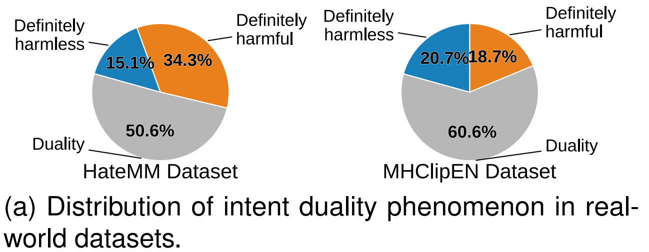
MATCH: Multi-Agentic Evidence Grounding for Explainable Hate Video Detection

Kaiju Li, Rongpei Hong, *Student Member, IEEE*, Jian Lang, *Student Member, IEEE*, Jin Wu, Fan Zhou, *Member, IEEE*, Jingkuan Song, *Member, IEEE*

***Abstract*—The growing prevalence of hate videos promoting intolerance, bigotry, and discrimination presents significant psychosocial threats to both individuals and society. Current detection methods often rely on black-box models, which lack interpretability — a crucial factor for fostering more reliable content moderation and trustworthy AI. To bridge this gap, we propose MATCH, the first attempt to achieve interpretable hate video detection via multiple Large Multimodal Model (LMM) agent collaboration. Our method facilitates a novel *diversely generate-then-verify* paradigm, where LMM agents work in tandem to generate diverse clues and verify them to yield more faithful explanations. MATCH begins by proposing a new Dual-Perspective Proposing paradigm, where two LMM agents are regarded as *Proposers* to independently identify evidential clues from opposing angles – hate and non-hate. Leveraging these comprehensive clues, we introduce an innovative Spatiotemporal Evidence-Grounded Verification mechanism. In this mechanism, a third LMM agent acts as a *Verifier*, rigorously validating and reconciling the proposed clues against spatiotemporal evidences directly extracted from video content, yielding coherent and faithful explanations. Finally, these explanations are integrated with video features, enabling accurate identification of complex and ambiguous hateful content. Extensive experiments conducted on three benchmark datasets demonstrate that MATCH not only achieves state-of-the-art performance, but also provides reliable and trustworthy rationales for the predictions.**

***Index Terms*—hate video detection, interpretability, LMM, agent collaboration**



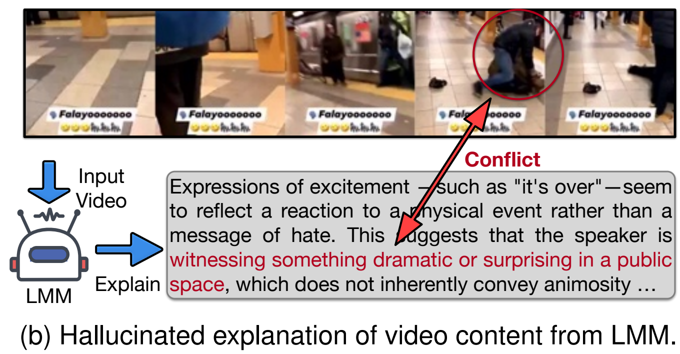


Figure 1. Challenges for vanilla LMMs in achieving interpretable hateful video detection. (a) Most videos in real-world datasets show intent duality, challenging vanilla LMM with single-perspective interpretation. (b) Vanilla LMMs struggle with long-range temporal modelling, often hallucinating interpretations that contradict the video, e.g., missing a fight and misreading the scene as a non-hate prank.

I. Introduction

The emergence of online video platforms, such as TikTok and Instagram Reels, offers an effective and efficient way for individuals to post and view videos, significantly influencing personal daily-life decisions [1]–[5]. Due to the unique characteristics of videos, *i.e.,* strong attraction and emotional appeal, they have become a highly effective medium for rapidly spreading hateful content, exacerbating discrimination based on race, religion, and gender [6], [7]. Consequently, developing effective hate video detection (HVD) approaches to efficiently identify hateful content in various videos has become an urgent necessity.

Existing methods in HVD mainly adopted pretrained vision-language model (e.g., CLIP [8]) to extract the multimodal content (i.e., visual, textual, and audio features) in videos and fuse them for prediction [9]–[11]. However, a significant *limitation* of these methods is their reliance on black-box models, which fails to provide interpretability regarding the rationale behind classifying a video as hateful. Such lack of transparency prevents viewers and platform moderators from understanding the basis for hate video classifications, thereby limiting the models' trustworthy and real-world application. Consequently, developing interpretable hateful video detection systems is essential, which facilitates more reliable content moderation and fosters user trust in the hate determination.

Nevertheless, designing an interpretable model for HVD is non-trivial, as it requires a comprehensive understanding

Manuscript received November 4, 2025.

This work was supported by the National Natural Science Foundation of China (Grant No.62572097 and No. U23A20315). *(Corresponding author: Jin Wu and Fan Zhou)*

Kaiju Li is with the University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: lkj2025@std.uestc.edu.cn).

Rongpei Hong is with the University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: rongpei.hong@std.uestc.edu.cn).

Jian Lang is with the University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: jian_lang@std.uestc.edu.cn).

Jin Wu is with the University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: wj@uestc.edu.cn).

Fan Zhou is with the University of Electronic Science and Technology of China, Chengdu 610054, China, the Intelligent Digital Media Technology Key Laboratory of Sichuan Province, Chengdu 611731, China and also with the Kash Institute of Electronics and Information Industry, Kashgar, 844000, China (e-mail: fan.zhou@uestc.edu.cn).

Jingkuan Song is with the University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: jingkuan.song@gmail.com).

Copyright © 2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

of multifaceted intentions and complex contexts conveyed by videos. Existing HVD methods often fall short in providing meaningful explanations for their predictions, primarily because they simply rely on surface-level features for classification rather than a holistic understanding of the intent and context of videos. The emergence of Large Multimodal Models (LMMs) [12], [13], with their rich pretrained knowledge and cross-modal reasoning capabilities, brings an encouraging path toward generating explanations for HVD. However, directly applying LMMs to realize interpretable hate detection in video domain remains insufficient, due to two critical challenges:

C1: Ambiguous and Multifaceted Video Intent. Creator intent in video content is often ambiguous and multifaceted, posing challenges for accurately interpreting or explaining the underlying motivations. We refer to this phenomenon as *intent duality*, which arises in cases where non-hateful videos may include hateful elements for narrative or educational purposes, such as historical illustration or critical commentary. Conversely, deliberately hateful videos may obscure their toxic intent through subtle or context-dependent expressions to evade content moderation. As illustrated in Figure 1(a), the pie charts present the distribution of intent duality in the HateMM [9] and MHClipEN [10] datasets, with results highlighting the high prevalence of such cases. These statistics were derived by analyzing each video's visual content, on-screen text, and transcript, followed by independent judgments from two annotators. However, vanilla LMMs typically predict the class of a given video before generating explanations, often resulting in one-sided interpretations that overlook the nuanced and multi-angle nature of video intent.

C2: Complex Temporal Dynamics of Video Content. The inherent temporal dynamics further exacerbate the challenges of vanilla LMMs in interpreting hateful content. Most existing LMMs are primarily pretrained on static images or short multi-image dialogues [14], [15], rendering them ill-suited for modelling long-range temporal dependencies. As a result, when applied to videos with long and complex context [16], these models often produce hallucinated or erroneous explanations for specific segments [17], [18]. For instance, in Figure 1(b), LMM misses violent segments in a long video, and misinterpreting it as a non-hate public prank. Such failures not only compromise the accuracy of hate video classification, but also critically undermine the reliability of the generated interpretations, posing a substantial threat to credible interpretability.

Consequently, relying solely on a *single* LMM is insufficient for effectively tackling the intricate challenges and achieving promising explainable hate video detection. To bridge this gap, we introduce a novel **M**ultiple LMM **A**gen**T** **C**ollaboration framework for interpretable **H**ate video detection (**MATCH**). To well overcome the two challenges, our framework draws inspiration from cognitive science research, which suggests that *humans handle multi-faced and ambiguous information by initially forming diverse hypotheses based on multi-angle observations, followed by rigorous verification and refinement through targeted analysis*. Aligned with this cognitive principle, we propose a *diversely generate-then-verify* paradigm to enhance the interpretability of hate video detection. In this paradigm, two LMM agents act as **proposers**, reasoning about potential evidential clues and providing preliminary justifications from two complementary angles, while the third agent serves as a **verifier**, critically assessing and refining the generated clues.

Specifically, to tackle **C1**, we introduce a Dual-Perspective Proposing paradigm that leverages the strong instruction-following capabilities [19] of LMMs to promote diverse hypothesis generation. This paradigm instantiates two LMM agents – the **Positive Proposer** and the **Negative Proposer**, who are explicitly instructed to adopt opposing viewpoints regarding whether the video conveys hate or non-hate intent. Both agents receive the same multimodal input but are guided to extract and reason over distinct clues that support their assigned stance. By compelling the agents to articulate reasoning from both hate and non-hate standpoints, our paradigm effectively surfaces both obvious and subtle hate indicators within the videos. This paradigm significantly reduces ambiguity and clarifies the multifaceted intents inherent in video content, leading to more comprehensive explanations. To address **C2**, we propose a Spatiotemporal Evidence-Grounded Verification mechanism. This mechanism first segments the video into a set of spatiotemporal units, each encapsulating detailed local context information crucial for accurate analysis. These units are then associated with candidate clues generated by the previous two agents through a multimodal retrieval process, resulting in well-aligned *clue-evidence* pairs. Subsequently, a third agent – termed the **Verifier** – is introduced to assess the reliability and authenticity of each clue-evidence pair. By explicitly providing the Verifier with the "retrieved evidences", it discriminates genuine clues supported by video content from hallucinated information, yielding more robust explanations and improving the trustworthiness of hate video detection. Finally, these insightful and faithful explanations are seamlessly integrated into the multimodal video features, enabling more precise hate video detection. Our main contributions are as follows:

- We propose MATCH, the first interpretable hate video detection framework via multiple LMM agents collaboration. We introduce a novel diversely generate-then-verify paradigm, effectively addressing the challenges of ambiguous intent and complex temporal dynamics in providing interpretable detection.
- We design a fresh Dual-Perspective Proposing paradigm, where two LMM agents are instructed to deliver contrasting clues from opposing stances, reducing ambiguity and clarifying the multifaceted intent inherent in video content.
- We develop a new Spatiotemporal Evidence-Grounded Verification mechanism, where a third LMM agent verifies speculative clues through retrieved spatiotemporal evidences, ensuring explanation faithfulness in long and temporally dynamic videos.

