LEAF: Towards Lightweight Explainable Hateful Video Detection via Self-Grounding CoT Guided Stage-Wise Distillation

Jian Lang[1], Rongpei Hong[1,2], Meihui Zhong[1], Kaiju Li[1], Ting Zhong[1], Qiang Gao[3], Fan Zhou[1,4,*],

[1]University of Electronic Science and Technology of China,
[2] Sichuan Youjianzhihui Technology Co., Ltd.,
[3]Southwestern University of Finance and Economics,
[4] Intelligent Digital Media Technology Key Laboratory of Sichuan Province,
jian_lang@std.uestc.edu.cn, fan.zhou@uestc.edu.cn *

Abstract

The rapid spread of hateful videos online has sparked growing social concerns, driving research efforts to detect and limit their dissemination. However, existing methods rely on opaque models that offer no insight into their decisions, eroding trust in detection systems. Large Multimodal Models (LMMs) provide a compelling alternative, thanks to their ability to generate free-text explanations for multimodal content. Yet, their high computational demands and pronounced bias toward benign predictions limit their practicality. We introduce **LEAF**, the first **L**ightweight, **E**xplainable h**A**teful video detection **F**ramework. At its core, LEAF distills the "explainability" from LMMs into efficient Smaller Multimodal Models (SMMs) through a controlled, de-biasing process, enabling lightweight yet interpretable Hateful Video Detection (HVD). We achieve this with a novel *Self-Grounding Chain-of-Thought mechanism* that guides LMMs to generate high-quality, unbiased explanatory supervision signals for videos. These signals then progressively train the SMM via a new *Stage-Wise Distillation paradigm*, resulting in faithful, human-readable natural language explanations for HVD. Extensive experiments on three video benchmarks demonstrate that LEAF not only outperforms prior methods in detection accuracy but also provides strong explainability — all with a lightweight design.

1 Introduction

The rise of video-sharing platforms like YouTube and TikTok has made video consumption a dominant form of online media (Lang et al., 2026b). Unfortunately, this surge also accelerates the dissemination of hateful content through these platforms, which targets race, gender, and religion, exacerbating severe social instability (Das et al., 2023; Hebert et al., 2024; Lang et al., 2026a). Consequently, developing robust methods for Hateful Video Detection (HVD) has become a critical research priority with significant implications.

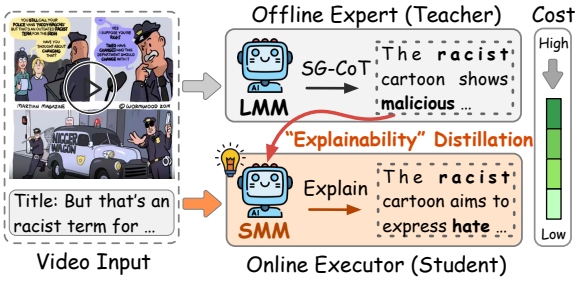


Figure 1: Concept diagram of our proposed LEAF. Explainability is distilled from LMM (**Teacher**) to SMM (**Student**) for lightweight yet interpretable HVD.

Existing HVD methods primarily extract and fuse shallow multimodal features from videos for prediction (Wang et al., 2024b; Lang et al., 2025; Hong et al., 2025b; Li et al., 2026). For instance, MHCL (Wang et al., 2024b) leverages pre-trained vision-language models to extract features from the videos and performs detection on these features. Despite their encouraging performance, current approaches lack *transparency*. They operate as black boxes, offering no clear evidence to justify their classifications (e.g., highlighting specific visual, audio, or textual elements indicative of hate). This opacity hinders content moderators and viewers from understanding the rationale behind decisions, ultimately undermining trust in the systems.

Explainable HVD fundamentally requires two abilities: (1) *understanding* deep multimodal semantics and subtle hateful cues in videos, often drawing on contextual or background knowledge (e.g., recognizing a neutral-seeming gesture as derogatory in certain cultures), and (2) *explaining* prediction rationales in human-understandable terms. Advanced Large Multimodal Models (LMMs) such as GPT-4o and Grok-4 excel at this, thanks to their vast pre-trained knowledge and skill in generating free-text explanations (Huang et al., 2024; Nguyen et al., 2024). *However*, their mas-

*corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 12420–12438

sive computational needs make them impractical for low-latency platforms. *Moreover*, since most LMMs are trained on benign data (Dubey et al., 2024; Bai et al., 2025), they exhibit a strong bias toward non-hateful predictions. Our empirical analysis shows that 54.4% and 56.9% of videos predicted as "benign" by LMMs are actually hateful on the MHClip-Y and MHClip-B datasets, respectively. This bias worsens when LMMs are tasked with both predicting and explaining: 81.0% and 80.4% of "benign" videos predictions on those datasets are hateful (see our Preliminary Experiment). As prior work notes (Guo et al., 2024; Liu et al., 2024), such bias often yields unfaithful explanations that mismatch or ignore the prediction — e.g., claiming hate without citing video-specific cues. Fine-tuning LMMs on task data could help, but it's computationally prohibitive (Mei et al., 2025).

In contrast, Small Multimodal Models (SMMs) like Gemma3-4B (Team et al., 2025) have substantially fewer parameters and lower overhead, making them ideal for fine-tuning and real-time deployment on video platforms (Marafioti et al., 2025). They can interpret simple multimodal inputs in natural language, *but* their limited capacity hinders handling complex hateful video semantics or producing reliable, human-aligned explanations.

In light of the above insights, we *combine the strengths of LMMs and SMMs to overcome their weaknesses*, proposing **LEAF**: the first **L**ightweight **E**xplainable h**A**teful video detection **F**ramework via Self-Grounding Chain-of-Thought (CoT (Wei et al., 2022)) Guided Stage-Wise Distillation. As illustrated in Figure 1, a large-scale LMM serves as an offline expert, generating high-quality explanatory supervision signals to train a lightweight SMM as the online executor, enabling strong explainability in HVD.

Specifically, to instill the explainability in the SMM, we introduce a novel three-step Self-Grounding CoT (SG-CoT) mechanism that prompts the LMM to create high-quality, unbiased supervision signals for videos. In the **Reason** step, the LMM analyzes multimodal content, supplying rich semantics and context to identify potential harmful elements. The **Explain** step then produces clear natural language explanations tied to this reasoning, plus a prediction, without conditioning on ground-truth labels to avoid *label-driven rationalization* (Turpin et al., 2023) and ensure the explanation reflects the model's genuine reasoning. Finally, the **Ground** step counters benign bias by comparing the prediction to the ground-truth label; if they match, it further utilizes the LMM's verification strengths (Pang et al., 2023) to self-assess whether the explanation adequately supports the prediction by grounding it in the video, yielding faithful rationales as teaching signals. Moreover, based on these fine-grained supervision signals, we design a new *Stage-Wise Distillation paradigm* to progressively train the SMM. Specifically, in **Understand Stage**, using Reason outputs, the SMM learns to detect nuanced hateful cues by absorbing distilled semantic and contextual knowledge from the LMM. In **Explain Stage**, leveraging Explain and Ground outputs, a multi-task distillation strategy guides the SMM to generate faithful, human-aligned explanations while predicting labels — boosting detection accuracy and ensuring explanation-prediction consistency. Our contributions are listed as follows:

- We present LEAF, the first lightweight explainable framework for hateful video detection, which pioneers in distilling explainability from LMMs into lightweight SMMs to provide faithful, human-readable explanations for HVD.
- We introduce a novel self-grounding CoT guided stage-wise distillation paradigm that extracts reliable, unbiased supervision signals from LMMs to enhance SMMs' explainability in HVD.
- Extensive experiments on three video datasets show LEAF achieves state-of-the-art detection performance and superior explainability with minimal computational costs.

Code and data are at `https://github.com/Jian-Lang/LEAF`.

2 Related Work

Hateful Video Detection. Hateful Video Detection (HVD) aims to identify harmful content in videos by analyzing their multimodal information. Recent studies employ deep learning methods and utilize multiple modalities to effectively detect hateful videos (Das et al., 2023; Wang et al., 2024b; Lang et al., 2025; Wang et al., 2025). For instance, MoRE (Lang et al., 2025) proposed a mixture-of-experts (Eigen et al., 2013) network that dynamically assigns sample-specific modality contributions for improved detection. *Despite* encouraging performance, existing methods perform black-box detection, which hinders users from understanding the prediction rationale and weakens credibility of the detection systems. To address this issue, we

propose LEAF, the first lightweight explainable framework for HVD, which distills the explainability of LMMs into SMMs to facilitate interpretable yet efficient video-based hateful detection.

Knowledge Distillation from LLM. Recent advances in prompting strategies, such as CoT reasoning (Wei et al., 2022), have significantly enhanced the reasoning capabilities of LLMs and LMMs by encouraging them to decompose complex problems into natural language reasoning steps (Wang et al., 2023; Zhang and Zhang, 2025), e.g., OpenAI o3 (OpenAI, 2025). Motivated by this, recent work explored transferring such strong reasoning abilities into SLMs via rationale-based distillation, where rationales from LLMs serve as supervision signals to guide the SLMs in mimicking their reasoning process (Hsieh et al., 2023; Zhuang et al., 2025). *Building upon these studies*, we further steer these natural language reasoning steps to human-aligned explanation supervision signals in HVD.

