

# TRIAGE: Label-Free Hateful Video Detection with Boundary Mapping and Resolution

Junyi Chen<sup>1</sup>, Qian Liu<sup>1\*</sup>, Jing Sun<sup>1</sup>, Yi Zhang<sup>2</sup>

<sup>1</sup>University of Auckland    <sup>2</sup>University of Technology Sydney

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

## Abstract

Hateful video detection is essential for maintaining safe and trustworthy online platforms. Current state-of-the-art methods are predominantly trained in a supervised manner, which incurs substantial annotation costs as on-line video content continues to grow rapidly. Although recent label-free approaches have shown promising progress, they still suffer from an effectiveness–efficiency trade-off, lagging behind supervised methods while requiring considerable computational resources. To address this limitation, we propose TRIAGE, a coarse-to-fine label-free detection framework that first localizes a boundary region and then progressively refines the cases within it. The framework introduces two core designs. The *Boundary Mapper* uses a lightweight MLLM to efficiently localize a boundary region, finalizing cases outside this region while retaining ambiguous hard cases for further reasoning. The *Adaptive Boundary Resolver* then queries diverse stronger MLLMs sequentially, resolves cases within the boundary region through Bayesian sequential updates, and terminates once sufficient evidence is accumulated. Across four public datasets, TRIAGE improves accuracy over the best label-free/few-shot baseline by 2.5–8.7 points, reduces parameter-weighted inference cost by at least 2.0 $\times$ , and outperforms the best supervised baseline on three datasets. Code is available at <https://github.com/JunyiChen-ai/TRIAGE>.

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

## 1 Introduction

Hateful videos refer to multimedia content that conveys or promotes harmful, offensive, or discriminatory messages toward individuals or groups based on attributes such as race, religion, gender, or nationality (Das et al., 2023; Wang et al., 2024;

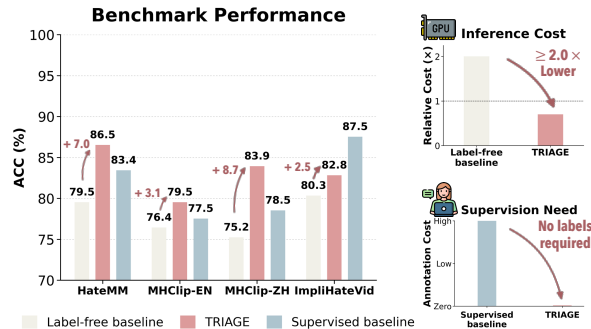


Figure 1: Comparison between TRIAGE and prior methods. Supervised approaches achieve strong performance but require labor-intensive annotated data, while existing label-free methods lag behind and still incur expensive reasoning costs. In contrast, TRIAGE substantially alleviates this effectiveness–efficiency bottleneck by consistently and significantly outperforming existing label-free baselines while maintaining high efficiency, and further surpasses supervised baselines on three out of four datasets without requiring any annotation cost.

Rehman et al., 2025). Detecting such content is critical for maintaining safe online environments and mitigating the spread of harmful information. Current state-of-the-art methods are supervised multimodal detectors, which fuse textual, visual, and acoustic features and are trained on annotated data to learn a discriminative decision boundary between hateful and normal videos (Das et al., 2023; Wang et al., 2024; Lang et al., 2025; Rehman et al., 2025; Li et al., 2026; Zhang et al., 2024). However, producing such annotations requires substantial human effort and may expose annotators to harmful or distressing content (Röttger et al., 2021; Kiela et al., 2020).

To address this issue, recent low-resource approaches attempt to leverage pretrained multimodal large language models (MLLMs) as zero-shot or few-shot moderators. In hateful video detection, MARS (Yang et al., 2026) represents a label-free multi-stage reasoning pipeline, while related

\*Corresponding author

harmful meme detection systems employ few-shot examples, retrieval, pseudo-labeling, or agentic self-improvement (Cao et al., 2024; Lang et al., 2026; Huang et al., 2024). Despite the significant progress achieved by these methods, a fundamental limitation remains: they rely on strong MLLMs to perform sophisticated reasoning for every input video, incurring substantial computational cost, yet their performance still lags considerably behind supervised approaches.

This motivates the following question: *can we design an MLLM-based label-free detector that achieves performance comparable to supervised methods while maintaining efficiency?* Intuitively, supervised approaches achieve strong performance because they learn a target-specific decision boundary from annotated data during training. At inference time, this boundary allows most samples to be confidently correct, leaving only a small number of uncertain and error-prone cases near the boundary. Therefore, in the label-free setting, if an analogous mechanism can derive a reliable boundary region that leaves most likely correct cases outside the region and concentrates refinement on the ambiguous cases within it, comparable performance may be achievable. To this end, two key challenges should be explicitly addressed.

C1: How to efficiently derive a reliable boundary region from MLLM outputs. In a label-free setting, textual MLLM outputs provide only discrete verdicts, so they cannot indicate how close a case is to the decision boundary. Next-token probabilities provide a more suitable continuous signal; in binary detection, the default 0.5 threshold naturally separates hateful and normal videos. However, raw MLLM probabilities are often poorly calibrated (Guo et al., 2017) and may be sensitive to distributional shifts. As a result, the boundary induced by this threshold can be unreliable, further making it unclear how to derive a reliable boundary region for additional action.

C2: How to allocate additional computation near the boundary. Once the boundary region is identified, the samples within it become the main opportunity for improving effectiveness. Prior work (Chen et al., 2023; Verga et al., 2024) suggests that different MLLMs can exhibit complementary strengths and instance-dependent failures, making diverse model panels useful for resolving hard cases. However, the benefit of additional computation remains case-dependent even within the boundary region (Geifman and El-Yaniv, 2017;

Schuster et al., 2021, 2022). Some cases can be resolved after querying only a few additional MLLMs, whereas others remain ambiguous and require more evidence from the panel. A label-free detector therefore needs an adaptive mechanism that queries additional MLLMs as needed and stops once the decision becomes sufficiently stable.

**Our approach.** We introduce a novel framework named TRIAGE for label-free hateful video detection, consisting of two key modules: the *Boundary Mapper* and the *Adaptive Boundary Resolver*. The *Boundary Mapper* first uses a lightweight MLLM to extract a continuous hate-saliency score for each video, then calibrates the raw scores to localize a boundary region with entropy guidance, allowing confident cases to be finalized immediately and ambiguous cases to receive additional computation. The *Adaptive Boundary Resolver* then refines cases within the boundary region by sequentially querying a set of diverse stronger MLLMs. It incorporates each verifier output through a Bayesian update in posterior probability space and stops once the accumulated evidence moves the case outside the boundary region. In this way, TRIAGE operates as a coarse-to-fine paradigm that first localizes a reliable boundary region and then adaptively refines only the ambiguous cases within it, rather than uniformly applying expensive reasoning to every video. Fig. 1 contrasts TRIAGE with the current landscape of hateful video detection, showing that our method avoids annotation cost while achieving strong performance with substantially less computation.

**Contributions.** We make four contributions:

- **A coarse-to-fine paradigm for label-free multimodal moderation.** We recast label-free hateful video detection as a boundary region mapping and resolution problem: a lightweight mapper localizes a reliable boundary region, and an adaptive resolver concentrates additional reasoning on cases within it.
- **A lightweight mapper for boundary region localization.** The Boundary Mapper extracts normalized next-token probabilities from a lightweight MLLM, calibrates the unlabeled score distribution, and localizes a reliable boundary region for selective refinement.
- **An adaptive resolver for boundary region refinement.** The Adaptive Boundary Resolver

refines cases within the boundary region by querying a diverse panel of stronger MLLMs, incorporating each verifier output through Bayesian sequential updates, and stopping once the decision becomes sufficiently stable.

- **Extensive empirical evaluation.** Across four public hateful video datasets, TRIAGE improves accuracy over the best label-free/few-shot baseline by 2.5–8.7 points, reduces parameter-weighted inference cost by at least  $2.0\times$ , and outperforms the best supervised baseline on three datasets.

## 2 Method

**Problem Formulation.** Let  $\mathcal{V}$  denote a collection of videos from a target moderation stream (also referred to as the test set). The task of hateful video detection is to assign each video  $v \in \mathcal{V}$  a binary prediction  $\hat{y}(v) \in \{0, 1\}$ , where 1 denotes hateful content and 0 denotes normal content. In the label-free setting, no sample from any set is associated with ground-truth annotations.

**Overview.** TRIAGE consists of two modules (Fig. 2): a *Boundary Mapper* and an *Adaptive Boundary Resolver*. The Boundary Mapper (§2.1) extracts a continuous hate-saliency score from each video with a lightweight MLLM, calibrates it into a posterior probability, and uses posterior entropy to localize a boundary region. Cases outside the boundary region are finalized by the mapper, whereas cases within the region retain their posterior proxy as an initial state and are passed to the Adaptive Boundary Resolver (§2.2), which validates them using a panel of stronger MLLMs until the posterior uncertainty is resolved.

### 2.1 Boundary Mapper

The Boundary Mapper is designed to make lightweight MLLM outputs usable for boundary region localization. Raw MLLM confidence may not provide a reliable decision signal under label-free deployment, so the mapper converts answer probabilities into a calibrated posterior signal. It then uses posterior entropy to identify the boundary region: cases outside this region are finalized immediately, while cases inside it are passed to the Adaptive Boundary Resolver.