Extensive experiments demonstrate that MATCH outperforms state-of-the-art baselines while providing well-reasoned explanations.

II. Related Work

A. Multimodal Hate Detection

Multimodal hate detection is a task aimed at identifying any possible hate content expressed through multimodal media. Existing studies in this field have predominantly focused on **hate meme detection**, owing to the extensive meme benchmarks [20]–[23]. These works utilized pretrained language and vision encoders to extract features, which are then integrated for prediction. For instance, Pro-Cap [22] leveraged vision-language models (i.e., BLIP [24]) to generate contextually enriched captions for enhanced detection. In contrast, the task of **hate video detection** (HVD) remains relatively underexplored, primarily due to the inherent complexity of jointly modelling three modalities (i.e., text, audio, and vision) for effective prediction. Early work in HVD mainly adopted machine learning techniques [25]–[27], such as SVM [28], Random Forests [29], and deep learning architectures [25], [27], [30]. Recently, the introduction of high-quality benchmarks, including HateMM [9] and MHClip [10], has advanced HVD by offering robust datasets and baselines that integrate multimodal features for prediction.

Beyond improving model performance in multimodal hate detection, offering interpretable and transparent predictions is also critical and meaningful. Nevertheless, only a few studies have attempted to develop interpretable solutions for hate meme detection (e.g., ExplainHM [31]), while no prior work has explored this direction in the video domain. Unlike ExplainHM, which focuses on image–text hateful meme detection and relies on debate-style reasoning between LLMs without explicit grounding, our MATCH is designed for video-based hateful content, where explanations are generated through a multi-agent generate-then-verify paradigm with spatiotemporal evidence grounding, explicitly addressing ambiguous intent and long-range temporal dynamics in videos.

B. Multi-Agent Collaboration

Significant advancements in LLMs have spurred substantial study aimed at enhancing their performance across various downstream tasks, such as software engineering and maths problem-solving [32], [33]. These efforts span from improving a single LLM's reasoning capabilities through Chain-of-Thought (CoT) prompting [34], [35] and self-improvement techniques [36] to leveraging multi-agent collaboration to simulate human-like behavior [37], [38]. Multi-agent collaboration paradigm uses multiple LLMs in a collaborative setting to solve complex tasks [39]–[42]. The motivation is that by cross-agent interaction, LLMs can collectively exhibit enhanced performance through aggregating their individual strengths [43], [44]. For instance, MapCoder [40] introduces multi-agent prompting to replicate the full cycle of program synthesis, achieving superior performance in code generation tasks across multiple programming languages. ChatEval [41] employs a multi-agent framework to assess the quality of LLM-generated answers across diverse domains, thereby enhancing the accuracy and reliability of evaluations.

In this work, we propose a novel multi LMM agent collaboration framework, in which multiple agents cooperate to address the limitations (Ambiguous video intent & Complex temporal dynamics) of a vanilla single LMM in achieving interpretable HVD. This collaborative design enables the generation of more comprehensive and persuasive explanations for hate video predictions.

III. Methodology

Problem Definition Let $\mathcal{S} = \{\mathcal{S}_1, \cdots, \mathcal{S}_N\}$ represent the set of videos, where N denotes the total number of videos. Each video $\mathcal{S}_i$ is characterized by its multimodal content, consisting of audio, visual, and textual modalities, represented as $\mathcal{S}_i = \{\mathcal{A}_i, \mathcal{V}_i, \mathcal{T}_i\}$. The objective of HVD is to classify each video as either *hate* or *non-hate*, formally expressed as $\hat{y}_i = f_\Phi(\mathcal{A}_i, \mathcal{V}_i, \mathcal{T}_i)$, where $\hat{y}_i$ is the predicted category of video $\mathcal{S}_i$, $f_\Phi(\cdot)$ denotes the detection model parameterized by Φ.

Our Paradigm for Interpretable HVD. Given the importance of fostering more reliable content moderation on video platforms and making hate content judgments more transparent to viewers, we extend the HVD to an interpretable setting. Specifically, we propose an innovative *diversely generate-then-verify* paradigm driven by the collaboration of multiple LMM agents. Our paradigm begins with two LMM agents acting as *Proposers*, each generating clues from opposing perspectives (positive proposer $\mathcal{M}_{\text{pro-p}}$ *vs.* negative one $\mathcal{M}_{\text{pro-n}}$) to uncover the comprehensive intent of the video from both sides, yielding positive and negative clues $\mathcal{C}$. To enhance the faithfulness of these clues, spatiotemporal units $\mathcal{X}$ extracted from the video are utilized as evidences and paired with the corresponding clues through a multimodal retrieval, forming clue-evidence pairs $\mathcal{R}$. These pairs are then passed to a third LMM agent acting as the *Verifier* $\mathcal{M}_{\text{vfy}}$, which performs cross-verification to assess the consistency and credibility of the claims, generating the clear and persuasive rationales $\mathcal{E}$. The insightful rationales $\mathcal{E}$ are then served as additional knowledge and input to the Predictor($\cdot$), enhancing the final detection of the video.

Methodology Layout. We first outline the processing pipeline for video data in Section III-A. Subsequently, we describe the framework MATCH in details: A *Dual-Perspective Proposing Paradigm* is proposed to encourage the generation of diverse explainable clues in Section III-B. Besides, a *Spatiotemporal Evidence-Grounded Verification Mechanism* is designed to reduce hallucinations in long-context video reasoning in Section III-C. Finally, we propose a *Rationale-Enhanced Predictor* to improve prediction accuracy of HVD in Section III-D. The overall architecture of our proposed framework MATCH is illustrated in Figure 2.

A. Video Data Preprocessing

In this section, we describe the video data preprocessing pipeline employed in our work. We begin by preprocessing the audio, textual, and visual modalities of the video data. For the audio modality, we extract the audio track from each video, denoted as $\mathcal{A}$. For the visual modality, we uniformly sample M frames from the video to construct the visual representation $\mathcal{V} = \{\mathcal{V}_m\}_{m=1}^{M}$. For the textual modality, direct text $\mathcal{T}_v$ is obtained from the video title. In addition, transcripts $\mathcal{T}_t$ are

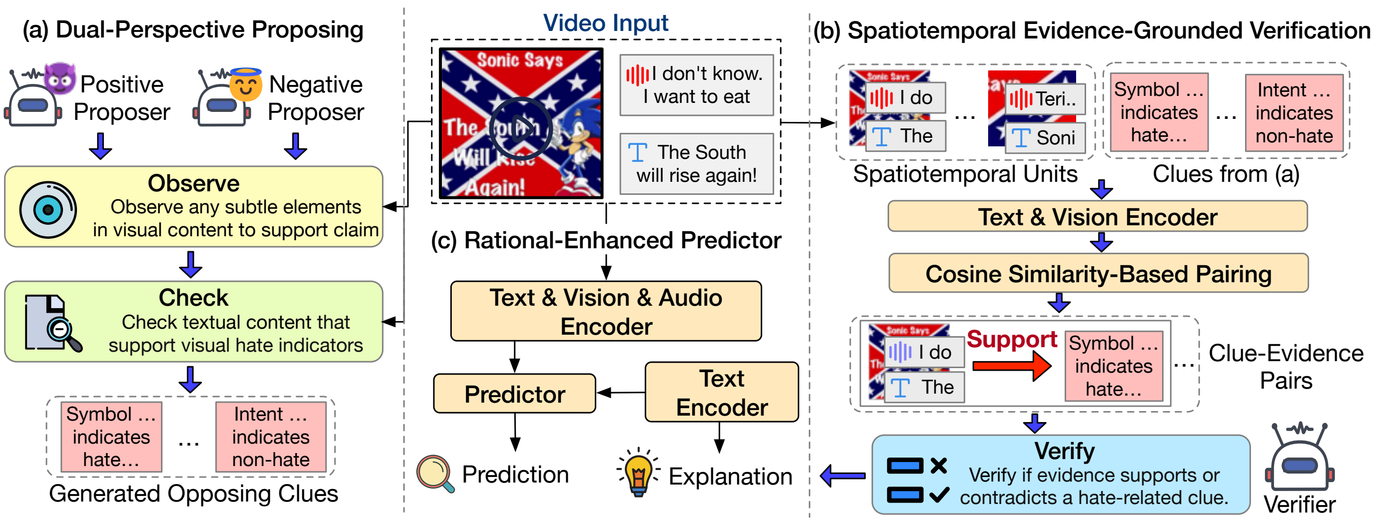


Figure 2. Overall framework of the proposed MATCH. (a) Dual-Perspective Proposing Paradigm generates opposing clues by positioning two proposer agents to inspect video content from contrasting viewpoints. (b) Spatiotemporal Evidence-Grounded Verification Mechanism first aligns fine-grained video segments with clues generated from proposers to form clue-evidence pairs. These pairs are then assessed by a verifier agent, yielding reliable interpretations. (c) Rationale-Enhanced Predictor integrates verified explanations and video features for final prediction.

generated from the audio using Whisper [45], and on-screen text $\mathcal{T}_o$ is extracted from sampled frames using PaddleOCR [46]. Consequently, we denote the textual information as $\mathcal{T} = \mathcal{T}_v \oplus \mathcal{T}_t \oplus \mathcal{T}_o$, where $\oplus$ is the concatenation operation.

