ImpliHateVid: A Benchmark Dataset and Two-stage Contrastive Learning Framework for Implicit Hate Speech Detection in Videos

Mohammad Zia Ur Rehman[1], Anukriti Bhatnagar[1], Omkar Kabde[2], Shubhi Bansal[1], Nagendra Kumar[1]
[1]Indian Institute of Technology Indore, Indore, India
[2]Chaitanya Bharathi Institute of Technology, Telangana, India
{phd2101201005, mt2302101007}@iiti.ac.in, ugs22058_cic.omkar@cbit.org.in, {phd2001201007, nagendra}@iiti.ac.in

Abstract

The existing research has primarily focused on text and image-based hate speech detection; video-based approaches remain underexplored. In this work, we introduce a novel dataset, ImpliHateVid, specifically curated for implicit hate speech detection in videos. ImpliHateVid consists of 2,009 videos comprising 509 implicit hate videos, 500 explicit hate videos, and 1,000 non-hate videos, making it one of the first large-scale video datasets dedicated to implicit hate detection. We also propose a novel two-stage contrastive learning framework for hate speech detection in videos. In the first stage, we train modality-specific encoders for audio, text, and image using contrastive loss by concatenating features from the three encoders. In the second stage, we train cross-encoders using contrastive learning to refine multimodal representations. Additionally, we incorporate sentiment, emotion, and caption-based features to enhance implicit hate detection. We evaluate our method on two datasets, ImpliHateVid for implicit hate speech detection and another dataset for general hate speech detection in videos, the HateMM dataset, demonstrating the effectiveness of the proposed multimodal contrastive learning for hateful content detection in videos and the significance of our dataset. The code and dataset will be made available on the GitHub repository[1].

1 Introduction

With approximately 66% of the world's population having access to the internet[2], online communication has become an integral part of daily life. However, the widespread accessibility of digital platforms has also facilitated the rapid dissemination of hate speech. Despite efforts by online platforms to regulate such content through AI-based detection, manual review, and community reporting (Hatano, 2023), hateful content remains a persistent challenge due to the vast amount of data generated every day (Ibañez et al., 2021; Das et al., 2023; Wu and Bhandary, 2020).

Hate speech is defined as public speech that expresses hate or encourages violence towards a person or group based on race, religion, sex, or caste[3]. It manifests in multiple forms, including texts, images, memes, gestures, and symbols, both online and offline[4]. Most existing research on hate speech detection has focused on textual content, such as tweets and comments (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017), or image-based hate speech, particularly in memes (Cao et al., 2023; **?**; Sharma et al., 2022; Nayak et al., 2022; Hee et al., 2023). While some studies have explored hate detection in videos (Alcântara et al., 2020; Das et al., 2023; Wang et al., 2024a), implicit hate speech detection in videos remains an underexplored area. To the best of our knowledge, we are the first to work in implicit hate speech detection in videos.

Implicit hate speech is defined as expressions that communicate discriminatory or prejudiced views indirectly, often through coded language, implied meanings, or contextual cues (ElSherief et al., 2021). Unlike overt hate speech, it subtly evades detection by adhering to platform guidelines and may appear innocuous superficially, yet still perpetuates harm or offense. Given the dominance of video content in digital communication, there is a strong need to develop specialized hate detection mechanisms for videos.

Figure 1 illustrates examples of both implicit and explicit hate speech in video content. In the implicit hate example, the hateful intent is conveyed through the underlying context and the associa-

[1]https://github.com/videohatespeech/Implicit_Video_Hate
[2]https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx
[3]https://dictionary.cambridge.org/us/dictionary/english/hate-speech
[4]https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17209–17221

Offensive content warning: The presented image is just for illustration. We do not support or promote the views presented in the image.

Video Frame	Transcript	Label
	That's even the best thing to make of it. He'll show exactly why liberals get away with political violence.	Implicit Hate
	... black people apparently are dark stained, love chickens. I am mortified ...	Explicit Hate

Figure 1: Implicit and explicit hate videos example.

tion of political ideology with aggression. Such indirect expressions make implicit hate speech particularly challenging to detect, as they require an understanding of both textual and visual cues in the video. Conversely, the explicit hate speech example features a cartoonish character who appears to be shouting a racial slur directed at Black individuals.

This distinction underscores the necessity of advanced multimodal methods for effective implicit hate speech detection in videos.

In this work, we address the critical challenge of detecting hateful content in videos by proposing a novel method with a particular focus on implicit hate speech, which is often subtle and context-dependent. Our two-stage contrastive learning method first extracts robust, modality-specific features from text, images, and audio, and then aligns them in a shared embedding space using cross-modal encoders with projection heads and supervised contrastive loss. This approach captures subtle cues of implicit hate speech that traditional unimodal or simple fusion methods often miss. By leveraging complementary multimodal information, our framework provides a richer, more nuanced representation for hate speech detection in videos. Our key contributions are as follows:

- We introduce a new dataset specifically curated for implicit hate speech detection in videos. The dataset consists of 2,009 videos and provides a valuable benchmark for future research on multimodal hate speech detection.
- We propose a two-stage contrastive learning approach to effectively model multimodal hateful content in videos. In the first stage, we train three modality-specific encoders (audio, text, and image) using contrastive loss, computed over concatenated feature representations. In the second stage, we train a cross-encoder using contrastive learning to refine multimodal representations further.
- We evaluate our approach on both our newly curated dataset and the publicly available HateMM dataset. The results demonstrate the effectiveness of our proposed multimodal contrastive learning framework in detecting hateful content in videos, particularly implicit hate speech.

2 Related Works

2.1 Implicit Hate Speech Detection

Recent studies on implicit hate speech detection have primarily focused on text-based approaches. For instance, ElSherief et al. (2021) and Kim et al. (2022) have advanced the field using linguistic analysis to capture subtle hate cues, while Ocampo et al. (2023) and Guo et al. (2023) further enhanced detection by integrating emotion, ambiguity, and multi-feature fusion techniques. Additionally, efforts such as ToxiGen (Hartvigsen et al., 2022) have leveraged machine-generated data to improve model robustness. Despite these advances, the majority of existing work relies solely on textual information, neglecting the rich, multimodal context inherent in video content. In contrast, our study is the first to explore implicit hate speech detection in videos, incorporating not only text but also visual and audio modalities to capture a more comprehensive spectrum of hateful content.