3 Methodology

HVD Task Definition. We define a hateful video detection dataset: $\mathcal{D} = \{\mathcal{S}_1, \cdots, \mathcal{S}_M\}$, where M is the number of videos. Each video $\mathcal{S}_i$ contains its textual, visual, and audio modalities: $\mathcal{S}_i = \{\mathcal{T}_i, \mathcal{V}_i, \mathcal{A}_i\}$, along with a ground-truth label $\hat{y}_i$. HVD aims to judge whether $\mathcal{S}_i$ is *hateful* or *benign* by considering its modalities: $f(\mathcal{T}_i, \mathcal{V}_i, \mathcal{A}_i) \rightarrow y_i$. To address the black-box limitation of prior methods, we extend the definition of HVD by introducing an explainability requirement, where the model is expected to justify its predictions in a human-aligned manner: $f(\mathcal{T}_i, \mathcal{V}_i, \mathcal{A}_i) \rightarrow (\mathcal{E}_i, y_i)$, where $\mathcal{E}_i$ is the explanation for prediction y_i.

Methodology Structure. We first present the video preprocessing pipeline. We then introduce our Self-Grounding CoT mechanism for explanation data annotation, followed by a Stage-Wise Distillation paradigm that supervises SMMs to acquire interpretability. Finally, we provide a complete inference process of the proposed LEAF. The overall framework of LEAF is presented in Figure 2.

3.1 Video Data Preprocessing

We convert each video $\mathcal{S}_i$ into an input format compatible with mainstream SMMs and LMMs. For visual modality, we uniformly sample a few frames from $\mathcal{S}_i$ and concatenate them to obtain the visual input $\mathcal{V}_i$. For textual modality, we combine the video title with on-screen text extracted from the sampled frames to construct the textual input $\mathcal{T}_i$. For audio modality, we transcribe the audio into text to form the audio input $\mathcal{A}_i$.

3.2 Self-Grounding CoT Mechanism

Endowing lightweight SMMs with explainable capabilities in HVD requires first fostering their understanding of complex multimodal semantics (i.e., subtle hateful cues) in videos, followed by guiding them to generate clear and human-aligned natural language explanations for their predictions. A naive approach is to follow prior work by prompting an LMM to directly process videos and generate corresponding annotated supervision signals (Hsieh et al., 2023). *However*, the benign prediction bias (Dubey et al., 2024; Bai et al., 2025) in current LMMs undermines the quality of the supervision data for explainability in HVD, as LMMs frequently misclassify hateful videos as benign and generate rationales that are either contradictory to or irrelevant to their own predictions. To tackle this issue, we propose a novel Self-Grounding CoT (SG-CoT) mechanism. As presented in Figure 2, SG-CoT facilitates LMMs to produce reliable and unbiased explanation-oriented supervision signals via a three-step multimodal CoT reasoning.

Step1: Reason. We first guide the LMM to **Reason** over the video sample $\mathcal{S}_i$, yielding rich semantic and contextual knowledge for objects, people or other elements that aids the understanding of potential hateful cues within $\mathcal{S}_i$ (e.g., a historical symbol which conveys hate when used in a specific context). This process can be written as:

$$\mathcal{R}_i = \Phi_{\text{LMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i], \mathcal{P}_{\text{reason}}), \quad (1)$$

where Φ_{LMM} denotes the LMM, $\mathcal{R}_i$ is the reasoning output, $\mathcal{P}_{\text{reason}}$ is the prompt for the Reason step.

Step2: Explain. Based on the reasoning output, we instruct the LMM to **Explain** its rationale for determining whether the video $\mathcal{S}_i$ is hateful or benign in clear natural language, by considering both the video input and the reasoning context. We do not provide the ground-truth labels to the LMM in this stage to avoid label-driven rationalization (Turpin et al., 2023) and ensure the explanation reflects its intrinsic reasoning process:

$$(\tilde{\mathcal{E}}_i, y_i) = \Phi_{\text{LMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i; \mathcal{R}_i], \mathcal{P}_{\text{explain}}), \quad (2)$$

where $\tilde{\mathcal{E}}_i$ is the vanilla explanation derived from LMM for its prediction y_i on video $\mathcal{S}_i$, $\mathcal{P}_{\text{explain}}$ denotes the prompt designed for the Explain step.

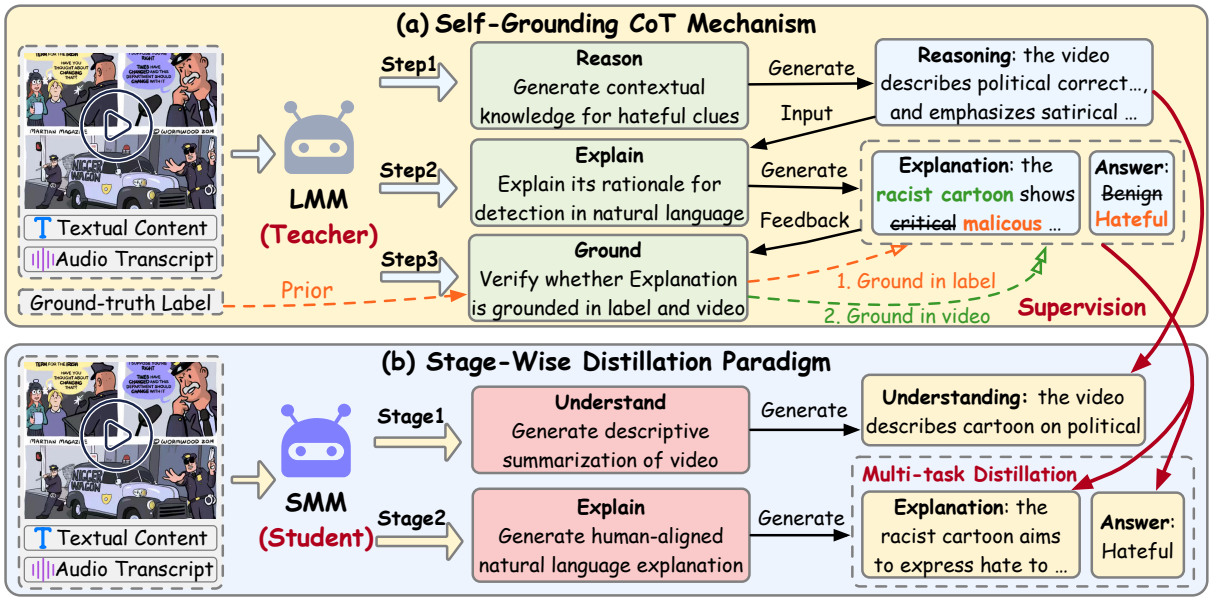


Figure 2: Overall framework of LEAF. (a) The LMM serves as the teacher and is guided by a three-step SG-CoT mechanism to produce high-quality and unbiased explanation supervisions. (b) The efficient SMM acts as the student, progressively acquiring explainability for HVD from the LMM through a Stage-Wise Distillation paradigm.

Step3: Ground. To mitigate the benign prediction bias in LMMs and guarantee high-quality explanation data annotation, we prompt the LMM to **Ground** its vanilla explanation. Specifically, we first verify the LMM's prediction for $\mathcal{S}_i$ against the ground-truth label to ensure that the explanation is oriented towards the correct direction. If the prediction is incorrect, we then guide the LMM to generate an explanation conditioned on the corrected label to tackle potential explanation errors. Formally, we unify the vanilla explanation generation in both the Explain and Ground steps:

$$\tilde{\mathcal{E}}_i \triangleq \begin{cases} \Phi_{\text{LMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i; \mathcal{R}_i; y_i], \mathcal{P}'_{\text{explain}}), y_i \neq \hat{y}_i \\ \Phi_{\text{LMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i; \mathcal{R}_i], \mathcal{P}_{\text{explain}}), \text{otherwise} \end{cases} \tag{3}$$

where the label y_i is explicitly incorporated as a "golden prior" to anchor the explanation toward the correct direction, and $\mathcal{P}'_{\text{explain}}$ is the prompt for label-guided explanation. Notably, if the prediction in the Explain step is correct, the original explanation $\tilde{\mathcal{E}}_i$ is retained without modification.

Subsequently, to further eliminate the benign bias and ensure that explanations are closely aligned with predictions (e.g., the explanation for a hateful prediction should clearly identify potential hateful elements), we instruct the LMM to ground both the prediction and the explanation in the video content, enabling verification and refinement of the vanilla explanation. This process can be written as:

$$\mathcal{E}_i = \Phi_{\text{LMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i; \tilde{\mathcal{E}}_i], \mathcal{P}_{\text{ground}}), \tag{4}$$

where $\mathcal{E}_i$ is the unbiased and faithful explanation related to hateful content detection on video $\mathcal{S}_i$, and $\mathcal{P}_{\text{ground}}$ is the prompt designed to guide the LMM in grounding the relevance and consistency between the prediction and the rationale and perform explanation refinement if necessary. *Importantly*, the LMM is instructed to autonomously decide whether refinement is needed, based on the alignment between the explanation and prediction. By instructing the LMM to perform the three-step SG-CoT reasoning on each video instance $\mathcal{S}_i$ in training set $\mathcal{D}_{\text{train}}$, we obtain an explainable-oriented annotated training set: $\mathcal{D}^{\text{exp}}_{\text{train}} = \{\mathcal{S}_i, \mathcal{R}_i, \mathcal{E}_i\}_{i=1}^{|\mathcal{D}_{\text{train}}|}$.

3.3 Stage-Wise Distillation Paradigm

Building upon high-quality annotated explanation data, we introduce a new Stage-Wise Distillation paradigm that incrementally guides the SMMs toward explainable HVD.

