

Hate Is More Than What We See: Precedent-Guided Structured Reasoning for Hateful Video Detection

Junyi Chen¹, Qian Liu^{1*}, Jing Sun¹, Yi Zhang²

¹University of Auckland ²University of Technology Sydney

{junyi.chen, liu.qian, jing.sun}@auckland.ac.nz, yi.zhang@uts.edu.au

Abstract

Hateful video detection is a critical task for maintaining a healthy online ecosystem. However, existing methods are limited in their ability to deeply comprehend hateful semantics in videos and lack an effective mechanism for handling ambiguous cases near the decision boundary, making implicit hate detection particularly challenging. In this work, we address these limitations through both internal meaning formation from the video itself and external evidence consultation, and propose a novel framework, HATELENS. Internally, we design perception-to-cognition reasoning to perceive multimodal cues, perform cognition-level deep understanding, and progressively construct structured hateful meaning. Externally, we propose a test-time retrieval and adaptation mechanism to provide richer local evidence to help disambiguate implicit hate, which operates on fused multimodal representations to exploit annotated precedents. Extensive experiments on four public datasets demonstrate that our method achieves state-of-the-art performance, with macro-precision improvements of up to 16%. The code of our project is available at <https://github.com/JunyiChen-ai/HATELENS>.

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

1 Introduction

The proliferation of short-form video platforms such as YouTube Shorts, TikTok, and Bilibili has created new vectors for the dissemination of hateful content (Das et al., 2023; Wang et al., 2024a; Lang et al., 2025; Jing et al., 2025). Unlike text-only hate speech, hateful videos embed harmful messages across multiple modalities (e.g., visual imagery, spoken language, on-screen text, and audio cues), making them particularly challenging to detect and moderate (Kiela et al., 2020; Lang et al., 2025; Mathew et al., 2021).

*Corresponding author

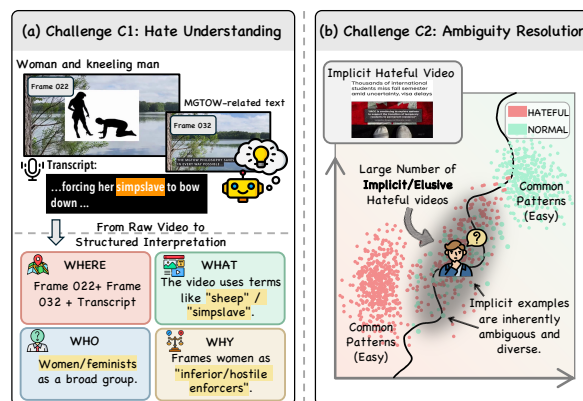


Figure 1: Illustration of two critical challenges in hateful video detection. (1) *Hate Understanding*: It requires structured semantic decomposition of hateful meaning in videos, instead of surface-level multimodal cue identification. (2) *Ambiguity Resolution*: Hate expressions in videos are often implicit, rendering a substantial portion of hateful samples ambiguous and distributed along the decision boundary.

Previous methods address hateful video detection through multimodal feature fusion, where representations from different modalities are extracted using pretrained encoders and combined via attention mechanisms (Das et al., 2023; Céspedes-Sarrias et al., 2025; Tsai et al., 2019), mixture-of-experts architectures (Lang et al., 2025), or graph-based fusion (Yue et al., 2025). More recently, MLLMs have been explored in hateful video detection (Jing et al., 2025; Yang et al., 2025), typically through chain-of-thought (CoT) prompting that first interprets visual and textual elements and then produces a final judgment. Despite these advances, two key limitations remain insufficiently addressed.

C1: Missing structural understanding of hateful meaning. Hateful meaning in videos is rarely conveyed by isolated modality cues. Instead, it often emerges from the contextual relations among multimodal cues, requiring structured semantic de-

composition for proper interpretation. As illustrated in Fig. 1 (a), understanding hate requires moving beyond surface-level cues to semantically understand *where* the supporting evidence appears, *what* hateful content is expressed, *who* is targeted, and *why* it is hateful. Existing multimodal fusion methods can encode multimodal cues, but they typically aggregate these cues into holistic representations directly without explicitly decomposing the hateful meaning they convey. As a result, the learned global representations tend to encode surface correlations rather than transferable hate semantics, limiting their generalizability to evolving hateful videos. Recent MLLM-based approaches (Jing et al., 2025; Yang et al., 2025, 2026) partially improve reasoning capability, yet they still mainly rely on surface-level descriptions and free-form reasoning rather than structured semantic decomposition for hate interpretation.

C2: Missing intervention mechanisms for resolving ambiguous cases. As illustrated in Fig. 1 (b), some implicit hateful videos lie near the decision boundary and are difficult to distinguish from normal content. This suggests that general global representations learned during training are insufficient to reliably resolve such cases without additional external intervention. Although retrieval-based methods (Lang et al., 2025) introduce external evidence from similar annotated cases, they typically rely on early multimodal features with manually designed modality weights, causing retrieved neighbors to be surface-similar rather than decision-relevant.

To alleviate these challenges, we propose HATE-LENS, a new method inspired by structured sense-making (Weick, 1995), which organizes fragmented observations into coherent interpretations, and case-based reasoning (Aamodt and Plaza, 1994), which resolves uncertain cases through similar annotated precedents. Specifically, HATELENS advanced hateful video detection through internal meaning formation and external ambiguity resolution. The *P2C Generator* leverages the MLLM to transform multimodal observations into a structured perception-to-cognition reasoning schema, summarizing *where* the evidence appears, *what* hateful content is expressed, *who* is targeted, *why* it is hateful, and *how* it may promote hostility. The *Schema-guided Modality Router* then fuses raw multimodal features with this structured reasoning schema to obtain discriminative feature representations. Finally, the *Whitened Neighbor In-*

terpolator performs test-time decision calibration for ambiguous cases by retrieving annotated precedents in a geometry-conditioned fused representation space and interpolating their label distribution with the classifier output. Compared with prior retrieval-based methods that rely on early multimodal features and manually specified modality weights, our interpolator retrieves neighbors that are more aligned with the final decision semantics, providing local evidence beyond surface-level similarity. Our contributions are summarized as:

- We propose a new paradigm for hateful video detection, formulating the task as evidence-grounded decision formation from both internal and external environments.
- We introduce the *P2C Generator*, which uses an MLLM to transform internal multimodal observations into a structured interpretive schema through perception-to-cognition reasoning.
- We propose the *Whitened Neighbor Interpolator*, a geometry-conditioned test-time retrieval module that refines ambiguous decisions with external annotated precedents that are more closely aligned with the final prediction space.
- Extensive experiments are conducted on four datasets across two languages, achieving state-of-the-art results with detailed analysis of both reasoning quality and retrieval effectiveness.

2 Method

Problem Formulation. Let q denote a video on online platforms. Each video q is represented by its multimodal content, including textual (t), visual (v), and audio (a) modalities. For each video, we extract textual, visual, and audio information. Specifically, for the textual modality t , we extract the video title and description; for the visual modality v , we uniformly sample n frames from the video, denoted as $v = \{v_1, \dots, v_n\}$; and for the audio modality a , we obtain a global transcript t_a using automatic speech recognition. The goal of hateful video detection is to determine whether a video q contains hateful content by jointly modeling the modalities t , v , and a .

Overview. The overall framework of HATELENS is illustrated in Fig. 2. Following the formulation in the introduction, HATELENS decomposes hateful video detection into internal meaning formation

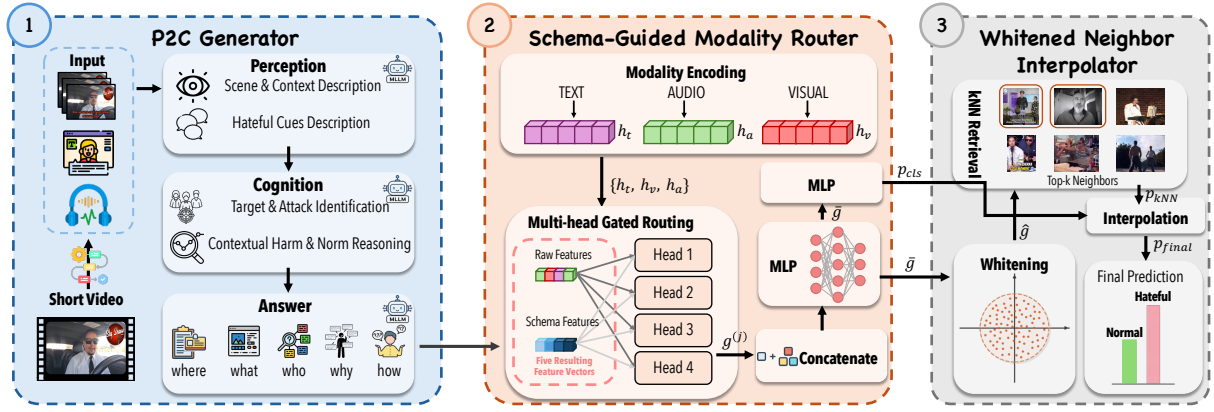


Figure 2: Overview of the HATELENS framework. Given a short video, (1) the **P2C Generator** prompts an MLLM through a perception-to-cognition CoT to produce a structured reasoning schema encoding *where* the evidence appears, *what* the hateful content is, *who* the target is, *why* it is harmful, and *how* it may cause harm; (2) the **Schema-guided Modality Router** fuses raw multimodal features with the interpretive schema via multi-head gated routing, yielding discriminative representations; (3) the **Whitenized Neighbor Interpolator** geometry-conditions these representations via covariance whitening and retrieves the k nearest annotated neighbors at inference time, interpolating their labels with the classifier output to resolve ambiguous predictions.

over the input video and external precedent consultation over annotated training cases. The pipeline follows a perception-cognition-response process: it first observes the video, organizes the observations into a structured interpretive schema, and then predicts the label with support from similar cases. The framework consists of three modules. (1) The *P2C Generator* prompts the MLLM to transform scene-level perception into a structured interpretive schema, including *where* the evidence appears, *what* hateful content is expressed, *who* is targeted, *why* it is harmful, and *how* it may promote hateful activity. (2) The *Schema-guided Modality Router* fuses multimodal features with the generated interpretive schema to produce discriminative representations. (3) The *Whitenized Neighbor Interpolator* performs gradient-free test-time adaptation by retrieving annotated nearest neighbors in a whitenized representation space and interpolating their label evidence with the classifier output to support ambiguous cases.

2.1 P2C Generator

To address the challenge of structurally interpreting hateful videos, we design the *P2C Generator* to mimic the evidence-building process of human annotators in hateful video assessment: from *perception* of multimodal content, to *cognition* about targets and harms, and then to an evidence-supported *answer*. This design provides a structured alternative to conventional multimodal fusion, which often maps observable features directly to

labels, and to generic MLLM prompting, which may stop at surface description or reason from an implicit judgment without explicit decomposition. Specifically, P2C prompts the MLLM to reason within `<think>...</think>` tags and summarize the perception-to-cognition process into structured interpretive fields in `<answer>...</answer>` tags. The full prompt is provided in Appendix D.5. The workflow contains the following three stages.

Stage 1: Perception. Given video information t, t_a, v , P2C first asks the MLLM to observe what is directly perceivable rather than make an immediate label judgment. It performs scene-level perception through *Scene and Context Description* (Step 1), which summarizes the environment, participants, visible text, symbols, and overall activity, and cue-level perception through *Hateful Cues Description* (Step 2), which identifies verbal, textual, visual, or symbolic cues that may indicate hateful expression.

Stage 2: Cognition. After acquiring sufficient surface-level information, humans conduct reasoning based on the perceived cues and infer possible consequences. In hateful video assessment, this process moves from shallow to deep understanding: it first identifies who or what is being targeted, if anyone, and explains the type of attack involved, such as insult, harassment, or humiliation, which we term *Target and Attack Identification* (Step 3); it concludes with *Contextual Harm and Norm Reasoning* (Step 4), which explains why the content should be considered hateful in context.

