of memory bank $\mathcal{B}$ is provided in Appendix C.3.

3.1.2 Query Construction. To fully capture the unique characteristics of each modality, we first encode the audio, textual, and visual features independently. Specifically, for each short video S_i, we first extract its audio transcription using Whisper [54], a pre-trained automatic speech recognition model. The transcription is then processed by a pre-trained BERT [17] model to generate the audio query vector $\mathbf{r}_i^a \in \mathbb{R}^{d_a}$. For textual retrieval, we use the BERT

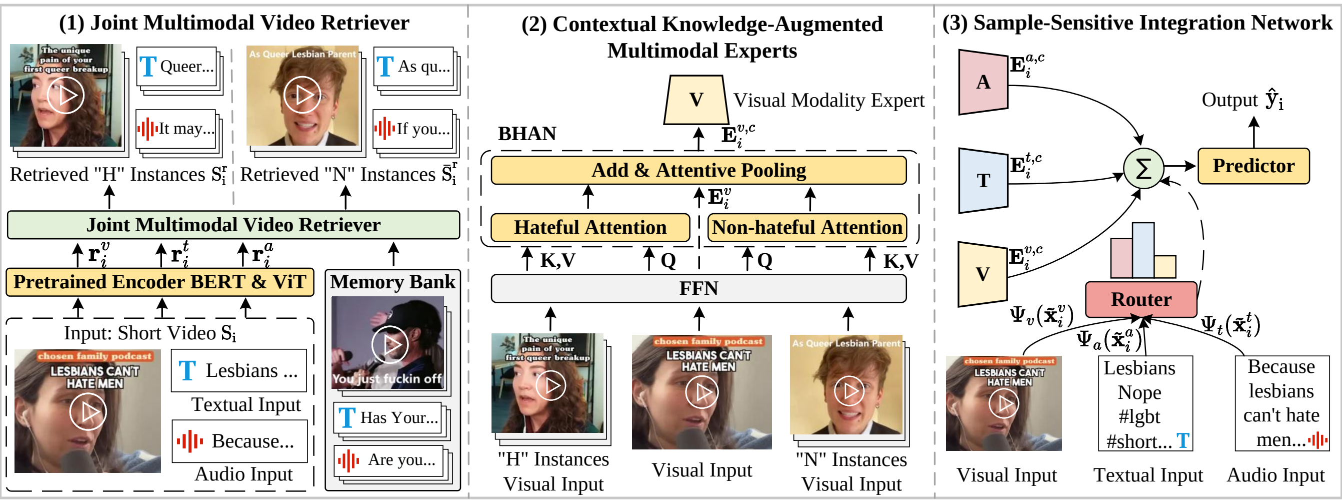


Fig. 2: Overall framework of MoRE. (1): The joint multimodal video retriever identifies similar instances by considering all the modalities. (2): The contextual knowledge-augmented multimodal experts are designed to utilize retrieved information from (1) to adapt to evolving hate expressions, while leveraging all the modalities for accurate detection. (3): The sample-sensitive integration network provides a flexible mixture to allocate weights to each expert in (2). "H": Hateful, "N": Non-hateful.

model to extract semantic features from the concatenated title and description of S_i, resulting in the textual query vector $\mathbf{r}_i^t \in \mathbb{R}^{d_t}$. Finally, for visual retrieval, we input the key frames of S_i into a pre-trained Vision Transformer (ViT) [20] and average the frame representations to generate the visual query vector $\mathbf{r}_i^v \in \mathbb{R}^{d_v}$.

3.1.3 Weighted Similarity-based Multimodal Retrieval. To effectively and comprehensively capture the relevance across audio, textual, and visual modalities, we propose a weighted similarity-based multimodal retrieval strategy. Specifically, given a short video S_i, we compute a weighted cosine similarity score that integrates the similarities from audio, textual, and visual queries. The similarity score between two videos S_i and S_j is computed as:

$$\text{Score} = w_a \cdot \text{sim}(\mathbf{r}_i^a, \mathbf{r}_j^a) + w_v \cdot \text{sim}(\mathbf{r}_i^v, \mathbf{r}_j^v) + w_t \cdot \text{sim}(\mathbf{r}_i^t, \mathbf{r}_j^t), \quad (1)$$

where w_a, w_v, and w_t are the weights assigned to the similarity of each modality; $\mathbf{r}_i^a$, $\mathbf{r}_i^v$, and $\mathbf{r}_i^t$ represent the audio, visual, and textual query vectors for video S_i, respectively. After calculating the similarity scores between S_i and each short video stored in $\mathcal{B}$, the top-K most similar hateful videos $S_i^r = \{S_i^{r_j}\}_{j=1}^K$ and the top-L most similar non-hateful videos $\bar{S}_i^r = \{\bar{S}_i^{r_j}\}_{j=1}^L$ are selected as retrieval results. These retrieved instances provide contextual knowledge, empowering modality experts to more effectively address the evolving nature of hateful content in short videos, which will be discussed in the next section.

3.2 Contextual Knowledge-Augmented Multimodal Experts

In the context of MoE, the experts represent neural networks designed to tackle particular types of tasks or data patterns. To begin with, we propose the multimodal experts networks, where each expert network is assigned to process a specific modality. Specifically, following the previous works [5, 68, 73], we simply define three modality experts, where each expert network adopts a feed-forward network (FFN) structure to capture modality-specific features,

$$\mathbf{E}_i^m = \text{FFN}(\mathbf{x}_i^m) = \left(\text{ReLU}(\mathbf{x}_i^m \mathbf{W}_1 + b_1)\right) \mathbf{W}_2 + b_2, \quad (2)$$

where $m \in \{a, t, v\}$ denotes the type of modality, $\mathbf{E}_i^m \in \mathbb{R}^{s \times d}$ is the representation of the modality expert for the short video S_i, s is the sequence length, and d is the feature dimension.

A **significant limitation** of the vanilla experts in prior works lies in their inability to adapt to the evolving nature of hateful content. To address this, we propose the Bipolar Hateful Attention Network (BHAN), which equips contextual knowledge from relevant videos to the vanilla experts to make them "up-to-date". Inspired by contrastive learning, BHAN utilizes hateful and non-hateful instances retrieved from the memory bank $\mathcal{B}$, equipping experts with contextual knowledge from both types of content. By leveraging these contrasting examples, BHAN empowers the experts to stay responsive to the ongoing shifts in hateful behavior and to capture the subtle distinctions between hateful and non-hateful content.