Specifically, we instruct the lightweight MLLM with a binary moderation prompt formulated as:

“Given the following platform policy<sup>1</sup>  $G$ , does the given video  $v$  violate any of these? Please answer with Yes or No only.” For each video  $v$ , the lightweight MLLM then produces logits  $\ell_{\text{Yes}}(v)$  and  $\ell_{\text{No}}(v)$  for the candidate answer tokens. We define the hate-saliency score as the locally normalized next-token probability of Yes:

$$s(v) = \frac{\exp(\ell_{\text{Yes}}(v))}{\exp(\ell_{\text{Yes}}(v)) + \exp(\ell_{\text{No}}(v))}. \quad (1)$$

Directly relying on  $s(v)$  for subsequent operations is suboptimal, as the scores produced by the MLLM are often poorly calibrated and may not reflect true posterior probabilities, especially under distributional shift and in the absence of labeled data (Guo et al., 2017). To address this issue, we fit a two-component Gaussian Mixture model ( $f_{\text{GMM}}(\cdot)$ ) to the transformed scores and use the resulting posterior as a more reliable proxy for decision making; the mixture fitting follows the standard incomplete-data maximum-likelihood view underlying EM-style estimation (Dempster et al., 1977). The procedure is formally defined as:

$$q(v) = f_{\text{GMM}}\left(\log \frac{s(v)}{1 - s(v)}\right), \quad (2)$$

where  $q(v)$  denotes the posterior probability of belonging to the hateful component. More generally, this calibration step can be instantiated with other unsupervised algorithms from the family of posterior-producing two-component density decomposition methods (Corduneanu and Bishop, 2001; Peel and McLachlan, 2000; Cord et al., 2006). By default, the model is fitted on the unlabeled training set.

Subsequently, to determine which predictions can be finalized immediately and which require an additional round of review, we leverage  $q(v)$  to compute the entropy for each sample and then obtain the mean entropy across the dataset (Shannon, 1948). Samples with higher entropy are regarded as belonging to the boundary region and are routed to the Adaptive Boundary Resolver. This process

<sup>1</sup>We use policies from YouTube for English videos and Bilibili for Chinese videos; the full policies are in Appendix C.

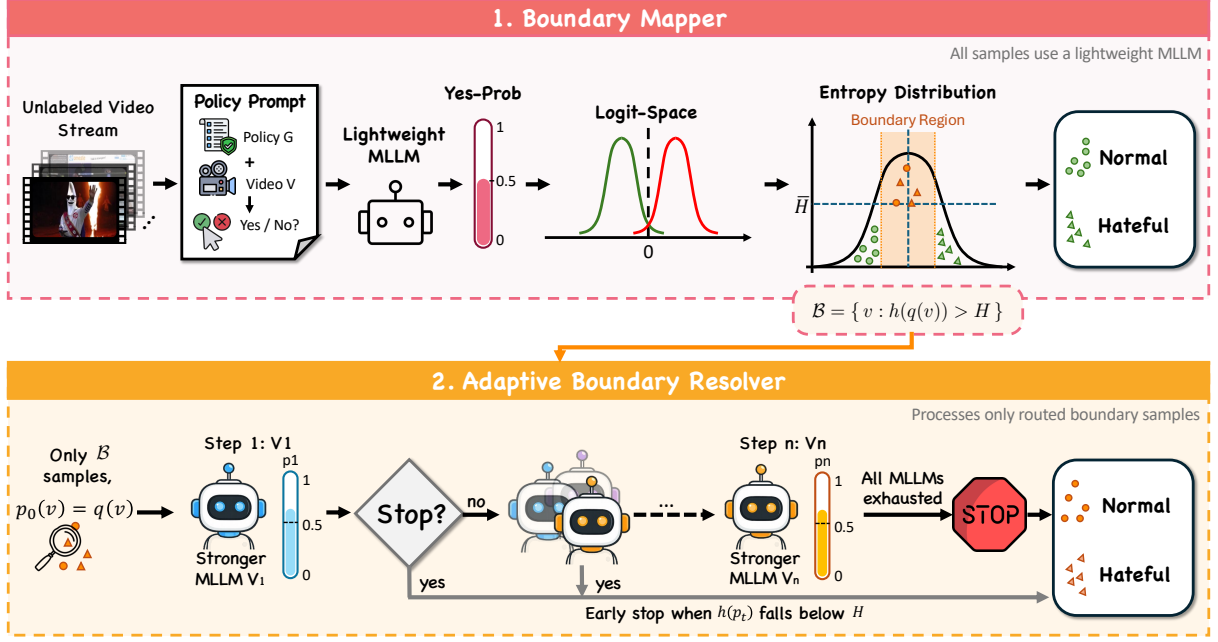


Figure 2: TRIAGE pipeline. The *Boundary Mapper* first utilizes a lightweight MLLM to produce calibrated initial predictions and identify a boundary region. Samples within this region are subsequently passed to the *Adaptive Boundary Resolver*, which queries a set of diverse and stronger MLLMs and performs Bayesian sequential updates to refine the posterior probability, terminating once sufficient evidence has been accumulated.

is formally defined as:

$$\begin{aligned}
 h(v) &= -q(v) \log q(v) \\
 &\quad - (1 - q(v)) \log(1 - q(v)), \\
 \bar{H} &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} h(v), \\
 \mathcal{B} &= \{v \in \mathcal{V} : h(v) > \bar{H}\}.
 \end{aligned} \tag{3}$$

where  $\mathcal{B}$  denotes the set of samples routed for further review. This boundary region captures highly uncertain samples because entropy is maximized when  $q(v)$  is close to 0.5, indicating that the model assigns comparable likelihood to both classes. In contrast, confident predictions (i.e.,  $q(v)$  close to 0 or 1) yield low entropy and are therefore excluded from the boundary region and directly finalized by the mapper using the 0.5 threshold on the posterior proxy, i.e.,  $\hat{y}^{(1)}(v) = \mathbb{I}[q(v) \geq 0.5]$ ,  $v \notin \mathcal{B}$ .

## 2.2 Adaptive Boundary Resolver

The Adaptive Boundary Resolver refines cases within the mapped boundary region, where additional model evidence can improve decisions but should not be spent uniformly. It selectively queries a diverse panel of stronger MLLMs, uses each verifier output to update the Boundary Mapper posterior, and stops once the posterior exits the boundary region or the panel is exhausted. The diverse

panel is intended to reduce correlated errors across MLLMs and exploit complementary model behaviors, following best practices in crowdsourced aggregation and recent panel-based LLM evaluation (Dawid and Skene, 1979; Verga et al., 2024).

**Verifier Effect.** The panel must specify the extent to which a single verifier’s judgment updates the posterior obtained from the Boundary Mapper. Since label-informed prior knowledge is unavailable, verifier-specific reliability cannot be directly estimated. Instead, we use a shared verifier effect to define a common evidence strength for all verifier responses. This effect should satisfy two requirements: (1) it should be strong enough for consistent responses to reinforce the current posterior direction; (2) it should be bounded enough that contradictory responses reduce posterior confidence without immediately overwriting it, leaving unresolved cases for further verification.

Accordingly, we derive a shared verifier effect based on the Boundary Mapper’s posterior entropy distribution. Specifically, we treat each verifier response as a unit of external evidence and set its strength so that a posterior at the decision boundary (i.e.,  $q(v) = 0.5$ ) is shifted to the edge of the boundary region, defined as:

$$\rho = \max\{u \in [0.5, 1] : h(u) = \bar{H}\}, \tag{4}$$

Table 1: Test set statistics of the datasets.

| Dataset      | Hateful/Normal | Dur. (s) | Lang. | Source   |
|--------------|----------------|----------|-------|----------|
| HateMM       | 86 / 129       | 150.0    | En    | BitChute |
| MHClip-EN    | 49 / 112       | 33.8     | En    | YouTube  |
| MHClip-ZH    | 45 / 104       | 31.8     | Zh    | Bilibili |
| ImpliHateVid | 200 / 201      | 155.1    | En    | BitChute |

where  $\rho$  denotes the largest posterior value whose entropy equals the entropy threshold defining the boundary region, and is used as the verifier effect for each MLLM in the resolver.

**Bayesian Sequential Update.** For each  $v \in \mathcal{B}$  at step  $t$ , an MLLM verifier in the panel receives the video together with a chain-of-thought prompt<sup>2</sup> to perform in-depth reasoning and outputs a binary verdict  $r_t(v) \in \{0, 1\}$  where 1 corresponds to hateful content. We define the Bayesian sequential update in a multiplicative odds form as:

$$a_t(v) = \begin{cases} \frac{\rho}{1-\rho}, & r_t(v) = 1, \\ \frac{1-\rho}{\rho}, & r_t(v) = 0, \end{cases} \quad (5)$$

$$p_t(v) = \frac{a_t(v)p_{t-1}(v)}{a_t(v)p_{t-1}(v) + 1 - p_{t-1}(v)},$$

where  $a_t(v)$  corresponds to the likelihood ratio under a symmetric noisy annotator model with reliability  $\rho$ , yielding a standard Bayesian odds update related to classical noisy-label aggregation (Dawid and Skene, 1979) that produces the updated posterior score  $p_t(v)$ . The initial posterior is set as  $p_0(v) = q(v)$ , where  $q(v)$  is obtained from the Boundary Mapper.

The process stops when the updated posterior exits the boundary region, i.e.,  $h(p_t(v)) < \bar{H}$ , or when all MLLMs in the panel have been exhausted. Otherwise, the same sample  $v$  is passed to the next MLLM in the panel to perform posterior updates according to Eq. 5. The final prediction for this sample is given by:

$$\hat{y}(v) = \mathbb{I}[p_T(v) \geq 0.5], \quad (6)$$

where  $T$  denotes the final stopping time step.