Stage1: Understand. In the first stage, we utilize the reasoning outputs from the Reason step of SG-CoT as supervision to guide the SMM in understanding complex semantics and nuanced hateful cues within videos. Specifically, we train the SMM to generate descriptive summarization of video content, focusing on elucidating specific elements that contribute to understanding its meaning:

$$\mathcal{L}_{\text{s1}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T_i} \log P_{\text{SMM}}(r^i_j \mid \mathcal{S}_i, \mathcal{P}_{s1}, r^i_{<j}), \tag{5}$$

where N is the batch size, r_j^i denotes the j-th token in the reasoning sequence $\mathcal{R}_i$ distilled from the LMM, and T_i is the length of that sequence, $\mathcal{S}_i = \{\mathcal{T}_i, \mathcal{V}_i, \mathcal{A}_i\}$ is content of video $\mathcal{S}_i$, and $\mathcal{P}_{s1}$ is the prompt designed to guide the SMM in describing the content of $\mathcal{S}_i$ in the first-stage distillation.

Stage2: Explain. After being endowed with rich contextual and background knowledge for understanding complex hateful semantics in videos, the SMM is further prompted to generate human-aligned natural language explanations for its predictions, supervised by the outputs of the Explain and Ground steps in SG-CoT. To facilitate the prediction-consistent explanation and improve the detection capability of SMM, we develop a multi-task distillation strategy, which jointly trains the SMM to yield both predictions and explanations:

$$\mathcal{L}_{\text{s2}} = \mathcal{L}_{\text{expl}} + \lambda \cdot \mathcal{L}_{\text{pred}}, \tag{6}$$

where λ is the balance ratio for the two objectives:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \hat{y}_i^{(c)} \log p_i^{(c)}, \tag{7}$$

$$\mathcal{L}_{\text{expl}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} \log P_{\text{SMM}}(e_j^i \mid \mathcal{S}_i, \mathcal{P}_{s2}, e_{<j}^i). \tag{8}$$

Here, $C = 2$ is the total number of video classes, e_j^i represents the j-th token in the explanation output $\mathcal{E}_i$ extracted from the LMM, and M_i is the length of the explanation sequence, and $\mathcal{P}_{s2}$ is the prompt designed to guide the SMM in explaining its prediction for $\mathcal{S}_i$ in the second-stage distillation. By progressively distilling valuable and fine-grained knowledge from the SG-CoT guided LMM, the SMM acquires not only enhanced detection capability but also the ability to generate clear and faithful explanations for its predictions in HVD.

3.4 Inference

We present the complete inference workflow of our LEAF for explainable HVD. Specifically, the multimodal content of each video $\mathcal{S}_i$ from test set $\mathcal{D}_{\text{test}}$ is fed into the SMM to produce both a prediction and its corresponding explanation.

$$(\mathcal{E}_i, y_i) = \Phi_{\text{SMM}}([\mathcal{T}_i; \mathcal{V}_i; \mathcal{A}_i]; \mathcal{P}_{\text{inference}}), \tag{9}$$

where $\mathcal{E}_i$ is the rationale for predicted result y_i, and $\mathcal{P}_{\text{inference}}$ is the inference prompt. The details of Stage-Wise Distillation paradigm are provided in the experiment section, while the prompts and training algorithm are in the Appendix A – B.

4 Experiments

4.1 Experimental Setup

We present a brief setup, with a detailed one and additional experimental results in the Appendix C.

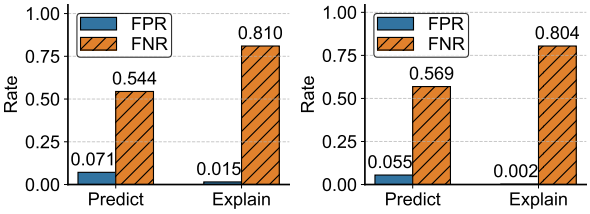
Datasets. We conduct experiments on three video benchmarks in HVD: MultiHateClip-YouTube (MHClip-Y), MultiHateClip-Bilibili (MHClip-B) (Wang et al., 2024b), and HateMM (Das et al., 2023). Each dataset is split into training, validation, and test sets in a 7:1:2 ratio, following prior work (Lang et al., 2025).

Baselines. We compare LEAF with 9 baselines with three groups: (1) *Multimodal detection methods*, including HTMM (Das et al., 2023), MHCL (Wang et al., 2024b), CMFusion (Zhang et al., 2024), and MoRE (Lang et al., 2025); (2) *LMM-based methods* that employ powerful LMMs with the same inference prompt as LEAF for zero-shot prediction, including GPT-4.1-mini (OpenAI, 2024), Gemma3-27B (Team et al., 2025), and Qwen2.5-VL-72B (Bai et al., 2025); (3) *SMM-based methods* that utilize lightweight SMMs with the same inference prompt as LEAF for zero-shot detection, including Gemma3-4B (Team et al., 2025) and Qwen2.5-VL-3B (Bai et al., 2025). Notably, these SMMs are also the backbone of LEAF for a fair evaluation.

Evaluation Metrics. For detection performance, following prior studies (Lang et al., 2025), we adopt four metrics: Accuracy (ACC), Macro-F1 score (M-F1), Macro-Precision (M-P), and Macro-Recall (M-R). Moreover, we quantitatively assess the explanations from LEAF using G-Eval (Liu et al., 2023), a reference-free, LLM-based evaluation framework. We utilize GPT-4.1-mini (OpenAI, 2024), an advanced and powerful LLM, and follow prior work (Wang et al., 2024a; Hong et al., 2025a) to score each rationale with five criteria: (1) Informativeness (**I**): the explanation offers new insights, such as additional contextual knowledge; (2) Soundness (**S**): the explanation is logical, valid, coherent, and aligns well with the video content; (3) Persuasiveness (**P**): the explanation is faithful to the prediction and compelling in justifying the decision; (4) Readability (**R**): the explanation follows standard grammar and structure; (5) Fluency (**F**): the explanation is smooth and natural. Each one is rated on a 5-point Likert scale (Joshi et al., 2015) (1 lowest and 5 highest quality) and the evaluation is repeated multiple times, and the final scores are reported as the average across runs. The evaluation

Method	MHClip-Y				MHClip-B				HateMM			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
HTMM	71.53	63.19	68.30	62.64	71.02	61.83	66.54	61.36	76.03	72.78	77.94	72.01
MHCL	71.03	65.47	67.22	64.86	76.50	73.11	73.20	73.02	77.41	76.54	76.49	76.59
CMFusion	67.50	63.39	63.44	63.34	70.50	63.36	65.96	64.98	78.80	77.23	78.62	76.60
MoRE	77.50	75.19	75.67	74.82	78.50	74.75	75.68	74.10	83.41	82.35	81.78	83.34
GPT-4.1-mini	71.36	55.82	77.99	58.20	77.39	66.90	83.44	65.67	81.02	80.16	80.22	80.10
Gemma3-27B	76.88	70.06	77.46	68.61	80.40	75.07	79.92	73.23	79.63	79.60	81.48	82.29
Qwen2.5-VL-72B	73.87	63.24	76.49	63.03	79.40	73.80	78.44	72.08	78.70	78.67	80.54	81.32
Gemma3-4B	71.36	64.39	66.55	63.69	71.86	66.27	67.32	65.70	75.00	74.99	77.46	77.85
+ **LEAF**	74.87	73.11	72.62	74.45	73.87	72.59	72.65	**75.81**	**85.65**	**85.21**	84.91	**85.72**
Qwen2.5-VL-3B	73.37	62.82	74.88	62.65	75.38	64.54	77.36	63.77	77.78	76.77	78.69	79.77
+ **LEAF**	**79.90**	**76.14**	**78.49**	**74.93**	**81.41**	**77.14**	**80.11**	75.61	84.72	83.30	**86.05**	82.19

Table 1: Detection results on the MHClip-Y, MHClip-B, and HateMM datasets. The best are in **bold** and the second underlined. Higher value of ACC, M-F1, M-P, and M-R indicate better performance.



(a) MHClip-Y Dataset. (b) MHClip-B Dataset.

Figure 3: False Positive Rate (FPR) and False Negative Rate (FNR) of the Qwen2.5-VL-72B under prediction-only and prediction-with-explanation settings.

prompts in prior work are modified to better align with HVD, and provided in the Appendix C.3.

Implementation Details. We primarily adopt the popular open-source `Qwen2.5-VL-72B` model (Bai et al., 2025) as the LMM teacher and its lightweight counterpart, `Qwen2.5-VL-3B`, as the SMM backbone. However, LEAF is model-agnostic and we also apply LEAF on `Gemma3-4B` model (Team et al., 2025) for detection evaluation to showcase its generalizability. The Stage-Wise Distillation is implemented using LoRA-based fine-tuning with the Unsloth (Daniel Han and team, 2023) framework, and the vision and language components are jointly optimized. Training is performed with a batch size of 8 using AdamW optimizer. All experiments use NVIDIA L40S GPUs.