2.2 Hate Speech Detection in Videos

Hate speech detection in videos is an emerging field, yet many existing works focus only on explicit hate. For example, HateMM (Das et al., 2023) introduced a dataset of 1,083 English BitChute videos aimed at binary hate classification only. Similarly, MultiHateClip (Wang et al., 2024a) provides a multilingual dataset of 2,000 YouTube and Bilibili videos with fine-grained labels in English and Chinese, but these labels predominantly target explicit hate content. Additionally, studies by Alcântara et al. (2020) and Wu and Bhandary (2020) have advanced video-based hate speech detection; however, they too tend to overlook the subtleties of implicit hate. This gap underscores the need for comprehensive approaches that not only detect overt hateful content in videos but also capture the nuanced and often overlooked manifestations of implicit hate speech.

Properties	Non Hate	Implicit Hate	Explicit Hate
Video count	1,000 (49.78%)	509 (25.36%)	500 (24.89%)
Total length	39 hours 26 mins 42 secs	18 hours 7 mins 51 secs	28 hours 58 mins 25 secs
Mean video length	2 mins 22 secs	2 mins 8 secs	2 mins 38 secs
Total number of frames	1,42,002	65,271	79,105
Mean number of frames	142.002	128.23	158.21
Mean number of words	175.404	85.166	80.326

Table 1: Dataset statistics.

3 Data Collection and Annotation

3.1 Platforms for Video Collection

For data collection we primarily used BitChute[5], a social video-hosting platform launched in 2017 with minimal content moderation. It has grown in prominence as an alternative to YouTube, hosting a significant amount of hateful content banned from mainstream platforms. In addition to BitChute, we also collected videos from Odysee[6], another alternative video-sharing platform.

3.2 Dataset Statistics

Our dataset consisted of 2,009 videos comprising 86.5 hours of English multimodal content collected from BitChute and Odysee. Among these, 1,000 videos were classified as non-hate, while the remaining 1,009 videos contained hateful content. The hate videos were further categorized into implicit hate and explicit hate, with 509 and 500 videos, respectively, to capture different degrees of hateful expressions. The details of the dataset can be seen in Table 1. The dataset is balanced with almost 50% hate and non-hate content by count. Duration-wise, we have 47 hours of hate content compared to 39.5 hours of non-hate content. Within the hate category, the number of samples of implicit and explicit hate is also roughly balanced. On average, all the videos are approximately 2 minutes long, with approximately 143 frames captured per video. Non-hate videos have a little more than double the number of words in transcripts than those in implicit and explicit hate videos.

3.3 Annotation Guidelines

The following labeling scheme provided the main framework for annotators, while a codebook ensured consistency in label interpretation as inspired by Das et al. (2023), Wang et al. (2024a), and Salles et al. (2025). Developed using YouTube's hate speech policy, the codebook contains detailed annotation guidelines. A video is considered hateful if:

"It promotes discrimination, disparages, or humiliates an individual or group based on characteristics such as race, ethnicity, nationality, religion, disability, age, veteran status, sexual orientation, or gender identity."

Moreover, the annotators were guided to pinpoint particular segments of a hate video (i.e., frame spans) they deemed hateful and indicate the specific communities targeted by the content.

3.4 Annotation Process

Annotator Training

The annotation process was supervised by one Professor and one PhD student with expertise in analyzing harmful content on social media, while the actual annotations were carried out by 1 postgraduate and three undergraduate students who were novice annotators. All annotators were computer science majors and participated voluntarily with full consent. As a token of appreciation, they were rewarded with free access to A100 GPU for 150 hours.

The annotators were trained by creating an initial gold-standard dataset. The expert annotators labeled 50 videos, comprising 30 hate videos and 20 non-hate videos which were then provided to the undergraduate annotators for labeling based on the annotation codebook. After completing their annotations, we reviewed and discussed the incorrect cases with them to refine their understanding and improve their annotation accuracy.

Annotation in Batch Mode

Following the initial training, we adopted a batch-mode approach, releasing a set of 50 videos per week for annotation. Given the potential negative psychological effects of annotating hate content (Ybarra et al., 2006), we advised annotators to take a minimum 10-15 minute break after labeling each video. Additionally, we imposed a strict limit of no more than 20 videos per day to prevent cognitive overload. To ensure the well-being of the annotators, we also conducted regular check-in meetings to monitor any potential adverse effects on their

[5]https://www.bitchute.com/
[6]https://odysee.com/

mental health.

4 Methodology

We propose a two-stage contrastive learning multimodal framework for detecting hateful content in videos as shown in Figure 2. Our approach comprises three key steps: (i) preprocessing to extract audio, text, and visual data; (ii) feature extraction using modality-specific encoders; and (iii) contrastive learning to align representations, culminating in a multimodal classifier.

4.1 Preprocessing

Videos are converted to WAV audio using FFmpeg and transcribed via speech-to-text conversion. For the visual modality, 100 frames are uniformly sampled (with padding for videos having fewer frames) using VideoCapture, ensuring consistent input dimensions across samples.

4.2 Feature Extraction

We extract features from image, text, and audio using a video-based contrastive learning method, ImageBind (Girdhar et al., 2023), that maps raw inputs into a shared 1024-dimensional space. We denote the extracted features with f_I, f_T, and f_A for the image, text and audio, respectively.

In addition to extracting the 1,024-dimensional audio, text, and video features, we extract complementary features to enrich our multimodal representation. For a text transcription x_T, we compute emotion features $e = \text{NRCLex}(x_T)$ (with $e \in \mathbb{R}^{d_e}$) and a sentiment score $s = \text{Vader}(x_T)$ ($s \in \mathbb{R}$); these are concatenated into a joint representation $f_{ES} = [e, s] \in \mathbb{R}^{d_e+1}$. Similarly, for an image x_I, we generate a caption $c = \text{caption\_gen}(x_I)$ and extract caption features $f_C = \text{BERT}(c)$ where $f_C \in \mathbb{R}^{d_c}$. Together, these features complement the primary modality representations to provide a comprehensive view of the content.