Specifically, for each modality expert, we first feed the retrieved hateful modality features $\mathbf{x}_i^{m,r} = \{\mathbf{x}_i^{m,r_j}\}_{j=1}^K$ and non-hateful modality features $\bar{\mathbf{x}}_i^{m,r} = \{\bar{\mathbf{x}}_i^{m,r_j}\}_{j=1}^L$ into the FFN to obtain the embeddings $\mathbf{E}_i^{m,r}$ and $\bar{\mathbf{E}}_i^{m,r}$, where $m \in \{a, t, v\}$. To equip the modality expert representation $\mathbf{E}_i^m$ with bipolar contextual knowledge, we introduce two attention mechanisms: Att_{Hat} for hateful and Att_{Non} for non-hateful attention. This process can be formalized as:

$$\tilde{\mathbf{E}}_i^{m,c} = \text{Att}_{\text{Hat}}(\mathbf{E}_i^m, \mathbf{E}_i^{m,r}, \mathbf{E}_i^{m,r}) + \text{Att}_{\text{Non}}(\mathbf{E}_i^m, \bar{\mathbf{E}}_i^{m,r}, \bar{\mathbf{E}}_i^{m,r}) + \mathbf{E}_i^m, \quad (3)$$

with the attention mechanisms Att_{Hat} and Att_{Non} defined as:

$$\text{Att}_{\text{Hat}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \alpha \cdot \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \quad (4)$$

$$\text{Att}_{\text{Non}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (1 - \alpha) \cdot \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \quad (5)$$

where α denotes the balance between the hateful and non-hateful attention contributions. We then apply an attentive pooling strategy [61] to $\tilde{\mathbf{E}}_i^{m,c} \in \mathbb{R}^{s \times d}$ across the sequence dimension to obtain the representation of the contextual knowledge-augmented multimodal experts $\mathbf{E}_i^{m,c} \in \mathbb{R}^d$ for the short video S_i.

3.3 Sample-Sensitive Integration Network

The considerable variability in modality characteristics across different short videos, along with the fact that the importance of each modality varies significantly for detecting hateful content in different videos, jointly pose a major challenge for traditional modal fusion techniques in SVHD. These methods typically apply equal weighting to all modalities, disregarding the variation in modal contributions across different samples during prediction. To address this, we propose a sample-sensitive integration network that adaptively assigns weights to each modality expert based on the unique characteristics of input video samples, prioritizing the most influential modalities for detecting hateful content in each video.

Specifically, we first employ a non-parametric strategy by applying average pooling to the original representations of each modality, resulting in comprehensive representations for the three modalities: $\tilde{\mathbf{x}}_i^a \in \mathbb{R}^{d_a}$, $\tilde{\mathbf{x}}_i^t \in \mathbb{R}^{d_t}$, and $\tilde{\mathbf{x}}_i^v \in \mathbb{R}^{d_v}$. Subsequently, we align the modal dimensions to a uniform size and design a Modality-mixture Soft Router (MSR) — i.e., a two-layer MLP — to generate dynamic weights for the fusion of the multimodal experts $\mathbf{E}_i^{a,c}$, $\mathbf{E}_i^{t,c}$, and $\mathbf{E}_i^{v,c}$ at the sample-level. This process yields the final representation for the short video S_i, which can be expressed as:

$$\tilde{\mathcal{W}}_i = [\tilde{w}_i^a, \tilde{w}_i^t, \tilde{w}_i^v] = \mathrm{MSR}([\Psi_a(\tilde{\mathbf{x}}_i^a), \Psi_t(\tilde{\mathbf{x}}_i^t), \Psi_v(\tilde{\mathbf{x}}_i^v)]), \tag{6}$$

$$w_i^m = \mathrm{Softmax}(\tilde{w}_i^m) = \frac{e^{\tilde{w}_i^m}}{\sum_{j \in \{a,t,v\}} e^{\tilde{w}_i^j}}, \tag{7}$$

$$\mathbf{E}_i = \sum_{m \in \{a,t,v\}} w_i^m \cdot \mathbf{E}_i^{m,c}, \tag{8}$$

where [,] is the concatenation operation, $\Psi_a(\cdot)$, $\Psi_t(\cdot)$ and $\Psi_v(\cdot)$ denote the linear mapping functions, w_i^m represents the weight assigned to each modality expert for short video S_i, and $\mathbf{E}_i \in \mathbb{R}^d$ is the final representation for prediction. $\mathbf{E}_i$ is then fed into a predictor (i.e., a two-layer MLP with an activation function) to generate the classification result for short video S_i: $\hat{y}_i = \mathrm{Predictor}(\mathbf{E}_i)$.

3.4 End-to-End Training

Previous MoE-based approaches [5, 68, 71] commonly employ a two-stage training paradigm. Each expert network is trained independently in the first stage, and in the second stage, these experts are integrated with a router network for joint optimization. While this approach allows for comprehensive expert training, it introduces considerable computational overhead by separate optimization phases, limiting its efficiency in real-world applications.

In contrast, we propose a more efficient and practically applicable **end-to-end training paradigm**, where the expert networks and the overall framework are optimized jointly, leading to greater computational efficiency. Specifically, we define the classification outputs from each modality expert as $\hat{y}_i^a$, $\hat{y}_i^t$, and $\hat{y}_i^v$. The joint training process is formulated as:

$$L_{\mathrm{joi}} = \min\{1 - f_{\mathrm{epo}}, 1 - \delta\} \cdot L_{\mathrm{exp}} + \max\{f_{\mathrm{epo}}, \delta\} \cdot L_{\mathrm{ovl}}, \tag{9}$$

$$L_{\mathrm{exp}} = \sum_{m \in \{a,t,v\}} L_{\mathrm{BCE}}(\hat{y}_i^m, y_i), \tag{10}$$

$$L_{\mathrm{ovl}} = L_{\mathrm{BCE}}(\hat{y}_i, y_i), \tag{11}$$

where L_{exp} represents the training loss for the expert networks and L_{ovl} denotes the loss for overall framework. δ represents a small positive constant (non-zero), used to ensure stability during training. L_{BCE} is the binary cross-entropy loss. The smoothly varying weight function $f_{\mathrm{epo}} = (\mathrm{epoch}_{\mathrm{current}}/\mathrm{epoch}_{\mathrm{total}})^2$ modulates the focus of the loss during training, placing greater emphasis on modality expert training during the early stages and gradually shifting toward optimizing the entire network in the later stages.