4.2 Preliminary Experiment: Benign Bias

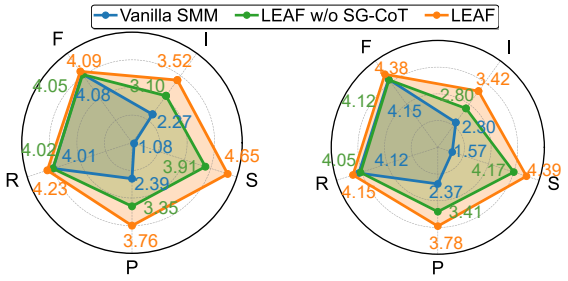
We empirically assess the benign prediction bias exhibited by current LMMs. Specifically, we prompt the Qwen2.5-VL-72B model to perform two prediction settings: prediction-only and prediction-with-explanation, on the MHClip-Y and MHClip-B datasets. To quantify the benign bias, we report the False Negative Rate (FNR), which reflects the ratio of hateful samples misclassified as benign. For comparison, we also report the False Positive Rate (FPR), which measures the ratio of benign samples predicted as hateful. As shown in Figure 3, the FNR is much higher than the FPR on both datasets, and this discrepancy is amplified when the LMM generates explanations alongside predictions, showcasing severe benign bias in current LMMs.

4.3 Detection Performance

We report the detection performance in Table 1, with observations as follows: **(O1): Multimodal detection methods** present strong capability on HVD, benefiting from their effective harnessing of the multimodal features with task fine-tuning. However, their performance remains limited when handling challenging samples that demand deep semantic understanding of hateful clues and broad contextual knowledge (i.e., interpreting the use of a monkey emoji as a racial slur in certain contexts). **LMM-based methods** alleviate this limitation by leveraging their extensive pre-trained knowledge to aid prediction. Nonetheless, their performance is hindered by the benign prediction bias, and debiasing through fine-tuning incurs substantial computational overhead. **(O2): SMM-based methods** exhibit lower zero-shot performance compared to their large-scale LMM counterparts due to their insufficient parametric knowledge. In contrast, our **LEAF** combines the strengths of both multimodal detection and LMM-based baselines by fine-tuning lightweight SMMs on HVD, while equipping them with valuable semantic and contextual knowledge distilled from LMMs, achieving superior detection performance with a lightweight structure.

Module	Variant	MHClip-Y M-F1	MHClip-B M-F1	HateMM M-F1
SG-CoT	w/o Ground	75.96	76.54	80.59
	w/o SG-CoT	72.91	74.09	77.31
Distillation	w/o Understand	72.33	71.57	80.38
	w/o Explain	67.85	67.50	71.75
	w/o Label	69.84	68.49	73.33
LEAF	**ALL**	**76.14**	**77.14**	**83.30**

Table 2: Ablation study on core components.



(a) MHClip-Y Dataset. (b) MHClip-B Dataset.

Figure 4: Quantitative comparisons of explanation.

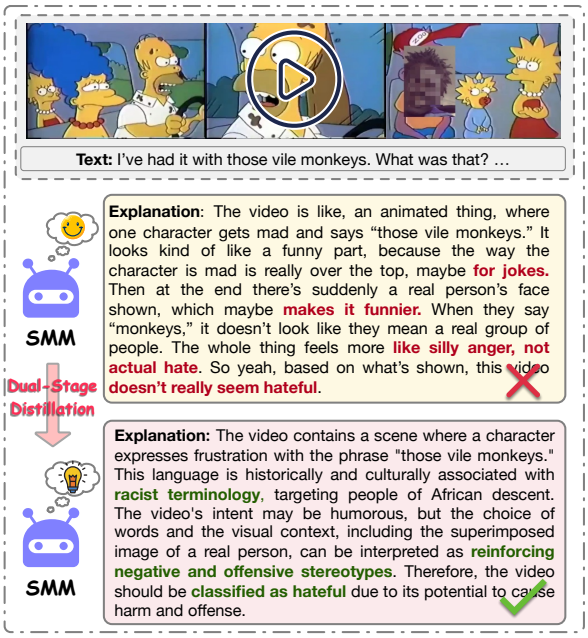


Figure 5: Case study of the explanations generated from the vanilla SMM and LEAF for a hateful video.

4.4 Ablation Study

We assess the role of component in LEAF, with the results in Table 2. To assess the *SG-CoT mechanism*, we develop variants: **(1) w/o Ground**, which drops the Ground step from the SG-CoT, and **(2) w/o SG-CoT**, which directly prompts the LMM to generate annotated explanations. Both variants result in degraded SMM performance, as the absence of proper guidance from SG-CoT compromises the quality of the distilled knowledge from the LMM. To validate the *Stage-Wise Distillation paradigm*, we design variants: **(1) w/o Understand**, and **(2) w/o Explain**, where each variant skips one of the distillation stages, and **(3) w/o Label**, which disables the multi-task distillation by removing HVD task fine-tuning in the Explain stage. The first two variants achieve degraded performance, highlighting the importance of both distillation stages in aiding the SMM to detect complex hateful clues in videos. The w/o Label variant also achieves suboptimal results, as lightweight SMMs, with limited model capacity and internalized knowledge, struggle to generalize to HVD without task-specific fine-tuning. We do not present the ablations of the Reason and Explain steps of SG-CoT, as removing them is equivalent to the ablations of the Understand and Explain distillation stages.

4.5 Explainability Evaluation

Quantitative Explainability Analysis. We quantitatively evaluate the explainability of LEAF against two baselines via G-Eval: (1) the vanilla SMM, and (2) the SMM that distills coarse explanation supervision data from LMM without the guidance of our proposed SG-CoT (dubbed as LEAF w/o SG-CoT) on five metrics defined in experimental setup: Informativeness **(I)**, Soundness **(S)**, Persuasiveness **(P)**, Readability **(R)**, and Fluency **(F)**, with results reported as average scores over the test set samples from both datasets. In Figure 4, LEAF achieves superior explainability compared to both baselines on all metrics (especially in **I**, **S**, and **P**), showcasing its strong capability to deliver faithful and compelling explanations for HVD.

Qualitative Explainability Analysis. To further validate the explainability of LEAF, we select a hateful video from the HateMM test set and qualitatively compare the relevant explanations from LEAF and the vanilla SMM. In Figure 5, the SMM produces a factually incorrect rationale, providing a misleading explanation for its benign decision due to its limited ability to understand and explain complex hateful cues (e.g., misinterpreting phrase "those vile monkeys" as humorous content). In contrast, LEAF distills the capabilities to *understand* and *explain* nuanced hateful elements in videos from SG-CoT guided LMMs, enabling it to generate clear and faithful natural language explanations that accurately identify specific hateful cues (e.g., recognizing "those vile monkeys" as a racist slur).

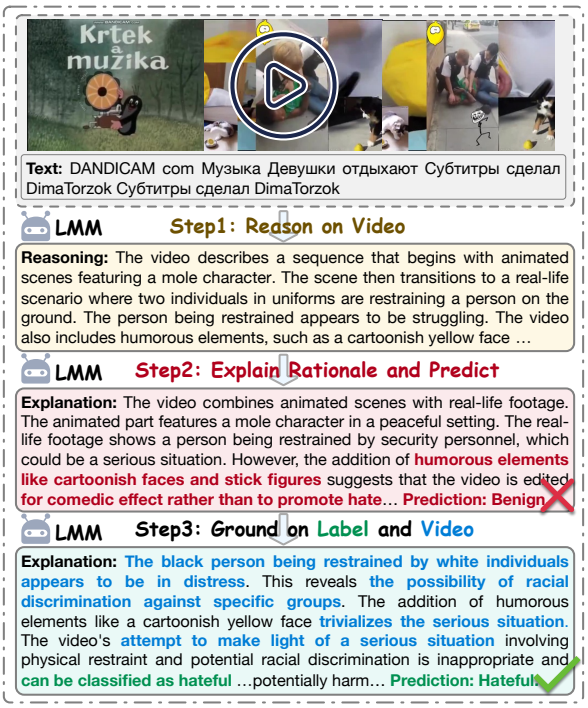
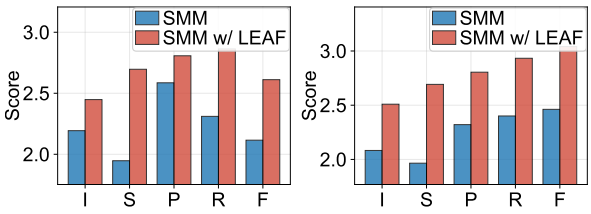


Figure 6: Case study of the explanation supervision signal generation guided by our SG-CoT mechanism.



(a) MHClip-Y Dataset. (b) HateMM Dataset.

Figure 7: Comparison of explainability on much smaller SMM with and without our proposed LEAF.

4.6 Exploration on Much Smaller SMMs

We further investigate the feasibility of achieving explainability in HVD with much more lightweight SMMs than those employed in our main experiments. Specifically, we select a SMM with particularly low model capacity, namely SmolVLM2-500M-Video-Instruct (Marafioti et al., 2025), with only 500M parameters. We then apply our proposed LEAF framework to endow it with explainability in HVD transferred from the LMM teacher, Qwen2.5-VL-72B. As shown in Figure 7, both SMMs exhibit noticeable improvements across all five explainability metrics, demonstrating the scalability of our framework to even more compact model scale. These results showcase the potential of LEAF to support highly lightweight and edge-adaptable models, making it particularly well-suited for latency-sensitive applications such as real-time video understanding on online platforms.

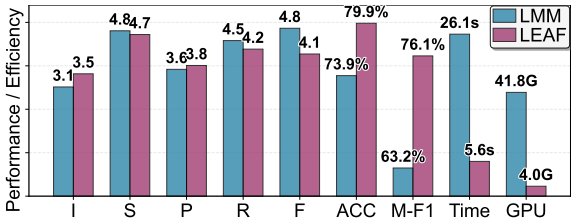


Figure 8: Inference efficiency and performance comparison between LEAF and the vanilla large-scale LMM.