Stage 3: Answer. After completing the reasoning

process, the MLLM is expected to provide a short structured summary related to the key evidential dimensions of hateful content in the video, including the classification label, the spatio-temporal-audio key evidence (**where**, e), a description of the hateful content (**what**, w), the target (**who**, τ), the reason (**why**, r) for this classification, and **how** (o) it may promote hateful activity and cause harm.

Note that all three stages are completed within a single prompt call.

2.2 Schema-Guided Modality Router

Modality routing has been shown to be an effective fusion paradigm for multimodal feature integration (Vaswani et al., 2017; Tsai et al., 2019). In this regard, we view the structural fields returned by the MLLM (i.e., e , w , τ , r , and o) as a reasoning schema that can more effectively guide multimodal fusion.

We map each modality input into latent features using the corresponding off-the-shelf encoders, including a pre-trained language model (LM), visual transformer (VT), and audio transformer (AT), as:

$$\begin{aligned} \mathbf{h}_t &= \text{LM}(t; t_a), \\ \mathbf{h}_v &= \text{VT}(v), \\ \mathbf{h}_a &= \text{AT}(a), \end{aligned} \quad (1)$$

where $\mathbf{h}_* \in \mathbb{R}^d$ denotes the extracted feature vector after applying mean pooling, and $;$ denotes concatenation. Accordingly, we also encode the textual structural fields separately using the same LM, resulting in $\mathbf{h}_{e/w/\tau/r/o}$.

Subsequently, we fuse the eight resulting feature vectors using a multi-head gated routing mechanism:

$$\begin{aligned} \alpha_m^{(j)} &= \frac{\exp(f_j(\mathbf{h}_m))}{\sum_{m' \in \mathcal{M}} \exp(f_j(\mathbf{h}_{m'}))}, \\ \mathbf{g}^{(j)} &= \sum_{m \in \mathcal{M}} \alpha_m^{(j)} \mathbf{h}_m, \end{aligned} \quad (2)$$

where $\mathcal{M} = \{t, v, a, e, w, \tau, r, o\}$. The final representation is formed by concatenating the outputs of all routing heads with a mean-pooled modality summary, and is then passed through a classification MLP:

$$\begin{aligned} \bar{\mathbf{g}} &= \text{MLP}([\mathbf{g}^{(1)}; \dots; \mathbf{g}^{(H)}]) \\ \mathbf{p}_{cls} &= \text{MLP}(\bar{\mathbf{g}}), \end{aligned} \quad (3)$$

where $\bar{\mathbf{g}}$ is the fused representation and $\mathbf{p}_{cls} \in \mathbb{R}^2$ denotes the predicted probability distribution over the hateful and normal classes for the input video. The model is trained with the cross-entropy loss.

2.3 Whitened Neighbor Interpolator

In the video setting, hateful expression can be highly diverse and implicit, making the general representations learned during training less reliable for capturing such cases at inference time. To address this, we propose leveraging training samples with similar fused features and annotated ground truth to sharpen the decision boundary. This design mirrors content moderation practice, where moderators learn general policies but consult prior precedent cases for difficult judgments (Klonick, 2018; Clarke et al., 2023; Snell et al., 2017). Specifically, we split this into two steps.

Geometric Conditioning. Previous methods (Lang et al., 2025; Lewis et al., 2020; Karpukhin et al., 2020) for multimodal retrieval typically operate directly on the initially extracted multimodal features with manually configured modality weights. This is problematic for two reasons. First, such designs are difficult to adapt to dynamically varying cases. Second, the initial features are not sufficiently discriminative for the final classification objective, which limits the usefulness of the retrieved cases for informing decisions. This motivates us to instead leverage the fused penultimate representations $\bar{\mathbf{g}}$. However, these representations combine heterogeneous sources, including raw multimodal features and LLM-derived reasoning, which results in an anisotropic embedding space where distances become unreliable (Yi et al., 2025; Tsukagoshi and Sasano, 2025; Gao et al., 2019). To enable meaningful decision semantic-based retrieval, we apply covariance-based whitening:

$$\hat{\mathbf{g}} = \frac{(\bar{\mathbf{g}} - \boldsymbol{\mu})\mathbf{W}_{\text{whiten}}}{\|(\bar{\mathbf{g}} - \boldsymbol{\mu})\mathbf{W}_{\text{whiten}}\|_2}, \quad (4)$$

where $\boldsymbol{\mu}$ is the training set mean and $\mathbf{W}_{\text{whiten}}$ is derived from shrinkage-regularized covariance estimation (Yi et al., 2025). This conditioning ensures that similarity computations reflect genuine semantic proximity rather than artifacts of the heterogeneous feature geometry.

Retrieval and Interpolation. At inference, we store whitened penultimate representations $\{\hat{\mathbf{g}}_i\}_{i=1}^N$ from the training set along with their labels. For a test sample with representation $\hat{\mathbf{g}}_{\text{test}}$, we retrieve its k nearest neighbors and compute a soft label distribution:

$$\mathbf{p}_{\text{kNN}}(y = c) = \frac{\sum_{i \in \mathcal{N}_k} \mathbb{1}[y_i = c] \cdot \kappa(\hat{\mathbf{g}}_{\text{test}}, \hat{\mathbf{g}}_i)}{\sum_{i \in \mathcal{N}_k} \kappa(\hat{\mathbf{g}}_{\text{test}}, \hat{\mathbf{g}}_i)}, \quad (5)$$

where $\mathcal{N}_k$ denotes the k nearest neighbors and $\kappa(\cdot, \cdot) = \exp(\text{sim}(\cdot, \cdot))$ is a similarity kernel, which enhances weights from more similar neighbors. The final prediction interpolates the classifier logits with the retrieval distribution:

$$\mathbf{p}_{\text{final}} = (1 - \alpha) \cdot \mathbf{p}_{\text{cls}} + \alpha \cdot \mathbf{p}_{\text{kNN}}, \quad (6)$$

where α is the trade-off parameter.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on four hateful video detection datasets: HateMM (Das et al., 2023), MultiHateClip-Youtube (MHClip-Y), MultiHateClip-Bilibili (MHClip-B) (Wang et al., 2024a), and ImpliHateVid (Rehman et al., 2025). We follow the data splits used in prior work (Lang et al., 2025; Rehman et al., 2025) for fair evaluation. Detailed dataset descriptions and statistics are provided in Appendix B.

Baselines. We compare against baselines from three categories: (1) *Unimodal*: BERT (Devlin et al., 2019), ViViT (Arnab et al., 2021), MFCC (Davis and Mermelstein, 1980); (2) *Multimodal fusion*: Pro-Cap (Cao et al., 2023), HateMM (Das et al., 2023), MHCL (Wang et al., 2024a), MoRE (Lang et al., 2025), ImpliHateVid (Rehman et al., 2025), CMFusion (Zhang et al., 2024); (3) *MLLM-based*: MiniCPM-V (Yao et al., 2024), LLaVA-OV (Li et al., 2024), Qwen2-VL (Wang et al., 2024b), HVGuard (Jing et al., 2025). Details of baselines are in Appendix C.

Metrics. Following previous work (Lang et al., 2025; Jing et al., 2025), we report Accuracy (ACC), Macro-F1 (M-F1), Macro-Precision (M-P), and Macro-Recall (M-R). The best checkpoint and hyperparameter configuration selected on the validation set is then evaluated on the test set for performance reporting.

Implementation Details. We use GPT-5.4-nano as the MLLM backbone, with frozen BERT (Devlin et al., 2019), ViT (Dosovitskiy et al., 2021), and WavLM (Chen et al., 2022) encoders for text, visual, and audio features, respectively. The fusion module uses $H=4$ routing heads with hidden dimension $h=192$. For the interpolator, we set $(k, \alpha) = (3, 0.3), (3, 0.5), (25, 0.1), (25, 0.5)$ for HateMM, MHClip-Y/B, and ImpliHateVid based on validation set. Further details and hyperparameter analysis are in Appendix D and E.2.

3.2 Comparison with SOTA

Table 1 presents the main results, from which we highlight three key observations. **(O1): Multimodal fusion outperforms unimodal approaches**, as hateful expression may emerge across different modalities that single-modality models cannot capture. **(O2): Incorporating MLLMs into multimodal modeling further improves performance**. MLLM-based methods (e.g., HVGuard) can outperform feature-fusion baselines, and our method extends this advantage further by guiding the MLLM through a structured perception-to-cognition process that explicitly models evidence, targets, and contextual harm. **(O3): HATELENS achieves the strongest performance across all four datasets on all metrics**. Compared with the best competitor, HATELENS yields absolute accuracy gains of 3.7%, 7.3%, 9.6%, and 6.8% respectively, while its higher precision and recall indicate more hateful content detected with fewer false positives. These results validate our paradigm of first examining videos through progressively deeper perception and cognition, then consulting external cases to support the final decision.

3.3 Ablation Study

Table 2 reports ablation results across four datasets. **Effect of P2C Generator.** At the answer level, removing individual fields (w/o *where/what/why/how*) consistently degrades performance, indicating that complete structural understanding is essential. At the prompting level, restricting the pipeline to surface-level observation (Perception only) or high-level reasoning (Cognition only) alone underperforms the full two-stage design, confirming their complementary roles; similarly, eliciting fields without the perception-to-cognition decomposition (Generic prompt) or replacing our prompt with a describe-analyze-judge pipeline (HVGuard CoT) both cause significant drops. **Effect of Schema-guided Router.** Substituting our router with prior advanced fusion methods (i.e., MoRE and HVGuard) leads to notable degradation, demonstrating that our schema-conditioned routing more effectively integrates LLM-derived reasoning with raw modality features. **Effect of Whitened Neighbor Interpolator.** Removing the interpolator entirely (w/o interpolation) confirms the value of consulting annotated cases at test time; removing only the whitening step (w/o whitening) also hurts, as anisotropic representa-

Table 1: Comparison with state-of-the-art methods on four hateful video detection datasets. Best results in **bold**, second best underlined. $\uparrow$ indicates the improvement of HATELENS over the best baseline. a_{-b} denotes $a \times 10^{-b}$;

Method	HateMM				MHClip-Y				MHClip-B				ImpliHateVid			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
<i>Unimodal Methods</i>																
BERT	69.1	63.7	70.1	64.0	65.5	49.1	55.2	52.2	72.5	67.7	68.4	62.8	75.2	75.2	75.3	75.2
ViViT	68.2	66.7	66.8	66.6	61.4	61.4	62.2	61.1	71.0	66.1	66.6	65.8	76.5	76.5	76.7	76.5
MFCC	65.4	60.3	64.1	60.7	46.5	47.2	58.8	52.2	63.1	52.5	54.1	53.0	70.0	70.0	70.1	70.0
<i>Multimodal Fusion Methods</i>																
Pro-Cap	64.5	63.3	63.4	63.2	70.1	66.3	66.3	66.3	72.5	66.8	66.1	68.3	82.3	82.3	82.4	82.3
HateMM	76.0	72.8	77.9	72.0	71.5	63.2	68.3	62.6	71.0	61.8	66.5	61.4	76.6	76.5	76.8	76.6
MHCL	77.4	76.5	77.5	76.6	65.4	65.4	67.2	64.9	76.5	73.1	73.2	73.0	85.2	85.2	85.4	85.2
MoRE	83.4	82.4	81.8	83.3	77.5	75.2	75.7	74.8	78.5	74.8	75.7	74.1	84.8	84.7	85.4	84.8
ImpliHateVid	82.8	82.0	82.1	82.0	76.1	71.7	71.6	71.8	78.3	72.8	74.4	71.8	<u>87.5</u>	<u>87.5</u>	<u>87.6</u>	<u>87.5</u>
CMFusion	75.6	74.1	75.3	73.7	72.2	68.8	69.5	68.2	73.8	70.1	71.0	69.4	<u>68.5</u>	<u>68.4</u>	<u>68.9</u>	<u>68.6</u>
<i>MLLM-based Methods</i>																
MiniCPM-V	72.4	72.3	77.8	76.4	69.1	67.4	69.3	67.4	73.5	73.2	72.0	73.0	77.5	77.5	77.6	77.5
LLaVA-OV	75.6	75.6	77.9	78.3	73.7	70.5	70.5	66.8	75.2	70.2	71.4	70.3	79.0	79.2	79.0	79.0
Qwen2-VL	73.7	73.7	78.1	77.3	70.5	66.8	66.8	66.7	73.3	74.0	73.9	72.9	73.4	74.9	73.5	73.4
HVGuard	<u>87.9</u>	<u>87.5</u>	<u>87.3</u>	<u>87.6</u>	<u>79.8</u>	<u>76.2</u>	<u>77.0</u>	<u>75.6</u>	<u>79.6</u>	<u>75.6</u>	<u>76.0</u>	<u>75.3</u>	86.0	86.0	86.1	86.0
HATELENS	91.6	91.4	91.0	91.9	87.1	84.0	85.9	82.6	89.2	85.6	92.0	82.5	94.3	94.3	94.4	94.3
$\uparrow$ Improv.	+3.7	+3.9	+3.7	+4.3	+7.3	+7.8	+8.9	+7.0	+9.6	+10.0	+16.0	+7.2	+6.8	+6.8	+6.8	+6.8
<i>p-val</i>	3.2 ₋₃	1.8 ₋₃	7.4 ₋₃	9.1 ₋₄	5.6 ₋₄	8.3 ₋₄	2.1 ₋₄	4.7 ₋₄	1.4 ₋₄	3.7 ₋₄	6.2 ₋₄	2.8 ₋₄	8.9 ₋₄	6.3 ₋₄	2.1 ₋₃	7.5 ₋₄