B. Dual-Perspective Proposing Paradigm

Video creators often express ambiguous and multifaceted intent in their published videos. This *intent duality* phenomenon causes the fact that hate videos may deliberately mask malicious intent with non-hate surface elements, while non-hate videos may unintentionally include symbols or cues that appear hate. To address the multi-angle intent inherent in videos for clearer detection interpretation, we propose a novel Dual-Perspective Proposing paradigm. It leverages the instruction-following capabilities of LMMs to reason about video intent from opposing perspectives, yielding comprehensive interpretable clues to well reflect the complex video intent.

Specifically, we instantiate two LMM agents: a positive proposer and a negative proposer. The positive proposer assumes the video contains hate content, interpreting all available clues as intentional expressions of hate. In contrast, the negative proposer adopts a non-hate stance, assuming that the creator has no hate intent and that any seemingly offensive symbols are used innocuously. To ensure grounded and faithful reasoning, we introduce a two-step CoT process. During the process, each agent first focuses on visual clues within the video content and then expands to cross-modal analysis by incorporating textual context. For clarity and conciseness, we only present the reasoning process of the positive proposer below. The reasoning process of the negative proposer follows a structurally similar approach.

For the first step, we prompt the proposer to **observe** the video content and actively identify potential hate visual signals, like offensive gestures, symbols, or scenes. The goal is to interpret these elements through a lens that highlights their hate implications. This yields the initial set of visual clues $\mathcal{C}_{\text{obs-p}}$, presented as:

$$\mathcal{C}_{\text{obs-p}} = \mathcal{M}_{\text{pro-p}}(\mathcal{V}, \mathcal{P}_{\text{obs-p}}), \tag{1}$$

where $\mathcal{M}_{\text{pro-p}}$ denotes the LMM-based positive proposer, and $\mathcal{P}_{\text{obs-p}}$ represents the prompt template for visual observation.

For the second step, we prompt the proposer to **check** for textual clues that either supports or amplifies the visual interpretation. By conditioning on both the visual clues $\mathcal{C}_{\text{obs-p}}$ and the video textual content $\mathcal{T}$, the model performs cross-modal reasoning to uncover deeper intent. This produces the final set of clues $\mathcal{C}_{\text{chk-p}}$:

$$\mathcal{C}_{\text{chk-p}} = \mathcal{M}_{\text{pro-p}}([\mathcal{C}_{\text{obs-p}}, \mathcal{T}], \mathcal{P}_{\text{chk-p}}), \tag{2}$$

where $\mathcal{P}_{\text{chk-p}}$ is the prompt for textual clue checking:. Similarly, the negative proposer are instructed to generate a set of clues from non-hate angle, represented as $\mathcal{C}_{\text{obs-n}}$ and $\mathcal{C}_{\text{chk-n}}$.

Then, we denote the set of all generated clues as $\mathcal{C} = \{\mathcal{C}_p\}_{p=1}^{P} = \mathcal{C}_{\text{obs-p}} \oplus \mathcal{C}_{\text{chk-p}} \oplus \mathcal{C}_{\text{obs-n}} \oplus \mathcal{C}_{\text{chk-n}}$, where each $\mathcal{C}_p$ represents an individual clue. These clues will be subsequently verified to ensure trustworthiness.

C. Spatiotemporal Evidence-Grounded Verification Mechanism

The video medium inherently exhibits complex spatiotemporal characteristics, where semantic cues are dispersed over time and highly context-dependent. These complex temporal dynamics in videos undermine the faithfulness and reliability of the clues generated by the two proposer agents, as they often contain temporal hallucinations and reasoning inconsistencies. To overcome these challenges, we introduce a new Spatiotemporal Evidence-Grounded Verification mechanism. The mechanism first segments each video into spatiotemporal units, each containing localized rich content that serves as representative evidence. These evidence segments are retrieved by the clues generated in the previous section to form clue-evidence pairs. Finally, these pairs are served as "retrieved evidences" and passed to a third LMM agent – verifier, yielding evidence-grounded and faithful explanations for HVD.

1) Clue-Evidence Pair Construction: To represent localized, semantically rich information within a video's temporal context, we segment the video into a sequence of spatiotemporal units. To begin with, we uniformly sample abundant frames $\{\mathcal{F}_l\}_{l=1}^{L}$ from the video. For each sampled frame $\mathcal{F}_l$, we then combine its on-screen text and the corresponding audio transcript fragment (i.e., the transcript from a few seconds around the frame), resulting in a localized textual context $\mathcal{W}_l$. Each spatiotemporal unit is then constructed as the following process:

$$\mathcal{X}_l = (\mathcal{F}_l, \mathcal{W}_l), \quad l = 1, \cdots, L. \tag{3}$$

These structured units serve as candidate evidences for interpreting the video's intent. To leverage these evidential units for verifying the clues generated by the proposers, we align each clue with its most semantically relevant spatiotemporal unit through a joint multimodal retrieval, forming a set of clue-evidence pairs. Specifically, we use CLIP [8] text encoder Φ_t to embed the clue $\mathcal{C}_p$ and the localized textual content $\mathcal{W}_l$, and the vision encoder Φ_v to embed the visual frame $\mathcal{F}_l$. The similarity score between each clue $\mathcal{C}_p \in \mathcal{C}$ and spatiotemporal unit $\mathcal{X}_l$ is computed as:

$$\alpha_{p,l} = \lambda \cdot \frac{\Phi_t(\mathcal{C}_p) \cdot \Phi_v(\mathcal{F}_l)}{\|\Phi_t(\mathcal{C}_p)\| \|\Phi_v(\mathcal{F}_l)\|} + (1-\lambda) \cdot \frac{\Phi_t(\mathcal{C}_p) \cdot \Phi_t(\mathcal{W}_l)}{\|\Phi_t(\mathcal{C}_p)\| \|\Phi_t(\mathcal{W}_l)\|}, \tag{4}$$

where $\lambda \in [0, 1]$ is a hyperparameter that balances the contribution of visual and textual similarity.

We select the unit with the highest $\alpha_{p,l}$ for each clue $\mathcal{C}_p$, resulting in a set of P clue-evidence pairs:

$$\mathcal{R} = \{(\mathcal{C}_p, \mathcal{X}_l) | \mathcal{C}_p \in \mathcal{C}, \mathcal{X}_l = \arg\max_{\mathcal{X}_l \in \mathcal{X}} \alpha_{p,l}\}. \tag{5}$$

2) Evidence Grounding: To verify and refine the clues provided by the proposers with associated evidence, we introduce a third LMM agent, dubbed as verifier. The verifier takes as input the clue-evidence pairs and assesses whether the selected evidence supports, contradicts, or refines the initial clue. This cross-verification process enhances explanation faithfulness by grounding each reasoning step in verifiable multimodal context. Formally, the final explanation is generated:

$$\mathcal{E} = \mathcal{M}_{\text{vfy}}(\mathcal{R}, \mathcal{P}_{\text{vfy}}), \tag{6}$$

where $\mathcal{P}_{\text{vfy}}$ is the verification prompt that guides the model to reason over the clue-evidence pairs and produce grounded explanations.

D. Rationale-Enhanced Predictor

In this section, we aim to leverage the insightful explanations from the collaborative agents to enhance the detection for more subtle and challenging hate patterns in videos.

Specifically, we employ a pretrained BERT [47] model to encode the rationale features derived from the evidence: $\boldsymbol{x}_e = \Psi_t(\mathcal{E})$, where Ψ_t is the BERT model. To fully exploit the multimodal nature of the video data, we extract features from audio, visual, and textual modalities. Specifically, we compute Mel-Frequency Cepstral Coefficients (MFCCs) [48] for audio features, denoted as $\boldsymbol{x}_a = \Psi_a(\mathcal{A})$. Textual features are extracted by BERT, represented as $\boldsymbol{x}_t = \Psi_t(\mathcal{T})$, while visual features are derived from the ViT [49], denoted as $\boldsymbol{x}_v = \Psi_v(\mathcal{V})$. Notably, for textual and visual modalities, we adopt encodings of the [CLS] tokens to represent the their features, and we average the features of each frames in $\mathcal{V}$ to form the visual representation $\boldsymbol{x}_v$. These modality features and rationale features are then processed through modality-specific Feed-Forward Networks (FFNs) to unify their dimensions to d, formalized as $\tilde{\boldsymbol{x}}_m \in \mathbb{R}^d$, where $m \in \{e, a, t, v\}$. The final prediction is achieved by combining the features of both multimodal content and the rationales:

$$\hat{y} = \text{Predictor}([\tilde{\boldsymbol{x}}_e \oplus \tilde{\boldsymbol{x}}_a \oplus \tilde{\boldsymbol{x}}_t \oplus \tilde{\boldsymbol{x}}_v]), \tag{7}$$

where the Predictor is implemented with a two-layer Multilayer Perceptron (MLP). For model training, we freeze the parameters of all pretrained modality encoders, and adopt Binary Cross-Entropy loss to optimize the model. The detailed prompts are provided in the source code files.

Table I
DATASET STATISTICS FOR THREE DATASETS.

Dataset	# Total	# Hate	# Non-Hate	Platform	Language
HateMM	1,083	431	652	Bitchute	English
MHClipEN	1,000	338	662	YouTube	English
MHClipZH	1,000	322	678	Bilibili	Chinese

IV. Experiments

A. Experimental Setup

- **Datasets.** To assess the effectiveness of MATCH, we conduct comprehensive experiments using three real-world video datasets, including **HateMM** [9], MultiHateClip YouTube (**MHClipEN**) [10], and MultiHateClip Bilibili (**MHClipZH**). Detailed statistics of three video datasets are presented in Table I, with splits consistent with the dataset creators. Notably, in the MHClipEN and MHClipZH datasets, we merge the offensive and hateful categories into a single hateful class to conduct binary classification.
- **Baselines.** We compare the proposed MATCH with 15 competitive baselines, which can be categorized into three classes: (1) Unimodal Detection Methods which utilize single modality information for hate detection: BERT [47], ViT [49], HuBERT [50], and TSformer [51]. (2) Multimodal Detection Methods which fuse multimodal information conduct detection: HTMM [9], MHCL [10], and CMFusion [52]. Notably, due to the scarcity of methods in HVD, we also include competitive baselines from hate meme detection for comparison, including Pro-Cap [22], RGCL [53], Mod-HATE [54], and ExplainHM [31]. (3) LMM-based Methods which leverage the advanced and powerful LMMs for prediction, including Qwen2.5-VL-7B-Instruct [13], Phi-4-multimodal-instruct-5.6B [55], InternVL2-8B [15], and GPT-4o-mini [56]. Remarkably, we provide these LMMs with in-context demonstrations to fully activate their performance.
- **Evaluation Metrics.** For prediction performance, following prior works [9], [10], we utilize four classification

Table II
EXPERIMENTAL RESULTS OF MATCH AND THE COMPETITIVE BASELINES ON THE HATEMM, MHCLIPEN, AND MHCLIPZH DATASETS. THE TOP RESULTS ARE IN **BOLD**, WHILE THE SECOND RESULTS ARE <u>UNDERLINED</u>. HIGHER ACC, M-F1, M-P, AND M-R INDICATE BETTER PERFORMANCE.