4.7 Case Study on Self-Grounding CoT

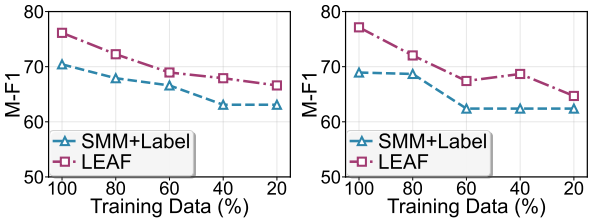
We conduct an in-depth analysis of how the proposed SG-CoT mechanism guides the LMM (i.e., Qwen2.5-VL-72B) to generate unbiased and faithful explanation supervisions for SMMs through a case study. We select a hateful video from the HateMM dataset, where the initial rationale from the Explain step is factually incorrect (i.e., overlooking violent acts of racial discrimination) due to the LMM's benign prediction bias. The Ground step first identifies this incorrect explanation trajectory and explicitly provides the ground-truth label "hateful" as a golden prior to guide the regeneration process. Besides, to ensure consistency and relevance between the prediction and the explanation, the Ground step further grounds both the rationale and the prediction in the video to produce a more faithful and compelling explanation (i.e., explicitly identifying the racial discrimination, where two white individuals violently restrain a black person).

4.8 Inference Efficiency Analysis

To showcase that LEAF achieves both cost efficiency and strong explainability on HVD, we compare its average per-sample inference cost on runtime and GPU memory, as well as its explainability in HVD on the MHClip-Y dataset, against the large-scale LMM (Qwen2.5-VL-72B). As shown in Figure 8, LEAF achieves comparable or even superior explainability to the vanilla LMM by alleviating its benign bias through the high-quality explanation supervised distillation. Meanwhile, LEAF incurs significantly lower resource overhead, making it compatible for online video platforms.

4.9 Low-Resource Dataset Evaluation

Beyond transferring explainability, the proposed LEAF also reduces the reliance of SMMs on large-scale training data for HVD owing to the knowledge distillation. To validate this, we conduct a data low-resource evaluation, where the vanilla SMM



(a) MHClip-Y Dataset. (b) MHClip-B Dataset.

Figure 9: Performance comparison between label fine-tuned SMM and LEAF on varying training size.

(with only detection label fine-tuning) and SMM with LEAF are evaluated under progressively reduced training set sizes (from 100% to 20%). As shown in Figure 9, LEAF achieves superior performance in low-resource settings, benefiting from the additional contextual and semantic knowledge from the LMM via the Stage-Wise Distillation.

5 Conclusion

In this study, we introduced LEAF, the first lightweight explainable framework for HVD. Specifically, we proposed a novel three-step SG-CoT mechanism to guide the LMM in generating high-quality explanation supervisions on videos. Supervised by these annotated signals, the SMM progressively acquires strong interpretability for HVD via a fresh Stage-Wise Distillation paradigm. Experiments presented that LEAF achieves superior detection performance and strong explainability in HVD, with lightweight overhead.

Limitations

While this work focuses on generating explanations in natural language form, we envision that future research can expand this by leveraging multiple modalities in the context of hateful video detection. For example, providing key visual, audio, and textual cues alongside natural language explanations could deliver more comprehensive and convincing justifications for detection outcomes. Nevertheless, LEAF is still powerful as the first lightweight yet effective explainable HVD framework.

Ethical Considerations

This work aims to mitigate the spread of hateful and discriminatory content in online videos by developing lightweight and explainable hateful video detection models that assist human moderation. Our proposed framework LEAF is designed solely for analysis and detection purposes and does not generate, promote, or endorse any harmful, hateful, or discriminatory content. While we acknowledge the potential risk that automated systems may be misused or circumvented by malicious actors, such behavior is strongly discouraged and beyond the intended scope of this research. LEAF is not intended to replace human judgment; instead, it provides grounded, human-readable explanations to support content moderators in understanding model decisions and decoding implicit hateful cues. We believe this approach contributes to more transparent, responsible, and human-aligned AI systems for online content moderation.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No.62572097 and No. U23A20315).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *arXiv.org*, abs/2502.13923.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

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

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit Optimizers via Block-wise Quantization. In *International Conference on Learning Representations (ICLR)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek et al. Kadian. 2024. The Llama 3 Herd of Models. *arXiv.org*, abs/2407.21783.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv.org*, abs/2411.10915.

Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2024. Multi-modal discussion

transformer: Integrating text, images and graph transformers to detect hate speech on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22096–22104.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025a. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference (WWW)*, pages 4684–4698.

Rongpei Hong, Jian Lang, Ting Zhong, and Fan Zhou. 2025b. 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)*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8003–8017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2024. Sida: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv.org*, abs/2412.04292.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

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 on Web Conference (WWW)*, pages 2763–2774.

Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. 2026a. 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*.

Jian Lang, Rongpei Hong, Ting Zhong, Yong Wang, and Fan Zhou. 2026b. 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*.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoyue Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. Pp-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System. *arXiv.org*, abs/2206.03001.

Kaiju Li, Rongpei Hong, Jian Lang, Jin Wu, Fan Zhou, and Jingkuan Song. 2026. Match: Multi-agentic evidence grounding for explainable hate video detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: Nlg Evaluation using Gpt-4 with Better Human Alignment. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522.

Ziyi Liu, Soumya Sanyal, Isabelle Lee, Yongkang Du, Rahul Gupta, Yang Liu, and Jieyu Zhao. 2024. Self-contradictory reasoning evaluation and detection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Andrés Marafioti, Orr Zohar, Miquel Farr'e, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, L. V. Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv*.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. *arXiv e-prints*, page arXiv:2502.13061.

Truong Thanh Hung Nguyen, Tobias Clement, Phuc Truong Loc Nguyen, Nils Kemmerzell, Van Binh Truong, Vo Thanh Khang Nguyen, Mohamed Abdelaal, and Hung Cao. 2024. Langxai: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8754–8758.

OpenAI. 2024. Gpt-4.1. Accessed: 2025-06-06.

OpenAI. 2025. Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card.

Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. 2023. Language model self-improvement by reinforcement learning contemplation. In *The Twelfth*

International Conference on Learning Representations.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pages 74952–74965.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable Fake News Detection with Large Language Model via Defense Among Competing Wisdom. In *The Web Conference (WWW)*, pages 2452–2463.

Han Wang, Rui Yang Tan, and Roy Ka-Wei Lee. 2025. Cross-modal transfer from memes to videos: Addressing data scarcity in hateful video detection. In *Proceedings of the ACM on Web Conference (WWW)*, pages 5255–5263.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024b. Multihateclip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502. ACM.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*.

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

Boxuan Zhang and Ruqi Zhang. 2025. Cot-UQ: Improving Response-wise Uncertainty Quantification in LLMs with Chain-of-Thought. *arXiv.org*, abs/2502.17214:arXiv:2502.17214.

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 (ICDMW)*, pages 183–190. IEEE.

Xianwei Zhuang, Zhihong Zhu, Zhichang Wang, Xuxin Cheng, and Yuexian Zou. 2025. Unicott: A Unified Framework for Structural Chain-of-Thought Distillation. In *The Thirteenth International Conference on Learning Representations*.

A Detailed Prompt Content for LEAF

A.1 Prompt for SG-CoT Mechanism

In this section, we provide the detailed content of prompts fed to Large Multimodal Models (LMMs) for yielding fine-grained and unbiased explanation supervisions via the three step multimodal Self-Grounding Chain-of-Thought (CoT (Wei et al., 2022)) mechanism, including the prompt for Reason, Explain, and Ground steps.

- **Prompt for Reason Step**: "You are a video content analyzer on a video platform. You are provided with the following inputs from a video: - 16 consecutive video frames (visual context), - the video title, - its full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [{transcript}] On-screen Text: [{ocr}] Your task is to consider all the inputs and write a detailed and logically complete analysis of the video that clearly conveys the main storyline. Try to understand the nuanced sentiment of the content, identify key people or objects, provide relevant background context, and describe important visual elements such as symbols or actions. Avoid using bullet points or headings. The goal is to produce an output that is clear, consistent, and easy for a smaller student model to learn from. Respond in the following format: "The video describes [analysis]".. "

- **Prompt for Explain Step**: " You are a moderator on a video platform. You are provided with the following inputs from a video: - 16 consecutive video frames (visual context), - the video title, - its full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [{transcript}] On-screen Text: [{ocr}] You are also provided with a detailed description of the video content which is generated by previous steps, along with relevant background knowledge. Video Description: [{description}] Hateful video definition: The video is classified as hateful as it contains hateful content targeting individuals or groups based on race, religion, gender, sexuality, or other identities. Your task is to consider all the inputs and incorporate contextual background knowledge of the video to understand the nuanced sentiment of the content. Based on this, judge whether the video is hateful or benign. Provide a clear and easy-to-understand reasoning for your judgment. Avoid using bullet points or headings. The goal is to produce an output that is clear, consistent, and easy for a smaller student model to learn from. Respond in the following format: Thought: [your reasoning in natural language] Answer: [hateful/benign] "