4.3 Two-stage Contrastive Learning

4.3.1 Stage 1: Modality-specific Encoder Training

To jointly optimize the image, text, and audio encoders from the first stage, denoted by encoder_{II}, encoder_{TT}, and encoder_{AA}, we merge their outputs via a projection head. Let the encoder outputs be defined as f_{II}, f_{TT}, and f_{AA} for the image, text, and audio modalities, respectively, from the first stage encoders. These features, each of dimension d (e.g., 1,024), are concatenated to form a joint representation:

$$f_{ITA} = \text{Concat}(f_{II}, f_{TT}, f_{AA}) \in \mathbb{R}^{3d} \quad (1)$$

A projection head $P : \mathbb{R}^{3d} \to \mathbb{R}^{d'}$ is then applied to map the concatenated features into a shared embedding space:

$$z_{ITA} = P(f_{ITA}) \in \mathbb{R}^{d'} \quad (2)$$

The merged encoder (denoted as ITA) is optimized using a supervised contrastive loss. Specifically, for a batch of N samples, the loss is defined as:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\sum_{j\in\mathcal{P}(i)} \exp\left(\frac{\text{sim}(z^i_{ITA}, z^j_{ITA})}{\tau}\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\frac{\text{sim}(z^i_{ITA}, z^k_{ITA})}{\tau}\right) + \epsilon} \quad (3)$$

where: $\text{sim}(z^i_{ITA}, z^j_{ITA})$ is the cosine similarity between embeddings, τ is the temperature parameter, ϵ is a small constant for numerical stability, $\mathcal{P}(i)$, and $\mathcal{N}(i)$ denote the sets of indices corresponding to positive (same class) and negative (different classes) pairs for sample i, respectively. In this way, the individual encoders encoder_{II}, encoder_{TT}, and encoder_{AA} are optimized jointly using the merged ITA representation, thereby aligning multimodal features in a unified embedding space. Moreover, specialized encoders for emotion-sentiment (ES) and image captions (CP) are trained in a similar manner.

4.3.2 Stage 2: Cross-Modal Encoder Training

In this stage, we align representations across modalities by training cross-modal encoders that merge outputs from modality-specific encoders via a projection head. We can understand this through an example of image_text cross encoders. The features from 1st-stage encoders are first processed by cross encoders encoder_{IT} and encoder_{TI} to produce refined features f_{IT} and f_{TI}:

$$f_{IT} = \text{encoder}_{IT}(f_{II}) \quad (4)$$

$$f_{TI} = \text{encoder}_{TI}(f_{TT}) \quad (5)$$

Next, a projection head $P(\cdot)$, implemented as a dense layer with ReLU activation, maps these to a shared embedding space:

$$z_{IT} = P(f_{IT}) \quad (6)$$

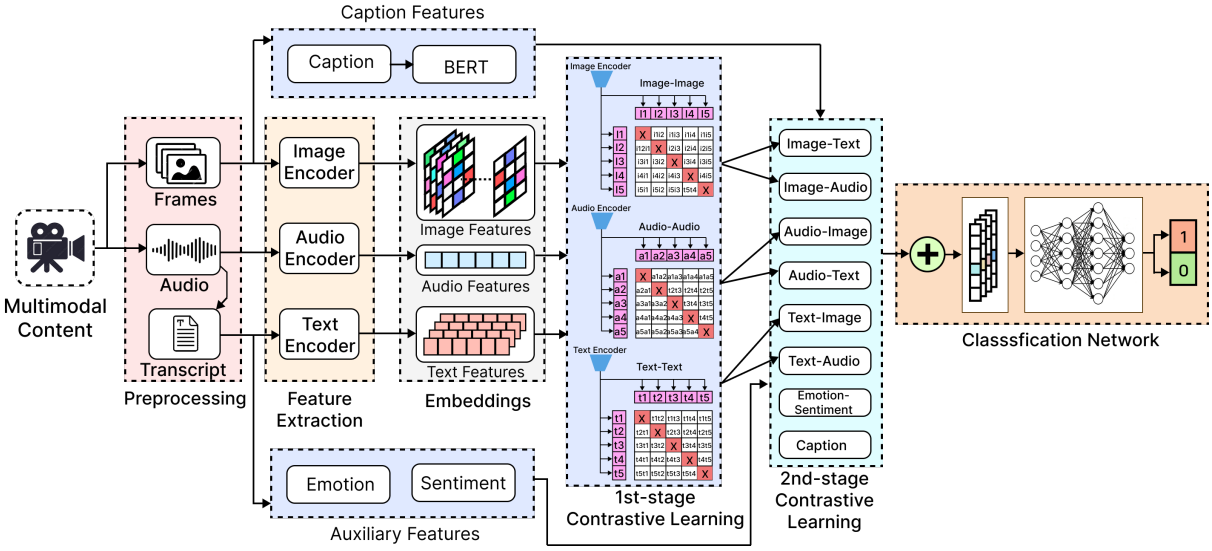

Figure 2: The proposed method first extracts modality-specific features from text, images, and audio using dedicated encoders. It then aligns these features in a shared embedding space through cross-modal encoders with projection heads and supervised contrastive loss. Additionally, emotion, sentiment and caption features are also extracted.

$$z_{TI} = P(f_{TI}) \tag{7}$$

The embeddings are concatenated to form the cross-modal representation:

$$z_{\text{cross}} = \text{Concat}(z_{IT}, z_{TI}) \tag{8}$$

Subsequently, the cross-modal encoder is optimized using a supervised contrastive loss. In addition to the image_text encoders, we similarly train cross encoders for image_audio (IA and AI) and text_audio (TA and AT) pairs.

The overall training objective for our system combines the losses from the first stage (modality-specific encoders), the cross-modal encoders, and the specialized ES and CP encoders:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{stage 1}} + \mathcal{L}_{\text{stage 2}} + \mathcal{L}_{\text{sup}}^{ES} + \mathcal{L}_{\text{sup}}^{CP} \tag{9}$$

By jointly optimizing these components, the framework effectively aligns and integrates multimodal features from image, text, and audio. This unified representation enhances the model's ability to capture complementary information, thereby improving performance in tasks such as detecting hateful content in videos.

4.4 Multimodal Classification

The learned representations f_{IT}, f_{IA}, f_{TI}, f_{TA}, f_{AI}, f_{AT}, f_{ES}, and f_{CP} are concatenated to form a unified feature F. This vector is passed through dense layers with ReLU activations and dropout regularization to produce the final prediction:

$$y = \text{softmax}\Big(W_4\,\text{Dropout}\big(\sigma(W_3\,\text{Dropout}(\sigma(W_2\,\text{Dropout}(\sigma(W_1 F + b_1)) + b_2)) + b_3)\big) + b_4\Big) \tag{10}$$

where $y \in \mathbb{R}^C$ represents the predicted probability distribution over C classes.