4 Experiments

In this section, we conduct extensive experiments to verify the efficacy of MoRE. Initially, we provide an overview of the datasets, baselines, metrics, and implementation details, with details available in Appendix C.

Datasets. To evaluate the efficacy of the proposed MoRE, we conduct comprehensive experiments on three real-world short video datasets, including HateMM [14], MultiHateClip-Youtube (MHClip-Y) and MultiHateClip-Bilibili (MHClip-B) [63].

Baselines. We compare MoRE with 9 competitive baselines, which can be categorized into three distinct groups: (1) *Unimodal hate detection methods*, which utilize a single modality for hate detection, including BERT [17], ViViT [3], and MFCC [16]. (2) *Multimodal hate detection methods*, which incorporate all available modalities within the short video to enhance the prediction accuracy, including Pro-Cap [7], HTMM [14], and MHCL [63]. (3) *Large Vision-Language Model (LVLM)-based methods*, which leverage task-agnostic multimodal pre-training and demonstrate superior performance in visual question answering and video captioning, including the recently released MiniCPM-V [70], LLaVA-OV [38], and Qwen2-VL [64].

Metrics. Following prior works [14, 63], we adopt four metrics in SVHD to comprehensively evaluate the model's performance: classification Accuracy (**ACC**), Macro-F1 score (**M-F1**), Macro Precision (**M-P**) and Macro Recall (**M-R**).

Implementation Details. During the retrieval, the default weight for each modality is set to equal. The number of retrieved videos K and L are set to 50, the bipolar attention balancing ratio α is set to 0.7, and the positive constant δ in end-to-end training is set to 0.2. We utilize the AdamW [42] optimizer with a learning rate of 5×10^{-4} and a weight decay of 5×10^{-5} for model parameters optimization. For baseline models, we strictly adhere to the settings specified in their original papers.

4.1 Overall Performance

To verify the superiority of our MoRE, we compare it with 9 competitive baselines on three datasets and the results are reported in Table 1. From these results, we have the following observations:

(O1): Multimodal hate detection methods generally outperform the unimodal methods. Unimodal methods only leverage single modality for prediction, which is prone to missing essential information and overlooking hateful content manifesting in other

Table 1: Experimental results of the competitive baseline models and the proposed MoRE on the HateMM, MHClip-Y and MHClip-B datasets. ACC: Accuracy, M-F1: Macro-F1 score, M-P: Macro Precision, M-R: Macro Recall. The best results are in red bold, while the second results are in black bold. Higher values of ACC, M-F1, M-P, and M-R indicate better performance.

	HateMM				MHClip-Y				MHClip-B			
Method	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**
BERT	0.6912	0.6368	0.7008	0.6396	0.6547	0.4909	0.5522	0.5220	0.7251	0.6771	0.6839	0.6279
ViViT	0.6820	0.6670	0.6682	0.6661	0.6705	0.6143	0.6215	0.6111	0.7099	0.6610	0.6661	0.6575
MFCC	0.6543	0.6031	0.6410	0.6069	0.6650	0.4715	0.5877	0.5222	0.6307	0.5250	0.5410	0.5304
Pro-Cap	0.6451	0.6326	0.6335	0.6321	0.7006	0.6633	0.6633	0.6633	0.7250	0.6677	0.6606	0.6832
HTMM	0.7603	0.7278	0.7794	0.7201	0.7153	0.6319	0.6830	0.6264	0.7102	0.6183	0.6654	0.6136
MHCL	**0.7741**	**0.7654**	0.7649	0.7659	0.7103	0.6547	0.6722	0.6486	**0.7650**	0.7311	0.7320	**0.7302**
MiniCPM-V	0.7235	0.7228	0.7781	0.7635	0.6910	0.6742	0.6929	**0.6740**	0.7157	0.7015	0.7359	0.7044
LLaVA-OV	0.7558	0.7557	0.7790	**0.7828**	**0.7350**	**0.6766**	**0.7045**	0.6674	0.7521	0.7078	0.7143	0.7031
Qwen2-VL	0.7373	0.7371	**0.7805**	0.7732	0.7050	0.6677	0.6684	0.6671	0.7601	**0.7326**	**0.7385**	0.7285
MoRE	**0.8341**	**0.8235**	**0.8178**	**0.8334**	**0.7750**	**0.7519**	**0.7567**	**0.7482**	**0.7850**	**0.7475**	**0.7568**	**0.7410**
Improv.	7.75%↑	7.59%↑	4.78%↑	6.46%↑	5.44%↑	11.13%↑	7.41%↑	11.01%↑	2.61%↑	2.03%↑	2.48 %↑	1.48%↑
p-val.	9.72e^{-3}	8.52e^{-3}	7.44e^{-3}	7.51e^{-3}	9.91e^{-4}	3.07e^{-4}	1.47e^{-3}	3.67e^{-4}	2.29e^{-4}	1.68e^{-3}	2.62e^{-4}	3.61e^{-3}

modalities, leading to weak performance. Multimodal detection methods leverage features from all the modalities to improve the precision of prediction. Moreover, MoRE performs best among multimodal methods, as these multimodal baselines typically overlook the evolving nature of hateful content, which requires the model to remain current. Furthermore, these methods often adopt a vanilla fusion strategy that treats modalities equally in modal fusion, overlooking the varying importance of each modality across different instances in SVHD, which requires a more flexible fusion approach.

(O2): **LVLM-based methods exhibit strong performance in SVHD.** LVLM-based methods have recently gained prominence due to their impressive performance across a wide range of multimodal tasks. These methods leverage the latest advanced LVLMs, whose effectiveness largely stems from extensive pre-training on large-scale vision-language corpora, enabling them to generalize well across many multimodal downstream tasks. Despite their strong capability in detecting hate in short videos, MoRE outperforms these models due to the lack of task-specific adaptation in LVLMs required for SVHD.

(O3): **MoRE outperforms all strong baseline models across three datasets.** Notably, MoRE achieves average improvements of 5.27% in ACC and 6.91% in M-F1 across all three datasets. To further validate MoRE's superiority, we compute the statistical differences between MoRE and the best-performing baseline by retraining both models five times. These performance gains demonstrate the effectiveness of incorporating expressive contextual knowledge from retrieved instances, which enables the experts to adapt to the evolving nature of hateful content and enhances their discriminative power. Moreover, the sample-sensitive integration network dynamically allocates contribution for each expert based on the characteristics of each video sample, leading to further improvements in SVHD.