- **Prompt for Ground on label**: You are a moderator on a video platform. You are provided with the following inputs from a {text_label} video: - 16 consecutive video frames (visual context), - the video title, - its full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [{transcript}] On-screen Text: [{ocr}] You are also provided with a description of the video for its hate prediction. Video Description: [{description}] The video has been classified as {text_label} by the standard of the platform. Your task is to consider the video content and provide a clear and easy-to-understand resoning for judging the video as {text_label}. Avoid using bullet points or headings. The goal is to produce an output that is clear, consistent, and easy for a smaller student model to learn from. Respond in the following format: Thought: [your reasoning in natural language] Answer: [{text_label}]

- **Prompt for Ground on Video**: You are a moderator on a video platform. Given the following information: - 16 consecutive video frames (visual context), - Video Title: [{title}] - Video Transcript: [{transcript}] - On-screen Text (OCR): [{ocr}] - Video Description: [{description}] - Your Previous Explanation: [{explanation}] The video has been classified as prediction by both the platform standard and your previous prediction. Hateful video definition: A video is hateful if it targets individuals or groups based on race, religion, gender, sexuality, or other identities. Your ONLY task is to verify if your previous explanation faithfully matches the prediction based on the video content. If your previous explanation accurately reflects the prediction, respond ONLY with: "OK" ONLY IF your previous explanation does NOT faithfully match the prediction, provide a revised explanation. In this case, clearly provide a brief, faithful, and easy-to-understand revised explanation WITHOUT changing the prediction. Do NOT provide any analysis, reasoning or other comments if your original explanation is already faithful. Your response must strictly follow ONE of these formats: If faithful: OK If not faithful: Respond in the following format: Re-

vised Explanation: [your brief and clear revised explanation] Answer: [{prediction}]

A.2 Prompt for Stage-Wise Distillation Paradigm

In this section, we provide the detailed content of prompts for Smaller Multimodal Models (SMMs) from LMMs in acquiring explainability for interpretable HVD via the Stage-Wise Distillation paradigm, including the prompts for Understand and Explain stages.

- **Prompt for Understand Step**: "You are a video content analyzer on a video platform. You are provided with the following inputs from a video: - 4 consecutive video frames (visual context), - the video title, - its full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [{transcript}] On-screen Text: [{ocr}] Your task is to consider all the inputs and write a detailed and logically complete analysis of the video that clearly conveys the main storyline. Try to understand the nuanced sentiment of the content, identify key people or objects, provide relevant background context, and describe important visual elements such as symbols or actions. Avoid using bullet points or headings. The goal is to produce an output that is clear, consistent, and easy for a smaller student model to learn from. Respond in the following format: "The video describes [analysis]". "

- **Prompt for Explain Stage**: " You are a moderator on a video platform. You are provided with the following inputs from a video: - 4 consecutive video frames (visual context), - the video title, - the full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [transcript] On-screen Text: [{ocr}] Hateful video definition: The video is classified as hateful as it contains hateful content targeting individuals or groups based on race, religion, gender, sexuality, or other identities. Your task is to consider all the inputs and incorporate contextual background knowledge of the video to understand the nuanced sentiment of the content. Based on this, judge whether the video is hateful or benign. Provide a clear and easy-to-understand reasoning for your judgment. Respond in the following format: Thought: [your reasoning in natural language] Answer: [hateful/benign] "

Characteristics	MHClip-Y	MHClip-B	HateMM
Total Videos	1,000	1,000	1,083
Hateful Videos	338	322	431
Benign Videos	662	678	652
Avg. Duration (s)	33.8	31.8	150.0
Languages	English	Chinese	English
Platforms	YouTube	Bilibili	BitChute

Table 3: Characteristics of three video datasets in HVD.

A.3 Prompt for Inference

In this section, we provide the detailed content of prompt for SMMs in delivering both prediction and human-aligned explanation for videos in the reference phase:

" You are a moderator on a video platform. You are provided with the following inputs from a video: - 4 consecutive video frames (visual context), - the video title, - the full transcript, - any on-screen text extracted via OCR. Video Title: [{title}] Video Transcript: [{transcript}] On-screen Text: [{ocr}] Hateful video definition: The video is classified as hateful as it contains hateful content targeting individuals or groups based on race, religion, gender, sexuality, or other identities. Your task is to consider all the inputs and incorporate contextual background knowledge of the video to understand the nuanced sentiment of the content. Based on this, judge whether the video is hateful or benign. Provide a clear and easy-to-understand reasoning for your judgment. Respond in the following format: Thought: [your reasoning in natural language] Answer: [hateful/benign] "

B Algorithm of Proposed LEAF

In this section, we present the detailed algorithms of our proposed framework LEAF, including both the SG-CoT mechanism guided explanation annotation process and the Stage-Wise Distillation paradigm. The algorithm for explanation data annotation is presented in Algorithm 1, while the distillation procedure is provided in Algorithm 2.

C Detailed Experimental Settings

C.1 Datasets

We conduct extensive experiments on three real-world public video datasets: MultiHateClip-YouTube (MHClip-Y), MultiHateClip-Bilibili (MHClip-B) (Wang et al., 2024b), and HateMM (Das et al., 2023). Each dataset is split into training, validation, and test sets in a

Algorithm 1 SG-CoT Guided Explanation Data Annotation

Require: Training HVD dataset $\mathcal{D}_{\text{train}} = \{\mathcal{S}_i\}_{i=1}^{M_t}$.
Ensure: Explanation-annotated HVD dataset $\mathcal{D}_{\text{train}}^{\text{exp}} = \{(\mathcal{S}_i, \mathcal{R}_i, \mathcal{E}_i)\}_{i=1}^{M_t}$.

for each sample $\mathcal{S}_i \in \mathcal{D}_{\text{train}}$ **do**
  // Video Preprocessing
  Sample frames to construct visual input $\mathcal{V}_i$.
  Extract on-screen text and combine with title to obtain textual input $\mathcal{T}_i$.
  Transcribe audio to obtain audio input $\mathcal{A}_i$.
  // Step 1: Reason
  Generate contextual reasoning output $\mathcal{R}_i$ via Eq.(1).
  // Step 2: Explain
  Generate vanilla explanation and LMM prediction via Eq.(2).
  // Step 3: Ground
  if $y_i \neq \hat{y}_i$ **then**
    Re-generate explanation $\widetilde{\mathcal{E}}_i$ using ground-truth label as golden prior via Eq.(3).
  end if
  Refine explanation with video content grounding via Eq.(4).
  Store $(\mathcal{S}_i, \mathcal{R}_i, \mathcal{E}_i)$ into $\mathcal{D}_{\text{train}}^{\text{exp}}$.
end for
return $\mathcal{D}_{\text{train}}^{\text{exp}}$.

Algorithm 2 Stage-Wise Distillation Paradigm

Require: Explanation-annotated dataset $\mathcal{D}_{\text{train}}^{\text{exp}} = \{(\mathcal{S}_i, \mathcal{R}_i, \mathcal{E}_i)\}_{i=1}^{M_t}$.
Ensure: Trained SMM model with explainability in HVD.

// Stage 1: Understand
for each $(\mathcal{S}_i, \mathcal{R}_i)$ in $\mathcal{D}_{\text{train}}^{\text{exp}}$ **do**
  Encode video content $\mathcal{S}_i = \{\mathcal{T}_i, \mathcal{V}_i, \mathcal{A}_i\}$ into SMM.
  Generate reasoning output $\hat{\mathcal{R}}_i$ using prompt $\mathcal{P}_{s1}$.
  Compute cross-entropy loss for reasoning supervision via Eq.(5).
end for
Update SMM parameters using reasoning loss $\mathcal{L}_{\text{s1}}$.
// Stage 2: Explain
for each $(\mathcal{S}_i, \mathcal{E}_i)$ in $\mathcal{D}_{\text{train}}^{\text{exp}}$ **do**
  Encode video content $\mathcal{S}_i$ with SMM.
  Predict class label $\hat{y}_i$ and explanation $\hat{\mathcal{E}}_i$ using prompt $\mathcal{P}_{s2}$.
  Compute classification loss via Eq.(7).
  Compute explanation distillation loss via Eq.(8).
  Combine both into multi-task objective via Eq.(6).
end for
Update SMM parameters using final loss $\mathcal{L}_{s2}$.
return Trained SMM model.

7:1:2 ratio, following prior work (Lang et al., 2025). The detailed dataset statistics are provided in Table 3. Below, we present the detailed dataset descriptions for each dataset.

- **MHClip-Y** and **MHClip-B** (Wang et al., 2024b): These two datasets are specifically constructed for the task of hateful video detection on YouTube and Bilibili, two of the most popular online video-sharing platforms. Each data instance comprises a video along with its corresponding metadata, including the title, transcript, and fine-grained human annotations. The annotations indicate whether the content is hateful, offensive, or benign for downstream analysis. Following prior work (Lang et al., 2025), we group hateful and offensive categories under a unified hateful label, thereby framing the task as a binary classification problem (*hateful* vs. *benign*).

- **HateMM** (Das et al., 2023):This dataset collects videos from BitChute, an alternative video-sharing platform known for its minimal content moderation. Each video sample is manually annotated by trained experts to ensure labeling quality. For each sample, the full video content is provided along with a binary label indicating whether the video is hateful or benign, enabling standard supervised learning for binary classification.