### 3 Experimental Evaluation

#### 3.1 Experiment Setup

**Datasets and Metrics.** We evaluate on four publicly released hateful video detection datasets: HateMM (Das et al., 2023), MHClip-EN, MHClip-ZH (Wang et al., 2024), and ImpliHateVid

(Rehman et al., 2025). We use the dataset splits released by previous work (Lang et al., 2025; Rehman et al., 2025) and report accuracy (ACC), macro-F1 (M-F1), macro precision (M-P), and macro recall (M-R). Detailed dataset statistics are provided in Table 1, and full dataset descriptions are provided in Appendix A.

**Baselines.** We compare our method against state-of-the-art baselines from two groups: 1) *Supervised Baselines* includes BERT (Devlin et al., 2019), ViViT (Arnab et al., 2021), MFCC (Davis and Mermelstein, 1980), Pro-Cap (Cao et al., 2023), HateMM (Das et al., 2023), MHCL (Wang et al., 2024), MoRE (Lang et al., 2025), ImpliHateVid (Rehman et al., 2025), and CMFusion (Zhang et al., 2024), and 2) *Label-free/Few-shot Baselines* includes MiniCPM-V (Yao et al., 2025), LLaVA-OV<sub>7B</sub> (Li et al., 2024), Qwen2.5-VL<sub>7B</sub> (Bai et al., 2025b), Qwen3-VL<sub>2B</sub> (Bai et al., 2025a), InternVL3.5<sub>8B</sub> (Wang et al., 2025), MARS (Yang et al., 2026), Mod-HATE (Cao et al., 2024), LOReHM (Huang et al., 2024), and ALARM (Lang et al., 2026). Detailed baseline descriptions and implementations are provided in Appendix B.

**Implementation details.** Unless otherwise stated, the Boundary Mapper uses Qwen3-VL<sub>2B</sub>. The default verifier order in the Adaptive Boundary Resolver is Gemma-3<sub>27B</sub>, Qwen2.5-VL<sub>32B-AWQ</sub>, and Qwen2.5-VL<sub>72B-AWQ</sub>, arranged in ascending order of computational cost. We set the temperature to zero for reproducibility. Further implementation details are provided in Appendix C.

#### 3.2 Comparison with SOTA

Table 2 reports the main comparison between TRIAGE and existing hateful video detection baselines on four benchmarks. Supervised multimodal models remain strong but require sizable annotated data on each target distribution, while label-free / few-shot baselines reduce this burden but still lag behind and vary across datasets. In contrast, TRIAGE is the strongest method within the label-free / few-shot regime on every dataset, improving over the best such baseline by +7.0 ACC on HateMM, +3.1 on MHClip-EN, +8.7 on MHClip-ZH, and +2.5 on ImpliHateVid. Compared with supervised baselines, TRIAGE achieves higher accuracy on HateMM, MHClip-EN, and MHClip-ZH, and remains comparable to most supervised baselines on ImpliHateVid. This robustness comes from

<sup>2</sup>Full prompt is shown in Fig. 8.

Table 2: Performance comparison on four hateful video detection datasets. Supervised-group best cells are shaded ; the best label-free / few-shot baseline cells are underlined. The last two rows report the improvement of TRIAGE over the best label-free / few-shot baseline and the best supervised method, respectively.

| Method                                   | HateMM      |             |             |             | MHClip-EN   |      |             |      | MHClip-ZH   |             |             |             | 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  |
| <b>① Supervised Baselines</b>            |             |             |             |             |             |      |             |      |             |             |             |             |              |      |      |      |
| 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 |
| 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        | 87.5         | 87.5 | 87.6 | 87.5 |
| 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        | 68.5         | 68.4 | 68.9 | 68.6 |
| <b>② Label-free / Few-shot Baselines</b> |             |             |             |             |             |      |             |      |             |             |             |             |              |      |      |      |
| 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 <sub>7B</sub>                   | 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 |
| Qwen2.5-VL <sub>7B</sub>                 | 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 |
| Qwen3-VL <sub>2B</sub>                   | 67.4        | 58.7        | 71.3        | 60.7        | 71.4        | 52.9 | 67.8        | 55.4 | 74.5        | 58.2        | 77.9        | 59.0        | 55.6         | 45.7 | 69.9 | 55.5 |
| InternVL3.5 <sub>8B</sub>                | 77.9        | 76.4        | 77.6        | 75.9        | 72.1        | 62.9 | 76.3        | 60.2 | 72.5        | 66.6        | 80.1        | 63.1        | 71.9         | 70.7 | 76.3 | 71.8 |
| MARS                                     | 69.8        | 69.6        | 75.9        | 74.0        | 65.8        | 62.8 | 62.7        | 64.5 | 73.8        | 46.8        | 46.7        | 47.0        | 80.3         | 80.3 | 80.4 | 80.3 |
| Mod-HATE                                 | 77.7        | 76.3        | 77.0        | 76.0        | 70.2        | 59.4 | 63.0        | 59.1 | 38.3        | 38.2        | 46.2        | 46.3        | 71.0         | 69.8 | 74.8 | 70.9 |
| LoReHM                                   | 79.1        | 78.9        | 79.4        | 80.6        | 76.4        | 67.3 | 74.0        | 65.8 | 74.5        | 62.2        | 71.9        | 61.6        | 77.0         | 76.9 | 77.6 | 77.0 |
| ALARM                                    | 79.5        | 79.4        | 79.8        | 81.0        | 69.6        | 60.2 | 62.4        | 59.8 | 71.1        | 47.5        | 73.0        | 52.9        | 71.8         | 69.7 | 79.2 | 71.6 |
| <b>TRIAGE (ours)</b>                     | <b>86.5</b> | <b>85.8</b> | <b>86.2</b> | <b>85.5</b> | <b>79.5</b> | 73.2 | <b>77.3</b> | 71.5 | <b>83.9</b> | <b>81.9</b> | <b>80.8</b> | <b>84.0</b> | 82.8         | 82.7 | 83.6 | 82.8 |
| $\Delta$ vs. ②                           | +7.0        | +6.4        | +6.4        | +4.5        | +3.1        | +2.7 | +1.0        | +4.1 | +8.7        | +7.9        | +0.7        | +11.0       | +2.5         | +2.4 | +3.2 | +2.5 |
| $\Delta$ vs. ①                           | +3.1        | +3.4        | +4.1        | +2.2        | +2.0        | -2.0 | +1.6        | -3.3 | +5.4        | +7.1        | +5.1        | +9.9        | -4.7         | -4.8 | -4.0 | -4.7 |

Table 3: Ablation study (%).  $\bar{\Delta}$  denotes the mean ACC drop against TRIAGE across four datasets.

|                                   | HateMM      |             | MH-EN       |             | MH-ZH       |             | ImpliHV     |             | $\bar{\Delta}$ |
|-----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------------|
|                                   | ACC         | M-F1        | ACC         | M-F1        | ACC         | M-F1        | ACC         | M-F1        |                |
| <b>TRIAGE</b>                     | <b>86.5</b> | <b>85.8</b> | <b>79.5</b> | <b>73.2</b> | <b>83.9</b> | <b>81.9</b> | <b>82.8</b> | <b>82.7</b> | —              |
| <i>Boundary Mapper</i>            |             |             |             |             |             |             |             |             |                |
| w/o Calibration                   | 84.2        | 83.1        | 72.0        | 58.5        | 80.5        | 70.8        | 70.8        | 67.7        | -6.3           |
| w/o Selection                     | 86.0        | 85.4        | 78.9        | 72.6        | 79.9        | 77.4        | 77.3        | 76.7        | -2.7           |
| w/ Random                         | 83.3        | 82.5        | 77.3        | 69.1        | 82.3        | 79.6        | 80.2        | 80.2        | -2.4           |
| <i>Adaptive Boundary Resolver</i> |             |             |             |             |             |             |             |             |                |
| w/o Verification                  | 80.5        | 79.3        | 76.4        | 65.3        | 79.2        | 75.8        | 81.8        | 81.8        | -3.7           |
| w/ Self-Con                       | 81.9        | 81.1        | 76.4        | 67.3        | 81.2        | 78.5        | 81.5        | 81.4        | -2.9           |
| w/ Majority Voting                | 86.0        | 85.3        | 77.0        | 69.0        | 81.9        | 79.6        | 82.8        | 82.7        | -1.3           |
| w/o Verifier Effect               | 83.0        | 82.3        | 77.6        | 70.1        | 82.6        | 80.4        | 80.8        | 80.7        | -2.2           |
| w/o Early Stop                    | 85.5        | 84.8        | 77.0        | 69.0        | 81.9        | 79.6        | 81.5        | 81.6        | -1.7           |

the coarse-to-fine design: the Boundary Mapper obtains a label-free posterior and passes samples within the boundary region to the Adaptive Boundary Resolver, which concentrates MLLM reasoning where the mapper is most likely to fail.

### 3.3 Ablation Study

Table 3 evaluates variants for both modules. For the *Boundary Mapper*, w/o *Selection* sends all samples to the verifier panel, w/o *Calibration* uses the raw output score, and w/ *Random* keeps the same selection budget but chooses samples randomly; the drops show that the mapper is not merely a cost-

saving device, since verifying all samples can hurt and calibration makes boundary-region selection reliable. For the *Adaptive Boundary Resolver*, w/o *Verification* uses the Boundary Mapper prediction, w/ *Self-Con* uses only Gemma-3<sub>27B</sub> with three self-consistency votes, w/ *Majority Voting* replaces posterior updating with fixed three-verifier voting, w/o *Verifier Effect* removes calibrated evidence scaling, and w/o *Early Stop* forces all boundary-region samples to use the full panel; their consistent degradation shows that heterogeneous verifiers, calibrated sequential updates, and early termination all contribute to performance.