4.2 Ablation Study

To further understand the roles of core components and multimodal experts in our proposed MoRE framework, comprehensive ablation studies are conducted.

Table 2: Ablation study on core components within MoRE. The best results are in black bold.

	HateMM		MHClip-Y		MHClip-B	
Variant	**ACC**	**M-F1**	**ACC**	**M-F1**	**ACC**	**M-F1**
Uni Retriever	0.7972	0.7744	0.7402	0.6810	0.7790	0.7303
w/o Retriever	0.7557	0.7355	0.6950	0.6637	0.7150	0.6836
BHAN-Att$_{\text{Hat}}$	0.8018	0.7887	0.7610	0.6881	0.7550	0.7009
BHAN-Att$_{\text{Non}}$	0.8110	0.7985	0.7550	0.7240	0.7750	0.7358
w/o BHAN	0.7880	0.7723	0.7315	0.7120	0.7001	0.6581
w/o Router	0.7882	0.7734	0.7302	0.6815	0.7211	0.6902
MoRE	**0.8341**	**0.8235**	**0.7750**	**0.7519**	**0.7850**	**0.7475**

4.2.1 Ablation Study on Core Components. We conduct an ablation study to analyze the role of each core component within MoRE, and the results are summarized in Table 2.

Effect of joint multimodal video retriever. To validate the efficacy of the joint multimodal video retriever, we designed two variant models: (1) **Uni Retriever**: replacing multimodal joint video retriever with an unimodal retriever, performing text-to-text retrieval, and (2) **w/o Retriever**: removing the retriever entirely by using random samples to replace the retrieved instances. The results demonstrate that unimodal retrieval, limited to a single modality, fails to capture the most relevant instances, leading to suboptimal performance. Furthermore, completely removing the retrieval process causes a substantial drop in performance, highlighting the crucial role of high-quality retrieved instances. In contrast, our multimodal joint video retriever, which incorporates information from all three modalities, consistently improves the retrieval quality and strengthens the overall framework performance.

Effect of contextual knowledge-augmented multimodal experts. To analyze the impact of contextual knowledge equipped to the modality experts, we design three variant models: (1) **BHAN-Att$_{\text{Hat}}$**: removing the non-hateful attention from the BHAN, (2) **BHAN-Att$_{\text{Non}}$**: removing the hateful attention from the BHAN, and (3) **w/o BHAN**: removing the BHAN entirely. The removal

Table 3: Ablation study on multimodal experts within MoRE. The best results are in black bold. A: Audio expert, T: Textual expert, V: Visual expert.

	HateMM		MHClip-Y		MHClip-B	
Expert(s)	ACC	M-F1	ACC	M-F1	ACC	M-F1
{ A }	0.6451	0.5826	0.6521	0.5132	0.6497	0.4531
{ T }	0.7188	0.6972	0.7350	0.6765	0.7201	0.6880
{ V }	0.6866	0.6415	0.7002	0.5888	0.7150	0.6557
{ A, T }	0.7281	0.7004	0.7250	0.6491	0.7305	0.6614
{ A, V }	0.7373	0.6935	0.6651	0.4132	0.6850	0.5771
{ T, V }	0.8110	0.7954	0.7402	0.6739	0.7502	0.7252
MoRE	**0.8341**	**0.8235**	**0.7750**	**0.7519**	**0.7850**	**0.7475**

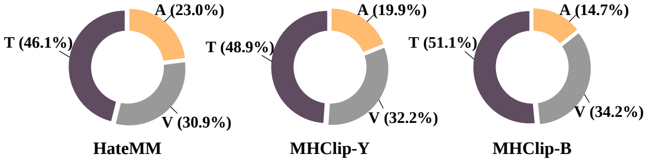
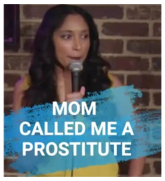
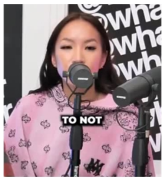


Fig. 3: Visualization of modality experts contribution allocation of MoRE across all three datasets. A: Audio expert, T: Textual expert, V: Visual expert.

Table 4: Presentation of the retrieval quality. H: Hateful, N: Non-hateful. "V / A / T" refers to the cosine similarity scores between the target video and the retrieved videos across visual, audio, and textual modalities.

	Target: H	Top-1: H	Top-1: N
Vision			
Audio	I'm a prostitute. I don't charge body for sex. I give man a way for free...	That is, I look like a prostitute and I am charging the man for sex...	In Greek legend, Phryne, famous prostitute, the god give the body...
Text	I give a way for free; I am a lady; Mom called me a prostitute...	I am a prostitute; I am a lady of the evening dropped pants...	In Ancient Greek Prostitute; famous prostitute avoid showing...
V / A / T	N/A	0.75 / 0.93 / 0.90	0.81 / 0.89 / 0.87

of each type of attention results in a notable performance drop, highlighting the importance of integrating contextual knowledge from both hateful and non-hateful relevant instances. Moreover, eliminating the entire BHAN leads to a substantial performance decline, underscoring the critical role of equipping modality experts with contextual insights, which facilitates the experts to adapt the ever-changing hate and improve their ability to distinguish the subtle difference between content of hate and non-hate.

Effect of sample-sensitive integration network. We evaluate the impact of the sample-sensitive integration network by designing the variant model: **w/o Router**: replacing the router network with a simple sum-based fusion method. The results indicate that equally fusing the modalities fails to accurately detect hate in short videos. In fact, the hateful content may manifest in different modalities, which necessitates a flexible fusion approach, the sample-sensitive integration network, to dynamically assign the modal contribution for each short video instance.