C.2 Baselines

In this study, we compare our proposed LEAF with 9 competitive baselines, which can be grouped into three classes: (1) *Multimodal detection methods* that model cross-modal relationships for detection, including HTMM (Das et al., 2023), MHCL (Wang et al., 2024b), CMFusion (Zhang et al., 2024), and MoRE (Lang et al., 2025); (2) *LMM-based methods* that employ powerful LMMs for prediction, including GPT-4.1-mini (Hurst et al., 2024), Gemma3-27B (Team et al., 2025), and Qwen2.5-VL-72B (Bai et al., 2025); (3) *SMM-based methods* that utilize lightweight SMMs for prediction,

including Gemma3-4B (Team et al., 2025) and Qwen2.5-VL-3B (Bai et al., 2025). Notably, the prompts for both LMM- and SMM-based baselines are the same as our LEAF utilized in inference stage. Below we provide detailed descriptions for each baseline model.

- **HTMM** (Das et al., 2023): HTMM encodes multimodal inputs, including transcripts, visual frames, and audio signals—into unified representations via modality-specific feature extractors. The concatenated features are then processed by an MLP-based classifier to identify hateful content in short-form videos.
- **MHCL** (Wang et al., 2024b): MHCL evaluates the contribution of each modality, audio, text, and vision, in detecting hateful content in videos. It employs LSTM-based encoders to extract and integrate temporal features across modalities for final hate classification.
- **CMFusion** (Zhang et al., 2024): CMFusion enhances multimodal hate video detection by introducing a channel-wise and modality-wise fusion mechanism. It first captures temporal dependencies between audio and visual streams via a temporal cross-attention module, then adaptively fuses text, audio, and video features to generate discriminative representations for classification.
- **MoRE** (Lang et al., 2025): MoRE introduces a retrieval-augmented mixture of multimodal experts framework to enhance HVD. It retrieves semantically relevant video instances to improve generalization to emerging hateful content and employs a multimodal MoE network to adaptively fuse modality-specific experts at the instance level.
- **GPT-4.1-mini** (OpenAI, 2024): GPT-4.1-mini is a lightweight member of the GPT-4.1 model family recently released by OpenAI. This model achieves strong performance on instruction following and code generation tasks, outperforming previous models such as GPT-4o and GPT-4o-mini. GPT-4.1-mini also supports long-context processing, with improved ability to utilize extended context for reasoning and generation. In this study, we choose it as a strong and competitive baseline with zero-shot setting.
- **Gemma-3** (Team et al., 2025): Gemma 3 is a latest released LMM that integrates visual and textual inputs within a unified architecture. It employs a SigLIP-based vision encoder to transform video frames into a fixed set of visual tokens, which are interleaved with text tokens and processed using alternating local and global self-attention layers to support long-context understanding (up to 128K tokens). This design enables Gemma 3 to perform a variety of multimodal reasoning and understanding tasks in a zero-shot setting through instruction tuning. In this study, we adopt two model sizes from the Gemma-3 family, Gemma-3-4B and Gemma-3-27B, as our SMM and LMM baselines, respectively. The 4B model also serves as our SMM backbone in the detection experiment.
- **Qwen2.5-VL** (Bai et al., 2025): Qwen2.5-VL is the latest flagship model in the Qwen vision-language series, showcasing substantial improvements in both core capabilities and advanced functionalities. It marks a significant step forward in visual understanding and interaction, excelling in fine-grained visual recognition, accurate object localization, robust document parsing, and comprehensive long-video analysis. In this work, we employ two model sizes from the Qwen2.5-VL family, Qwen2.5-VL-3B and Qwen2.5-VL-72B, as our SMM and LMM baselines, respectively. The 3B model also serves as our primary SMM backbone in most experiments.

Notably, both the LMM- and SMM-based baselines are evaluated using the same inference prompt as our proposed LEAF to ensure a fair comparison.

C.3 Quantitative Explainability Evaluation Details

As mentioned in the main paper, we adopt five metrics for quantitative evaluation of explainability, including:

- **Informativeness** (**I**): the explanation offers new insights, such as additional contextual knowledge.
- **Soundness** (**S**): the explanation is logical, valid, coherent, and aligns well with the video content.
- **Persuasiveness** (**P**): the explanation is faithful to the prediction and compelling in justifying the decision.
- **Readability** (**R**): the explanation follows standard grammar and structure.

- **Fluency** (**F**): the explanation is smooth and natural.

Each criterion is rated on a 5-point Likert scale (Joshi et al., 2015), where 1 denotes the lowest and 5 the highest quality. To facilitate evaluation, we employ GPT-4.1 mini (gpt-4.1-mini-2025-04-14) as the assessment LLM and construct tailored prompts corresponding to each evaluation criterion. The prompt for each criterion are provided below:

- **Prompt for Informativeness Evaluation**: You will be given a video including title and textual content. You will be provided with an explanation regarding whether this video contains hateful content. Your task is to rate the explanation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Informativeness (1-5) - assesses whether the explanation offers new insights and additional contextual knowledge beyond the basic classification. A high score indicates that the explanation provides valuable contextual information, background knowledge, or deeper insights that enhance understanding of why the content may or may not be hateful. Evaluation Steps: 1. Carefully review the video's title and textual content. 2. Read the provided explanation about whether the video contains hateful content. 3. Assess how much new insights and contextual knowledge the explanation provides, including background details and additional context that helps understand the situation better. 4. Assign a score for Informativeness on a scale of 1 to 5, where 1 is the lowest (providing little to no new insights or contextual knowledge) and 5 is the highest (offering extensive new insights and contextual knowledge). Example: video: Title: {Title} Content: {Content} Explanation: {Explanation} Evaluation Form (score ONLY): - Informativeness:

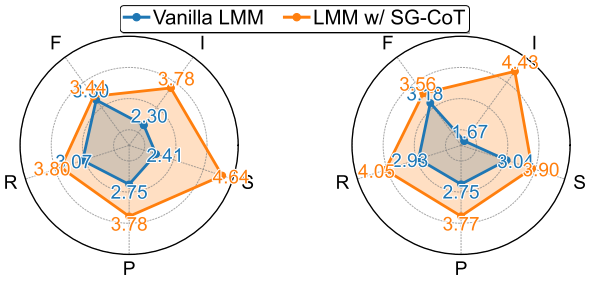
- **Prompt for Readability Evaluation**: You will be given a video including title and textual content. You will be provided with an explanation regarding whether this video contains hateful content. Your task is to rate the explanation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Readability (1-5) - evaluates whether the explanation follows standard grammar and structure. A high score indicates that the explanation adheres to proper grammatical rules, maintains consistent structure, and is well-organized according to standard writing conventions. Evaluation Steps: 1. Carefully review the video's title and textual content. 2. Read the provided explanation about whether the video contains hateful content. 3. Assess the explanation's readability by considering: - Adherence to standard grammar rules - Consistency in structural organization - Proper use of punctuation and sentence construction - Overall conformity to standard writing conventions 4. Assign a score for Readability on a scale of 1 to 5, where 1 is the lowest (poor grammar and structure) and 5 is the highest (excellent grammar and structure). Example: video: Title: {Title} Content: {Content} Explanation: {Explanation} Evaluation Form (score ONLY): - Readability:

- **Prompt for Soundness Evaluation**: You will be given a video including title and textual content. You will be provided with an explanation regarding whether this video contains hateful content. Your task is to rate the explanation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Soundness (1-5) - evaluates whether the explanation is logical, valid, coherent, and aligns well with the video content. A high score indicates that the explanation demonstrates logical consistency, valid reasoning, coherent structure, and strong alignment with the actual video content. A low score suggests that the explanation lacks logical validity, contains inconsistencies, or poorly aligns with the video content. Evaluation Steps: 1. Carefully review the video's title and textual content. 2. Read the provided explanation about whether the video contains hateful content. 3. Assess the explanation's soundness by considering: - Logical consistency and validity of the reasoning - Coherence of the overall structure and arguments - Alignment between the explanation and the actual video content - Absence of logical fallacies or contradictions 4. Assign a score for Soundness on a scale of 1 to 5, where 1 is the lowest (not sound, poor alignment) and 5 is the highest (very sound, excellent alignment). Example: video: Title: {Title} Content: {Content} Explanation:

{Explanation} Evaluation Form (score ONLY): - Soundness:

- **Prompt for Persuasiveness Evaluation**: You will be given a video including title and textual content. You will be provided with an explanation regarding whether this video contains hateful content. Your task is to rate the explanation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Persuasiveness (1-5) - evaluates whether the explanation is faithful to the prediction and compelling in justifying the decision. A high score indicates that the explanation accurately reflects the model's prediction and provides compelling justification for the classification decision. A low score suggests that the explanation is unfaithful to the prediction or lacks compelling justification for the decision. Evaluation Steps: 1. Carefully review the video's title and textual content. 2. Read the provided explanation about whether the video contains hateful content. 3. Assess the explanation's persuasiveness by considering: - Faithfulness to the actual prediction/classification - Compelling nature of the justification provided - Effectiveness in supporting the classification decision - Overall convincingness in explaining why the decision was made 4. Assign a score for Persuasiveness on a scale of 1 to 5, where 1 is the lowest (unfaithful and unconvincing) and 5 is the highest (highly faithful and compelling). Example: video: Title: {Title} Content: {Content} Explanation: {Explanation} Evaluation Form (score ONLY): - Persuasiveness:

- **Prompt for Fluency Evaluation**: You will be given a video including title and textual content. You will be provided with an explanation regarding whether this video contains hateful content. Your task is to rate the explanation on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Fluency (1-5) - evaluates whether the explanation is smooth and natural. A high score indicates that the explanation flows smoothly, uses natural language patterns, and reads effortlessly with seamless transitions between ideas. A low score suggests that the explanation is choppy, unnatural, or difficult to read smoothly. Evaluation Steps: 1. Carefully review the video's title and textual content. 2. Read the provided explanation about whether the video contains hateful content. 3. Assess the explanation's fluency by considering: - Smoothness and naturalness of language flow - Natural rhythm and pacing in sentence construction - Seamless transitions between ideas and sentences - Overall ease and comfort in reading the text 4. Assign a score for Fluency on a scale of 1 to 5, where 1 is the lowest (choppy and unnatural) and 5 is the highest (smooth and natural). Example: video: Title: {Title} Content: {Content} Explanation: {Explanation} Evaluation Form (score ONLY): - Fluency:



(a) MHClip-Y Dataset. (b) MHClip-B Dataset.