### 3.4 Boundary Region Analysis

We analyze the boundary region produced by the Boundary Mapper by comparing its size, the erroneous predictions it captures, and inside-region versus outside-region accuracy in Fig. 3. The boundary region captures a substantially larger proportion of errors than its size, indicating a high signal-to-noise ratio for difficult samples. Its inside-region accuracy is also much lower than outside-region accuracy, showing that the decision boundary is well localized and that the boundary region provides a reliable signal for allocating downstream computation.

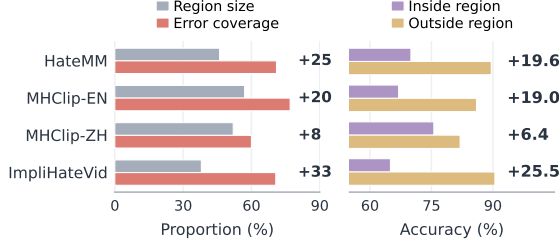


Figure 3: Boundary region analysis. Left: Boundary region sample share and mapper-error coverage. Right: Inside- versus outside-region accuracy.

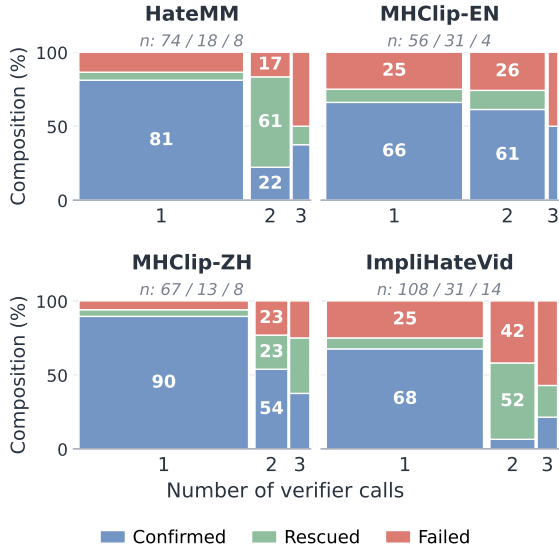


Figure 4: Boundary resolving analysis. Each bar shows the outcome composition of samples stopping after one, two, or three verifier calls: confirmed (Boundary Mapper and final predictions are correct), rescued (verification corrects a mapper error), and failed (final prediction remains incorrect).

### 3.5 Boundary Resolver Analysis

We next examine whether the Adaptive Boundary Resolver allocates the minimal sufficient computation to boundary region samples.

Fig. 4 decomposes each stopping bucket into three outcomes. Most boundary region samples stop after a single verifier and are dominated by confirmed-correct decisions, indicating that easy agreements are finalized early. Two-call buckets contain a larger share of rescued cases, showing that extra computation helps resolve conflicting boundary samples. Samples requiring all three verifiers have higher failure rates, suggesting genuinely hard cases that may require manual review. This pattern is consistent across datasets and supports the intended adaptive behavior.

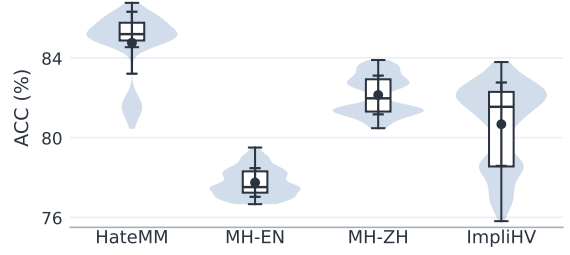


Figure 5: Generalizability analysis across 600 model configurations.

### 3.6 Generalizability

To verify that the proposed paradigm is generalizable and not overly sensitive to specific model choices, we conduct a model-space robustness study across both modules. For the **Boundary Mapper**, the candidate pool contains Qwen3-VL<sub>2B</sub>, Qwen2.5-VL<sub>7B</sub>, MiniCPM-V<sub>2.6</sub>, Pixtral<sub>12B</sub>, and Gemma-3<sub>12B</sub>. For the **Adaptive Boundary Resolver**, the verifier pool contains Gemma-3<sub>27B</sub>, Qwen2.5-VL<sub>32B</sub>, Qwen2.5-VL<sub>72B</sub>, Qwen3-VL<sub>8B</sub>, InternVL3.5<sub>8B</sub>, and Gemma-3<sub>12B</sub>. Crossing the five mapper candidates with all ordered three-verifier panels from this pool yields 600 model configurations.

Fig. 5 reports the final accuracy distribution across these configurations. The distributions are compact and the mean performance remains consistently high, suggesting that TRIAGE does not rely heavily on a specific MLLM choice or ordering. The mapper-resolver paradigm remains robust as long as the Boundary Mapper provides a usable label-free saliency distribution and the Adaptive Boundary Resolver offers diverse verifier judgments for samples within the boundary region.

### 3.7 Probabilistic Reliability Analysis

Beyond hard-label accuracy, we examine whether TRIAGE can work reliably in probability space. We use Brier score (Brier, 1950), the squared error between predicted probability and binary ground truth, where lower is better. Fig. 6 shows that TRIAGE obtains the lowest Brier score on all four datasets, indicating that its final posterior scores are well aligned with ground truth. This observation also suggests that most early-finalized cases receive reliable confidence, while harder cases benefit from accumulated verifier evidence and obtain more trustworthy posterior estimates.

|             | HateMM       | MH-EN        | MH-ZH        | ImpliHV      |
|-------------|--------------|--------------|--------------|--------------|
| BERT        | 0.320        | 0.369        | 0.356        | 0.214        |
| ViViT       | 0.389        | 0.397        | 0.395        | 0.342        |
| MFCC        | 0.344        | 0.485        | 0.420        | 0.342        |
| Pro-Cap     | 0.297        | 0.335        | 0.323        | 0.198        |
| MHCL        | 0.302        | 0.457        | 0.403        | 0.227        |
| HateMM      | 0.276        | 0.355        | 0.336        | 0.216        |
| CMFusion    | 0.289        | 0.343        | 0.323        | 0.212        |
| <b>Ours</b> | <b>0.126</b> | <b>0.214</b> | <b>0.153</b> | <b>0.159</b> |

Figure 6: Probabilistic reliability analysis. Lower Brier scores indicate better probability estimates.

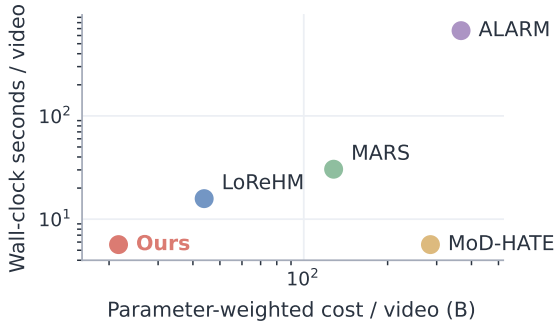


Figure 7: Efficiency analysis. Each point denotes the end-to-end cost per test video; both axes are log-scaled.

### 3.8 Efficiency Analysis

Finally, we evaluate efficiency against label-free and few-shot LLM baselines using parameter-weighted model cost and wall-clock time. The former sums the parameter sizes of invoked models and normalizes by the number of test videos, while the latter measures average end-to-end runtime per sample. Fig. 7 shows that uniform multi-stage methods occupy the high-cost region because they apply expensive reasoning broadly. In contrast, TRIAGE stays near the lower-left region: every video first receives one lightweight Boundary Mapper call, and only samples within the boundary region consume additional verifier calls with early stopping.

## 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 hateful video detection. Current state-of-the-art performance in this task is predominantly achieved by supervised multimodal hate detection methods

(Das et al., 2023; Wang et al., 2024; Lang et al., 2025; Rehman et al., 2025; Li et al., 2026; Zhang et al., 2024). These approaches typically design fusion networks to integrate textual, visual, and acoustic features from annotated training data to learn a discriminative decision boundary. While highly effective, their reliance on large-scale annotated data incurs substantial annotation cost. To alleviate this issue, the community has explored hateful video detection under low-resource settings. Notably, such work remains limited in the video domain, and related tasks such as harmful meme detection are often considered. With the emergence of LLMs, their strong reasoning capability has been successfully applied across diverse domains (Chen et al., 2026b, 2025, 2026a; Lu et al., 2026b,a; Jia et al., 2026; Liu et al., 2025), motivating low-resource hate detection methods built on MLLM reasoning. For example, MARS (Yang et al., 2026) formulates the problem as a multi-stage reasoning pipeline in a label-free setting. LoReHM (Huang et al., 2024) and ALARM (Lang et al., 2026) employ MLLMs as agents to perform iterative self-evolving reasoning with accumulated evidence, while Mod-HATE (Cao et al., 2024) adopts  $k$ -shot learning with LoRA-based fine-tuning. Despite these advances, such methods typically incur substantial computational overhead and still fall significantly short of supervised performance. In this work, we address this effectiveness-efficiency bottleneck by proposing a coarse-to-fine detection framework. Our approach first targets the decision boundary in a label-free manner, analogous to the role of annotated data in supervised learning, through calibrated thresholding and an associated boundary region (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; Guo et al., 2017), and then resolves ambiguous cases using a diverse panel of MLLMs with minimal sufficient computation (Dawid and Skene, 1979; Verga et al., 2024).