4.2.2 Ablation Study on Multimodal Experts. The second ablation study evaluates the contribution of each modality expert in detecting hateful content. It employs various combinations of modality experts in MoRE, with the results presented in Table 3. Based on these results, we have the following observations:

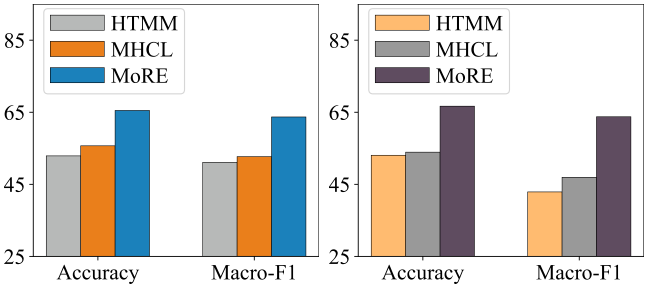
(O1): **Different modal experts have significantly different impacts.** Across all three datasets, we observe significant variability in the impact of each expert. The textual expert consistently plays a more crucial role in SVHD compared to the visual and audio experts, with the audio expert contributing the least. This observation also aligns with the distribution of hateful content across each modality in the dataset, as exemplified by the MHClip-B dataset shown in the blue donut chart of Fig. 1(b).

(O2): **Effectively combining all experts brings better performance.** We observe that combining multiple experts consistently improves performance compared to using a single expert. In particular, combining textual and visual experts outperforms combining audio with either modality expert, reinforcing the relative weakness of the audio expert. Notably, our proposed MoRE effectively integrates all three experts through the sample-sensitive integration network to achieve optimal performance in SVHD.

To provide further insight into how MoRE leverages three modality experts, we present the average weight assigned by the router network in MoRE to each expert across different datasets in Fig. 3. The router consistently assigns the highest weight to the textual expert, followed by the visual expert, with the audio expert receiving the lowest weight. It demonstrates that the router can effectively adapt to the strengths of each expert, thereby providing an intuitive explanation for the observed improvements in MoRE performance.

4.3 Retrieval Quality Presentation

To validate the effectiveness of the proposed joint multimodal video retriever, we randomly select a hateful video from the test set of the MHClip-Y dataset with the retrieved results. As illustrated in Table 4, we observe that both hateful (H) and non-hateful (N) instances exhibit similar backgrounds and subjects, specifically featuring a woman speaking, which closely aligns with the target video's visual information. Furthermore, the texts and audio transcriptions of the retrieved instances show significant content overlap with the target video, including keywords such as "prostitute", "body", and "sex". This observation underscores the efficacy of our multimodal retrieval strategy, which seamlessly integrates all three modalities to retrieve the most relevant instances. Notably, the



(a) From HateMM to MHClip-Y. **(b) From MHClip-Y to HateMM.**

Fig. 4: Generalizability between the baselines: HTMM, MHCL, and our MoRE on the HateMM and MHClip-Y datasets

content within the text and audio transcriptions in retrieved instances shares overlapping keywords with the target video, such as "prostitute". However, the hateful instance employs harmful and offensive language (e.g., "charging", "sex", "lady"), in stark contrast to the non-hateful instance, which engages in neutral discourse, such as the historical story of the prostitute in "Ancient Greece". Consequently, by effectively learning the nuanced distinctions between hateful and non-hateful instances, the modality experts are endowed with enhanced discriminative capabilities.

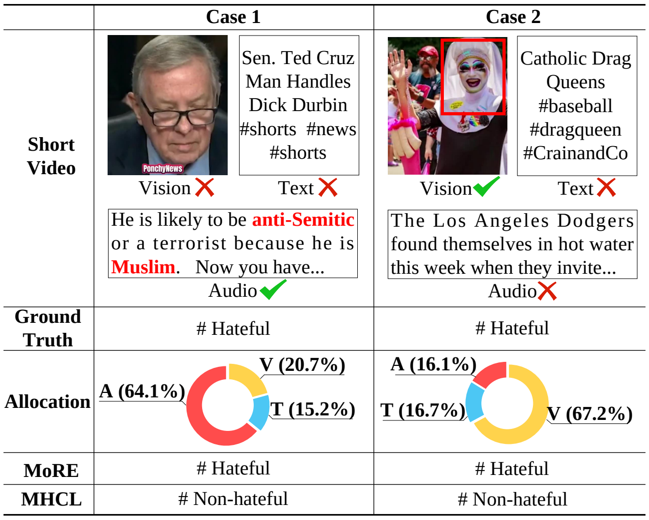
4.4 Model Generalizability

To investigate the generalizability of MoRE and two competitive baseline models, particularly their ability to adapt to the new form of hateful content, we conduct experiments where the models are trained on one dataset and tested on the other. The HateMM and MHClip-Y datasets are selected due to the significant differences in their video content, stemming from their origins on entirely distinct online platforms. In these experiments, the memory bank $\mathcal{M}$ of MoRE is constructed using the training set of the target dataset. Initially, the models are trained on HateMM and tested on MHClip-Y, and subsequently, this setup is reversed to train on MHClip-Y and test on HateMM. The results are presented in Fig. 4.

Both baseline models demonstrate extremely weak performance when confronted with previously unseen hateful content, primarily due to their lack of design for handling generalization. In contrast, the proposed MoRE exhibits remarkable adaptability to these new forms of hate, as it leverages contextual knowledge from retrieved instances in the target dataset to enable the multimodal experts to effectively detect "unencountered" hateful content. These findings further confirm the superiority of MoRE in adapting to the evolving nature of hate in short videos and its ability to meet real-world demands by training on one platform and generalizing across multiple platforms.

4.5 Case Study: Model Explainability

In this section, we explore the explainability of MoRE by conducting a case study on two randomly selected hateful short videos from the test set of the MHClip-Y dataset. This case study aims to elucidate how MoRE adaptively assigns weights to multimodal experts to achieve accurate predictions for different video samples.



	Case 1	Case 2
Short Video	Sen. Ted Cruz Man Handles Dick Durbin #shorts #news #shorts Vision ✗ Text ✗ He is likely to be **anti-Semitic** or a terrorist because he is **Muslim**. Now you have... Audio ✓	Catholic Drag Queens #baseball #dragqueen #CrainandCo Vision ✓ Text ✗ The Los Angeles Dodgers found themselves in hot water this week when they invite... Audio ✗
Ground Truth	# Hateful	# Hateful
Allocation	A (64.1%) V (20.7%) T (15.2%)	A (16.1%) T (16.7%) V (67.2%)
MoRE	# Hateful	# Hateful
MHCL	# Non-hateful	# Non-hateful

Fig. 5: Case study of the MoRE's explainability on dynamically assigning weights to modality experts for each video instance. A: Audio expert, T: Textual expert, V: Visual expert.