Table 2: Ablation study (%). $\bar{\Delta}$: mean ACC drop across four datasets; darker shading indicates larger drop.

	HateMM		MH-Y		MH-B		ImpliHV		$\bar{\Delta}$
	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	
Full	91.6	91.4	87.1	84.0	89.2	85.6	94.3	94.3	—
<i>P2C Generator</i>									
w/o what	87.4	86.7	82.1	77.0	86.6	83.0	91.3	91.3	−3.7
w/o where	87.9	87.3	82.8	79.8	84.7	83.1	92.8	92.8	−3.5
w/o why	87.6	87.0	82.2	78.4	85.4	84.0	92.3	92.3	−3.7
w/o how	87.1	86.2	81.9	76.6	84.7	82.2	92.8	92.8	−3.9
Perception only	87.9	87.4	80.9	75.1	83.5	80.9	91.8	91.8	−4.5
Cognition only	86.0	85.5	80.2	74.3	82.1	81.1	89.5	89.5	−6.1
Generic prompt	86.4	85.8	80.5	74.8	83.0	81.4	90.2	90.2	−5.5
HVGuard CoT	85.1	84.3	79.1	73.2	78.3	73.2	89.3	89.3	−7.6
<i>Schema-guided Router</i>									
MoRE fusion	85.6	84.8	77.9	70.7	81.5	79.1	91.3	91.3	−6.5
HVGuard fusion	87.0	86.2	80.4	75.1	80.9	76.9	92.0	92.0	−5.5
<i>Whitened Neighbor Interp.</i>									
w/o interpolation	87.0	86.4	80.4	75.4	86.0	82.9	91.0	91.0	−4.5
w/o whitening	88.4	87.7	81.0	74.9	85.4	82.2	93.5	93.5	−3.5
Pre-fusion retrieval	85.6	84.4	82.8	77.2	82.2	73.7	91.0	91.0	−5.2

tions yield unreliable neighbors. Finally, performing retrieval on raw pre-fusion features (Pre-fusion retrieval) further degrades results, showing that retrieval quality critically depends on the discriminativeness of the embedding space.

3.4 P2C CoT Analysis

An effective MLLM-based CoT should exhibit two properties: (1) the output reasoning should provide informative evidence that aids human decision-making, and (2) the CoT design should generalize across different MLLMs with consistent perfor-

Table 3: Explanation quality evaluated by GPT-4o (5-point Likert).

Method	Info.	Sou.	Per.	Rea.	Flu.	Avg.
HVGuard	3.42	3.51	3.28	4.15	4.20	3.71
RAMF	3.38	3.45	3.35	4.02	4.21	3.68
MARS	3.41	3.34	3.57	3.96	4.18	3.69
Ours	4.12	4.05	3.92	4.28	4.21	4.12

mance gains.

Explanation Quality. We randomly sample 100 videos of each dataset and evaluate the reasoning outputs from four methods using GPT-4o as an expert judge across five criteria: (1) *Informativeness*: whether the explanation provides background and context beyond surface description; (2) *Soundness*: whether the reasoning is valid and logically coherent; (3) *Persuasiveness*: whether the conclusion is convincing; (4) *Readability*: whether the text follows proper grammar and structure; and (5) *Fluency*: whether the explanation flows smoothly with well-connected ideas. Each criterion is rated on a 5-point Likert scale. The detailed judge prompt is provided in Table 6 in the Appendix. We evaluate this by comparing against HVGuard (Jing et al., 2025), MARS (Yang et al., 2026) and RAMF (Yang et al., 2025).

As shown in Table 3, our P2C reasoning achieves the highest scores across all five criteria. This suggests that the structured perception-to-cognition

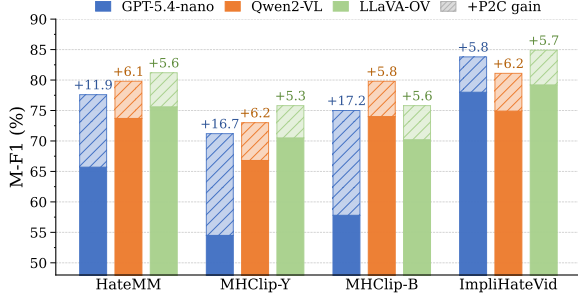


Figure 3: Generalizability of P2C CoT across different MLLMs. Applying our CoT design consistently improves performance over each MLLM’s default prompting across all four datasets.

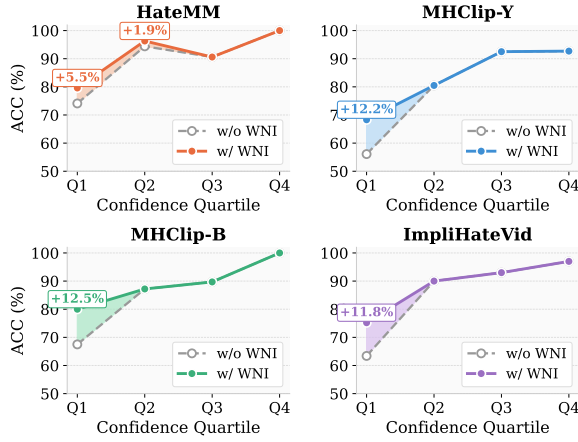


Figure 4: Interpolator gain by classifier confidence quartile across four datasets. Gains are concentrated in low-confidence bins.

decomposition produces more thorough and well-grounded explanations compared to the describe-analyze-judge paradigm.

Generalizability across MLLMs. To assess whether our CoT design transfers across different MLLMs, we apply the P2C prompt to three models of varying scale: GPT-5.4-nano, Qwen2-VL (Wang et al., 2024b), and LLaVA-OV (Li et al., 2024). As shown in Figure 3, applying our P2C CoT consistently improves each MLLM’s performance. This confirms that the perception-to-cognition decomposition is a generalizable design principle.

3.5 Interpolator Analysis

Correcting Hard Samples. To validate that the Whitened Neighbor Interpolator addresses the decision challenge on ambiguous content, we analyze *where* interpolator helps by binning test samples into quartiles based on the classifier’s prediction confidence. As shown in Figure 4, interpolator gains are overwhelmingly concentrated in low-

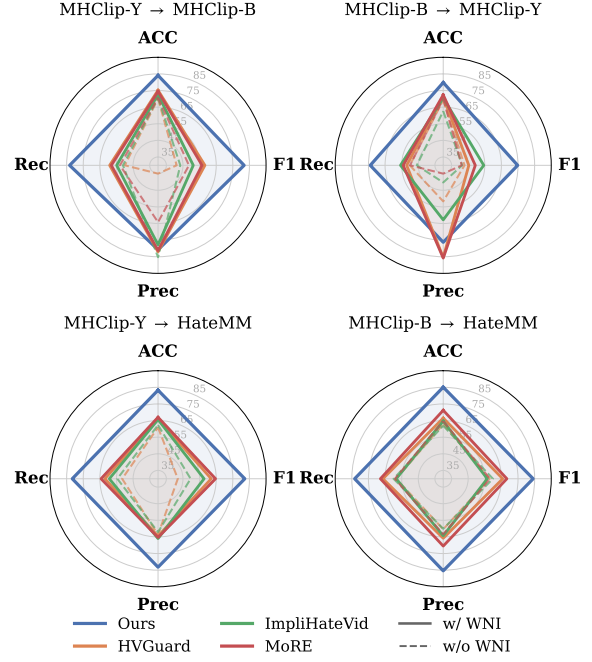


Figure 5: Cross-dataset transferability comparison. Solid/dashed lines show baselines with and without our Whitened Neighbor Interpolator.

confidence bins. This pattern is consistent across all four datasets, confirming that retrieval acts as targeted decision support for ambiguous cases rather than a uniform post-hoc adjustment.

Enhancement of Transferability. Hateful video detection models are often trained on one platform but deployed on another (Lu et al., 2019). To evaluate cross-domain robustness, we conduct exhaustive transfer experiments: for each of the four datasets, we train on one (source) and directly evaluate on the test set of another (target). We mainly focus on the performance of the baselines after plug-and-play integration of our interpolator module, where training instances from the source domain are used as the neighbor bank.

Figure 5 shows the transfer results for HateMM ↔ MHClip, with exhaustive transfer directions provided in Appendix E.5. We observe that after applying our interpolator module, the performance of all baselines improves substantially. This provides strong evidence that the interpolator can serve as an effective test-time adaptation module across varied contexts. Furthermore, when combined with P2C reasoning, our method achieves the strongest performance, benefiting from the complementary strengths of both components and yielding a more transferable decision boundary.

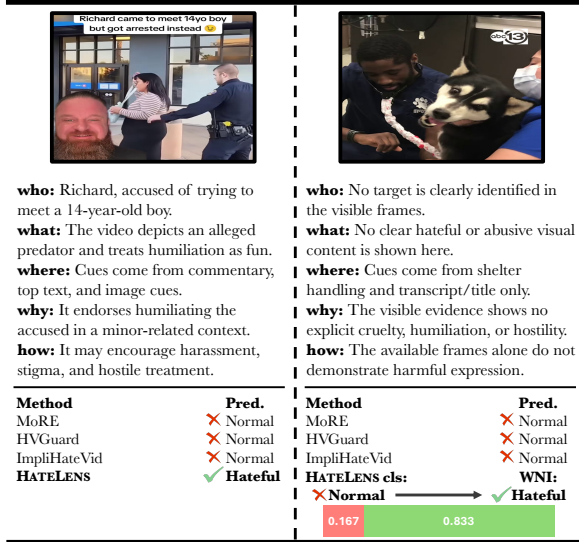


Figure 6: Case study of two hateful videos. Reasoning fields have been simplified for demonstration purpose. The colored bar chart shows the KNN voting result.

3.6 Case Study

Figure 6 presents predictions for two representative cases. In the first, the hateful expression is implicit and all baselines fail to identify it. Our method succeeds by grounding the prediction in a structured interpretive schema, which captures discriminative multimodal representations that surface-level analysis misses. The second case is more challenging: all methods including ours initially predict non-hateful, as the hateful cues are too subtle to be captured by MLLM-based reasoning or multimodal-based feature fusion. Nevertheless, the Whitened Neighbor Interpolator recovers the correct prediction at inference time by retrieving semantically aligned annotated neighbors in the geometrically conditioned embedding space to provide local precedent evidence, whose majority vote overrides the classifier output. More case studies are provided in Appendix E.8.