## 5 Conclusion

We introduced TRIAGE, a coarse-to-fine label-free framework for hateful video detection. We show that effective label-free moderation does not require expensive reasoning for every input to achieve competitive performance. Three key findings are summarized as follows: (1) the next-token probabilities of a lightweight MLLM, after unsupervised calibration, provide a reliable decision

signal whose boundary region captures a disproportionate share of errors; (2) concentrating additional reasoning exclusively on boundary-region samples is superior to uniform reasoning in terms of both effectiveness and efficiency; and (3) cross-model Bayesian sequential verification with early stopping can effectively resolve hard cases without complex prompt engineering.

## Limitations

Although TRIAGE improves label-free hateful video detection by localizing and refining uncertain cases, it still depends on the underlying perception and reasoning ability of MLLMs. When hateful intent is expressed through highly implicit visual cues, coded language, sarcasm, cultural references, or external context that is not observable from the video itself, the framework may still produce incorrect predictions. Such cases are especially challenging because additional verifier queries cannot recover evidence that is missing from the model inputs or beyond the models' understanding.

TRIAGE also assumes that the initial MLLM scores provide a meaningful uncertainty structure for boundary localization. This assumption may be less reliable under extremely small, highly imbalanced, or adversarial target streams, where the score distribution may not form a clear separable pattern.

## Ethics Statement

This work studies hateful video detection, a task that aims to support safer online environments by identifying harmful or discriminatory content. All experiments are conducted on publicly released benchmarks, and we do not collect new user data or introduce additional human annotation. Since the evaluated datasets contain hateful, offensive, or otherwise disturbing content, researchers who reproduce or extend this work should follow appropriate data-handling practices and avoid unnecessary exposure to such material.

Our framework is intended to reduce the annotation and computation costs of hateful video moderation, rather than to serve as a fully autonomous moderation system. Automated detectors may produce false positives that over-suppress legitimate expression or false negatives that fail to identify harmful content. Therefore, we recommend using TRIAGE as a decision-support tool, especially for prioritizing uncertain or difficult cases for further

review, rather than as the sole basis for enforcement decisions in high-stakes settings.

There is also a potential risk that moderation technologies could be misused for overly broad content control or applied under policies that are not transparent to users. To mitigate this risk, deployment should be accompanied by clear platform policies, human oversight, appeal mechanisms, and regular auditing across different languages, communities, and content distributions. We hope this work contributes to more efficient and responsible moderation while acknowledging that technical systems alone cannot resolve the broader social and policy challenges involved in online harm detection.

## References

- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. [ViViT: A video vision transformer](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6816–6826.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 1 others. 2025a. [Qwen3-VL technical report](#). *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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 9 others. 2025b. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Glenn W. Brier. 1950. [Verification of forecasts expressed in terms of probability](#). *Monthly Weather Review*, 78(1):1–3.
- Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. [ProCap: Leveraging a frozen vision-language model for hateful meme detection](#). In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252. ACM.
- Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. [Modularized networks for few-shot hateful meme detection](#). In *Proceedings of the ACM Web Conference 2024*, pages 4575–4584. ACM.
- 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.
- Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2023. Large language models are visual reasoning coordinators. In *Advances in Neural Information Processing Systems*, volume 36, pages 70115–70140.
- Aurélien Cord, Christophe Ambroise, and Jean-Pierre Cocquerez. 2006. Feature selection in robust clustering based on Laplace mixture. *Pattern Recognition Letters*, 27(6):627–635.
- Adrian Corduneanu and Christopher M. Bishop. 2001. Hyperparameters for soft Bayesian model selection. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pages 63–70. PMLR.
- 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*, 17:1014–1023.
- 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*, 28(4):357–366.
- A. Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- 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: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.
- 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.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Jianzhao Huang, Hongzhan Lin, Liu Ziyang, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with LMM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- 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*, volume 33, pages 2611–2624.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pages 611–626. ACM.
- 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 2025*, pages 2763–2774. ACM.
- Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. 2026. From shallow humor to metaphor: Towards label-free harmful meme detection via LMM agent self-improvement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pages 588–599. ACM.
- 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 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*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- 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*, pages 1–1.
- 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.
- 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.
- D. Peel and G. J. McLachlan. 2000. [Robust mixture modelling using the  \$t\$  distribution](#). *Statistics and Computing*, 10(4):339–348.
- 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- 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*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Dr. 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 (Volume 1: Long Papers)*, pages 17209–17221, Vienna, Austria. Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58. Association for Computational Linguistics.
- Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. [Confident adaptive language modeling](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. 2021. [Consistent accelerated inference via confident adaptive transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4962–4979, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. [Replacing judges with juries: Evaluating LLM generations with a panel of diverse models](#). *arXiv preprint arXiv:2404.18796*.
- Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024. [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.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, and 1 others. 2025. [InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *arXiv preprint arXiv:2508.18265*.

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, Chi Chen, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Ronghua Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, and 5 others. 2025. [Efficient GPT-4V level multimodal large language model for deployment on edge devices](#). *Nature Communications*, 16(1):5509.

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.

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.

## A Dataset Details

**HateMM** (Das et al., 2023) is an English hateful video dataset sourced from BitChute, with each video annotated as Hate or Non-Hate.

**MHClip-EN and MHClip-ZH** (Wang et al., 2024) together form a multilingual short-clip dataset drawn from YouTube (English) and Bilibili (Chinese), with a shared three-way annotation scheme (Hateful, Offensive, Normal). Following common practice on this dataset (Lang et al., 2025; Li et al., 2026), we collapse the three-way labels into a binary setting by mapping Offensive to Hateful, and report results on this binary task. It is also noteworthy that the authors of these datasets do not directly provide a complete downloadable video package; instead, users are required to retrieve the original videos from the source platforms. However, a portion of the videos has become unavailable since the datasets were released.

**ImpliHateVid** (Rehman et al., 2025) is an English dataset designed to emphasize implicit hate, where the hateful intent is not explicitly stated and

must be inferred from context. The dataset is released with binary labels.

For all four datasets, we evaluate exclusively on the released test splits from prior work (Das et al., 2023; Lang et al., 2025; Wang et al., 2024; Rehman et al., 2025).

## B Baseline Descriptions

We compare TRIAGE against two categories of baselines: a supervised group that requires hateful video labels for training, and a label-free / few-shot group that requires no or only a small number of such labels.

### B.1 Supervised 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 the auto-transcribed speech of each video.
- **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 factorised 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, apply 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 of each video.
- **HateMM** (Das et al., 2023) is the supervised detector released alongside the HateMM benchmark. It encodes text with BERT, visual

content with a 3D-CNN, and audio with a pre-trained acoustic encoder, and fuses the three streams via cross-modal attention.

- **MHCL** (Wang et al., 2024) is the supervised baseline introduced alongside the MultiHate-Clip dataset. It aligns the 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 sharpen the 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.

## B.2 Label-free / Few-shot Baselines

### MLLM-based Baselines.

- **MiniCPM-V** (Yao et al., 2025) is a series of efficient MLLMs designed for end-side deployment.
- **LLaVA-OV<sub>7B</sub>** (Li et al., 2024) (LLaVA-OneVision) supports single-image, multi-image, and video understanding through a unified architecture.
- **Qwen2.5-VL<sub>7B</sub>** (Bai et al., 2025b) encodes images and videos at their native resolution via a dynamic resolution module and Multimodal Rotary Position Embedding (M-RoPE).
- **Qwen3-VL<sub>2B</sub>** (Bai et al., 2025a) is the latest open-weights model in the Qwen-VL family, providing improved video and long-context understanding.
- **InternVL3.5<sub>8B</sub>** (Wang et al., 2025) is an open-weights multimodal model with a video-aware pre-training stage and progressive context-length scaling.

### Multi-stage MLLM Pipelines.

- **MARS** (Yang et al., 2026) performs four sequential reasoning stages per video, in which an MLLM is repeatedly prompted to observe the content, generate evidence for and against the hateful hypothesis, and finalize the verdict.
- **Mod-HATE** (Cao et al., 2024) composes three task-specific LoRA modules over a frozen language-model backbone, with the composition weights optimized on a  $K$ -shot labeled support set. The original method targets memes; we adapt it to video by feeding sampled frame-level captions and the transcript as the textual surrogate.
- **LoReHM** (Huang et al., 2024) augments a base multimodal language model with retrieval over a labeled support pool and a disagreement-triggered re-ask pattern. The original method targets memes; we adapt it to video by feeding a uniformly sampled set of frames as the visual input and corresponding transcript as textual input.
- **ALARM** (Lang et al., 2026) iteratively self-improves an MLLM agent by pseudo-labeling confidently-explicit training samples and distill general hate-related knowledge that can be used at test time. It is developed for harmful meme detection, and we adapt it to the video setting.

All baselines strictly follow the train/validation/test splits released by previous work (Lang et al., 2025; Rehman et al., 2025) and adopt their published optimal configurations.

## C Implementation Details

**Data Preprocessing.** Following previous conventions, we uniformly sample 16 frames from each video and obtain the global transcript using the Whisper model (Radford et al., 2023). For the input to the MLLM, we use the sampled frames together with the concatenated video title and corresponding transcript.

**Backbones and Framework.** For *Boundary Mapper*, we use Qwen3-VL<sub>2B</sub>, while for *Adaptive Boundary Resolver*, we use Gemma-3<sub>27B</sub>, Qwen2.5-VL<sub>32B-AWQ</sub>, and Qwen2.5-VL<sub>72B-AWQ</sub> by default. All MLLMs are loaded from officially released pre-trained checkpoints and invoked sequentially with zero temperature. We run inference with

vLLM (Kwon et al., 2023) on a single NVIDIA A100 GPU with 80 GB memory.