The first case illustrated on the left in Fig. 5 involves a short video where hateful content, specifically "anti-Semitic" and "Muslim", is presented solely in the audio modality. Our MoRE successfully captures the hateful evidence by prioritizing the audio expert, assigning it the highest weight (64.1%). The second case is more challenging, as neither the text nor audio contains hateful content. However, some frames in video show a group of men dressed as Catholic nuns mocking Christianity, which constitutes the hateful element. In this instance, MoRE effectively allocates the highest contribution (67.2%) to the visual expert, resulting in a correct identification of the hateful content. In contrast, the baseline model MHCL, which treats each modality equally, fails to detect the hateful content in these cases, leading to incorrect prediction.

5 Conclusion

In this work, we propose a novel MoRE framework to address SVHD. This multimodal framework leverages features from all modalities to enhance the precision of SVHD. A multimodal joint video retriever is developed to identify the most relevant instances for the target video. Multimodal experts gain contextual knowledge from these retrieved hateful and non-hateful instances, enhancing their ability to adapt to the dynamic evolution of hateful content. Additionally, a sample-sensitive integration network within MoRE adaptively adjusts the contributions of each expert based on different samples, further improving performance in SVHD. Furthermore, an end-to-end training paradigm is introduced to enhance the practical applicability of MoRE in real-world large-scale SVHD applications. Our extensive experiments conducted on three real-world datasets demonstrate the effectiveness of the proposed MoRE for SVHD.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62176043, No.62072077, and No.U22A2097).

References

[1] Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. SharedCon: Implicit Hate Speech Detection using Shared Semantics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 10444–10455.

[2] Rana Al-Maroof, Kevin Ayoubi, Khadija Alhumaid, Ahmad Aburayya, Muhammad Alshurideh, Raghad Alfaisal, and Said Salloum. 2021. The acceptance of social media video for knowledge acquisition, sharing and application: A comparative study among YouYube users and TikTok users' for medical purposes. *International Journal of Data and Network Science* 5, 3 (2021), 197.

[3] 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 (ICCV)*. 6836–6846.

[4] Kalyanaswamy Banuroopa and D Shanmuga Priyaa. 2021. MFCC based hybrid fingerprinting method for audio classification through LSTM. *International Journal of Nonlinear Analysis and Applications* 12, Special Issue (2021), 2125–2136.

[5] Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *International Conference on Information and Knowledge Management (CIKM)*. 110–119.

[6] Christina Bucknell Bossen and Rita Kottasz. 2020. Uses and gratifications sought by pre-adolescent and adolescent TikTok consumers. *Young consumers* 21, 4 (2020), 463–478.

[7] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection. In *Proceedings of the ACM International Conference on Multimedia (MM)*. ACM, 5244–5252.

[8] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification.. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 321–332.

[9] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized Networks for Few-shot Hateful Meme Detection. In *Proceedings of the ACM Web Conference (WWW)*. 4575–4584.

[10] Rochana Chaturvedi, Sugat Chaturvedi, and Elena Zheleva. 2024. Bridging or Breaking: Impact of Intergroup Interactions on Religious Polarization. In *Proceedings of the ACM Web Conference (WWW)*. 2672–2683.

[11] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10638–10647.

[12] Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 445–455.

[13] Zhangtao Cheng, Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and S Yu Philip. 2024. Information Cascade Popularity Prediction via Probabilistic Diffusion. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2024).

[14] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)* 17 (2023), 1014–1023.

[15] Thomas Davidson, Dana Warmsley, Michael W. 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 (ICWSM)*. 512–515.

[16] S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1980).

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

[18] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9346–9355.

[19] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4065–4080.

[20] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

[21] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18166–18176.

[22] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. In *International Conference on Learning Representations (ICLR)*.

[23] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference (BMVC)*. 12.

[24] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

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

[26] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems (Neurips)* 26 (2013).

[27] Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. [n. d.]. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[28] Nicola J Hodges, A Mark Williams, Spencer J Hayes, and Gavin Breslin. 2007. What is modelled during observational learning? *Journal of Sports Sciences* 25, 5 (2007), 531–545.

[29] Douglas R Hofstadter. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

[30] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12976–12985.

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

[32] M Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in human behavior* 66 (2017), 236–247.

[33] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems (Neurips)*, Vol. abs/2005.04790.

[34] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* 27, 1 (2013), 1621–1622.

[35] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning. *arXiv preprint arXiv:2501.01120* (2025).

[36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[37] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations (ICLR)*.

[38] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).

[39] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2592–2607.

[40] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the ACM International Conference on Multimedia (MM)*. 1786–1794.

[41] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models.. In *Proceedings of the ACM Web Conference (WWW)*. 2359–2370.

[42] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

[43] Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. 2021. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008* (2021).

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (Neurips)* 32 (2019).

[45] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019), e0221152.

[46] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2023. Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 5333–5347.

[47] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning. In *Proceedings of the Annual Meeting of the Association for Computational*

Linguistics (ACL). 5333–5347.

[48] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS*. Springer, 928–940.

[49] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE* 15, 8 (2020), e0237861.

[50] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems (Neurips)* 35 (2022), 9564–9576.

[51] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 299–307.

[52] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. 2022. Highly accurate dichotomous image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 38–56.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

[54] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, Vol. 202. PMLR, 28492–28518.

[55] Kurniawan Nur Ramadhani, Rinaldi Munir, and Nugraha Priya Utama. 2024. Improving Video Vision Transformer for Deepfake Video Detection using Facial Landmark, Depthwise Separable Convolution and Self Attention. *IEEE Access* (2024).

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016), 1137–1149.

[57] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems (Neurips)* 34 (2021), 8583–8595.

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

[59] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2016. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*.

[60] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling Vision-Language Models with Sparse Mixture of Experts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 11329–11344.

[61] Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 3418–3428.

[62] B Vimal, Muthyam Surya, VS Sridhar, Asha Ashok, et al. 2021. Mfcc based audio classification using machine learning. In *International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–4.

[63] Han Wang, Rui Yang Tan, Usman Naseem, and Roy Ka-Wei Lee. 2024. MultiHate-Clip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv* (2024).

[65] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. *Proceedings of the Workshop on Language in Social Media* (2012), 19–26.

[66] Zheng Wei, Yixuan Xie, Danyun Xiao, Simin Zhang, Pan Hui, and Muzhi Zhou. 2024. Social Media Discourses on Interracial Intimacy: Tracking Racism and Sexism through Chinese Geo-located Social Media Data. In *Proceedings of the ACM Web Conference (WWW)*. 2337–2346.