This streamlined framework effectively integrates multimodal data and contrastive learning to improve the detection of hateful content in videos.

5 Results

5.1 Datasets

In addition to the proposed ImpliHateVid dataset, we used HateMM (Das et al., 2023), another multimodal dataset, to evaluate the performance of our model. HateMM consists of 1,083 videos in total,l comprising 43.26 hours of multimodal content in English collected from BitChute and Odysee platforms. Out of the 1,083 videos, 431 videos have been labeled as hate videos and 652 videos have been labeled as non-hate videos. However, we used only 1,035 videos out of 1,083 for our study.

5.2 Experimental Setup

We split both datasets into train, validation, and test sets containing the total videos, respectively. The HateMM dataset had 662, 166, and 210 samples in training, validation, and test sets, respectively, while our dataset, ImpliHateVid, had 1,283 samples in the training set, 325 samples in the validation set,

Modality	Method	ImpliHateVid Dataset				HateMM Dataset			
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Text	BERT	0.6907	0.6884	0.6907	0.6907	0.7350	0.6640	0.6750	0.6670
	GPT-4o	0.5312	0.1132	**1.0000**	0.0600	0.5652	0.1964	0.3793	0.1325
	Llama 3.1-8b	0.5312	0.2034	0.6667	0.1200	0.5459	0.2540	0.3721	0.1928
Image	ViT	0.7655	0.7684	0.7658	0.7656	0.7480	0.6720	0.6950	0.6560
	ViViT	0.4912	0.5255	0.4912	0.4914	0.5293	0.5176	0.5172	0.5182
Audio	MFCC	0.4987	0.6655	0.2493	0.5000	0.6750	0.6220	0.5930	0.6790
	Wav2Vec2	0.7531	0.7724	0.7610	0.7533	0.5810	0.5810	0.5270	0.5160
Video	GPT-4	0.4988	0.6656	0.4988	**1.0000**	0.4010	0.5724	0.4010	**1.0000**
	LlamaVL	0.4010	0.5724	0.4010	**1.0000**	0.3800	0.5300	0.3700	0.9500
Multimodal	DeepCNN	0.7623	0.7800	0.7481	0.7933	0.5622	0.3065	0.4565	0.2307
	CMHFM	0.7922	0.7921	0.7860	0.7980	0.6057	0.5629	0.6184	0.6184
	CSID	0.8150	0.8154	0.8082	0.8233	0.7320	0.7140	0.7200	0.7230
	MCMF	0.8224	0.8220	0.8200	0.8240	0.5769	0.0435	0.5000	0.2422
	MulT	0.8352	0.8352	0.8320	0.8380	0.6571	0.5212	0.4318	0.6571
	Proposed Method	**0.8753**	**0.8773**	0.8796	0.8752	**0.9758**	**0.9758**	**0.9745**	0.9710

Table 2: Effectiveness comparison for binary classification across different methods and datasets.

Modality	Method	Non Hate Videos				Implicit Hate Videos				Explicit Hate Videos				Overall
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
Text	BERT	0.7195	0.7192	0.7122	0.7264	0.7107	0.2927	0.4138	0.2264	0.7207	0.5172	0.4348	0.6383	0.5907
	GPT-4o	0.5362	0.6804	0.5197	**0.9851**	0.7282	0.0684	0.3636	0.0377	**0.7880**	0.1748	**1.0000**	0.0957	0.3078
	Llama 3.1-8b	0.5771	0.5066	0.4514	0.5771	0.0189	0.0231	0.0299	0.0189	0.4043	0.4444	0.4935	0.4043	0.3247
Image	ViT	0.7805	0.7854	0.7703	0.8010	0.7307	0.4906	0.4906	0.4906	0.7706	0.4889	0.5116	0.4681	0.5883
	ViViT	0.5012	0.5745	0.5019	0.6716	0.6559	0.1786	0.2419	0.1415	0.6708	0.1951	0.2286	0.1702	0.3161
Audio	MFCC	0.5262	0.6769	0.5142	0.9900	**0.7357**	0.2563	0.3180	0.2146	0.7506	0.0741	0.2857	0.0426	0.2503
	Wav2Vec2	0.7781	0.7963	0.7357	0.8657	**0.7357**	0.3117	0.5000	0.2264	0.7930	**0.6066**	0.5470	**0.6809**	0.5716
Video	GPT-4	0.4938	0.6381	0.4972	0.8905	0.7082	0.0488	0.1765	0.0283	0.7556	0.1695	0.4167	0.1064	0.2855
	Llama-VL	0.4250	0.4800	0.4000	0.7800	0.2500	0.0250	0.1000	0.0150	0.3800	0.1500	0.3800	0.1000	0.2180
Multimodal	DeepCNN	0.7623	0.7512	0.7634	0.7398	0.6785	0.6345	0.6189	0.6512	0.6612	0.5803	0.5692	0.5917	0.6587
	CMHFM	0.7645	0.7534	0.7701	0.7405	0.6802	0.6372	0.6241	0.6509	0.6634	0.5814	0.5723	0.5922	0.6604
	CSID	0.7658	0.7556	0.7714	0.7437	0.6814	0.6394	0.6273	0.6534	0.6649	0.5834	0.5745	0.5938	0.6621
	MCMF	0.7661	0.7568	0.7735	0.7452	0.6819	0.6401	0.6289	0.6541	0.6652	0.5845	0.5751	0.5942	0.6625
	MulT	0.7667	0.7574	0.7741	0.7459	0.6823	0.6408	0.6296	0.6548	0.6658	0.5849	0.5756	0.5948	0.6627
	Proposed Method	**0.8955**	**0.8448**	**0.8000**	0.8955	0.6698	**0.6605**	**0.6513**	**0.6698**	0.4894	0.5702	0.6866	0.4894	**0.6918**

Table 3: Effectiveness comparison for multiclass classification across different methods on ImpliHateVid dataset.

and 401 samples in the test set. We experimented with several combinations of hyperparameters. 32 and 64 were used as batch sizes. Our learning rate was in the range 1e-3,1e-4,1e-5, and we trained our model for 30, 50, 75, and 100 epochs. Adam optimizer (Diederik, 2014) was used. We used Accuracy (Acc), Precision (Prec), Recall (Rec), F1-score (F1), and Macro-F1 metrics to evaluate the performance.