4 Related Work

The proliferation of harmful multimodal content, such as manipulated media and multimodal misinformation, has attracted growing research attention (Zhang et al., 2026a; Lian et al., 2026; Zhang et al., 2026b). Among these threats, this work focuses on hate speech conveyed through videos. Early hate speech detection research mainly focused on text-based settings, typically using traditional machine learning classifiers (Davidson et al., 2017; Mathew et al., 2021). With the rise of deep learning, increas-

ing attention has been devoted to hate detection in multimodal settings, from the introduction of image-text meme datasets (Kiela et al., 2020; Hee et al., 2022) to more recent studies on hateful video detection (Das et al., 2023; Wang et al., 2024a; Rehman et al., 2025). For videos, most existing work extends conventional visual-textual feature fusion to three modalities, i.e., visual, textual, and audio fusion, through different fusion mechanisms (Céspedes-Sarrias et al., 2025; Zhang et al., 2024; Lang et al., 2025; Yue et al., 2025; Rehman et al., 2025; Zadeh et al., 2017). With the emergence of LLMs, their strong reasoning capability has been successfully applied across diverse domains (Chen et al., 2025, 2026a; Lu et al., 2026b; Liu et al., 2025; Jia et al., 2026; Chen et al., 2026b; Lu et al., 2026a). Specifically, in the area of hateful video detection, some studies have begun to explore different prompting techniques with MLLMs, either by generating rationales to assist the training of downstream conventional models (Jing et al., 2025; Cao et al., 2023), or by directly performing zero-shot inference (Yang et al., 2025, 2026). However, most existing methods share two common limitations: they do not provide a well-grounded understanding of video content, but mainly model *what* is in the video, and they lack an effective mechanism for handling ambiguous cases involving implicit and diverse hateful expression. In this work, we first propose to internally understand videos through a perception-to-cognition reasoning process, and then resolve ambiguity by consulting annotated precedents that provide meaningful local precedent evidence.

5 Conclusion

In this work, we presented HATELENS, a dual-perspective framework for hateful video detection. In particular, we identify two key challenges: existing methods model surface-level features without understanding how hateful meaning is formed, and lack effective mechanisms for resolving the inherent ambiguity of implicit hate. By addressing these challenges, we arrive at two findings: (1) introducing a structured reasoning schema grounded in explicit evidence about content, targets, and contextual harm, into latent multimodal feature modeling yields substantially more discriminative representations and largely boosts detection performance, and (2) consulting annotated precedent samples at inference time through geometry-conditioned re-

trieval provides targeted correction for ambiguous predictions near the decision boundary, a complementary benefit that neither parametric modeling nor reasoning alone can achieve. Together, these findings suggest that structured internal reasoning combined with external precedent consultation offers a productive path toward more nuanced and trustworthy hateful video detection.

Limitations

While HATELENS achieves strong empirical results, several limitations remain. First, hateful content evolves rapidly on real-world platforms, with new memes, slang, and cultural references emerging continuously; our method trained on static datasets may struggle to generalize to unseen expressions without periodic retraining. Second, the retrieval bank is populated from a fixed training set, meaning the framework provides limited benefit in truly zero-shot scenarios where no annotated precedents exist for a given domain or language. Third, our evaluation covers English and Chinese content; how well the perception-to-cognition reasoning transfers to other languages and cultural contexts remains unclear. Finally, hateful content in the wild increasingly appears in longer-form videos where harmful intent may be sparsely distributed or gradually escalated across extended temporal spans, yet our evaluation is conducted primarily on short-video datasets, and whether our framework generalizes to such settings remains an open question.

Ethics Statement

This work involves detecting hateful content in online videos, which necessarily requires engaging with offensive material. All datasets used are publicly available and were collected with appropriate ethical oversight by their original creators. We note several ethical considerations: (1) Hate speech annotations are inherently subjective and may reflect annotator biases and our method inherits these biases through training data. (2) Automated detection systems risk both false positives (censoring legitimate speech) and false negatives (failing to flag harmful content), with disparate impacts across communities. (3) Our system should be used as a *decision-support tool* alongside human review, not as an autonomous moderator. (4) The LLM component operates on potentially sensitive content; researchers and practitioners should be aware that

prompting MLLMs with hateful material may produce harmful outputs as a byproduct of the reasoning process. (5) Deploying hate detection systems without community consultation risks disproportionate harm to marginalized groups whose speech patterns may be misclassified and we encourage practitioners to conduct community-specific evaluations before deployment. (6) We open-source our code and release the LLM-generated reasoning outputs to facilitate reproducibility, while committing to responsible access practices to prevent misuse of the released resources.

References

- Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, pages 39–59.
- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6816–6826.
- 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 ACM International Conference on Multimedia*, pages 5244–5252.
- Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. 2025. MM-HSD: Multi-modal hate speech detection in videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 2546–2555.
- Junyi Chen, Leyuan Liu, and Fan Zhou. 2025. Do not wait: Preemptive rumor detection with cooperative llms and accessible social context. *Information Processing & Management*, 62(3):103995.
- Junyi Chen, Qian Liu, Jing Sun, and Yi Zhang. 2026a. Truelens: video fake news detection with dual level evidence gathering and consolidation. In *Proceedings of the ACM Web Conference 2026*, pages 7178–7188.
- Junyi Chen, Mengjia Wu, Qian Liu, and Yi Zhang. 2026b. Explainable prediction of knowledge recombination: A synergized method with heterogeneous hypergraph learning and large language models. *Information Processing & Management*, 63(1):104336.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022.

- Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, pages 1505–1518.
- Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 364–376.
- 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*, pages 1014–1023.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 512–515.
- Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 357–366.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.
- 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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6201–6210.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.
- Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655.
- Runsong Jia, Mengjia Wu, Ying Ding, Jie Lu, and Yi Zhang. 2026. Agent-enhanced heterogeneous graph rag for academic question answering. In *Proceedings of the ACM Web Conference 2026*, pages 8765–8768.
- Yiheng Jing, Mingming Zhang, Yong Zhuang, Jiacheng Guo, Juan Wang, Xiaoyang Xu, Wenzhe Yi, Keyan Guo, and Hongxin Hu. 2025. Hvguard: Utilizing multimodal large language models for hateful video detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8993–9006.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.
- Kate Klonick. 2018. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6):1598–1670.
- Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 2025. Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection. In *Proceedings of the ACM Web Conference 2025*, pages 2763–2774.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900.
- Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. 2026. Generating attribution reports for manipulated facial images: A dataset and baseline. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30455–30473.
- Chenghao Liu, Qian Liu, Ziqin Zhu, Hao Fei, and Aniket Mahanti. 2025. David vs. goliath: Cost-efficient financial QA via cascaded multi-agent reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

- Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3192–3201.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
- Mingfei Lu, Mengjia Wu, Jiawei Xu, Weikai Li, Feng Liu, Ying Ding, Yizhou Sun, Jie Lu, and Yi Zhang. 2026a. From newborn to impact: Bias-aware citation prediction. In *Proceedings of the ACM Web Conference 2026*, pages 7600–7609.
- Mingfei Lu, Yi Zhang, Mengjia Wu, and Yue Feng. 2026b. From query to counsel: Structured reasoning with a multi-agent framework and dataset for legal consultation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17064–17078.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14867–14875.
- 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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.
- Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Nagendra Kumar. 2025. Implihatevid: A benchmark dataset and two-stage contrastive learning framework for implicit hate speech detection in videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 17209–17221.
- Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.
- 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 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.
- Alexander Tsesis. 2002. *Destructive messages: How hate speech paves the way for harmful social movements*. NYU Press.
- Hayato Tsukagoshi and Ryohei Sasano. 2025. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Jeremy Waldron. 2012. *The harm in hate speech*. Harvard University Press.
- 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 ACM International Conference on Multimedia*, pages 7493–7502.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Karl E. Weick. 1995. *Sensemaking in Organizations*, volume 3 of *Foundations for Organizational Science*. Sage Publications, Thousand Oaks, CA.
- Shuonan Yang, Tailin Chen, Jiangbei Yue, Guangliang Cheng, Jianbo Jiao, and Zeyu Fu. 2025. Reasoning-aware multimodal fusion for hateful video detection. *arXiv preprint arXiv:2512.02743*.
- Shuonan Yang, Yuchen Zhang, and Zeyu Fu. 2026. Training-free and interpretable hateful video detection via multi-stage adversarial reasoning. *arXiv preprint arXiv:2601.15115*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Lingjie Yi, Michael Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. 2025. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *International Conference on Learning Representations*.
- Jiangbei Yue, Shuonan Yang, Tailin Chen, Jianbo Jiao, and Zeyu Fu. 2025. Multimodal hate detection using dual-stream graph neural networks. In *British Machine Vision Conference*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Yinghui Zhang, Tailin Chen, Yuchen Zhang, and Zeyu Fu. 2024. Enhanced multimodal hate video detection via channel-wise and modality-wise fusion. In *2024 IEEE International Conference on Data Mining Workshops*, pages 183–190.

Yuchen Zhang, Yaxiong Wang, Kecheng Han, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. 2026a. Cultivating forensic reasoning for generalizable multimodal manipulation detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28545–28560.

Yuchen Zhang, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. 2026b. The coherence trap: When mllm-crafted narratives exploit manipulated visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8760–8769.

Wenjia Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. WhitenedCSE: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12135–12148.

A Definition of Hateful Video

Hateful video can be understood as a multimodal vehicle for the propagation of hate speech and prejudicial views. The precise definition of hate speech varies across platforms and cultural contexts, but can be coherently summarized from multiple sources (Das et al., 2023; Rehman et al., 2025; Wang et al., 2024a; Davidson et al., 2017) as *speech that promotes discrimination or hostility, or that disparages, abuses, or humiliates individuals or groups based on characteristics such as race, ethnicity, nationality, religion, disability, age, veteran status, sexual orientation, or gender identity*. Beyond the content-level characterization, scholars (Waldron, 2012; Tsesis, 2002) have also defined hate speech from a consequentialist perspective, emphasizing the tangible harms it inflicts on targeted groups’ dignity, social standing, and sense of belonging.

B Dataset Descriptions

We use four public datasets for evaluation, adhering to official splits where available, or otherwise

Table 4: Dataset statistics. Abbreviations: Norm. denotes Normal, Dist. denotes Distribution, Off. denotes Offensive, and Hat. denotes Hateful. Note that Offensive is mapped to Hateful in binary setting.

Dataset	Total	Split			Class Dist.		
		Train	Val	Test	Norm.	Off.	Hat.
HateMM	1,066	744	107	215	639	—	427
MHClip-Y	891	618	91	182	601	218	72
MHClip-B	897	633	88	176	605	180	112
ImpliHateVid	2,009	1,283	325	401	1,000	—	1,009

following the splits adopted by (Lang et al., 2025) for fair comparison. We note that MultiHateClip (Wang et al., 2024a) does not provide downloadable videos directly, but only raw metadata required to manually retrieve content from the source platform. However, due to platform content moderation policies, a significant portion of hateful videos have since been removed and are no longer accessible. This issue persists despite contacting the original authors. We summarize the detailed dataset statistics in Table 4, followed by detailed descriptions below.

HateMM (Das et al., 2023) is the first large-scale multimodal hateful video dataset. It comprises approximately 43 hours of videos collected from BitChute using hate lexicon-based keyword search, with 1,083 videos manually annotated as hate or non-hate. Each video is also annotated with frame-level rationale spans to explain labeling decisions. The dataset covers English-language content and employs a binary classification scheme.