[67] Fan Wu, Sanyam Lakhanpal, Qian Li, Kookjin Lee, Doowon Kim, Heewon Chae, and Kyounghee Hazel Kwon. 2024. Not All Asians are the Same: A Disaggregated Approach to Identifying Anti-Asian Racism in Social Media. In *Proceedings of the ACM Web Conference (WWW)*. 2615–2626.

[68] Wenxin Xu, Hexin Jiang, et al. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[69] Shuai Yang, Yuzhen Zhao, and Yifang Ma. 2019. Analysis of the reasons and development of short video application-Taking Tik Tok as an example. In *Proceedings of the International Conference on Information and Social Science (ICISS)*. 12–14.

[70] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800* (2024).

[71] Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023. MMOE: Mixture of Multimodal Interaction Experts. *arXiv preprint arXiv:2311.09580* (2023).

[72] Hongchun Yuan, Zhenyu Cai, Hui Zhou, Yue Wang, and Xiangzhi Chen. 2021. Transanomaly: Video anomaly detection using video vision transformer. *IEEE Access* 9 (2021), 123977–123986.

[73] Cong Zhang, Dongyang Liu, Lin Zuo, Junlan Feng, Chao Deng, Jian Sun, Haitao Zeng, and Yaohong Zhao. 2023. Multi-gate Mixture-of-Contrastive-Experts with Graph-based Gating Mechanism for TV Recommendation. In *International Conference on Information and Knowledge Management (CIKM)*. 4938–4944.

[74] Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. 2024. Predicting Micro-video Popularity via Multi-modal Retrieval Augmentation. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*. 9–16.

[75] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[76] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. 2024. Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond. *arXiv* abs/2405.03520 (2024).

A Additional Literature Review

A.1 Multimodal Retrieval

Multimodal retrieval aims to retrieve the most relevant instances by leveraging information across different modalities, such as text, vision, and audio. Previous studies have primarily focused on text-image retrieval, with the objective of retrieving images that correspond to a given text query or text that corresponds to a given image query [23, 26, 35, 36, 51, 56]. These earlier studies typically relied on models that did not employ pre-training, such as Convolutional Neural Networks (CNNs) [23] and Faster R-CNN [56], to extract representations from both image and text data. The introduction of powerful vision-language pre-trained models [21, 30, 39, 44, 53] has enabled researchers to develop methods that jointly encode text and image representations for more accurate retrieval. These models have demonstrated significant improvements in the quality of text-image retrieval tasks. With the growing popularity of short video content, video retrieval has become an increasingly important area of study. Many studies in video retrieval have focused on text-to-video retrieval, where a text query is used to retrieve relevant video content from large video collections [11, 18, 19, 40]. These approaches leverage pre-trained models to generate a common embedding space, facilitating the alignment of video and text representations. Despite advancements in text-to-video retrieval, limited research addresses video-to-video retrieval, where the goal is to find the most relevant video content given a video query. In this work, we propose a novel joint multimodal video retriever that integrates audio, textual, and visual modalities to enable comprehensive and precise video-to-video retrieval.

A.2 Mixture of Experts

The Mixture of Experts (MoE) was first proposed by Jacob et al. [31] as a method to combine multiple experts, each trained on different subsets of data, into a single powerful model. Eigen et al. [22] extended the MoE concept to neural networks by incorporating a

Table 5: Characteristics of three short video datasets.

Dataset Characteristic	HateMM	MHClip-Y	MHClip-B
Total Videos	1,083	1,000	1,000
Hateful Videos	431	82	128
Offensive Videos	N/A	256	194
Non-Hateful Videos	652	662	678
Avg. Duration (s)	150.0	33.8	31.8
Languages	English	English	Chinese
Platforms	BitChute	YouTube	Bilibili

layer consisting of expert networks and a trainable gating mechanism. This gating mechanism assigns weights to the experts on a per-example basis, enabling MoE to produce a weighted combination of the experts' outputs. Recently, MoE has been extensively studied as a technique to enhance the model's capacity in terms of parameter size without incurring additional computational cost, particularly in the fields of natural language processing [24, 37, 59] and computer vision [27, 43, 50, 57, 60]. Switch Transformer [24] developed a sparse MoE architecture that improves sample efficiency in training by minimizing communication and computational overhead, making it effective for natural language processing tasks. In the multimodal learning domain, LIMoE [50] presented a sparse MoE model that allows for the simultaneous processing of both image and text using a contrastive loss during training. Much of the current work primarily focuses on using the sparsity of MoE to augment model parameters, overlooking one of the key strengths of MoE: the ability to dynamically adjust outputs based on the input data through expert routing. In contrast, our work first time introduces MoE into the task of video-based hate detection by designing contextual knowledge-augmented multimodal experts to tackle different modalities of the short video. Furthermore, a sample-sensitive integration network is proposed to identify the specific contributions of each modality expert's features to hate detection in each video.

B Feature Extraction

For the short video S_i, we start by extracting its initial information from each modality. Specifically, we isolate the audio component from the video, resulting in the audio representation s_i^a. Additionally, we uniformly sample m key frames from the video, which contribute to the visual content information denoted as $s_i^v = \{\mathbf{v}_i^1, \mathbf{v}_i^2, \ldots, \mathbf{v}_i^m\}$. The textual information s_i^t incorporates the title and description of the short video S_i.

To ensure alignment with prior research [14, 63] for fair comparison, we utilize the pre-trained BERT [17] and ViT [20] as textual and visual feature extractors. This allows us to derive the text features $\mathbf{x}_i^{\mathrm{t}} \in \mathbb{R}^{n \times d_t}$ and visual features $\mathbf{x}_i^{\mathrm{v}} \in \mathbb{R}^{m \times d_v}$, where n is the number of word tokens, while d_t and d_v denote the dimensions of the textual and visual embeddings, respectively. Specifically, the visual embedding for each key frame is derived from the classification token in the last hidden states of the Vision Transformer (ViT), which serves as the global representation of the frame. For audio feature extraction, we compute the Mel Frequency Cepstral Coefficients (MFCC), resulting in audio features $\mathbf{x}_i^{\mathrm{a}} \in \mathbb{R}^{l \times d_a}$, where l denotes the number of audio frames, and d_a represents the number of MFCC coefficients extracted from each audio frame.