5.3 Compared Methods

We have compared our proposed method with several unimodal and multimodal methods to demonstrate the effectiveness of our approach. For the textual modality, we evaluated the performance of **BERT** (Kenton and Toutanova, 2019) alongside large language models (LLMs), including **GPT-4o** (Radford, 2018) and **Llama 3.1-8b** (Touvron et al., 2023). For the vision modality, we utilized **ViT** (Dosovitskiy, 2020) and **ViViT** (Arnab et al., 2021), while for the audio modality, we employed **MFCC** (Jung et al., 2021) and **Wav2Vec2** (Baevski et al., 2020). To assess video performance holistically, we also incorporated vision-language models such as **GPT-4**[7] and **Llama-VL** (Zhang et al., 2023). Features extracted from these models (excluding GPT and Llama) were fed into a feedforward neural network for classification, comprising four dense layers with 512, 256, 128, and 64 neurons, respectively, and a dropout rate of 0.3. GPT and Llama models were used using APIs with zero-shot prompting. Additionally, we compared our proposed method against other multimodal approaches, including **DeepCNN** (Dixit and Satapathy, 2024) and **CMHFM** (Wang et al., 2024b). The model proposed by Li et al. (2024) is denoted as **CSID**, while the one by Li et al. (2023) is labeled as **MCMF** in our results. We also examined the performance of the Transformer-based model **MulT** (Tsai et al., 2019) across both datasets.

5.4 Effectiveness Comparison

5.4.1 Binary Classification

Table 2 highlights the performance improvements of our proposed multimodal method over the best performing approaches in each category on the ImpliHateVid and HateMM datasets. On the Impli-

[7]https://cdn.openai.com/papers/GPTV_System_Card.pdf

Modality	Non Hate Videos				Implicit Hate Videos				Explicit Hate Videos				Overall
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
Text	0.8905	0.8424	0.7991	0.8905	0.2264	0.3038	0.4615	0.2264	0.7447	0.6393	0.5600	0.7447	0.5951
Image	0.8607	0.8317	0.8046	0.8607	0.2453	0.3077	0.4127	0.2453	0.7234	0.6267	0.5587	0.7234	0.5587
Audio	0.7811	0.7677	0.7548	0.7811	0.4151	0.4583	0.5116	0.4151	0.6064	0.5672	0.5327	0.6064	0.5977
Text + Audio	0.8657	0.8208	0.7803	0.8657	0.3113	0.3750	0.4714	0.3113	0.6915	0.6435	0.6019	0.6915	0.6131
Audio + Image	0.8706	0.8413	**0.8140**	0.8706	0.2642	0.3394	0.4746	0.2642	**0.7872**	**0.6697**	0.5827	**0.7872**	0.6168
Text + Image	0.6816	0.6903	0.6995	0.6816	0.1604	0.2716	**0.8500**	0.1604	0.5532	0.3728	0.2811	0.5532	0.4448
Proposed Method	**0.8955**	**0.8448**	0.8000	**0.8955**	**0.6698**	**0.6605**	0.6513	**0.6698**	0.4894	0.5702	**0.6866**	0.4894	**0.6918**

Table 4: Impact of different modalities on multiclass classification on ImpliHateVid dataset.

Features	Non Hate Videos				Implicit Hate Videos				Explicit Hate Videos				Overall
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
w/o Emotions	0.8209	0.8312	**0.8418**	0.8209	0.6415	0.6507	0.6602	0.6415	**0.5851**	0.5609	0.5392	**0.5851**	0.6809
w/o Captions	0.8258	0.8217	0.8177	0.8258	0.6509	0.6448	0.6389	0.6509	0.5425	0.5542	0.5667	0.5425	0.6736
w/o Sentiments	0.8408	0.8325	0.8244	0.8408	0.6226	0.6438	**0.6667**	0.6226	0.5638	0.5548	0.5463	0.5638	0.6770
Proposed Method	**0.8955**	**0.8448**	0.8000	**0.8955**	**0.6698**	**0.6605**	0.6513	**0.6698**	0.4894	**0.5702**	**0.6866**	0.4894	**0.6918**

Table 5: Impact of different features on multiclass classification on ImpliHateVid dataset.

HateVid dataset, our method achieves an F1-score of 87.73%, which is approximately 10.5% higher than the best unimodal method, Wav2Vec2, which attains an F1-score of 77.24%. When compared to the leading multimodal baseline, MulT, with an F1-score of 83.52%, our approach shows a 4.21% improvement.

On the HateMM dataset, the differences are even more pronounced. Our method reaches an F1-score of 97.58%, outperforming the best text-based model, BERT, which records an F1-score of 66.40%, by roughly 31.18%. Similarly, against the top multimodal method in this category, CSID, with an F1-score of 71.40%, our proposed approach gains about 26.18%. These substantial gains in F1-score clearly demonstrate the effectiveness of our multimodal framework in leveraging complementary cues from text, image, and audio modalities, thereby significantly enhancing implicit hate speech detection across different datasets.

5.4.2 Multiclass Classification

Table 3 presents the multiclass classification results across non-hate, implicit hate, and explicit hate video categories. Our proposed method achieves an overall macro-F1 of 69.18%, which represents a significant improvement over the best unimodal approaches and existing multimodal baselines. For instance, among unimodal models, the text-based BERT attains a macro-F1 of 59.07%, while the image-based ViT and audio-based Wav2Vec2 achieve 58.83% and 57.16%, respectively. This indicates that our model improves the macro-F1 score by roughly 10.1% points over the best unimodal method. GPT and Llama fail to identify instances of implicit hate accurately due to noisy transcriptions. When compared to multimodal baselines, our approach also demonstrates clear superiority. The strongest competing multimodal method, MulT, records a macro-F1 of 66.27%; hence, our proposed model exhibits an absolute improvement of approximately 2.91%. Additionally, our model consistently outperforms other multimodal methods such as DeepCNN, CMHFM, CSID, and MCMF across individual categories such as non-hate, implicit hate, and explicit hate, highlighting its balanced performance. These results underscore the effectiveness of our two-stage contrastive learning framework in integrating complementary textual, visual, and audio cues, thereby establishing a new state-of-the-art for multiclass hate speech detection in videos.