MultiHateClip-YouTube (MHClip-Y) and MultiHateClip-Bilibili (MHClip-B) (Wang et al., 2024a) are two subsets of MultiHateClip, a multilingual benchmark introduced by Wang et al. at ACM Multimedia 2024. The original dataset contains 2,000 short video clips (1,000 per language) curated from YouTube (English) and Bilibili (Chinese) via gender-based hate lexicons, and annotated into three categories: Hateful, Offensive, and Normal. Annotations additionally identify hateful segments, target victims, and contributing modalities (text, visual, audio). The dataset provides a cross-cultural perspective on gender-based hate speech. In the main body of the paper, we evaluate this dataset under a binary setting by mapping both Offensive and Hateful labels to Hateful, following prior work (Jing et al., 2025; Lang et al., 2025). We additionally report

results for the original three-way classification setting in Table 5 to provide a more comprehensive benchmark comparison.

ImpliHateVid (Rehman et al., 2025) is a benchmark dataset specifically curated for implicit hate speech detection in videos, published at ACL 2025. It contains 2,009 videos annotated into three classes: Non-hate (1,000), Implicit Hate (509), and Explicit Hate (500), making it one of the first video datasets to explicitly distinguish implicit from explicit hateful content. The videos average approximately 2 minutes in length. Unlike prior datasets that primarily focus on explicit or overt hate, ImpliHateVid targets veiled, context-dependent expressions of hate that lack explicit slurs or insults, posing a substantially greater detection challenge.

C Baseline Descriptions

We compare HATELENS against three categories of baselines.

Unimodal Baselines.

- **BERT** (Devlin et al., 2019) is a bidirectional transformer pre-trained via masked language modeling. We fine-tune bert-base-uncased on the concatenation of the video title and ASR transcript.
- **ViViT** (Arnab et al., 2021) is a pure-transformer model for video classification that extracts spatio-temporal tokens from uniformly sampled frames (Feichtenhofer et al., 2019; Liu et al., 2022). We use the factorized encoder variant pre-trained on Kinetics-400 and fine-tune it on the visual stream of each dataset.
- **MFCC** (Davis and Mermelstein, 1980) is a classical audio feature based on Mel-frequency cepstral coefficients. We extract 40-dimensional MFCC features, compute mean and standard deviation pooling, and train an MLP classifier on the resulting representation.

Multimodal Fusion Baselines.

- **Pro-Cap** (Cao et al., 2023) prompts a frozen vision-language model (Radford et al., 2021; Li et al., 2022) with hateful-content-related questions in a zero-shot VQA manner and uses the answers as informative captions for a downstream text classifier. We adapt this

pipeline to video by applying it to a representative frame.

- **HateMM** (Das et al., 2023) encodes text with BERT, visual content with a 3D-CNN, and audio with a pre-trained audio model, then fuses the three streams via cross-modal attention.
- **MHCL** (Wang et al., 2024a) is the baseline introduced alongside the MultiHateClip dataset. It aligns all three modality representations through contrastive learning objectives for cross-lingual hate-video detection.
- **MoRE** (Lang et al., 2025) is a retrieval-augmented mixture-of-experts model that employs modality-specific experts and retrieves annotated neighbors using early-fusion representations with manually specified modality weights.
- **ImpliHateVid** (Rehman et al., 2025) proposes a two-stage contrastive learning framework for implicit hate video detection. The first stage aligns modality representations, and the second stage contrasts hateful with non-hateful examples to improve detection of subtle cues.
- **CMFusion** (Zhang et al., 2024) applies separate attention mechanisms along both the channel and modality dimensions for multimodal feature fusion in hate video detection.

MLLM-based Baselines.

- **MiniCPM-V** (Yao et al., 2024) is a series of efficient MLLMs designed for end-side deployment. We use the MiniCPM-Llama3-V-2.5 variant and apply it in a zero-shot prompting setting with a representative frame grid and video metadata.
- **LLaVA-OV** (Li et al., 2024) (LLaVA-OneVision) supports single-image, multi-image, and video understanding via a unified architecture. We prompt it in a zero-shot chain-of-thought setting to describe the content and produce a binary judgment.
- **Qwen2-VL** (Wang et al., 2024b) encodes images and videos at their native resolution via a Naive Dynamic Resolution module and Multimodal Rotary Position Embedding (M-RoPE). We evaluate it under the same zero-shot prompting setting as LLaVA-OV.

- **HVGuard** (Jing et al., 2025) prompts an MLLM with a describe-then-analyze chain-of-thought to generate rationales, and then fine-tunes a downstream fusion classifier on the rationales together with multimodal features.

All baselines strictly follow the released data splits from (Lang et al., 2025) and (Rehman et al., 2025) and adopt their optimal settings. For fair comparison, all MLLM-based baselines are evaluated under two settings: using the backbone specified in their original papers, and using the same backbone adopted in this work. In both cases, each method follows its optimal configuration.

D Implementation Details

D.1 Preprocessing

Following (Jing et al., 2025), we uniformly sample $n=32$ frames from each video and arrange them into eight 2×2 quad images (four frames per quad) in chronological order for MLLM input, which substantially reduces API cost compared to processing individual frames. The raw 32 frames are retained as input to the visual encoder. We apply Whisper (Radford et al., 2023) to obtain the full spoken transcript of each video.

D.2 LLM Configuration

We use GPT-5.4-nano as the backbone MLLM for the P2C Generator. The model receives eight 2×2 quad images, obtained by arranging the 32 uniformly sampled frames in chronological order with four frames per quad, together with the video title and ASR transcript. Both the `<think>` and `<answer>` stages are completed within a single prompt call, with structured JSON parsing applied to extract the six answer fields. The average cost per video is approximately \$0.002.

D.3 Encoder Backbones and Fusion Module

Text features are encoded with bert-base-uncased (Devlin et al., 2019), visual features with vit-base-patch16-224 (Dosovitskiy et al., 2021), and audio features with WavLM-Base+ (Chen et al., 2022). All three encoders are kept frozen throughout training. The Schema-guided Modality Router uses $H=4$ routing heads with hidden dimension $h=192$, followed by a two-layer MLP classifier with ReLU activations.

D.4 Training Details

All models are trained with AdamW ($\beta_1=0.9$, $\beta_2=0.999$, weight decay 10^{-4}) for up to 50 epochs with early stopping based on validation Macro-F1, using a cosine learning-rate schedule with linear warm-up over 5 epochs and peak learning rate 5×10^{-4} . Batch size is 32. For the Whitened Neighbor Interpolator, covariance whitening follows (Yi et al., 2025) with the shrinkage coefficient determined by the Ledoit-Wolf estimator.

D.5 P2C Prompt

The full prompt template used by the P2C Generator is shown in Figure 16. Notably, the definition of hateful content varies across datasets, which also reflects real-world moderation settings, where different platforms may adopt different definitions of what constitutes hateful content. For each benchmark, we therefore summarize the definition of hateful content and the annotation criteria described in the corresponding dataset paper (Das et al., 2023; Wang et al., 2024a; Rehman et al., 2025), and insert the resulting dataset-specific definition into the prompt. The prompt instructs the MLLM to reason through four structured steps enclosed in `<think>` tags, progressing from scene and context description to hateful cues description to target and attack identification to contextual harm and norm reasoning, before producing a six-field structured answer in `<answer>` tags.

E Additional Experiments

E.1 Three-Way Classification

Table 5 reports results under the full three-way classification setting (Normal / Offensive / Hateful) on MHClip-Y and MHClip-B. This is a strictly harder task than the binary setting in the main paper, as the model must additionally distinguish offensive content from hateful content, a boundary that is inherently subjective and often ambiguous. HATE-LENS achieves the best performance on all metrics across both datasets. The gains in M-P are particularly large, suggesting that both the structured P2C reasoning and the Whitened Neighbor Interpolator contribute to suppressing false positives: the former grounds predictions in an explicit interpretive schema about targets, intent, and contextual harm, while the latter retrieves annotated precedents that help resolve ambiguous cases near the decision boundary. The improvement in M-F1 over the best baseline is more modest, reflecting the inherent dif-

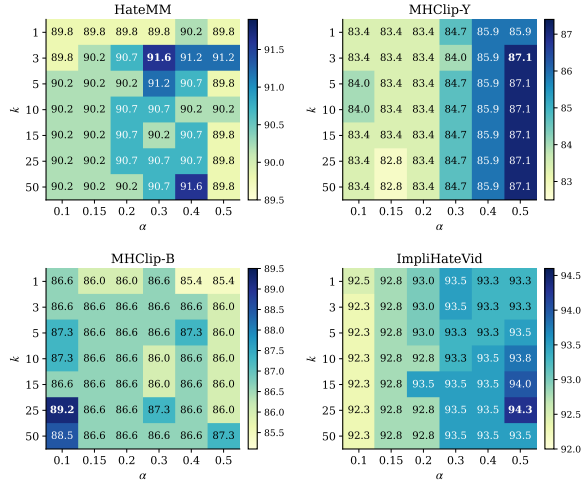


Figure 7: Sensitivity of test accuracy to the number of retrieved neighbors k and the interpolation weight α across all four datasets. Darker cells indicate higher accuracy.

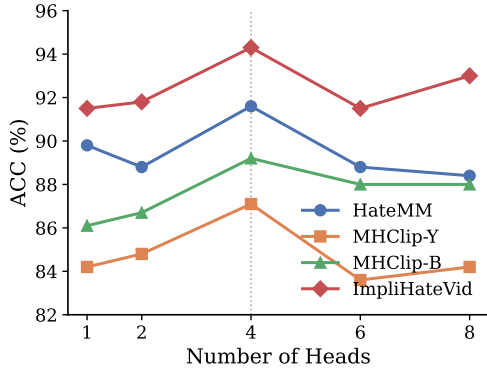


Figure 8: Effect of the number of routing heads in the Schema-guided Modality Router. The dashed vertical line marks the default setting of $H=4$.

ficulty of the offensive class, which is small and semantically close to both Normal and Hateful.

E.2 Hyperparameter Sensitivity

Interpolator. Figure 7 reports the accuracy of HATELENS as a function of the two key hyperparameters of the Whitened Neighbor Interpolator: the number of retrieved neighbors $k \in \{1, 3, 5, 10, 15, 25, 50\}$ and the interpolation weight $\alpha \in \{0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$. Across all four datasets the method is broadly stable. Performance improves as more neighbors are included, and is further boosted when α is set to a moderate value that balances the parametric classifier signal with the local retrieval evidence. The tuning of hyperparameters is searched on validation set and then evaluated on test set.

Number of Routing Heads. Figure 8 shows the

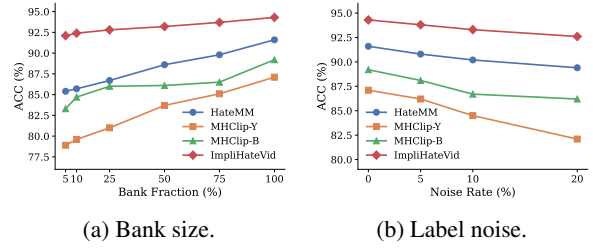


Figure 9: Retrieval robustness analysis.

effect of varying the number of routing heads $H \in \{1, 2, 4, 6, 8\}$ in the Schema-guided Modality Router. Performance consistently peaks at $H=4$ across all four datasets, suggesting that four heads provide sufficient capacity to capture diverse modality interactions without overfitting. Fewer heads underperform as a single routing perspective is insufficient to model the heterogeneous multimodal and schema-derived inputs, while more heads yield diminishing returns and slight degradation, likely due to increased optimization difficulty with a fixed hidden dimension.

E.3 Retrieval Robustness

Figure 9a varies the fraction of the training set used to populate the retrieval bank from 5% to 100%. Performance improves monotonically and plateaus around 75%–100%, confirming that the interpolator benefits from a larger and more diverse evidence pool. Notably, even with only 25% of the bank the model retains most of the improvement from full-bank retrieval, showing robustness to partial-bank settings.

Figure 9b evaluates robustness by randomly flipping a fraction of training-set labels (0%–20% noise rate) before populating the retrieval bank. Accuracy degrades gracefully: even at 20% label noise the model retains strong performance across all datasets. This is because the whitening step geometrically conditions the embedding space, which reduces the influence of outlier neighbors that are likely to correspond to mislabeled instances.