Baseline	**HateMM**				**MHClipEN**				**MHClipZH**			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
BERT	0.6543	0.6030	0.6410	0.6069	0.6499	0.4901	0.5520	0.5221	0.7250	0.6771	0.6839	0.6726
ViT	0.7004	0.6415	0.7238	0.6454	0.7200	0.5985	0.7283	0.6043	0.6899	0.6408	0.6435	0.6387
HuBERT	0.6501	0.6343	0.7189	0.6396	0.6550	0.5227	0.5498	0.5329	0.5898	0.5838	0.5892	0.5859
TSformer	0.7327	0.7063	0.7299	0.7008	0.6499	0.5706	0.5869	0.5701	0.7200	0.6667	0.6771	0.6609
HTMM	0.7787	0.7679	0.7705	0.7660	0.7150	0.6320	0.6831	0.6265	0.7100	0.6184	0.6654	0.6137
MHCL	0.7788	0.7584	0.7822	0.7508	0.7100	0.6548	0.6700	0.6486	<u>0.7650</u>	<u>0.7311</u>	0.7321	<u>0.7302</u>
CMFusion	<u>0.7880</u>	<u>0.7723</u>	<u>0.7862</u>	0.7660	0.6750	0.6339	0.6344	0.6334	0.7050	0.6336	0.6596	0.6498
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
RGCL	0.7558	0.7355	0.7296	0.7524	0.7133	0.6322	0.6264	0.6728	0.7250	0.7103	<u>0.7365</u>	0.7096
Mod-HATE	0.6866	0.6536	0.6510	0.6760	0.6774	0.6388	0.6376	0.6702	0.6445	0.5107	0.5233	0.5046
ExplainHM	0.7315	0.6888	0.6822	0.7013	0.7250	0.6700	0.6890	0.6737	0.7500	0.7295	0.7318	0.7249
Qwen2.5-VL	0.7072	0.7070	0.7525	0.7447	0.7130	0.6804	0.6798	0.6811	0.7160	0.6161	0.6777	0.6123
Phi-4	0.6668	0.6680	0.7133	0.7054	0.6910	0.5415	0.6767	0.5653	0.6860	0.5165	0.6285	0.5451
InternVL2	0.6620	0.6317	0.6440	0.6297	0.7000	<u>0.6832</u>	0.6813	<u>0.6988</u>	0.6710	0.6521	0.6529	0.6723
GPT-4o-m	0.7452	0.7451	0.7791	<u>0.7790</u>	<u>0.7320</u>	0.6388	<u>0.7301</u>	0.6340	0.7160	0.5765	0.7072	0.5866
MATCH	**0.8203**	**0.8158**	**0.8130**	**0.8215**	**0.7575**	**0.7030**	**0.7379**	**0.6996**	**0.8050**	**0.7689**	**0.7827**	**0.7598**

metrics: Accuracy (**ACC**), Macro-F1 score (**M-F1**), Macro Precision (**M-P**), and Macro Recall (**M-R**). For explanation quality, we quantitatively assess the rationales using G-Eval [57], a reference-free LLM-based evaluation approach. Specifically, we employ two powerful LLMs — the open-source `Qwen2.5-72B-Instruct` [58] and the commercial `GPT-5` [59] to evaluate each rationale along five widely adopted criteria [60], [61]: Informativeness (**I**) — the rationale offers new insights, such as contextual or background explanations; Soundness (**S**) — the reasoning is logical, valid, and coherent; Persuasiveness (**P**) — the explanation is convincing and compelling; Readability (**R**) — the rationale follows standard grammar and structure; Fluency (**F**) — the rationale is smooth and naturally flowing. Each criterion is rated on a 5-point Likert scale [62], where 1 indicates the lowest and 5 the highest quality. - **Implementation Details.** We employ Qwen2.5-VL (i.e., `Qwen2.5-VL-7B-Instruct`) as the backbone LMM for all three agents. During video data processing, we evenly sample $M = 16$ frames for each video and set the maximum text sequence length to 512. For spatiotemporal units segmentation, we segment $L = 64$ units for each video. All experiments are conducted on two NVIDIA RTX 4090 GPUs. In the multimodal retrieval for clue-evidence pairs matching, we utilize CLIP [8] as text encoder Φ_t and vision encoder Φ_v. Specifically, we use the pretrained version `jina-clip-v2`, a general-purpose CLIP embedding model for text and images. In the multimodal feature encoding process, we employ MFCCs implemented by librosa as audio feature extractor Ψ_a, and employ ViT (i.e., `vit-base-patch16-224`) as visual model Ψ_v, and BERT (i.e., `bert-base-uncased` for English and `bert-base-chinese` for Chinese) as Ψ_t.We freeze all the parameters of the pretrained encoders. We utilize the AdamW optimizer with learning rate of 1×10^{-4}. Both the baseline models and the proposed MATCH are trained and tested five times, with the average results reported as the final result. We set the same random seed for LMM invocation to ensure fair comparison. To ensure fair comparisons, we strictly adhere to the parameter configurations specified in the original papers of baseline methods.

B. Hate Detection Performance

To evaluate the capability of MATCH in detecting hate videos, we compare it with 15 baseline methods, with results presented in Table II. From the results, we have the following analysis.

Our proposed MATCH consistently outperforms all competitive baselines, demonstrating its effectiveness in detecting implicit hate within videos. This performance gain stems from the informative rationales generated by our multi-agent collaboration paradigm, which provides valuable supplementary knowledge to uncover deeper hate patterns and enhance prediction accuracy. Moreover, our MATCH can provide clear and well-reasoned explanations behind its predictions, largely improving the transparency in HVD.

Unimodal detection methods perform the worst, underscoring the importance of leveraging all available modalities for HVD. Multimodal approaches demonstrate strong performance, benefiting from carefully designed structures. For example, RGCL employs a retrieval-based framework that learns hatefulness-aware representations through an auxiliary contrastive objective and dynamically retrieved examples. However, these multimodal methods still struggle to capture the implicit and complex hate signals embedded in videos, which often require a deeper semantic understanding. In contrast, LMM-based methods have the potential to play well in this regard, owing to their extensive pretrained knowledge from large-scale multimodal data. Nevertheless, their performance remains limited due to the lack of task-specific fine-tuning, which is often impractical because of the substantial computational cost. Our proposed framework MATCH *combines the strengths of both multimodal and LMM methods while avoiding their limitations*. It integrates

Table III
ABLATION STUDY ON MULTI-AGENT COLLABORATION MECHANISM.

Variant	HateMM		MHClipEN		MHClipZH	
	ACC	M-F1	ACC	M-F1	ACC	M-F1
w/o P-Proposer	0.7972	0.7904	0.7400	0.6987	0.7749	0.7281
w/o N-Proposer	0.7926	0.7828	0.7250	0.6570	0.6750	0.5576
w/o Proposer	0.8064	0.7898	0.7225	0.6571	0.7950	0.7614
w/ Vanilla Prompt	0.8065	0.8007	0.7500	0.6988	0.7950	0.7462
LMM w/ Continual	0.7742	0.7663	0.7100	0.6853	0.7550	0.7394
LMM w/ Single	0.7512	0.7453	0.6950	0.6928	0.7350	0.7195
MATCH	**0.8203**	**0.8158**	**0.7575**	**0.7030**	**0.8050**	**0.7689**

Table IV
ABLATION STUDY ON EVIDENCE RETRIEVAL STRATEGY.

Variant	HateMM		MHClipEN		MHClipZH	
	ACC	M-F1	ACC	M-F1	ACC	M-F1
w/o Evidence	0.7788	0.7469	0.7250	0.6532	0.7749	0.7129
w/ Random	0.7788	0.7542	0.7250	0.6670	0.7750	0.7461
w/ Top-2	0.8118	0.7987	0.7400	0.6983	0.7950	0.7483
w/ Text-Only	0.7926	0.7706	0.7300	0.7013	0.7800	0.7373
w/ Vision-Only	0.7834	0.7628	0.7200	0.6509	0.7750	0.7292
w/o Verifier	0.7696	0.7572	0.7000	0.6000	0.7800	0.7366
w/ Twice Verify	0.8111	0.7969	0.7500	0.6952	0.7950	0.7624
MATCH	**0.8203**	**0.8158**	**0.7575**	**0.7030**	**0.8050**	**0.7689**

valuable knowledge (i.e., detection rationales) generated by LMMs and multimodal video features into a lightweight model, which then undergoes task-specific training, achieving outstanding performance with lower resource overhead.

C. Ablation Study on Main Components

To assess the effect of main components within our framework MATCH in identifying hate videos, we conduct comprehensive ablation studies with results reported in Tables III to V.