5.5 Ablation Analysis

5.5.1 Impact of Different Modalities

We compared the performance of our model, considering all combinations of the three modalities on the ImpliHateVid dataset. The results of binary classification can be seen in Figure 3 and those of three-class classification have been highlighted in Table 4.

Examining the table for unimodal results, the text-only model achieves strong performance for non-hate videos with an F1 of 84.24% but struggles with implicit hate content with F1 of 30.38%. Similarly, the image-only and audio-only models show moderate performance, with audio achieving a relatively higher F1 of 45.83% on implicit hate videos compared to text and image modalities.

The text + audio combination boosts performance on implicit hate videos with an of F1 of

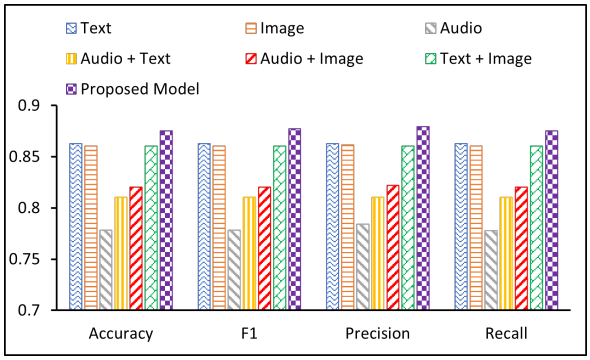

Figure 3: Impact of different modality combinations on binary classification on ImpliHateVid dataset.

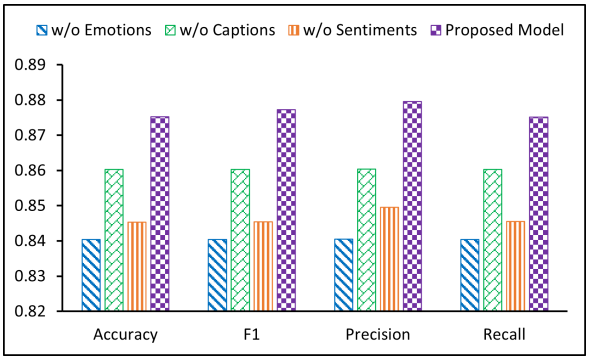

Figure 4: Impact of different features on binary classification on ImpliHateVid dataset.

37.50% compared to using text alone. The audio + image configuration yields a notable improvement for explicit hate videos with an F1 of 66.97%, highlighting the benefit of integrating complementary visual and auditory cues. Conversely, the text + image combination, despite showing promise in precision, falls short in overall effectiveness, as indicated by a lower macro-F1 of 44.48%.

Our proposed method, which integrates textual, visual, and audio features simultaneously, outperforms all the ablated configurations. It achieves an overall accuracy of 89.55%, an F1-score of 84.48% for non-hate videos, and significantly higher performance for implicit hate with an F1 of 66.05% and explicit hate videos with an F1 of 57.02%. The overall macro-F1 of 69.18% clearly demonstrates the advantage of fully leveraging multimodal information. Figure 3 further illustrates this performance gain, where the proposed model outperforms all other configurations in binary classification. These results emphasize that combining all three modalities leads to a more robust and balanced classification outcome across all categories.

5.5.2 Impact of Emotion, Caption, and Sentiment Features

Table 5 summarizes the impact of omitting specific features on the three-class classification performance. For each ablation, removing emotions, captions, or sentiments, the table reports accuracy, F1-score, precision, and recall for each video category, along with the overall macro-F1 score and overall accuracy. Notably, the full proposed model, which uses all features, achieves a balanced performance with an overall accuracy of 71.32% and a macro-F1 of 66.29%, outperforming the ablated versions. The binary classification results are shown in Figure 4.

6 Conclusion

This work makes two significant contributions. First, we introduce ImpliHateVid, a novel dataset specifically curated for implicit hate speech detection in videos. Comprising 2,009 videos, ImpliHateVid represents one of the first large-scale benchmarks for video-based implicit hate detection. This dataset not only fills a critical gap in the literature but also provides a valuable resource for advancing research in multimodal hate speech analysis. Second, we propose a two-stage contrastive learning framework that effectively integrates textual, visual, and audio modalities. In the first stage, dedicated encoders extract robust, modality-specific features. These features are then aligned into a unified embedding space via cross-modal encoders in the second stage, using a supervised contrastive loss that clusters similar samples and separates dissimilar ones. Additionally, our model is enhanced by incorporating sentiment, emotion, and image caption features, which capture subtle cues associated with hate speech. Extensive experiments on ImpliHateVid and the HateMM dataset demonstrate that our approach significantly outperforms all the baselines. These findings underscore the effectiveness of leveraging cross-modal information to capture the full context of hateful content in videos. Future work will focus on extending the framework to incorporate multilingualism and exploring real-time detection applications. Overall, our contributions pave the way for more accurate systems in combating online hate in videos.

Limitations

Despite the promising results, our approach has several limitations. First, the performance of the multimodal framework depends heavily on the quality

and alignment of data across modalities. Noisy or misaligned text, image, or audio inputs can adversely affect the joint embedding space and, consequently, the overall classification accuracy. Second, the reliance on pre-trained encoders means that any shortcomings in these models, such as domain mismatch or insufficient representation of hate speech nuances, can limit the effectiveness of our system.

Third, the supervised contrastive loss, while effective, is sensitive to hyperparameter settings such as the temperature parameter and the strategy for selecting positive and negative pairs. Improper tuning can lead to suboptimal clustering of similar samples and separation of dissimilar ones. Future research should focus on developing more robust and adaptive models and expanding the dataset to cover a broader range of hate speech phenomena.

Ethical Considerations

Data Collection and Terms of Service

Our dataset consists of videos collected from publicly accessible platforms BitChute and Odysee. The collection process adhered to standard practices established in prior research on video-based hate speech detection (Wang et al., 2024a; Das et al., 2023), ensuring consistency with ethical norms in the field. Only publicly available videos were gathered; no restricted, private, or login-gated content was included. We ensured that our data collection methods were fully compliant with the terms of service of both platforms.