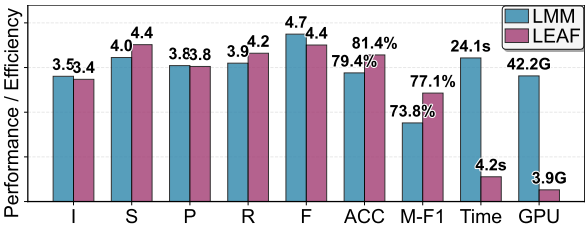
Figure 10: Comparison of quality of the annotated explanation data between the Vanilla LLM and the LMM with our proposed SG-CoT mechanism.

In practice, for evaluation generation, we set the temperature to 2.0 and the n (the number of completions) to 20. The final result is reported as the average over these completions.

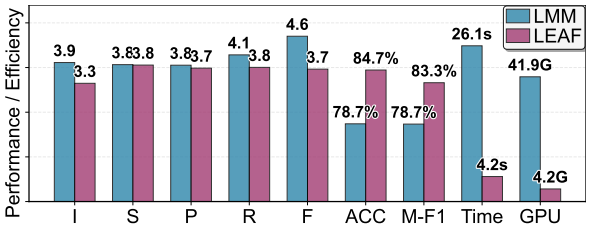
C.4 Implementation Details

In this section, we provide implementation details related to our work, including the video data preprocessing, the LMM and SMM backbone, the explanation data annotation, the Stage-Wise Distillation, and the running environment.

- **Video Data Preprocessing**: We detail our video data preprocessing for both visual and textual modalities. For the visual modality, we sample 16 frames per video when annotating explanation data to ensure sufficient temporal coverage for reasoning. In contrast, we sample 4 frames for Stage-Wise Distillation to stabilize training and reduce inference overhead. For the textual modality, we use PaddleOCR (Li et al., 2022) to extract on-screen text and Whisper (Radford et al., 2023) to transcribe audio.



(a) Inference efficiency and performance comparison between LEAF and vanilla LMM on the MHClip-B dataset.



(b) Inference efficiency and performance comparison between LEAF and vanilla LMM on the HateMM dataset.

Figure 11: Inference efficiency and performance comparison between LEAF and vanilla LMM on the MHClip-B and HateMM datasets.

- **LMM and SMM backbone**: We mainly adaopt the popular open-source `Qwen2.5-VL-72B` model (Bai et al., 2025) as the LMM teacher and its lightweight counterpart, `Qwen2.5-VL-3B` model, as the SMM backbone. However, LEAF is model-agnostic and we also apply LEAF on `Gemma3-4B` model (Team et al., 2025) for detection evaluation to showcase its generalizability.

- **Explanation Data Annotation**: We employ `Qwen2.5-VL-72B-unsloth-bnb-4bit` as the teacher model, and set the temperature to 0.000001.

- **Stage-Wise Distillation**: We adopt Unsloth (Daniel Han and team, 2023) as our fine-tuning framework and utilize two SMM: `Qwen2.5-VL-3B-unsloth-bnb-4bit` and `gemma-3-4b-it-unsloth-bnb-4bit`. To mitigate category imbalance in the training data, we apply negative sampling during data preprocessing, ensuring a more balanced label distribution. For LoRA (Hu et al., 2022) fine-tuning, we configure the rank r to 16 and set lora alpha to 16, enabling joint fine-tuning of both vision and language modules. Our training setup accommodates a maximum sequence length of 32,768 tokens and a batch size of 8. We employ the AdamW 8-bit optimizer (Dettmers et al., 2022), with a warmup cosine scheduler, using a weight decay of 0.01 and a warmup ratio of 0.1. Learning rates are adjusted in a stage-specific manner: during the understand stage, we explore learning rates in the range of 5e-5 to 9e-5, while in the explain stage, we increase this range from 6e-4 to 9e-4.

- **Running Environment**: All experiments are conducted on a system equipped with an AMD EPYC 7K62 CPU, an NVIDIA L40s GPU, and 128 GB of system RAM.

D Quantitative Evaluation on SG-CoT Mechanism

In the main paper, we have demonstrated the effectiveness of SG-CoT through both detection performance and a quantitative explainability analysis within the LEAF framework. However, this evaluation is indirect. In this section, we present a more straightforward assessment by evaluating the quality of the annotated explanation-oriented data generated by SG-CoT, providing further evidence of its efficacy. Specifically, we compare the quality of annotated data generated by the vanilla LMM (Qwen2.5-VL-72B) and by the same LMM guided by our SG-CoT mechanism (LLM w/ SG-CoT), using five pre-defined explainable evaluation criteria. As illustrated in Figure 10, the annotations produced by the vanilla LMM exhibit compromised quality across almost all metrics, with particularly low quality in Informativeness, Soundness and Persuasiveness due to its severe benign bias. In contrast, the annotations generated under the guidance of SG-CoT demonstrate significantly enhanced quality, showcasing its efficacy in steering the LMM to produce more faithful and unbiased explanation supervision signals.

E Additional Efficiency Analysis

In this section, we additionally provide the inference efficiency comparison on the MHClip-B and HateMM datasets to complement the results in the main paper, which were omitted due to space constraints. As illustrated in Figure 11a and Figure 11b, LEAF achieves comparable and even better explainability than the vanilla LMM on some metrics, as LEAF alleviates the severe benign bias of the LMM through the Stage-Wise Distillation. Meanwhile,

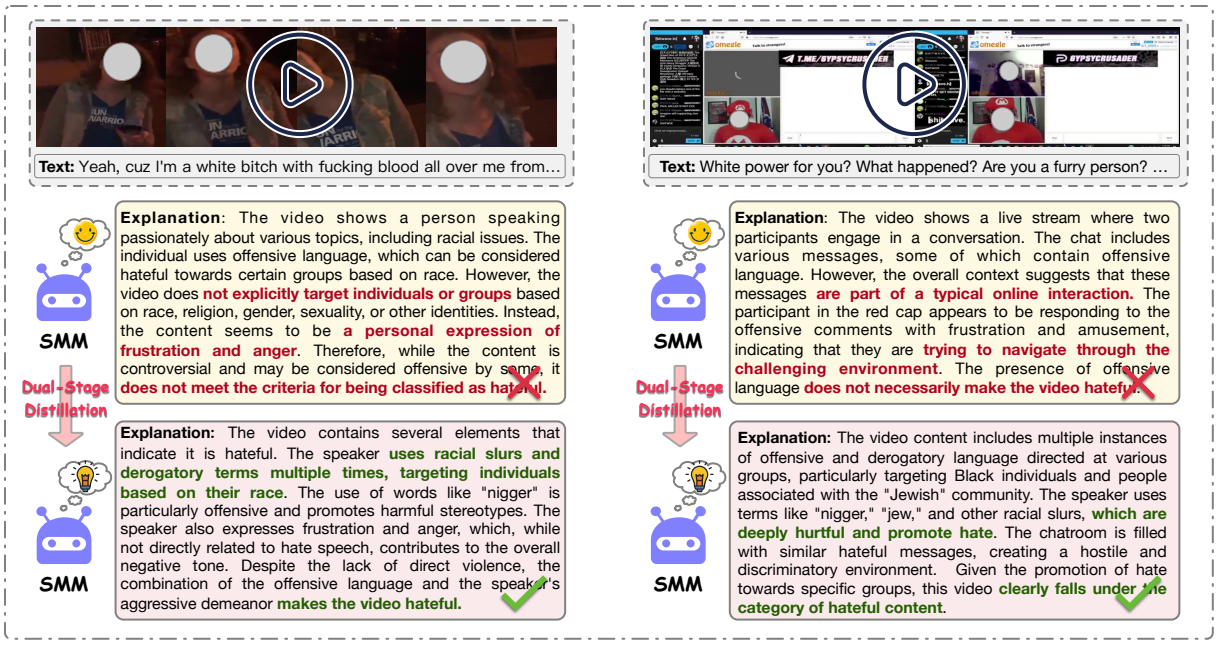


Figure 12: Additional cases of the explanations generated from the vanilla SMM and LEAF for a hateful video.

Figure 13: Cases of the explanation supervision signal generation process guided by our SG-CoT mechanism.

LEAF incurs significantly lower resource overhead compared to the LMM in terms of runtime latency and GPU memory consumption, making it more suitable for real-world video platforms.

F Additional Case Study on Explainability

In this section, we provide more cases from HateMM dataset to qualitatively demonstrate the explainability of our proposed LEAF in Figure 12.

G Additional Case Study on SG-CoT Mechanism

As illustrated in Figure 13, we present more cases from HateMM dataset to showcase the effectiveness of our proposed SG-CoT mechanism in guiding LMMs to provide high-quality explanation supervisions for videos.