Effect of Multi-Agent Collaboration Mechanism. As shown in Table III, we evaluate the effectiveness of the multi-agent collaboration mechanism by designing several variants: (i) *w/o P-Proposer*: the positive proposer agent is removed from the collaboration; (ii) *w/o N-Proposer*: the negative proposer is excluded; (iii) *w/o Proposer*: both positive and negative proposers are removed, and a vanilla LMM is employed to direct generate the explanation from one angle based on its prediction; (iv) *w/ Vanilla Prompt*: using concise and degraded prompts without the carefully designed instructions; (v) *LMM w/ Continual*: a single LMM generates all clues through continual conversation turns instead of using separate agents; (vi) *LMM w/ Single*: a single LMM directly generates the final rationale in one pass without the propose-verify process. From the results, we observe that each variant meets significant performance drops. These drops stem from the limited clues from only one-side, which incurs compromised quality of the rationales, providing limited and even noise information for the prediction. The LMM w/ Continual and LMM w/ Single variants demonstrate that our multi-agent design is superior to single-model alternatives, as multiple specialized agents can explore diverse perspectives more effectively. Notably, the w/ Vanilla Prompt variant exhibits only slight degradation, demonstrating the robustness of MATCH to suboptimal prompt designs, as the multi-agent collaboration mechanism itself provides strong guidance even with simpler instructions.

Table V
ABLATION STUDY ON RATIONALE FUSION STRATEGY.

Variant	HateMM		MHClipEN		MHClipZH	
	ACC	M-F1	ACC	M-F1	ACC	M-F1
w/o Rationale	0.7741	0.7654	0.7099	0.6547	0.7649	0.7310
w/ Attention	0.8111	0.8030	0.7450	0.6893	0.8000	0.7642
w/ Gated	0.8065	0.7907	0.7400	0.6816	0.7900	0.7615
MATCH	**0.8203**	**0.8158**	**0.7575**	**0.7030**	**0.8050**	**0.7689**

Effect of Spatiotemporal Evidence-Grounded Verification Mechanism. As shown in Table IV, we assess the effectiveness of the Spatiotemporal Evidence-Grounded Verification Mechanism by designing several variants: (i) *w/o Evidence*: the spatiotemporal evidences from videos are removed and the verifier agent only leverages its pretrained knowledge for clue verification; (ii) *w/ Random*: evidences are randomly sampled from the video instead of using similarity-based retrieval; (iii) *w/ Top-2*: the top-2 most similar evidences are retrieved instead of only the top-1; (iv) *w/ Text-Only*: only text similarity is used for evidence retrieval without visual similarity; (v) *w/ Vision-Only*: only visual similarity is used for evidence retrieval without text similarity; (vi) *w/o Verifier*: the verifier agent is eliminated and the opposing clues from the proposer agents are combined to augment the prediction; (vii) *w/ Twice Verify*: the verifier agent performs two rounds of verification on each clue. We observe a drastic performance declination in *w/o Evidence* variant, since the verifier agent can also generate hallucinated content like the proposers without outside evidences, which undermines the quality of the final rationales and the detection performance. The w/ Random variant shows that our similarity-based retrieval is crucial for finding relevant evidences. The w/ Top-2 variant exhibits minor performance degradation, demonstrating the robustness of MATCH to suboptimal evidence matching, as incorporating additional less-relevant evidences introduces slight noise but does not drastically compromise the verification quality. The w/ Text-Only and w/ Vision-Only variants demonstrate the necessity of combining both modalities for comprehensive evidence matching. Moreover, *w/o Verifier* variant also achieves suboptimal performance, as clues only without any verification provide limited guidance in the prediction. The w/ Twice Verify variant achieves comparable performance to our method, indicating that additional verification rounds maintain effectiveness but do not provide further gains.

Effect of Rationale-Enhanced Predictor. As shown in Table V, we validate the effectiveness of the Rationale-Enhanced Predictor by designing several variants: (i) *w/o Rationale*: the rationales are eliminated from the prediction process; (ii) *w/*

Table VI
SENSITIVITY ANALYSIS ON THE MODALITY BALANCE WEIGHT λ.

λ	HateMM		MHClipEN		MHClipZH	
	ACC	M-F1	ACC	M-F1	ACC	M-F1
0.1	0.7880	0.7722	0.7450	0.6940	0.7750	0.7386
0.3	0.8065	0.7808	0.7250	0.6790	0.8000	0.7621
0.5	**0.8203**	**0.8146**	**0.7600**	**0.7013**	**0.8150**	**0.7640**
0.7	0.7834	0.7572	0.7400	0.6909	0.7700	0.7426
0.9	0.8018	0.7781	0.7200	0.6668	0.7950	0.7543

Table VII
SENSITIVITY ANALYSIS ON THE NUMBER OF SPATIOTEMPORAL UNITS L.

L	HateMM		MHClipEN		MHClipZH	
	ACC	M-F1	ACC	M-F1	ACC	M-F1
16	0.7972	0.7773	0.7200	0.6790	0.7700	0.7223
32	0.7834	0.7682	0.7350	0.6751	0.7900	0.7572
64	**0.8203**	**0.8126**	**0.7600**	**0.6989**	**0.8150**	**0.7643**
128	0.8157	0.7937	0.7500	0.6823	0.8000	0.7610
256	0.8111	0.8002	0.7550	0.6922	0.8100	0.7515

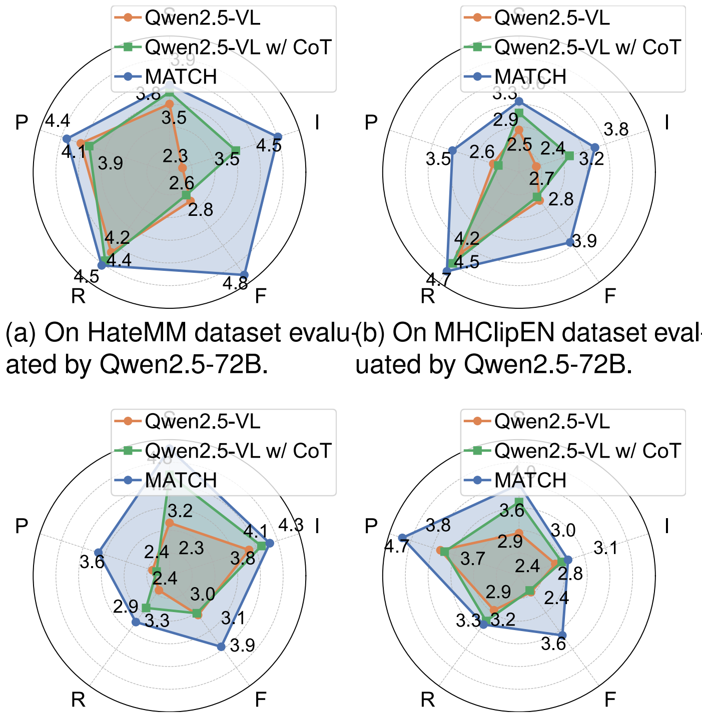
Attention: using attention mech to fuse rationale features with multimodal features; (iii) *w/ Gated*: using gated mechanism to control the contribution of rationale features. The w/o Rationale variant meets a substantial performance drop, highlighting the efficacy of the rationales in improving the detection. The w/ Attention and w/ Gated variants explore alternative fusion strategies but exhibit slight performance degradation, as these more complex mechanisms tend to overfit on the training data, while our concatenation-based fusion achieves the best performance by preserving the complete information from both sources without introducing additional learnable complexity.

D. Hyperparameter Sensitivity Analysis

To investigate the sensitivity of our framework MATCH to key hyperparameters, we conduct experiments on the modality balance weight λ and the number of spatiotemporal units L.

Effect of Modality Balance Weight λ. The hyperparameter λ controls the balance between visual and textual similarity in clue-evidence pair matching (see Equation (4)). As shown in Table VI, we vary λ from 0.1 to 0.9 and observe that $\lambda = 0.5$ achieves the best performance across all three datasets. When λ is too small (e.g., 0.1), the model over-relies on textual similarity, which may miss important visual cues that are crucial for hate detection. Conversely, when λ is too large (e.g., 0.9), the model overemphasizes visual similarity, potentially neglecting contextual textual information. The optimal $\lambda = 0.5$ indicates that balanced contribution from both modalities yields the most effective evidence retrieval for clue verification.

Effect of Spatiotemporal Units L. The hyperparameter L determines the granularity of spatiotemporal segmentation for evidence extraction. As shown in Table VII, we evaluate $L \in \{16, 32, 64, 128, 256\}$ and find that $L = 64$ achieves optimal performance. When L is too small (e.g., 16), the coarse segmentation fails to capture fine-grained temporal dynamics, resulting in suboptimal evidence matching. When L is too large (e.g., 256), the excessive segmentation introduces redundant units that may dilute the semantic coherence of evidences and increase computational overhead without corresponding performance gains. The setting $L = 64$ provides an effective balance between temporal granularity and semantic coherence.



(a) On HateMM dataset evaluated by Qwen2.5-72B. (b) On MHClipEN dataset evaluated by Qwen2.5-72B.

(c) On HateMM dataset evaluated by GPT-5. (d) On MHClipEN dataset evaluated by GPT-5.

Figure 3. Comparison of rationale quality between MATCH and its vanilla LMM backbone Qwen2.5-VL-7B (w/ and w/o CoT).