E.4 Router Weight Analysis

Figure 10 visualizes the learned routing weights across the $H=4$ attention heads for each of the eight modality inputs (Txt, Aud, Frm, What, Who, Whr, Why, How) on all four datasets. The weights are fairly distributed across modalities, indicating that all inputs contribute to the fused representation. On ImphiHateVid, frame features (Frm) and content-description schema fields (What) receive

Table 5: Three-way classification results (Normal / Offensive / Hateful) on MHClip-Y and MHClip-B. Best results in **bold**, second best underlined. ↑ indicates the improvement of HATELENS over the best baseline.

Method	MHClip-Y (3-way)				MHClip-B (3-way)			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
<i>Unimodal Methods</i>								
BERT	69.8	41.9	38.9	40.1	69.3	34.5	36.9	36.5
ViViT	73.0	40.5	57.9	38.3	70.1	27.5	49.5	34.2
MFCC	69.9	27.4	24.5	33.1	70.1	27.5	23.7	33.2
<i>Multimodal Fusion Methods</i>								
Pro-Cap	<u>73.6</u>	46.9	58.4	44.1	68.8	43.5	48.9	41.7
HateMM	73.0	47.0	59.1	46.6	70.1	44.7	48.6	43.4
MHCL	<u>73.6</u>	<u>53.9</u>	51.8	<u>47.8</u>	72.6	<u>51.6</u>	49.7	46.5
MoRE	71.8	39.7	41.6	40.6	72.0	33.7	44.1	36.7
CMFusion	<u>73.6</u>	47.9	59.9	44.8	71.3	47.1	<u>50.5</u>	43.0
ImpliHateVid	69.9	27.4	23.3	33.3	<u>73.3</u>	45.9	<u>42.8</u>	<u>50.2</u>
<i>MLLM-based Methods</i>								
HVGuard	<u>73.6</u>	41.8	<u>66.3</u>	41.2	71.3	36.7	47.3	38.0
HATELENS	79.1	57.0	86.2	52.4	77.1	58.2	63.5	55.5
↑ Improv.	+5.5	+3.1	+19.9	+4.6	+3.8	+6.6	+13.0	+5.3

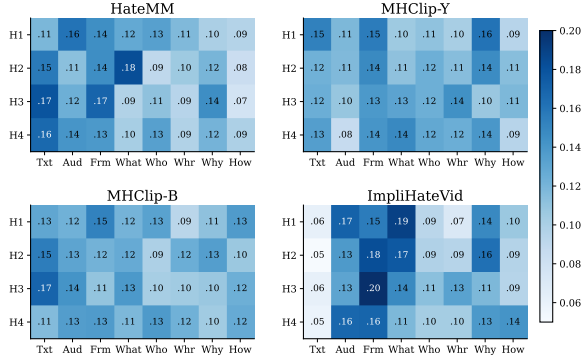


Figure 10: Mean routing weights per head and modality input on the four test sets.

notably higher average weight, consistent with the visually implicit nature of that dataset. Across datasets, the schema-derived fields (What, Who, Whr, Why, How) collectively account for roughly 35–45% of the total routing mass, confirming that the P2C reasoning schema provides complementary discriminative signal beyond the raw multimodal features.

E.5 Full Cross-Dataset Transfer Results

Figure 11 presents the remaining eight source→target directions of the cross-dataset transfer experiment, complementing the four directions in Figure 5 and together covering all twelve pairs. The Whitened Neighbor Interpolator (w/ WNI) consistently improves over the no-retrieval baseline (w/o WNI) across all methods

and directions. HATELENS achieves the best or second-best result in all twelve settings, with the largest gains on directions involving ImpliHateVid as the target domain, where the implicit nature of the hate expressions makes retrieval-based disambiguation most valuable.

Table 6: GPT-4o judge prompt for explanation quality evaluation.

System: You are an expert annotator evaluating the quality of machine-generated explanations for hateful video detection.

User: Given the following video metadata, frames and transcript, and the reasoning produced by a detection system, please rate the explanation on five criteria, each on a scale of 1 (very poor) to 5 (excellent):

- 1. Informativeness:** Does the explanation provide background, context, or cultural knowledge beyond surface-level description?
- 2. Soundness:** Is the reasoning logically coherent and free of contradictions?
- 3. Persuasiveness:** Is the conclusion convincing given the provided evidence?
- 4. Readability:** Does the text follow proper grammar and sentence structure?
- 5. Fluency:** Does the explanation flow smoothly with well-connected ideas?

Return your ratings as a JSON object:
{"informativeness": X, "soundness": X,
"persuasiveness": X, "readability": X,
"fluency": X}

E.6 Feature Visualization

Figure 12 visualizes the fused representations after the Schema-guided Modality Router via t-SNE.

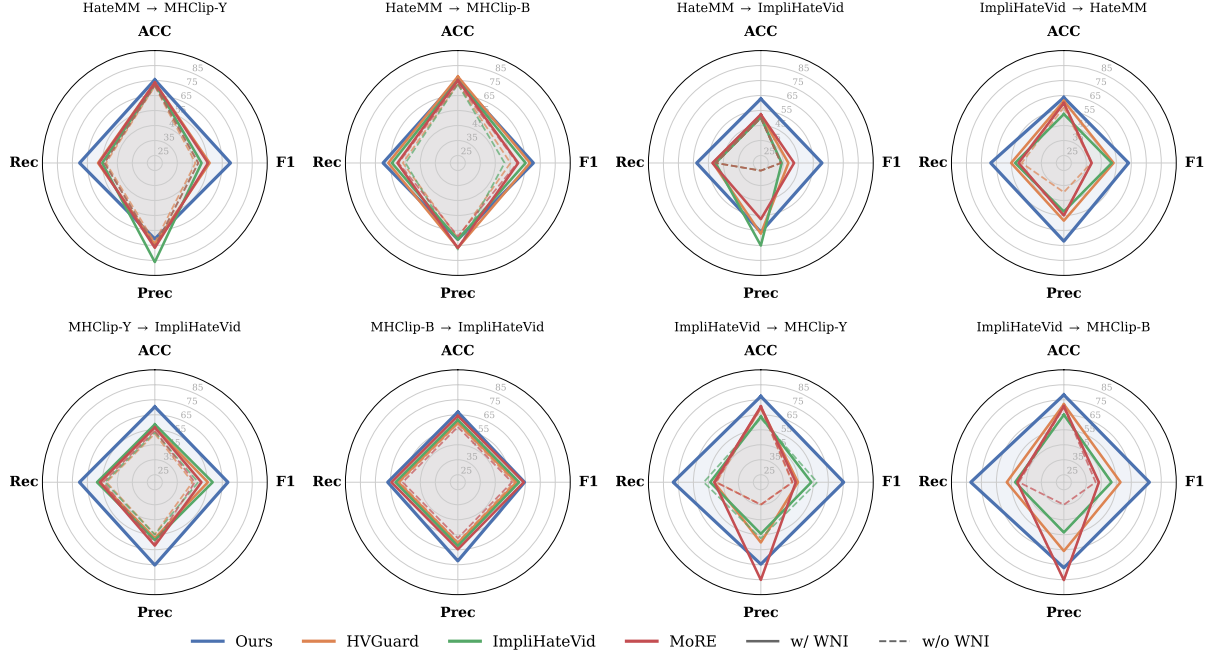


Figure 11: Cross-dataset transfer results for the remaining eight source→target pairs. Radar plots show ACC, M-F1, M-P, and M-R. Colors denote methods: HATELENS (blue), HVGuard (orange), ImpliHateVid (green), and MoRE (red); solid and dashed lines denote each method with (w/ WNI) and without (w/o WNI) our Whiten Neighbor Interpolator.

Across all four datasets, HATELENS produces markedly better-separated clusters than the baselines, with Normal and Hateful samples forming more compact and distinct groups. This separation stems from the P2C Generator, which supplies a structured interpretive schema about targets, intent, and contextual harm; the Schema-guided Modality Router then models the interaction between this schema and the raw multimodal features in the latent space, yielding representations that reflect the semantic structure of hateful content rather than surface-level feature correlations. Notably, the advantage is most pronounced on MHClip-B and ImpliHateVid, where implicit and diverse expressions render competing methods’ clusters heavily entangled. The already-discriminative fused representations shown here are further geometry-conditioned by the whitening step in the Whiten Neighbor Interpolator, which corrects the anisotropy introduced by heterogeneous feature sources (Gao et al., 2019; Zhuo et al., 2023) and ensures that similarity-based retrieval operates in a metrically meaningful space, enabling the kNN search to identify semantically relevant precedents rather than spurious nearest neighbors, and ultimately contributing to the significant performance gains reported in Table 1.

E.7 Modality Robustness

Figures 13–15 evaluate robustness by randomly corrupting one modality (Text, Audio, or Visual) at rates of 25%–100% under three settings: train-time only, test-time only, and both simultaneously. Specifically, text corruption replaces sample tokens with random words, audio corruption substitutes the waveform of sampled clips with Gaussian noise, and visual corruption randomly masks or shuffles sampled frames. HATELENS consistently maintains the highest accuracy across all corruption settings, modalities, and datasets, with a notably flatter degradation curve compared to all baselines. This robustness stems from two complementary components. The P2C Generator produces a structured interpretive schema grounded in explicit reasoning about targets, intent, and contextual harm, which means HATELENS does not over-rely on any single raw modality stream; even when one modality is heavily corrupted, this schema provides a stable signal that the Schema-guided Modality Router can leverage to compensate, a property that pure feature-fusion baselines fundamentally lack. The margin over MLLM-based baselines is particularly large under text corruption, where competing methods degrade more sharply because they rely on surface-level textual analysis rather than higher-

level semantic judgments. Beyond the parametric model, the Whitened Neighbor Interpolator provides an additional layer of robustness at test time: by consulting geometrically well-conditioned nearest neighbors from the training bank, it can correct uncertain predictions even when the input representation is partially distorted by corruption, further stabilizing performance at high corruption rates.

E.8 More Case Studies

Figures 17 and 18 present two representative cases where HATELENS succeeds while the strongest baseline fails. Figures 19 and 20 examine error cases of our full model.

Comparison with Baseline. In Case 1 (Figure 17), the baseline’s visual and textual analyses identify violent wording and shooting-related visuals, yet its final MLLM judgment interprets these observations as gaming-style dark humor and predicts Normal. Our method instead connects the transcript’s celebratory killing language with the visual gunfire, identifies the attacked group, and interprets the speaker’s stance as endorsement of targeted lethal violence. In Case 2 (Figure 18), the baseline similarly recognizes media manipulation and mind-control rhetoric, but its final judgment treats them as generic anti-media commentary. Our structured interpretation identifies the target as “the media/TV” and implied elites, grounds the evidence in repeated subtitles and transcript cues, and explains why portraying this target as deceptive and mind-controlling constitutes hostile delegitimization. These cases show that the key difference is not cue perception, but the mechanism used to interpret those cues. The baseline first performs modality-wise analysis and then passes the resulting visual and textual evidence to an MLLM for free-form hate judgment, leaving the model to decide on its own whether the observations jointly support a hateful interpretation. In contrast, our P2C Generator forces the model to organize the same observations into structured interpretive fields before prediction, explicitly specifying where the evidence appears, what hostile content is expressed, who is targeted, why the content is hateful, and how the framing may cause harm. This structured interpretation links evidence, target, stance, and implication into decision-relevant semantic relations, reducing the chance that suspicious cues are absorbed into benign surface frames such as gaming humor or general conspiracy commentary.