Confidentiality and Anonymization

Respecting user privacy is a core principle of our research. While the dataset comprises only publicly accessible content, we implemented rigorous measures to anonymize any potentially identifying information. No usernames, channel metadata, or personally identifiable information were retained in the dataset. Our procedures align with ethical guidelines such as those outlined by Rivers and Lewis (2014) and mirror the practices adopted in recent benchmarks such as MultiHateClip (Wang et al., 2024a) and HateMM (Das et al., 2023).

We recognize that achieving complete anonymization in video data is a complex challenge, especially given the visual nature of the content. We made no attempt to de-anonymize users or track individuals across platforms. Our analysis focuses solely on the content of the videos, not the individuals or channels behind them.

Intended Use and Responsible AI Considerations

The dataset is strictly intended for academic research on hate speech detection and content moderation. Access will be granted only to researchers with a demonstrated interest in this domain and will be contingent upon ethical review approval. Dataset access will be provided exclusively for research purposes and is not licensed for commercial use or any application that could be deemed harmful. We encourage researchers to engage with this resource in a manner consistent with the ethical standards of the community.

Our database comprises videos annotated with hate speech labels but contains no personally identifiable information. We reiterate that we analyzed only publicly available data and followed established ethical norms, making no effort to track or identify individual users.

Biases

While every effort was made to ensure fairness, we acknowledge the possibility of biases within the dataset. The classification of content as hateful or non-hateful can be subjective, and unintentional biases may be present in label distribution or annotation judgments. However, our annotations exhibit high inter-annotator agreement, which provides confidence in the overall reliability of the labels. We stress that no part of the dataset is intended to malign any individual or community, and any bias present is unintended.

Acknowledgments

We would like to thank all the reviewers for their valuable feedback and constructive suggestions, which helped improve the quality of this work. We also thank Bitchute and Odysee as the sources from which we collected publicly available video data for our research on hate speech detection. The data was used in accordance with the platforms' terms of service and solely for academic and non-commercial research purposes. The authors are also thankful for the Young Faculty Research Catalysing Grant (YFRCG) by the Indian Institute of Technology Indore for providing the resources to the project (Project ID: IITI/YFRCG/2023-24/03).

References

Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: dataset and base-

line results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4309–4319.

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

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

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

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

P Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)*.

Chhavi Dixit and Shashank Mouli Satapathy. 2024. Deep cnn with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications*, 240:122579.

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

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Tengda Guo, Lianxin Lin, Hang Liu, Chengping Zheng, Zhijian Tu, and Haizhou Wang. 2023. Implicit offensive speech detection based on multi-feature fusion. In *International Conference on Knowledge Science, Engineering and Management*, pages 27–38. Springer.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Ayako Hatano. 2023. Regulating online hate speech through the prism of human rights law: The potential of localised content moderation. *The Australian Year Book of International Law Online*, 41(1):127 – 156.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramón Rodriguez. 2021. Audio-based hate speech classification from online short-form videos. In *2021 International Conference on Asian Language Processing (IALP)*, pages 72–77. IEEE.

Shing-Yun Jung, Chia-Hung Liao, Yu-Sheng Wu, Shyan-Ming Yuan, and Chuen-Tsai Sun. 2021. Efficiently classifying lung sounds through depthwise separable cnn models with fused stft and mfcc features. *Diagnostics*, 11(4):732.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Yangyang Li, Yuelin Li, Shihuai Zhang, Guangyuan Liu, Yanqiao Chen, Ronghua Shang, and Licheng Jiao. 2024. An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency. *Knowledge-Based Systems*, 287:111457.

Zuhe Li, Qingbing Guo, Yushan Pan, Weiping Ding, Jun Yu, Yazhou Zhang, Weihua Liu, Haoran Chen, Hao Wang, and Ying Xie. 2023. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. *Information Fusion*, 99:101891.

Roshan Nayak, BS Ullas Kannantha, C Gururaj, and 1 others. 2022. Multimodal offensive meme classification u sing transformers and bilstm. *Int. J. Eng. Adv. Technol.*, 11(3):96–102.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate

speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3:38.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. Hatebrxplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in brazilian portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, and 1 others. 2022. Detecting and understanding harmful memes: A survey. In *31st International Joint Conference on Artificial Intelligence*, pages 5597–5606.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

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

Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, and 1 others. 2024b. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61(3):103675.

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

Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, 118(4):e1169–e1177.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

Appendix

A Data Curation and Annotation

A.1 Data Sources and Collection Timeframe

Videos were sourced from the platforms *BitChute* and *Odysee*, known for hosting user-generated content with diverse viewpoints. To mitigate temporal bias, we did not confine our data collection to a specific timeframe. This approach ensured a diverse representation of content, capturing various socio-political contexts and minimizing the risk of overrepresenting events from a particular period.

A.2 Search Terms and Filtering Criteria

To identify relevant videos, we employed a set of predefined search terms associated with hate speech and extremist content. These included terms such as, *"hate speech," "racism," "hate speech against black poeple," "hate speech against women," "misogyny," "anti-white hate speech,"* and *"abuse"*

The following filtering criteria were applied:

- **Language:** Videos in English.
- **Content Type:** Videos containing spoken content, ensuring the presence of audio for transcription and analysis.
- **Availability:** Publicly accessible videos without age restrictions or login requirements.

A.3 Annotation Categories and Guidelines

A.3.1 Annotation Categories

The annotation process involved categorizing content into the following labels:

- **Explicit Hate Speech**: Content that overtly promotes hatred or discrimination against a protected group.
- **Implicit Hate Speech**: Content that subtly conveys hateful messages, often through sarcasm, humor, or coded language.

- **Non-Hate Speech**: Content that does not contain hate speech or offensive language.

To handle complexities such as sarcasm, humor, and cultural context differences, annotators followed predefined criteria:

- **Sarcasm & Humor:** Annotators were instructed to identify cases where sarcasm or humor masked hateful intent by analyzing tone, visual cues, and context rather than relying solely on textual content.
- **Cultural Context Differences:** Given the subjective nature of implicit hate speech, we ensured that annotators were trained to recognize culturally specific expressions that could carry hateful undertones.
- **Ambiguous Cases:** Edge cases were reviewed through weekly discussions where disagreements were resolved collaboratively with the supervising professor and PhD researcher, ensuring that annotations reflected consensus-based labeling rather than individual biases.