E. Analysis of Model Interpretability

1) Quantitative Analysis of Interpretability: In this section, we assess the quality of the rationales generated by our MATCH from five perspectives defined in our evaluation metrics: Informativeness (I), Soundness (S), Persuasiveness (P), Readability (R), Fluency (F). To validate the effectiveness of MATCH, we compare it with its vanilla LMM backbone – Qwen2.5-VL-72B, as well as Qwen2.5-VL w/ CoT, which augments the vanilla LMM with Chain-of-Thought (CoT) prompting [34] to elicit step-by-step reasoning during rationale generation. As shown in Figure 3, MATCH demonstrates significant improvements in rationale quality over both baselines across all criteria, as consistently confirmed by both Qwen2.5-72B-Instruct and GPT-5 evaluators. While Qwen2.5-VL w/ CoT achieves higher scores than the vanilla Qwen2.5-VL by encouraging explicit reasoning steps, it still falls short of our MATCH across most metrics. This indicates that CoT prompting alone cannot fully address the limitations of LMMs in interpretable hate video detection, whereas our multi-agent collaboration paradigm with spatiotemporal evidence grounding provides more comprehensive and faithful explanations. These gains stem from our novel diversely generate-then-verify paradigm, which effectively addresses the inherent limitations in vanillar

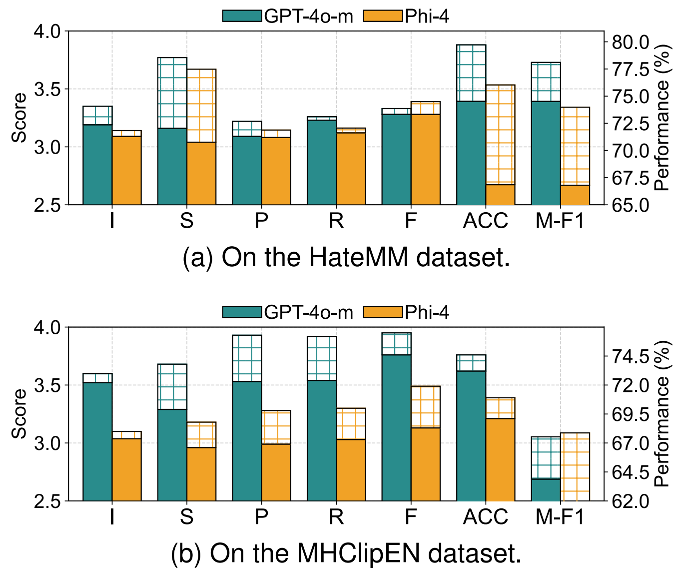


Figure 4. Compatibility with different LMM backbones.

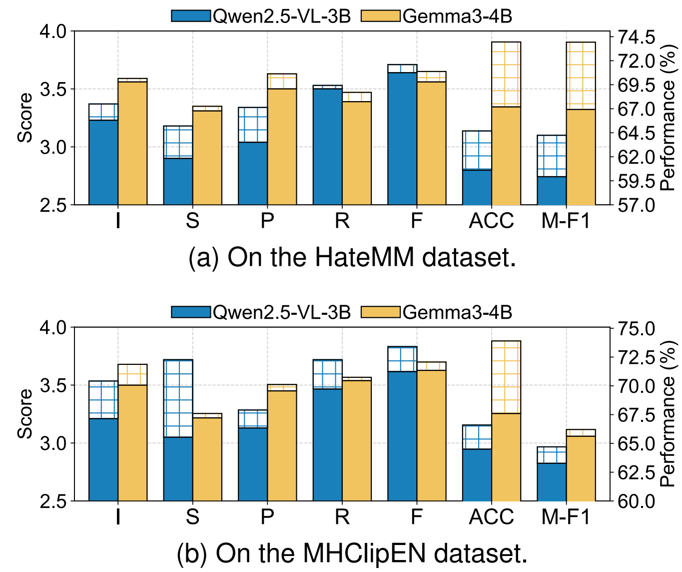


Figure 5. Scalability of interpretability to smaller models.

LMMs in achieving interpretable HVD. In this paradigm, two LMM agents proposes comprehensive clues and a third agent verifies these clues via spatiotemporal evidences, yielding clear and faithful explanations. In addition, we introduce a new metric, termed **Fidelity**, which measures the extent to which the generated explanations are *faithfully grounded in the input video evidence* and do not introduce unsupported or hallucinated content. A higher Fidelity score (also in [1-5]) indicates stronger consistency between the explanation and the underlying visual evidence. And we empirically observe that MATCH achieves 4.3 in Fidelity, outperforming the vanilla Qwen2.5-VL (2.5) and Qwen2.5VL w/ CoT (3.5).

2) Qualitative Analysis on Interpretability: To qualitatively evaluate the interpretability of MATCH, we randomly select a hate video from the HateMM dataset for a case study. We present both one hate and one non-hate clues generated from the proposer agents, along with their corresponding retrieved evidences and the final rationales from the verifier agent. As shown in Figure 8, the positive and negative proposer agents in MATCH first uncover the complex intent of the video by generating the hate clue — Confederate flag, and the non-hate clue — friendly and casual conversation. The hate clue is then paired with retrieved spatiotemporal evidence, which is verified as valid by the verifier agent. In contrast, while the non-hate clue seems reasonable from the retrieved transcript evidence alone, the retrieved visual evidence uncovers hateful symbols (e.g., the Confederate flag and skull-and-crossbones) that invalidate the clue. This inconsistency is effectively detected by the verifier agent, demonstrating how visual grounding corrects textual hallucinations and strengthens the overall explanation. This example showcases how the collaboration among the proposer and verifier agents enables MATCH to produce accurate and comprehensive explanations.

F. Analysis of Model Broad Applicability

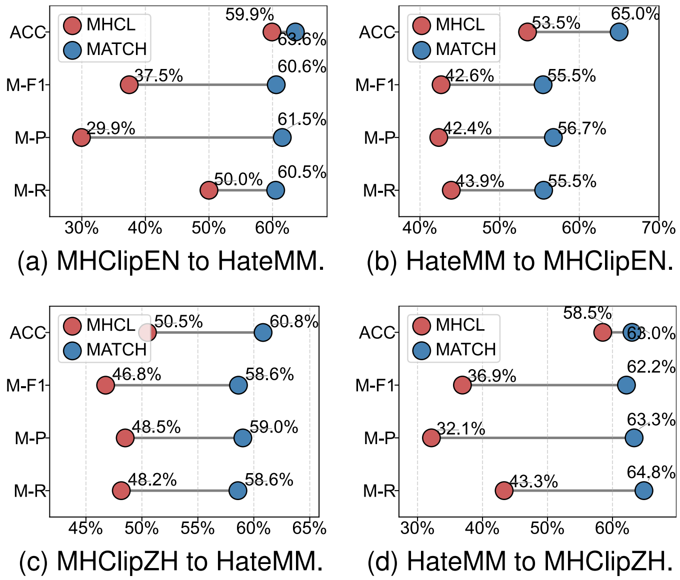
To further investigate the broad applicability and versatility of our framework MATCH, we examine its compatibility with various LMM backbones and its ability to scale down to lightweight models.

1) Compatibility with Different LMM Backbones: To validate the versatility of MATCH, we apply it to several popular and widely-used LMMs and evaluate the resulting interpretability and detection performance. Specifically, we adopt both close-source GPT-4o-mini [56] and open-source Phi-4 [55] as backbones, and compare their performance and rationale quality before and after integrating our MATCH framework on the HateMM and MHClipEN datasets. As shown in Figure 4, each LMM backbone demonstrates substantial improvements across all the rationale quality metrics and detection accuracy when equipping with our MATCH, highlighting the strong compatibility of MATCH across different LMM backbones.

2) Scalability of Interpretability to Smaller Models: To explore the potential of achieving a lightweight while explainable framework in HVD, we conduct an experiment to transfer the interpretability of MATCH to smaller models. Specifically, we leverage MATCH to yield explanations for the predictions on the HateMM and MHClipEN datasets. These explanations are then utilized to teach the smaller multimodal models, including Qwen2.5-VL-3B [13] and Gemma3-4B [63], to realize the interpretable predictions. As illustrated in Figure 5, we observe that each smaller model achieves notable improvements in both interpretability and detection performance, demonstrating the high quality of the rationales generated by our MATCH. Furthermore, the strong interpretability in smaller LMMs highlights the practical value of our proposed MATCH in facilitating both explainable and lightweight detection, which is crucial for limited computational resources scenarios.

G. t-SNE Visualization

We utilize t-SNE [64] to visualize the embedding space of hateful and non-hateful categories, as illustrated in Figure 7. The embeddings are extracted from the last layer of each model's classifier. While MHCL demonstrates partially entangled features and faces difficulty in distinguishing between



(a) MHClipEN to HateMM. (b) HateMM to MHClipEN.

(c) MHClipZH to HateMM. (d) HateMM to MHClipZH.

Figure 6. Cross-Platform detection comparison between MATCH and baseline MHCL.

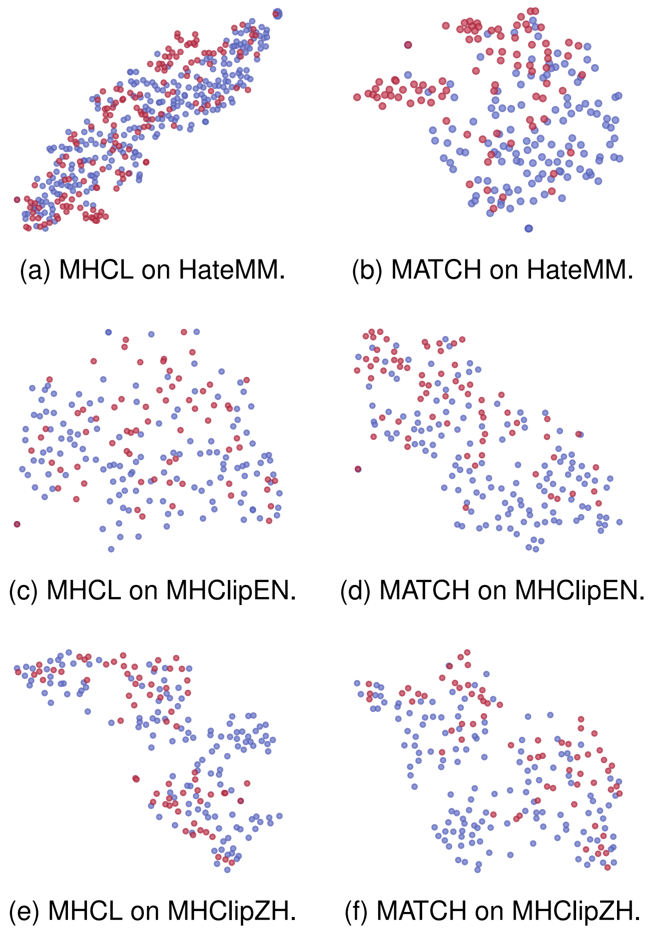
the categories, MATCH produces clearer and more distinct boundaries. This highlights its ability to detect subtle differences in hateful content. Such performance is attributed to the integration of insightful explanations generated through multi-agent collaborations into its prediction process, resulting in more precise predictions.