Error Case Analysis. For this diagnostic analysis, we report kNN votes over an enlarged neighborhood of $k=40$ rather than the per-dataset k used for prediction, which provides a more complete and balanced view of the local precedent distribution around each error case. Figure 19 shows two false positive cases whose benchmark labels are Normal. In the first case, the transcript contains repeated racial slurs, dehumanizing language, and an explicit exclusionary plan, while the visual frame also contains racially loaded signage. After manual inspection, we find that this sample is highly harmful despite its Normal label, suggesting a possible annotation inconsistency in the benchmark. The strong kNN vote (23:17) is also consistent with this observation. The second case presents a challenging reporting-versus-endorsement distinction: the speaker discusses antisemitic tropes such as “Jews run the banks” and “Jews dominate the media” in an ostensibly analytical context, making the boundary between neutral discussion and propagation of hateful claims difficult to determine. The near-balanced kNN vote (21:19) further reflects this uncertainty.

Figure 20 shows two false negative cases where the hateful implication is subtle and context-dependent. The first case hinges on a single ambiguous utterance (“*up yours, Diger?*”), whose harmful implication is difficult to decode without knowing the intended reference and cultural context. As a result, the structured interpretation treats it as a minor personal remark rather than group-directed hate. The second case is a single-sentence microaggression (“*why didn’t you stay in Mexico*”) whose xenophobic implication is expressed very indirectly. Apart from this sentence, the remaining visual frames and transcript appear largely normal, making the hateful signal sparse and easy for an MLLM to miss. The overwhelmingly Normal kNN vote (34:6) suggests that similar implicit attacks are also underrepresented in the retrieval set. Overall, these cases show that the remaining errors mainly arise from challenging annotation boundaries, reporting-versus-endorsement distinctions, coded language, and highly implicit microaggressions.

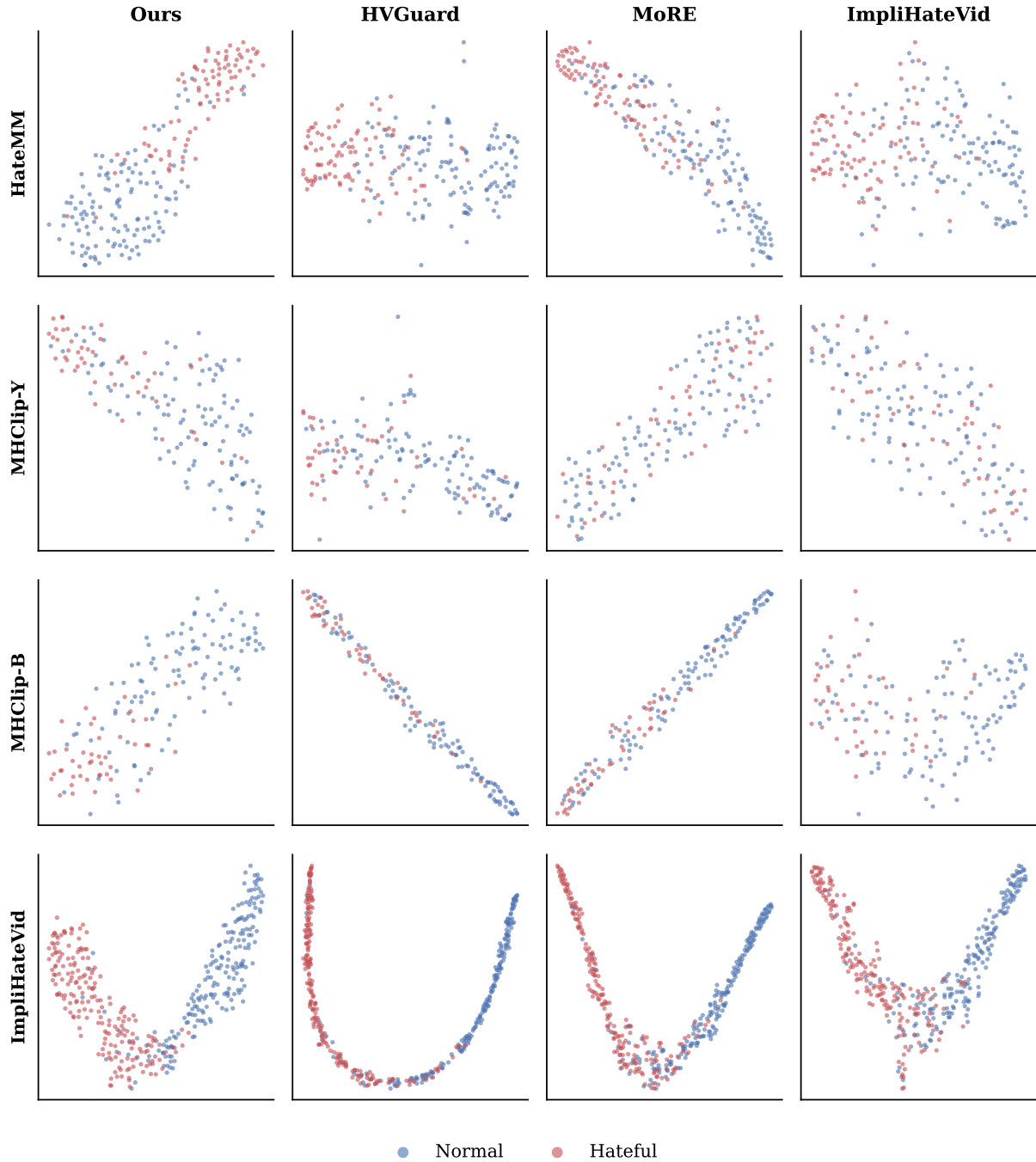


Figure 12: t-SNE visualization of learned representations on the test sets of all four datasets. Rows correspond to datasets; columns correspond to methods. Blue and red dots denote Normal and Hateful samples respectively.

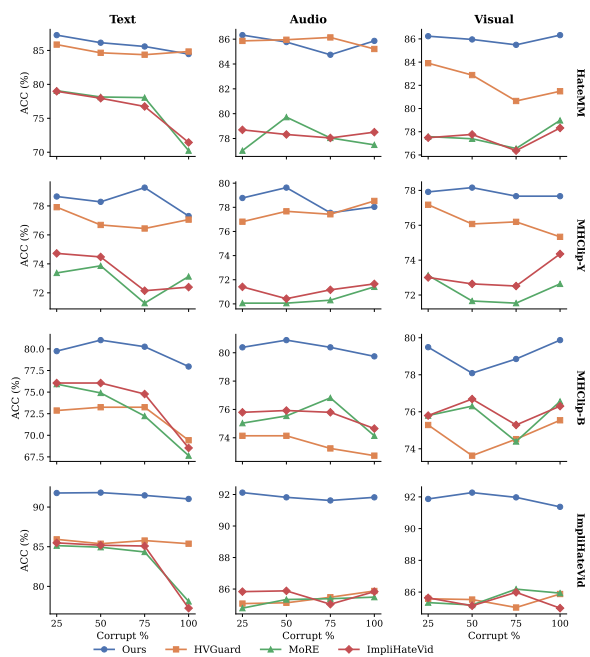


Figure 13: Modality robustness under train-time corruption. Each column corrupts one modality (Text / Audio / Visual) at increasing rates; rows correspond to datasets.

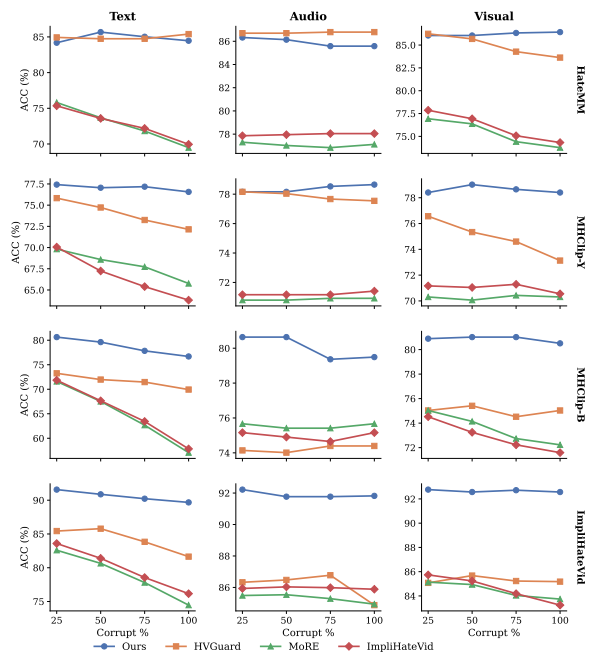


Figure 14: Modality robustness under test-time corruption.

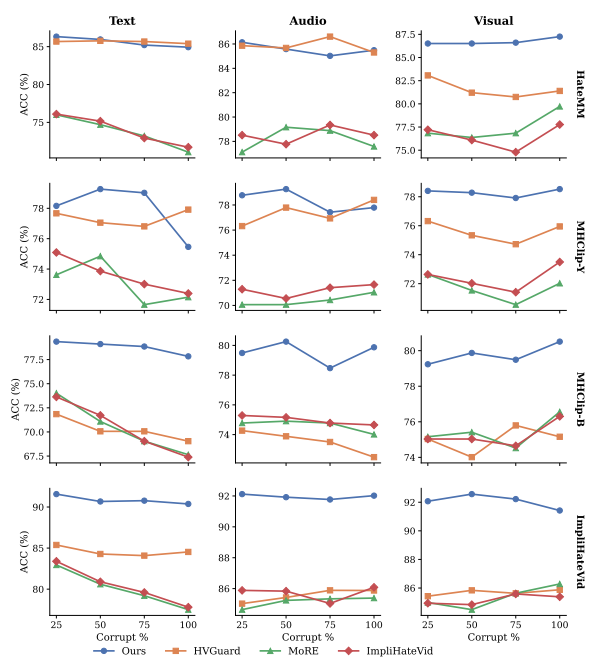


Figure 15: Modality robustness under train-and-test-time corruption.

P2C Generator Prompt Template

System: You are a content moderation expert. You analyze videos represented by multiple 2×2 quad images, where each quad contains four consecutive frames in temporal order.

User: Your task is to determine whether the given video should be classified as Hateful or Normal. A video is Hateful when {benchmark-specific definition}. Consider all available signals in the video, including visual scenes, on-screen text, subtitles, spoken language, gestures, and symbols. Generate two components in order:

1. A structured reasoning process enclosed in <think></think> tags.
2. A final explanation enclosed in <answer></answer> tags.

Thinking <think>: For all videos, the reasoning should follow this structured 4-step process enclosed in corresponding tags:

<step1> *Scene and Context Description*: Summarize the visible scene, participants, actions, on-screen text, and spoken or subtitle content.

<step2> *Hateful Cues Description*: Describe specific abusive, offensive, or hostile cues in the video, if any, and note where they appear.

<step3> *Target and Attack Identification*: Identify who or what is targeted and the type of attack involved. Distinguish endorsement from quotation, criticism, or neutral discussion.

<step4> *Contextual Harm and Norm Reasoning*: Explain why the content should be considered hateful in context, based on severity, target, intent, and norm violations.

Answer <answer>: Write a single coherent paragraph containing exactly the following tags in order:

<which> Hateful or Normal
<what> content description
<who> who is targeted
<where> key evidence location
<why> classification reason
<how> potential consequence

Input: Video Title: {title} Video Transcript: {transcript} Video Frame: {frames}

Figure 16: P2C Generator prompt template.



Transcript:

On my way to the clan rally to celebrate my highest kill tally, my GPS miscalculated directions into the wrong neighborhood. I glanced over and there they stood — a group of emaciated men. I walked behind one and blew out his brains, trampled over the dead remains and fired into the backs of the others. Despite not arriving at the rally on time, it was splendid. Next time I'll get here a bit earlier. San Bernardino will look like child's play after the moon rises.

HVGuard LLM Output — Wrong (Normal)

Text Analysis:

The transcript appears to be a distorted, possibly humorous or satirical narrative that uses dark imagery, slang, and violent wordplay. It interprets cues like "clan rally," "kill tally," "ammo," and "killing street" as part of a gaming/combat-style scenario rather than direct hateful endorsement. Homophonic puns and euphemisms such as "pack of flags," "sending lead flags," and "killing street" are treated as wordplay rather than targeted slurs. The references to "emaciated men" and "underweight" are read as character descriptions in a fictional scenario rather than dehumanizing language toward a real group.