**Prompt Template.** For *Boundary Mapper* with Qwen3-VL<sub>2B</sub>, we use the following prompt for fast binary scoring:

*“Given the following platform policy  $\mathcal{G}$ , does the given video  $\mathcal{V}$  violate any of these? Please answer with Yes or No only.”*

The platform policy is collected with respect to the source of each video: we use YouTube’s policy for English videos and Bilibili’s policy for Chinese videos. Their policies are summarized below.

### YouTube Policy Rules.

- YT-R1 **Encourage violence.** Encourage violence against individuals or groups based on their protected group status. This includes threats on YouTube, and implied calls for violence are treated as real threats.
- YT-R2 **Incite hatred.** Incite hatred against individuals or groups based on their protected group status.
- YT-R3 **Dehumanisation.** Dehumanise individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity based on their protected group status.
- YT-R4 **Praise or glorification of violence.** Praise or glorify violence against individuals or groups based on their protected group status.
- YT-R5 **Slurs and stereotypes.** Use racial, religious, or other slurs and stereotypes that incite or promote hatred based on protected group status. This can take the form of speech, text, or imagery promoting these stereotypes or treating them as factual.
- YT-R6 **Claims of inferiority.** Claim that individuals or groups are physically or mentally inferior, deficient, or diseased based on their protected group status. This includes statements that one group is less than another, calling them less intelligent or less capable, as well as calls for the subjugation or domination over individuals or groups based on their protected group status.

YT-R7 **Hateful supremacism.** Promote hateful supremacism by alleging the superiority of a group over those with protected group status to justify violence, discrimination, segregation, or exclusion. This includes content containing hateful supremacist propaganda, such as recruitment of new members or requests for financial support, and music videos promoting hateful supremacism in the lyrics, metadata, or imagery.

YT-R8 **Conspiratorial claims.** Make conspiratorial claims that individuals or groups are evil, corrupt, or malicious based on their protected group status.

YT-R9 **Denial of violent events.** Deny or minimise a well-documented, major violent event or the victimhood of such an event.

### Bilibili Policy Rules.

- BL-R1 **Discriminatory content.** Any content that includes discriminatory material or disparages any religion, belief system, race, sexual orientation, or protected group.
- BL-R2 **Attacks and disparagement.** Attacks, disparagement, or belittlement targeting a specific individual or group based on race, religion, gender, age, nationality, disability, sexual orientation, etc.
- BL-R3 **Verbal abuse and personal attacks.** Verbal abuse or personal attacks against a specific individual or group, antagonising others and inciting hatred.
- BL-R4 **Hateful ideologies.** Promoting hateful ideologies via active discussion or display of signs, symbols, flags, or slogans representing hateful organisations.
- BL-R5 **Mocking death and disability.** Content mocking death, sickness, or disability.
- BL-R6 **Disaster jokes.** Jokes about disasters and tragic social events.
- BL-R7 **Gender-based hate.** Hateful content directed toward a particular gender group, including misogynistic and misandrist content.
- BL-R8 **Inciting confrontation.** Content that maliciously provokes wars or conflicts, incites confrontations, or spreads hatred.

```

You are reviewing a video for content moderation.
Watch the video, read its title, and read any provided
transcript.
Title: {title}
Transcript: {transcript}
Hate definition: {dataset-specific hate
definition}
Examine the video carefully. Then output your deci-
sion in EXACTLY this two-field format on two lines,
with no additional fields, no extra prose, and no mark-
down:
rationale: <one paragraph (3–6 sentences)>
answer: <Yes or No>

```

Figure 8: Prompt template used by each MLLM verifier in the Adaptive Boundary Resolver.

For the *Adaptive Boundary Resolver*, we design a straightforward chain-of-thought prompt that instructs the MLLM to first provide a rationale justifying its assessment, followed by a final binary verdict. For each benchmark, we summarize the definition of hateful content and the annotation criteria described in the corresponding dataset paper (Das et al., 2023; Wang et al., 2024; Rehman et al., 2025), and insert the resulting dataset-specific definition into the verifier prompt. This setup also reflects real-world moderation settings, where different platforms may adopt different definitions of what constitutes hateful content. The prompt template is shown in Fig. 8.

## D Additional Experiments

### D.1 Robustness

#### D.1.1 Subsampling Robustness

The Boundary Mapper relies on an unlabeled score distribution to fit a calibration model and identify the corresponding boundary region. We therefore examine whether TRIAGE remains stable when this unlabeled fitting pool is reduced. For each dataset, we subsample the unlabeled training scores at ratios from 10% to 80%, fit the Boundary Mapper on the sampled pool, and evaluate the full pipeline on the fixed test set. We repeat each subsampling ratio with 30 random seeds.

Fig. 9 reports accuracy under different unlabeled-pool sizes. Although smaller pools introduce performance drop, especially at 10%, the full pipeline remains generally robust: performance does not collapse under aggressive subsampling, and the curves approach the full-pool reference as the unlabeled pool increases. This suggests that the Boundary Mapper does not require a large unlabeled cali-

bration set to provide useful routing signals for the Adaptive Boundary Resolver.

#### D.1.2 Calibration and Source Robustness

The Boundary Mapper uses an unlabeled score distribution to convert continuous MLLM saliency scores into a posterior decision signal. We evaluate whether this calibration step remains effective across different posterior-producing density decompositions and unlabeled fitting sources. Specifically, we compare several two-component calibration choices, including Bayesian Gaussian, Student-t, and Laplace mixture families (Corduneanu and Bishop, 2001; Peel and McLachlan, 2000; Cord et al., 2006), under both train set and test set fitting using only unlabeled scores. Otsu and Li-Lee are included as threshold-only variants: they set the Boundary Mapper hard threshold, while the boundary region is still computed from the GMM posterior.

The results are shown in Table 4. Across different calibration families, fitting sources, and thresholding variants, the full pipeline remains effective, showing that TRIAGE is not tied to a single implementation choice. This robustness supports boundary mapping and resolution as a general paradigm for label-free hateful video detection.

### D.2 Transferability Analysis

We further evaluate whether the overall TRIAGE framework can transfer across datasets under two unlabeled fitting settings. In the source-fit setting, for each source–target pair, we fit TRIAGE on the source training scores and apply the resulting pipeline directly to the target test set. In the target-fit setting, we fit TRIAGE directly on the unlabeled target test scores. We compare both settings with the strongest label-free/few-shot baseline for each target dataset.

Fig. 10 shows two findings. First, even without target test fitting, transferred configurations outperform the best label-free/few-shot baseline in most source–target directions, indicating that the framework generalizes beyond the dataset used for fitting. Second, target-fit TRIAGE further improves performance, showing that the framework can also exploit target-side score distributions without using labels. Together, these results highlight the transferability and generalizability of boundary mapping and resolution.

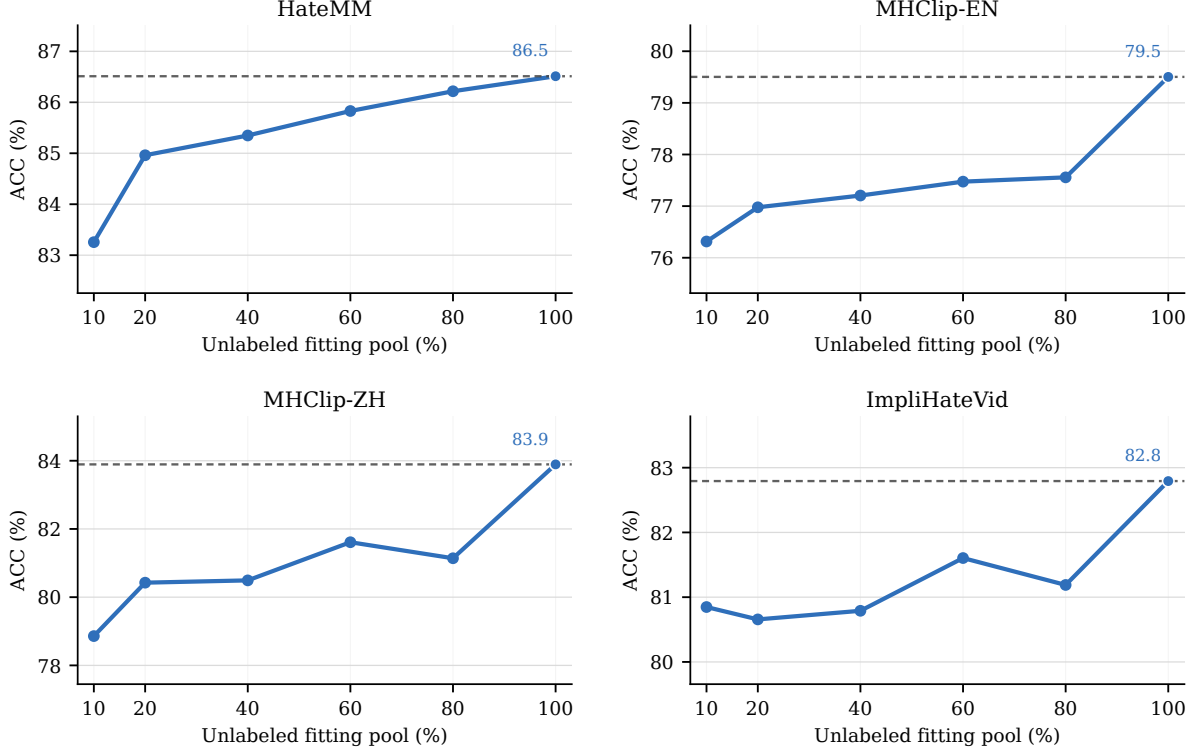


Figure 9: Unlabeled training subsampling robustness. Each point reports mean accuracy over 30 subsampling seeds, and the shaded region denotes  $\pm 1$  standard deviation.