H. Cross-Platform Generalization

In this section, we evaluate the cross-platform detection generalizability of our framework MATCH compared to the competitive baseline model MHCL. Specifically, we conduct experiments on two pairs of distinct datasets collected from different platforms: (1) HateMM and MHClipEN, and (2) HateMM and MHClipZH. Notably, the second pair presents a more challenging cross-lingual scenario due to the language barrier and cultural differences. To enable cross-lingual text encoding, we employ `bert-base-multilingual-uncased` as the text encoder Ψ_t in these cross-lingual experiments. To assess generalization, both MATCH and MHCL are trained on one dataset and evaluated on the other. As shown in Figure 6, MATCH consistently outperforms MHCL in cross-platform settings, demonstrating stronger generalization across varied content sources and languages. This superior performance lies in the fact that MATCH explicitly leverages the extensive multimodal knowledge and reasoning capabilities of LMMs (i.e., the high-quality rationales) to enhance detection. These external knowledge sources exhibit strong generalizability, enabling MATCH to achieve robust performance across different platforms. In contrast, MHCL relies solely on raw multimodal features without explicit reasoning, resulting in limited transferability across platforms.

I. Human Evaluation on Explainability

To provide a more rigorous and definitive assessment of explanation quality beyond automatic evaluation by G-eval,



(a) MHCL on HateMM. (b) MATCH on HateMM.

(c) MHCL on MHClipEN. (d) MATCH on MHClipEN.

(e) MHCL on MHClipZH. (f) MATCH on MHClipZH.

Figure 7. t-SNE visualization of MATCH alongside most competitive baseline model MHCL. Red points denote hate samples while blue points indicate non-hate samples. (a), (c), (e) show the baseline's performance, while (b), (d), (f) show our model's superior class separation on three datasets respectively.

Table VIII
COMPARISON OF RATIONAL QUALITY UNDER HUMAN EVALUATION.

	HateMM					MHClipEN				
Metrics	I	S	P	R	F	I	S	P	R	F
Qwen2.5-VL	2.6	2.8	2.4	3.7	3.3	2.8	2.2	2.0	2.5	3.8
MATCH	3.8	3.5	3.8	4.2	3.5	3.9	3.5	2.7	3.9	4.0

we conducted a comprehensive human-based evaluation. Our evaluation is to directly compare the explanations generated by our proposed framework, MATCH, against those from the LMM backbone Qwen2.5-VL.

We collect a diverse evaluation set by randomly sampling 200 hate videos from the HateMM dataset and 200 from MHClipEN. We engaged ten expert annotators, each with substantial experience in hateful content analysis, to independently score the generated explanations. The evaluation was conducted using the same five fine-grained metrics as the G-Eval assessment detailed in the main paper. The final scores, averaged across all annotators, are presented in Table VIII. To validate the consistency of the human judgments, we calculated the inter-annotator agreement. Following established guidelines [65], [66], we report an average Spearman's correlation of 0.71

Figure 8. Case study of correctly explained hate video.

and an Intraclass Correlation Coefficient (ICC) of 0.65. These values indicate substantial and moderate-to-good agreement among the annotators, respectively, confirming the reliability of our evaluation. The results yield two primary conclusions. First, the human-assessed scores align closely with the findings from our LLM-based G-Eval, lending further credibility to the automated evaluation framework. Second, and more importantly, the human evaluation confirms that MATCH significantly outperforms the LMM backbone across all five metrics, demonstrating its superior capability in generating high-quality and human-centric explanations.

J. Case Study on Interpretability

In this section, we randomly selected two more videos from the HateMM dataset to evaluate the interpretability of our proposed MATCH. The corresponding results are shown in Figure 8, which highlights MATCH's ability to deliver clear and reliable explanations for hate and non-hate video detection.

K. Efficiency Analysis

To evaluate the computational efficiency of our framework MATCH, we compare it with two LMM-based alternatives: (1) *LMM w/ Single*, where a single LMM directly generates the final rationale in one pass, and (2) *LMM w/ Continual*, where a single LMM generates all clues through continual conversation turns. We measure the per-video inference time, peak GPU memory usage during inference, and detection performance (ACC and M-F1) on the HateMM dataset. As shown in Table IX, our MATCH achieves significantly better detection performance while maintaining competitive computational costs.

Table IX
EFFICIENCY COMPARISON ON THE HATEMM DATASET.

Method	Time (s/video)	Memory (GB)	Performance ACC	Performance M-F1
LMM w/ Single	8.2	19.5	0.7512	0.7453
LMM w/ Continual	22.5	21.2	0.7742	0.7663
MATCH	24.8	22.4	**0.8203**	**0.8158**

Although our multi-agent collaboration paradigm requires multiple LMM invocations for clue generation and verification, the overhead is justified by the substantial performance gains (6.9% and 4.6% absolute improvement in ACC over LMM w/ Single and LMM w/ Continual, respectively). The total inference time for processing the entire HateMM test set (217 videos) is 1.5 hours for MATCH, compared to 0.5 hours for LMM w/ Single and 1.4 hours for LMM w/ Continual. These results demonstrate that our framework achieves a favorable trade-off between computational efficiency and detection accuracy, making it practical for real-world deployment scenarios.

L. Failure Case on Verifier Agent

In this section, we analyze the failure cases of our framework and observe that the main source of errors comes from the verifier agent, which may occasionally produce incorrect judgments when facing ambiguous content. To illustrate this, we randomly sample a failure case from the HateMM dataset, as shown in Figure 9. The video features a warrior/knight holding a sword and shield with a red cross, overlaid with the bold text "DEUS VULT," while the audio contains lyrics repeatedly using the term "coon" in phrases such as "It's the Coon Shootin' Boogie." The positive proposer correctly identifies the video as hate-related, noting that "coon" is widely recognized as a racial slur and that the crusader imagery with "DEUS VULT" amplifies a militant, hostile tone. However, the negative proposer argues a benign interpretation: the knight imagery represents protection and defense, and "coon" may refer to raccoon hunting slang in a casual outing context. The verifier agent, after analyzing both perspectives with their paired evidences, incorrectly concludes that the video is non-hateful. It is misled by the surface-level plausibility of the non-hate interpretation, such as the casual audio framing and the generic warrior imagery, while underweighting the well-established offensive connotation of the racial slur and the provocative crusader symbolism. This case suggests that when the retrieved evidence supports both interpretations at a superficial level, the verifier may not fully capture the deeper cultural and contextual signals necessary for accurate judgment, which could lead to error propagation from ambiguous clue-evidence pairs.

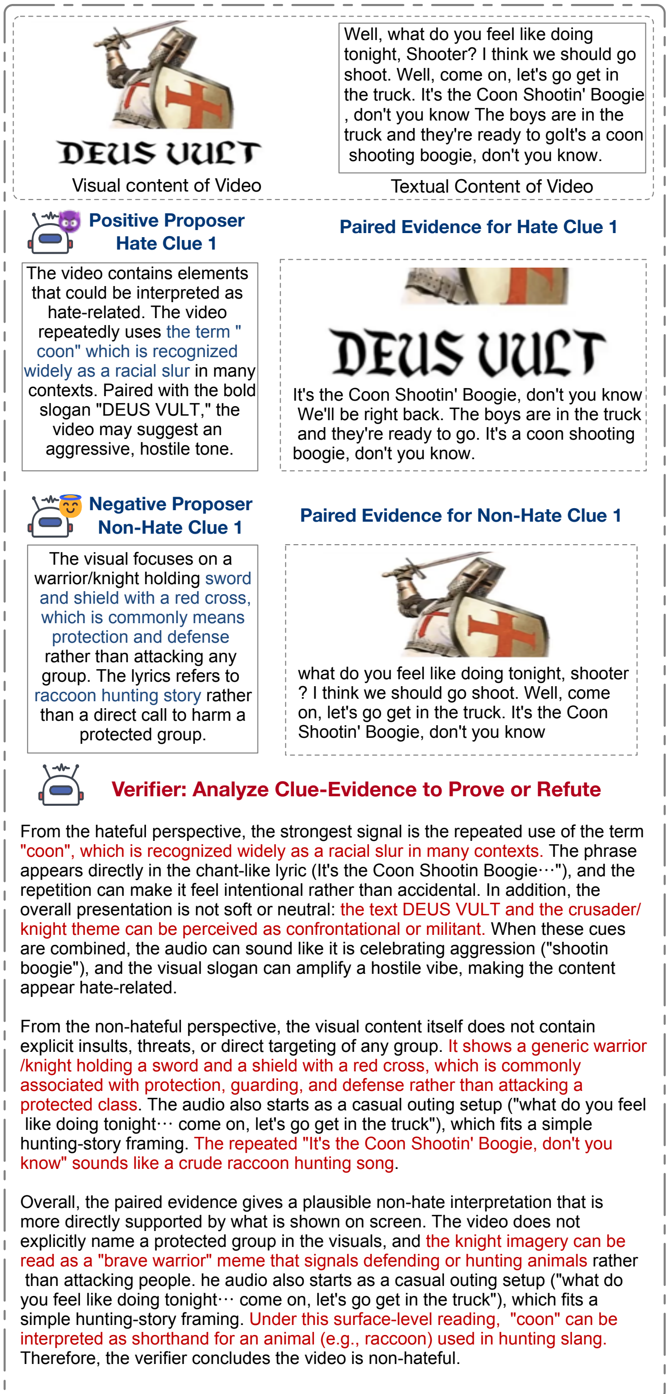


Figure 9. Failure case where the verifier agent incorrectly concludes the video is non-hateful, misled by the surface-level plausibility of a benign raccoon-hunting interpretation despite the presence of a racial slur and crusader symbolism.