Visual Analysis:

The video depicts a scene viewed through the reflection in a person's sunglasses, showing two men standing next to a campfire or an outdoor cooking setup with nearby containers. One man is shirtless; the other wears a brown shirt with white pants. There is visible movement, smoke, and sparks from the fire, but no clear special symbols, hate markers, or readable text in the visual content. The scene appears to be a casual outdoor gathering with no indicators of hatred or targeted hostility.

Final Judgment:

The content portrays murder and chaos through euphemism and wordplay, mimicking gaming or military scenarios in a darkly humorous tone. It does not promote real violence or hatred toward any specific group. The tone is read as satirical rather than genuinely threatening, and no explicit hate markers are identified in either the visual or textual content. The overall narrative is interpreted as dark entertainment rather than hateful endorsement. Therefore, the prediction is **Normal**.

Our P2C CoT Output — Correct (Hateful)

Step 1 (Scene Description):

The frames show an animated first-person shooter scene with weapons, ammunition, explosions, and gunfire, while the transcript repeatedly describes shooting and killing a group as celebratory violence.

Step 2 (Hateful Cue Identification):

The transcript explicitly celebrates mass violence and uses dehumanizing, hostile framing toward a targeted group, reinforced by visuals of active shooting and explosions.

Step 3 (Target Recognition):

The target is a specific group implied by the "clan rally" and "pack of flags," and the attack is direct physical violence presented with enjoyment and endorsement.

Step 4 (Classification Reasoning):

Because the video endorses targeted killing of a group through celebratory framing and graphic violence, it meets the criteria for Hateful content.

Answer:

<which> Hateful

<what> The video depicts an animated shooting attack; the transcript celebrates targeted killing using dehumanizing language and graphic violence.

<who> A specific group associated with the "clan rally" and "pack of flags."

<where> Foreground gunfire in the frames and the transcript describing the "killing street."

<why> The message endorses targeted lethal violence ("celebrate my highest kill tally," "wipe out"), not condemnation or neutral reporting.

<how> By narrating celebratory violence toward a targeted group, it normalizes hostility and encourages real-world harm.

Figure 17: Case study 1: A video celebrating targeted lethal violence against a specific group.

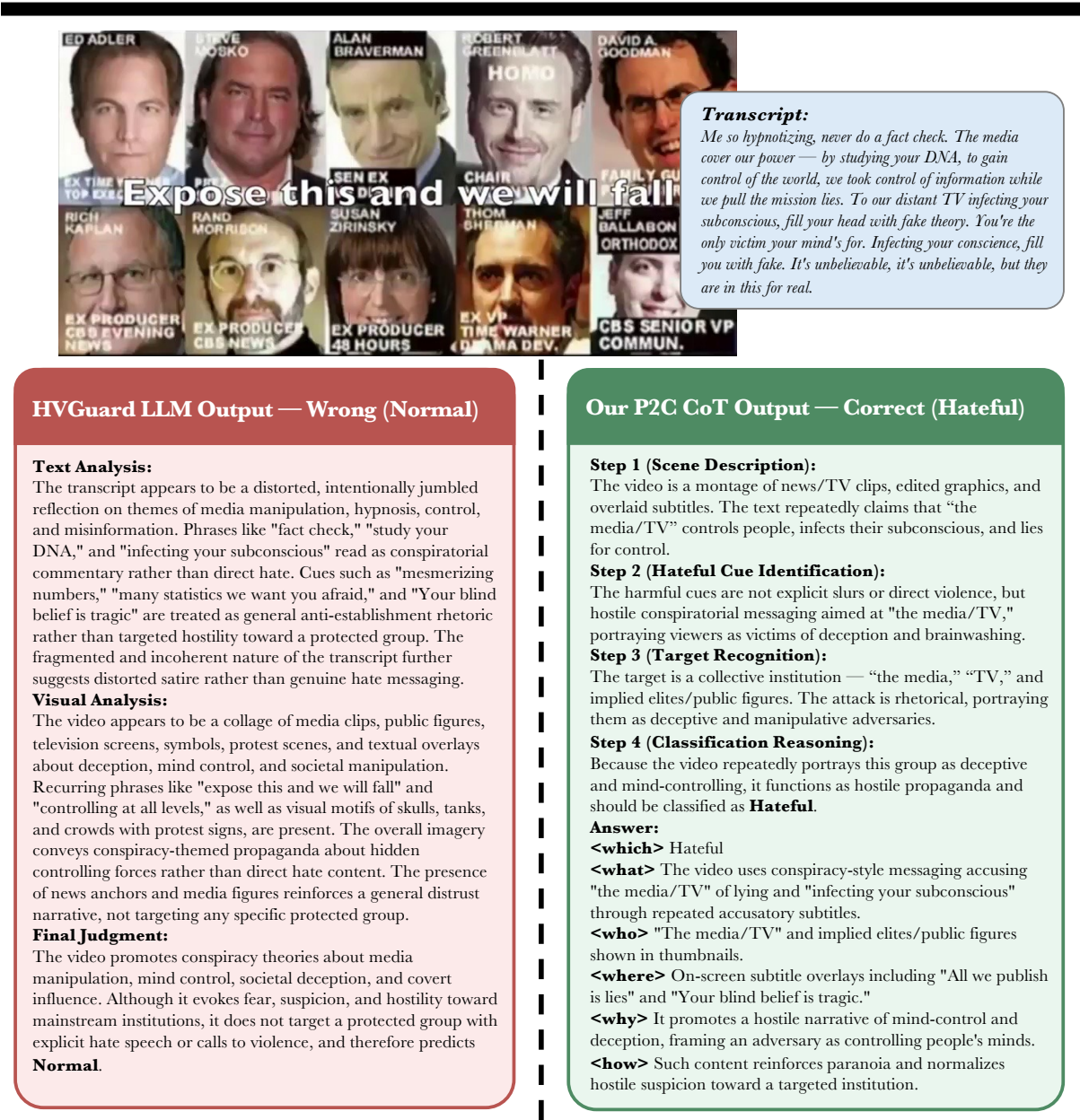


Figure 18: Case study 2: A video framing media institutions as deceptive mind-controlling entities.

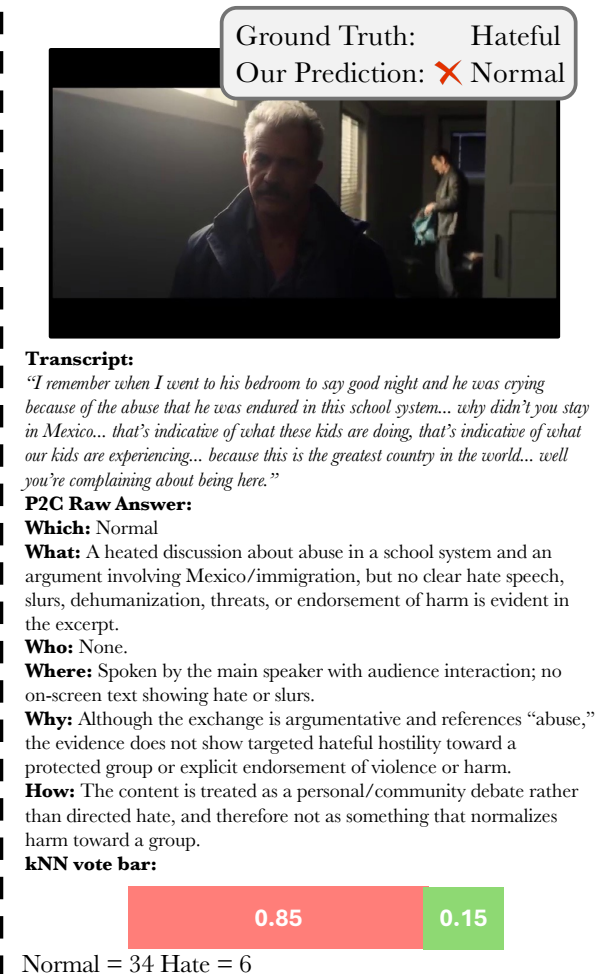
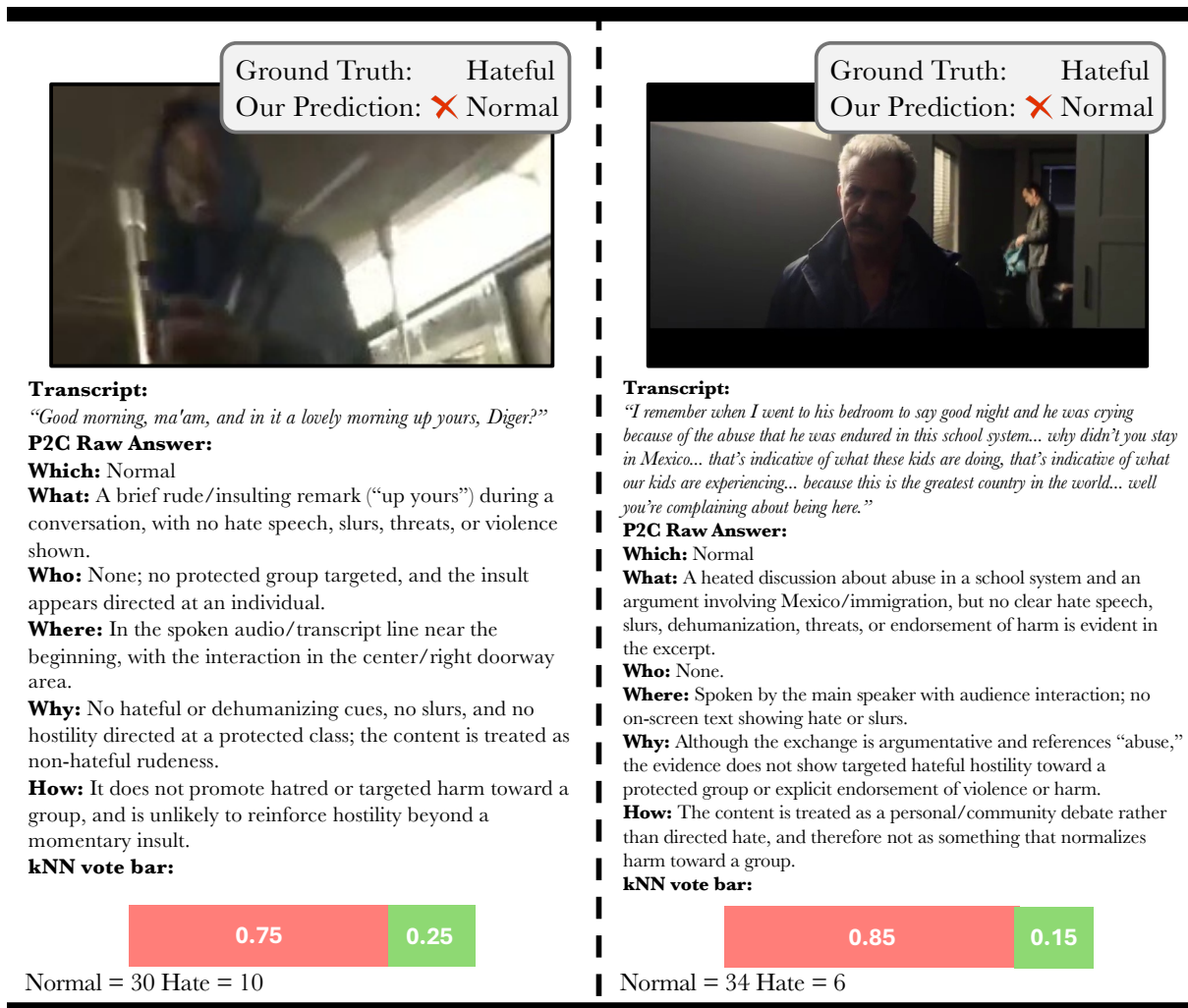


Figure 20: False negative error cases.