A.4 Annotator Compensation

To encourage participation and support the annotators' research endeavors, we provided 150 GPU hours for any project of their choice. This approach aimed to promote academic growth and practical experience without offering monetary incentives. Annotators had complete freedom to utilize the GPU resources for hackathons, competitions, project research, or any other purpose.

A.5 Annotator Agreement

Annotators	Cohen's Kappa (Binary)	Cohen's Kappa (Multiclass)
A1–A2	0.8855	0.8516
A1–A3	0.8706	0.8374
A1–A4	0.8417	0.7988
A2–A3	0.8338	0.8056
A2–A4	0.8069	0.7695
A3–A4	0.7939	0.7533
Fleiss' Kappa (Overall)	0.8387	0.8027

Table 6: Inter-annotator agreement using Cohen's and Fleiss' Kappa across binary and multiclass settings.

Table 6 presents inter-annotator agreement measured using Cohen's Kappa for pairwise comparisons and Fleiss' Kappa for overall agreement among four annotators. Results are reported separately for binary and multiclass classification settings. The agreement scores in the binary case are consistently high, with all pairwise Cohen's Kappa values above 0.79 and overall Fleiss' Kappa at 0.8387, indicating substantial agreement. In the multiclass setting, a slight decrease in agreement is observed, with Fleiss' Kappa at 0.8027. These values demonstrate strong annotation consistency across both scenarios.

B Explainanbility Analysis

To further interpret model predictions and validate the effectiveness of our two-stage contrastive learning method, we conducted an explainability analysis on video transcripts by visualizing token-wise attribution scores for the "hate" and "non-hate" classes. This means we analyzed how the model makes decisions by highlighting which words in the transcripts influenced its classification, helping us understand its reasoning.

Figure 5 illustrates two examples: the first instance is classified as hate speech with a high probability of 0.95, and key contributing tokens such as "faggots", "faggot", "woman", "men", and "jail" are highlighted with strong attribution to the hate class. Here, the figure demonstrates that the model correctly identifies hateful content by focusing on offensive or contextually harmful words, which receive high importance scores. These keywords align with overtly offensive and identity-targeted language, reflecting the model's sensitivity to hateful lexicons. This sentence confirms that the model is picking up on language commonly associated with hate speech, indicating it is learning meaningful patterns.

In contrast, the second instance is classified as non-hate with 93% confidence, and the highlighted tokens such as "we", "think", and "future" are aligned with neutral or positive discourse. The non-hate example shows that the model distinguishes safe, generic language and assigns low attribution to hate, reinforcing its discrimination ability. The near-zero attribution scores confirm the model's correct suppression of false positives. This means the model does not mistakenly flag neutral language as hateful, which is important for precision.

This token-level saliency validates that the fused multimodal representations, enhanced through our contrastive learning framework, are not only effective for classification but also interpretable, attribut-

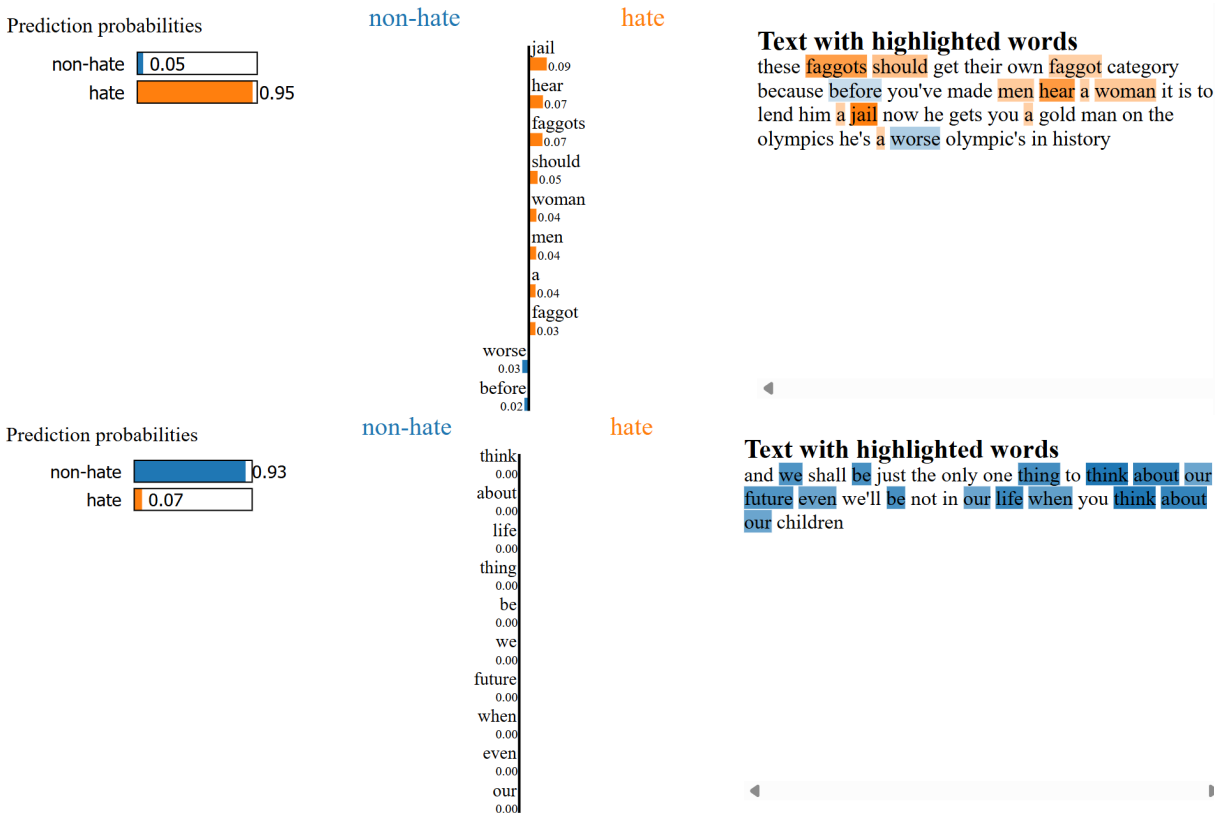

Figure 5: Explainability analysis using token-level saliency visualization on two video transcripts. The left figure corresponds to an instance of hate speech, while the right shows a non-hate instance.

ing decisions to contextually meaningful elements in the textual modality.