### D.3 Generalizability to Label-Free/Few-Shot Verifier Backends

The generalizability test in the main text varies the MLLM verifier pool. Here we further ask whether the Adaptive Boundary Resolver can also use prior label-free and few-shot methods as verifier backends. We replace the verifier response with the hard judgment from MARS, Mod-HATE, LoReHM, or ALARM, while keeping the Boundary Mapper, boundary region selection, and posterior update unchanged. This setting tests whether TRIAGE can exploit external detector judgments in a backend-agnostic way and adapt to stronger detectors.

Fig. 11 reports accuracy on the four datasets, where we find that TRIAGE consistently improves prior label-free/few-shot backends. For each baseline backend, the dashed gray mark denotes its standalone performance, while the blue mark denotes the performance after wrapping the same backend with TRIAGE’s boundary region resolver. The orange reference further uses all four baselines as a sequential verifier pool; we exhaustively enumerate all 24 backend orders and report the mean with a one-standard-deviation band. Across all 16 dataset-backend pairs, the wrapped configuration improves over the corresponding standalone base-

line. The gains are especially clear for weaker or less stable baselines, such as MARS on HateMM and MHClip-EN, Mod-HATE on MHClip-ZH and ImpliHateVid, and ALARM on MHClip-ZH and ImpliHateVid. These results show that the Adaptive Boundary Resolver is not restricted to a fixed set of base MLLM verifiers: it can also turn existing label-free/few-shot detectors into stronger verifier backends.

### D.4 Verifier Pool Size in Adaptive Boundary Resolver

We next examine whether the Adaptive Boundary Resolver is sensitive to the number of verifier models. Starting from a six-model verifier pool as mentioned in Section 3.6, we vary the pool size  $K$  from 1 to 6. For each  $K$ , we exhaustively enumerate the verifier-pool configurations and their ordered verifier sequences. We compare TRIAGE with two voting references: MAJORITY VOTE directly ensembles the verifier pool over the test set, while BOUNDARY-REGION MAJORITY VOTE applies the vote only to instances within the boundary region and keeps the Boundary Mapper prediction otherwise.

We report the results in Fig. 12. Our design con-

Table 4: Full-pipeline robustness to the Boundary Mapper posterior transform, thresholding rule, and unlabeled source.

| Method                        | Source    | HateMM |      |      |      | MHClip-EN |      |      |      | MHClip-ZH |      |      |      | 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  |
| TRIAGE                        | train set | 86.5   | 85.8 | 86.2 | 85.5 | 79.5      | 73.2 | 77.3 | 71.5 | 83.9      | 81.9 | 80.8 | 84.0 | 82.8         | 82.7 | 83.6 | 82.8 |
|                               | test set  | 86.5   | 85.7 | 86.4 | 85.3 | 75.8      | 67.3 | 72.3 | 65.9 | 80.5      | 75.0 | 78.1 | 73.5 | 81.8         | 81.6 | 82.9 | 81.8 |
| TRIAGE <sub>BayesGMM</sub>    | train set | 83.7   | 83.4 | 83.1 | 84.1 | 75.8      | 68.8 | 71.5 | 67.7 | 81.9      | 77.5 | 79.2 | 76.3 | 84.3         | 84.3 | 84.3 | 84.3 |
|                               | test set  | 86.5   | 85.7 | 86.4 | 85.3 | 75.8      | 67.3 | 72.3 | 65.9 | 83.2      | 80.0 | 80.2 | 79.8 | 84.0         | 84.0 | 84.2 | 84.0 |
| TRIAGE <sub>Student-t</sub>   | train set | 84.7   | 84.0 | 84.0 | 84.1 | 73.3      | 62.6 | 68.5 | 61.9 | 81.9      | 80.1 | 79.1 | 83.2 | 84.3         | 84.3 | 84.5 | 84.3 |
|                               | test set  | 85.1   | 84.4 | 84.6 | 84.3 | 75.8      | 68.8 | 71.5 | 67.7 | 81.2      | 74.8 | 80.5 | 72.7 | 83.8         | 83.7 | 84.4 | 83.8 |
| TRIAGE <sub>Laplace</sub>     | train set | 84.7   | 84.0 | 84.0 | 83.9 | 77.0      | 70.9 | 73.1 | 69.7 | 81.9      | 80.1 | 79.1 | 83.2 | 83.8         | 83.7 | 84.4 | 83.8 |
|                               | test set  | 84.7   | 84.0 | 84.0 | 83.9 | 74.5      | 68.6 | 69.7 | 67.9 | 81.2      | 74.8 | 80.5 | 72.7 | 81.5         | 81.4 | 82.7 | 81.5 |
| TRIAGE <sub>Otsu-thr.</sub>   | train set | 86.5   | 85.7 | 86.4 | 85.3 | 79.2      | 73.6 | 77.5 | 71.9 | 83.2      | 81.1 | 80.0 | 82.9 | 83.8         | 83.7 | 84.4 | 83.8 |
|                               | test set  | 86.5   | 85.7 | 86.4 | 85.3 | 75.8      | 67.3 | 72.3 | 65.9 | 80.5      | 75.0 | 78.1 | 73.5 | 81.8         | 81.6 | 82.9 | 81.8 |
| TRIAGE <sub>Li-Lee-thr.</sub> | train set | 86.5   | 85.7 | 86.4 | 85.3 | 75.2      | 69.6 | 70.5 | 68.9 | 83.2      | 81.1 | 80.0 | 82.9 | 83.8         | 83.7 | 84.4 | 83.8 |
|                               | test set  | 86.5   | 85.7 | 86.4 | 85.3 | 75.8      | 67.3 | 72.3 | 65.9 | 80.5      | 75.0 | 78.1 | 73.5 | 81.8         | 81.6 | 82.9 | 81.8 |

sistently achieves better performance than the other two settings. Specifically, directly applying majority voting to all samples yields the worst performance, as hard voting neither accounts for the ambiguity of individual cases nor models the variation of evidence across verifier decisions. In contrast, applying majority voting only to samples within the boundary region leads to noticeable improvement, since confident cases remain unaffected while hard voting can still provide limited benefits for ambiguous cases. Finally, for our method, the performance curve becomes nearly flat once  $K$  reaches 3, suggesting that the default three-verifier configuration is already sufficient to stabilize the resolver decision process. Increasing the verifier pool beyond this point yields only marginal improvements due to the absence of label-informed knowledge for calibrating verifier reliability.

## D.5 Deployment Collaboration

The preceding experiments assume the strict label-free setting, where all components must be fixed without labeled validation data. In practical deployments, however, a small annotated validation subset may be available. We therefore study a deployment collaboration setting in which the same subset from the target validation split (not test set) is used to jointly configure the mapper, boundary region, verifier evidence, and verifier order. We vary both the validation-label budget and the verifier pool size. The candidate pool contains eight offline MLLM judges—Gemma-3<sub>27B</sub>, Gemma-3<sub>12B</sub>, Qwen2.5-VL<sub>32B-AWQ</sub>, Qwen2.5-VL<sub>72B-AWQ</sub>, Qwen3-VL<sub>8B</sub>, InternVL3.5<sub>8B</sub>, MiniCPM-V, and LLaVA-OV<sub>7B</sub>—

together with the four label-free/few-shot methods considered above: MARS, Mod-HATE, LoReHM, and ALARM.

**Calibration procedure.** All calibration steps use the same annotated subset. First, we calibrate the mapper posterior with a Platt-style logistic model fitted on standardized mapper logits. Second, we select the boundary region by choosing an entropy threshold from a small validation grid, which also serves as the stopping threshold after verifier updates. Third, we estimate verifier evidence strengths with Laplace-smoothed class-conditional response likelihoods, yielding asymmetric log-odds updates for hateful and normal verifier judgments. Finally, for each verifier pool size  $K$ , we select an ordered pool by beam search on the calibration subset, ranking candidates by accuracy, macro-F1, and then lower verifier-call cost. Together, these steps jointly configure the mapper posterior, boundary region, verifier evidence model, and ordered verifier pool.

Fig. 13 reports the resulting deployment surface. In most settings, calibrated TRIAGE outperforms both majority voting over the same selected pool and the default label-free TRIAGE configuration, showing that the framework benefits from probabilistic evidence aggregation and label-informed calibration. Performance also improves steadily as the validation-label budget increases. Notably, with only a small amount of labeled validation data, calibrated TRIAGE exceeds the best supervised baseline in most settings, suggesting that deployment-time labels can be used effectively without retraining a supervised detector.