C Detailed Experimental Settings

C.1 Datasets

We conduct comprehensive experiments to evaluate the performance of the proposed MoRE framework compared to baseline models on three real-world short video datasets: HateMM [14], MultiHateClip-YouTube (MHClip-Y), and MultiHateClip-Bilibili (MHClip-B) [63]. The characteristics of these datasets are outlined in terms of the total number of videos, the counts of hateful and offensive videos, non-hateful videos, average video duration, languages, and the platforms from which the videos were sourced, as shown in Table 5.

- **HateMM**: This dataset is a hateful video detection dataset, collected from *BitChute*, an alternative video-sharing platform with minimal content moderation. The English-language videos were manually annotated by trained annotators. Each entry contains the full video, and hate/non-hate label, with additional annotations including frame spans indicating hateful content and targeted communities.
- **MHClip-Y, MHClip-B**: These two datasets are benchmark datasets designed for hateful video detection on *YouTube* and *Bilibili*, respectively. Each entry in these two datasets includes the video, its title, transcript, and detailed annotations. The annotations provide rich information, including the video's classification (hateful, offensive, or non-hateful), specific hateful/offensive segments with timestamps, the target victim group (e.g., Woman, Man, LGBTQ+), and the contributing modalities (audio, textual, and visual).

Notably, we present the binary classification experimental results in the main paper by merging the "offensive" and "hateful" categories into a single "hateful" class.

C.2 Baselines

To validate the efficacy of MoRE, we compare our framework with competitive baseline models, which can be classified into three distinct groups: (1) *Unimodal hate detection methods*; (2) *Multimodal hate detection methods*; and *(3) Large Vision-Language Model (LVLM)-based methods.* Below, we provide detailed descriptions of each baseline.

(1) *Unimodal hate detection methods*:

- **BERT** [17]: Given the efficacy of BERT in hate speech detection [48], we employ BERT as a competitive unimodal baseline. The text data, including the video title, description, and audio transcription, is passed through BERT to extract features (i.e., the [CLS] token) represented in a 768-dimensional space. These features are subsequently fed into two fully connected (FC) layers to yield the final prediction results.
- **ViViT** [3]: The Video Vision Transformer is the video version of ViT [20], which is effective in video understanding and classification [55, 72]. We utilize ViViT to extract a 768-dimensional feature vector from 32 sampled frames for each video. The features are then input into two FC layers to generate the final output.

Table 6: Example of prompt for hateful detection applied in LVLM-based methods.

Prompt: Now your task is to determine whether a short video is hateful or non-hateful based on its title, description, audio transcription and raw video content. If the video is hateful, output 1; otherwise, output 0.
Video Title: { Title }
Video Description: { Description }
Audio Transcription: { Transcription }
Raw Video Content: { Raw video content (in MP4 format) }
Now give your prediction (no need analysis, return 0 or 1 only).

- **MFCC**: MFCC plays a pivotal role in audio signal processing and has been widely used in audio classification [4, 62]. For each video, we generate a 128-dimensional MFCC vector, which is then processed through two fully connected (FC) layers to obtain the final prediction results.

(2) *Multimodal hate detection methods*:

- **Pro-Cap** [7]: Pro-Cap utilizes prompting techniques to guide pre-trained vision-language models in generating image captions associated with hateful content. It subsequently combines these generated captions with textual information to enhance the detection of hateful memes.
- **HTMM** [14]: HTMM extracts features from transcripts, video frames, and audio frames. These features are then concatenated and input into an MLP-based classifier to detect hateful content in short videos.
- **MHCL** [63]: MHCL analyzes the significance of each modality in the detection of hateful content within videos. It then leverages the audio, textual, and visual features with LSTM-based feature encoders to perform hateful video detection.

(3) *LVLM-based methods*:

- **MiniCPM-V** [70]: MiniCPM-V is a series of end-to-end VLLMs designed for vision-language understanding. These models accept text, images, and videos as inputs, generating high-quality text outputs. In this study, we adopt the latest and most advanced model in the MiniCPM-V series, MiniCPM-V 2.6, as our competitive baseline.
- **LLaVA-OV** [38]: LLaVA-OneVision (LLaVA-OV) is the newest family of open MLLMs in the LLaVA series, which achieves new state-of-the-art performance across single-image, multi-image, and video benchmarks.
- **Qwen2-VL** [64]: Qwen2-VL is the latest version of the vision language models in the Qwen model families. Qwen2-VL has the abilities of complex reasoning and decision making and achieves state-of-the-art performance on visual understanding benchmarks.

Notably, for LVLM-based methods, we provide the text and raw video content along with a specifically designed prompt to guide the output generation. An example of the prompt is presented in Table 6.

C.3 Implementation Details

In this section, we provide detailed implementation specifications for our proposed MoRE along with a comprehensive overview of the experimental setup.

- **Data processing.** We uniformly extract 16 key frames from each short video across all datasets to ensure consistent visual representation. To extract audio features, we employ the open-source library Librosa to compute the MFCC. For audio transcription, we employ two versions of the pre-trained Whisper [54] automatic speech recognition model, each separately fine-tuned for Chinese and English audio. To generate descriptions of the video content, we employ the pre-trained BLIP2 model, specifically the opt-2.7b version, to caption the extracted key frames. Additionally, we apply a chotomous image segmentation model IS-Net [52] fine-tuned in background removal task to separate the background from the subjects in the key frames.
- **Details of memory bank construction.** In this work, the memory bank $\mathcal{B}$ is composed of short videos from the training and validation sets, thereby preventing data leakage during model testing. However, in real-world applications, the memory bank must be continuously updated to reflect temporal changes, ensuring that the model can adapt to the rapidly evolving nature of hateful content.
- **Training configuration.** During the retrieval, the default weight for each modality is set to equal. For text, we set the maximum sequence length to 512 for all datasets. For key frames, we resize the images into 224×224. The number of retrieved hateful videos K and non-hateful videos L are selected from the set {10, 20, 30, 40, 50}, respectively. And the bipolar attention balancing ratio α is chosen from the range [0, 1]. The positive constant δ in end-to-end training is set to 0.2. We utilize the AdamW [42] optimizer with a learning rate of 5×10^{-4} and a weight decay of 5×10^{-5} for model parameters optimization. We set the random seed to 2024. For statistical testing, where each model is run five times, we use random seeds ranging from 2024 to 2028 and report the mean value as experimental results. For baseline models, we strictly adhere to the settings specified in their original papers.
- **Implementation environment.** All experiments are conducted on a system equipped with an Intel(R) Core(TM) i9-14900KF processor, an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, and 128 GB of system RAM.