V. Conclusion

In this study, we aim to make hate detection in video domain more transparent and trustworthy for both moderators and viewers. To this end, we propose MATCH, the first interpretable hate video detection framework based on multiple LMM agent collaboration. MATCH introduces a novel Dual-Perspective Proposing Paradigm, where two LMM agents act as *Proposers* to generate complementary clues from opposing perspectives, effectively addressing the inherent ambiguity and multifaceted nature of video intent. To mitigate hallucinations often introduced by LMMs when handling complex temporal dynamics in videos, we propose a Spatiotemporal Evidence-Grounded Verification Mechanism to improve the factuality of the generated clues from the proposers. In this stage, spatiotemporal evidence extracted from the video is paired with the clues via multimodal retrieval, and passed to a third LMM agent acting as the *Verifier*, which assesses their consistency and credibility, ultimately forming faithful and interpretable rationales for the prediction. These high-quality rationales distilled from LMMs provide rich and deep semantic cues, which are then combined with video features to achieve superior detection performance. Future work will focus on promoting MATCH to real-world video-sharing platforms, delivering both hate detection and explanation services.

VI. Ethical Statement

This work focuses on hateful video detection, a task with important societal implications. We acknowledge several ethical considerations associated with this line of research. Like all automated content moderation systems, our approach may produce false positives, which could lead to unjustified content removal or user penalties. We emphasize that MATCH is intended to support human moderators rather than replace them, and the provided explanations are designed to improve transparency and facilitate informed decision-making. Multi-agent reasoning systems may inherit or amplify biases present in large multimodal models or training data. To mitigate this risk, our framework incorporates evidence-grounded verification, which grounds model reasoning in observable video content instead of relying solely on parametric knowledge. Our method operates on existing video content and does not require access to personal metadata beyond the video itself. Nevertheless, we recognize that video analysis may raise privacy concerns, and we advocate that any real-world deployment should comply with relevant data protection regulations and platform policies.

References

[1] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li, "Fakingrecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process," in *ACM International Conference on Multimedia (MM)*, 2024, pp. 1351–1360.

[2] Rishikesh Devanathan, Apoorva Singh, A. S. Poornash, and Sriparna Saha, "Seeing Beyond Words: Multimodal Aspect-Level Complaint Detection in Ecommerce Videos.," in *ACM International Conference on Multimedia (MM)*, 2024, pp. 243–252.

[3] Weilong Chen, Wenhao Hu, Xiaolu Chen, Weimin Yuan, Yan Wang, Yanru Zhang, and Zhu Han, "Tri-modal transformers with mixture-of-modality-experts for social media prediction," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 35, no. 2, pp. 1897–1909, 2025.

[4] Penggang Qin, Shiwei Wu, Tong Xu, Yanbin Hao, Fuli Feng, Chen Zhu, and Enhong Chen, "When I fall in love: Capturing video-oriented social relationship evolution via attentive GNN," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 34, no. 6, pp. 5160–5175, 2024.

[5] Jian Lang, Rongpei Hong, Ting Zhong, Yong Wang, and Fan Zhou, "Nip rumors in the bud: Retrieval-guided topic-level adaptation for test-time fake news video detection," in *Proceedings of the 32st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2026.

[6] Shuo Niu, Keegan Veazey, Phoenix Pagan, and Abhisan Ghimire, "Understanding Hate Group Videos on YouTube.," in *Conference on Computer Supported Cooperative Work (CSCW)*, 2022, pp. 32–36.

[7] Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou, "Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection," in *The Web Conference (WWW)*. ACM, 2025.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision.," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[9] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee, "Hatemm: A Multi-Modal Dataset for Hate Video Classification," in *International Conference on Web and Social Media (ICWSM)*, 2023, pp. 1014–1023.

[10] Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee, "Multihateclip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili," in *ACM International Conference on Multimedia (MM)*, 2024, pp. 7493–7502.

[11] Han Wang, Tan Rui Yang, and Roy Ka-Wei Lee, "Cross-Modal Transfer from Memes to Videos: Addressing Data Scarcity in Hateful Video Detection," *arXiv*, 2025.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual Instruction Tuning.," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[13] Shuai Bai, Keqin Chen, Xuejing Liu, and Wang et al., "Qwen2.5-VL Technical Report," *arXiv.org*, vol. abs/2502.13923, 2025.

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li, "Llava-OneVision: Easy Visual Task Transfer," *Transactions on Machine Learning Research*, 2025.

[15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qing-Long Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai, "Internvl: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Rongpei Hong, Jian Lang, Ting Zhong, and Fan Zhou, "Borrowing eyes for the blind spot: Overcoming data scarcity in malicious video detection via cross-domain retrieval augmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2025.

[17] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha, "A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 11709–11724.

[18] Yiwei Sun, Zhihang Liu, Chuanbin Liu, Bowei Pu, Zhihan Zhang, and Hongtao Xie, "Hallucination Mitigation Prompts Long-term Video Understanding," *arXiv.org*, vol. abs/2406.11333, 2024.

[19] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, "Large Language Models are Zero-Shot Reasoners.," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[21] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty, "Detecting Harmful Memes and Their Targets.," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 2783–2796.

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

[23] Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou, "From shallow humor to metaphor: Towards label-free harmful meme detection via lmm agent self-improvement," in *Proceedings of the 32st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2026.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi, "Blip-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.," in *International Conference on Machine Learning (ICML)*, 2023, pp. 19730–19742.

[25] Mohd. Istiaq Hossain Junaid, Faisal Hossain, and Rashedur M. Rahman, "Bangla Hate Speech Detection in Videos Using Machine Learning," *Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021.

[26] Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramon Rodriguez, "Audio-Based Hate Speech Classification from Online Short-Form Videos," in *International Conference on Asian Language Processing (IALP)*, 2021, pp. 72–77.

[27] Ching Seh Wu and Unnathi Bhandary, "Detection of Hate Speech in Videos Using Machine Learning," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2020.

[28] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, 1998.

[29] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang, "Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models.," in *The Web Conference (WWW)*, 2024, pp. 2359–2370.

[32] Yue Hu, Yuzhu Cai, Yaxin Du, Xinyu Zhu, Xiangrui Liu, Zijie Yu, Yuchen Hou, Shuo Tang, and Siheng Chen, "Self-Evolving Multi-Agent Networks for Software Development," in *The Thirteenth International Conference on Learning Representations*, 2025, vol. abs/2410.16946.

[33] Riyaz Ahuja, Jeremy Avigad, Prasad Tetali, and Sean Welleck, "Improver: Agent-Based Automated Proof Optimization," in *The Thirteenth International Conference on Learning Representations*, 2025, vol. abs/2410.04753.

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[35] Penggang Qin, Tong Xu, Chao Zhang, Heda Wang, Yao Hu, and Enhong Chen, "Scenario-aware multimodal chain-of-thought prompting for rationales of video social relations," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025.

[36] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark, "Self-Refine: Iterative Refinement with Self-Feedback.," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[37] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu, "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 17889–17904.

[38] Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Weijuan Tu, Zhaofeng He, and Lijie Wen, "Llmarena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, vol. abs/2402.16499.

[39] Shih-Chia Huang, Ming-Kai Jiau, and Chih-An Hsu, "A high-efficiency and high-accuracy fully automatic collaborative face annotation system for distributed online social networks," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 24, no. 10, pp. 1800–1813, 2014.

[40] Md. Ashraful Islam, Mohammed Eunus Ali, and Md. Rizwan Parvez, "Mapcoder: Multi-Agent Code Generation for Competitive Problem Solving," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 4912–4944.

[41] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu, "Chateval: Towards Better LLM-based Evaluators through Multi-Agent Debate," in *International Conference on Learning Representations (ICLR)*, 2024.

[42] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al., "Metagpt: Meta programming for a multi-agent collaborative framework," in *The Twelfth International Conference on Learning Representations*.

[43] Yashar Talebirad and Amirhossein Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent llm agents," *arXiv preprint arXiv:2306.03314*, 2023.

[44] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun, "Communicative agents for software development," *arXiv preprint arXiv:2307.07924*, vol. 6, no. 3, 2023.

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

[46] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma, "Pp-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System," *arXiv*, vol. abs/2206.03001, 2022.
[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.
[48] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.," in *International Conference on Learning Representations (ICLR)*, 2021.
[50] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 3451–3460, 2021.
[51] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is Space-Time Attention All You Need for Video Understanding?," in *International Conference on Machine Learning (ICML)*, 2021, pp. 813–824.
[52] Yinghui Zhang, Tailin Chen, Yuchen Zhang, and Zeyu Fu, "Enhanced multimodal hate video detection via channel-wise and modality-wise fusion," in *IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2025, pp. 183–190.
[53] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin, "Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning," in *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2024, pp. 5333–5347.
[54] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang, "Modularized Networks for Few-shot Hateful Meme Detection.," in *The Web Conference (WWW)*, 2024, pp. 4575–4584.
[55] Abdelrahman Abouelenin, Atabak Ashfaq, and Adam et al. Atkinson, "Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs," *arXiv*, 2025.
[56] OpenAI, "Gpt-4o System Card," *arXiv.org*, vol. abs/2410.21276, 2024.
[57] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu, "G-Eval: Nlg Evaluation using Gpt-4 with Better Human Alignment.," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 2511–2522.
[58] Qwen An Yang, Baosong Yang, Beichen Zhang, and Hui et al., "Qwen2.5 Technical Report," *arXiv.org*, vol. abs/2412.15115, 2024.
[59] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, and Adi et al Ganesh, "Gpt-5 System Card," *arXiv*, 2025.
[60] Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang, "Explainable Fake News Detection with Large Language Model via Defense Among Competing Wisdom," in *The Web Conference (WWW)*, 2024, pp. 2452–2463.
[61] Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee, "Evaluating GPT-3 Generated Explanations for Hateful Content Moderation.," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 6255–6263.
[62] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal, "Likert scale: Explored and explained," *British journal of applied science & technology*, vol. 7, no. 4, pp. 396–403, 2015.
[63] Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, and Nino et al. Vieillard, "Gemma 3 Technical Report," *arXiv*, 2025.
[64] L. Maaten and Geoffrey E. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[65] Terry K Koo and Mae Y Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.
[66] Qiujie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang, "Human simulacra: Benchmarking the personification of large language models," in *The Thirteenth International Conference on Learning Representations*, 2025.