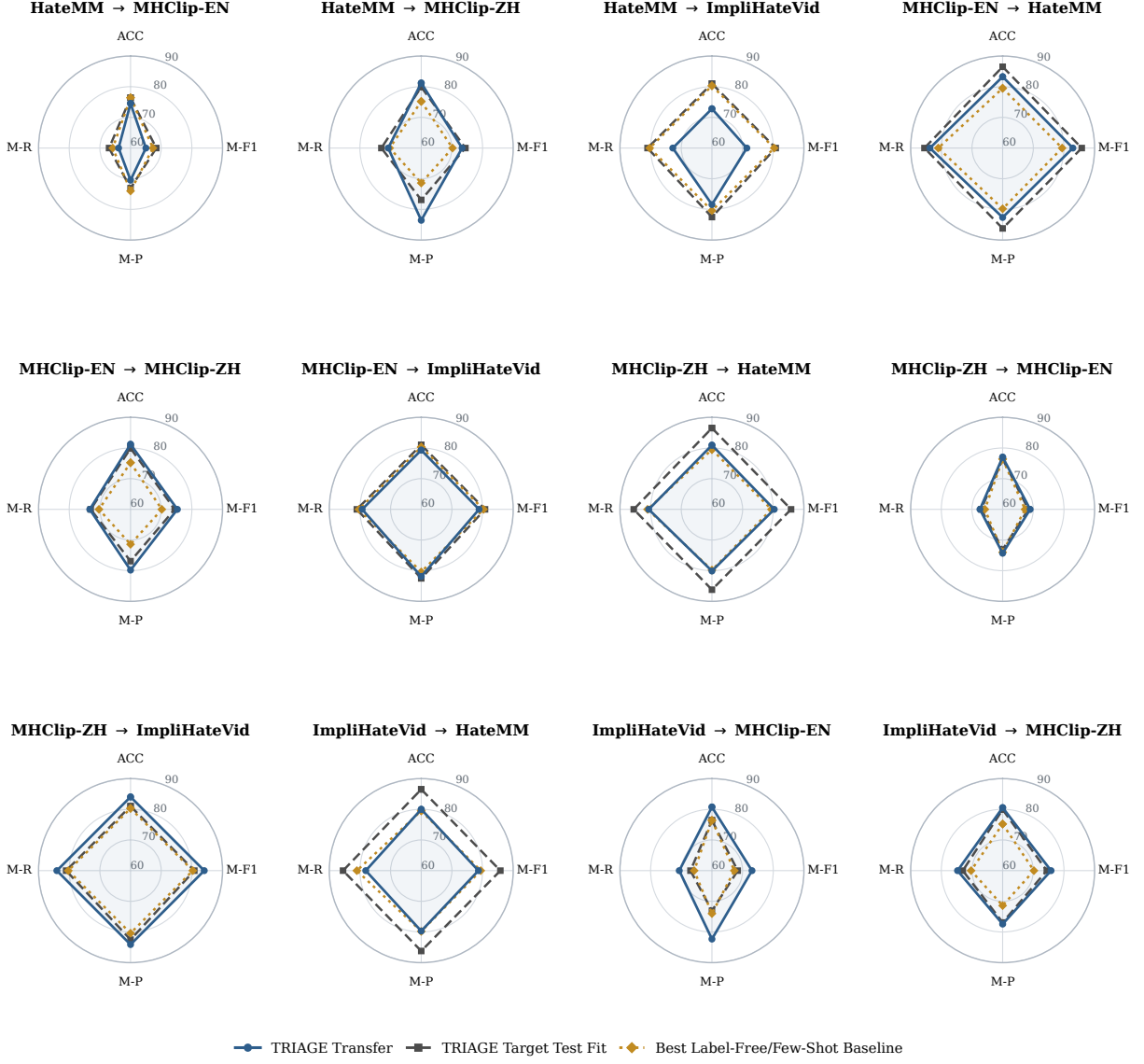


Figure 10: Transferability analysis across source-target dataset pairs. Each radar plot corresponds to one transfer direction and reports accuracy, macro-F1, macro-precision, and macro-recall on the target dataset.

## D.6 Adaptive Boundary Resolver: Class-wise Decision Shifts

Table 5 compares how the Adaptive Boundary Resolver changes class-specific precision and recall relative to the Boundary Mapper. Each value is a percentage-point change after verification. On HateMM, MHClip-EN, and MHClip-ZH, the resolver mainly expands hateful-class recall, indicating that many corrected cases are initially missed hateful videos near the boundary. ImphiHateVid behaves differently: the full pipeline becomes more precision-oriented for the hateful class, which is consistent with the dataset’s emphasis on implicit hate and the greater risk of over-interpreting ambiguous cues.

Table 5: Class-wise precision and recall shifts induced by the Adaptive Boundary Resolver. Values are percentage-point changes from the Boundary Mapper to the full pipeline.

| Dataset      | Hateful   |        | Normal    |        |
|--------------|-----------|--------|-----------|--------|
|              | Precision | Recall | Precision | Recall |
| HateMM       | +7.0      | +9.3   | +5.6      | +3.9   |
| MHClip-EN    | -2.7      | +18.4  | +4.7      | -3.6   |
| MHClip-ZH    | +4.5      | +15.6  | +6.4      | +0.0   |
| ImphiHateVid | +6.8      | -6.5   | -3.2      | +8.5   |

## D.7 Qualitative Cases

We finally examine several qualitative cases to provide further insights into the behavior of our frame-

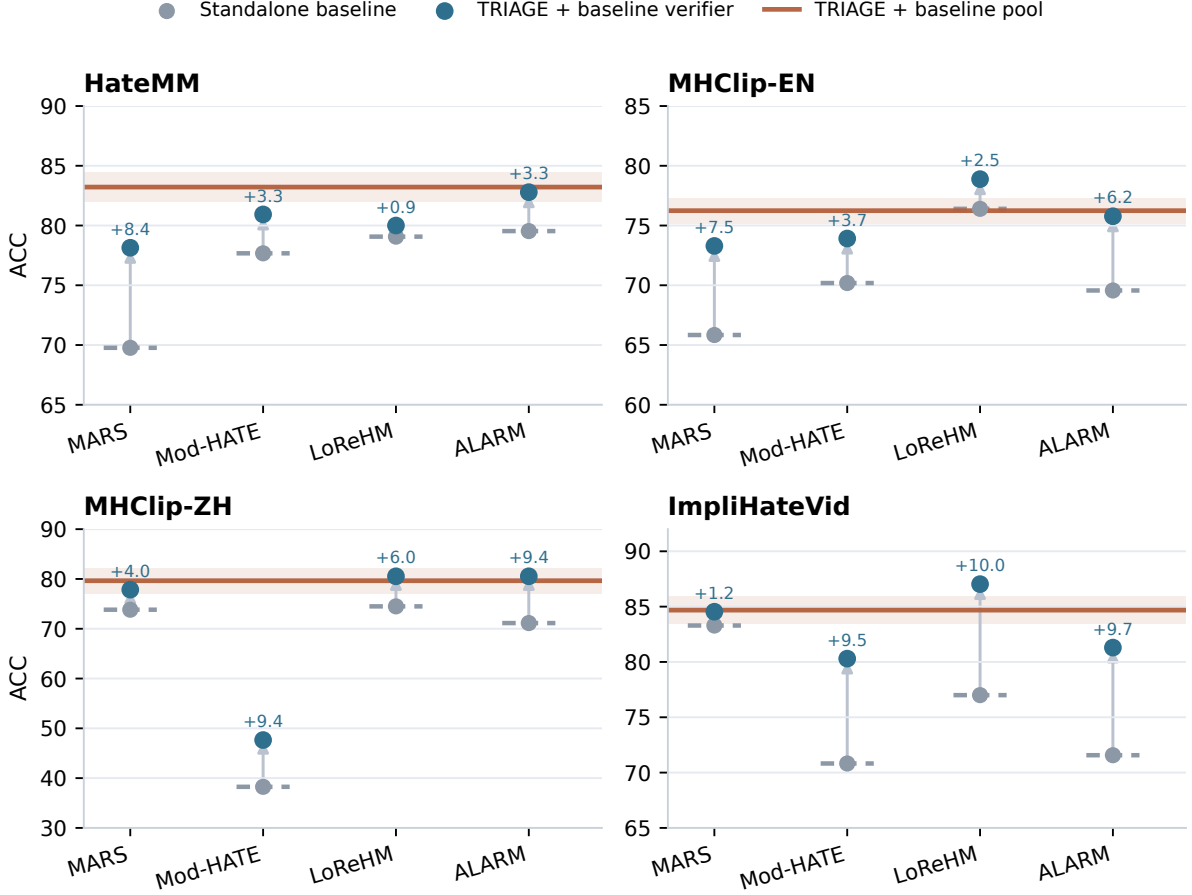


Figure 11: Generalizability to prior label-free/few-shot verifier backends. Dashed gray marks show the standalone baseline accuracy; blue marks show the same backend used inside TRIAGE; the orange line uses all four baselines as a sequential verifier pool averaged over all 24 backend orders, with the shaded band denoting  $\pm 1$  standard deviation.

work and its potential utility in real-world hateful video detection scenarios.

Fig. 14 presents three representative successful cases. S1 illustrates an early stopping case, where the first verifier in the Adaptive Boundary Resolver confirms the initial prediction produced by the Boundary Mapper, quickly pushing the sample outside the boundary region and enabling early termination, thereby avoiding unnecessary computational cost. S2 demonstrates how the three verifiers collaboratively resolve a difficult case through conflicting opinions. Such disagreement allows the framework to gather evidence from multiple perspectives and mitigates homogeneous errors that commonly arise in a single MLLM. This case is finalized after exhausting all verifiers and ultimately exits the boundary region. S3 follows a similar pattern to S2 and is also correctly classified after all verifiers are exhausted; however, it remains within the boundary region. This indicates that the framework can additionally identify cases that remain

uncertain even after the full verification process. Such cases may benefit from additional MLLM verifiers when computational resources permit, or be escalated to manual review.

Fig. 15 presents representative erroneous cases produced by the method. These cases commonly involve low-salience hateful evidence or require contextual cultural knowledge to interpret specific and coded expressions. In such cases, only one verifier may successfully recognize the hateful implication, while the others observe the relevant cues but fail to interpret them correctly, leading to evidence accumulation in the wrong direction. We attribute this to the varying levels of multimodal hate understanding across different MLLMs, where certain hate-related knowledge may be missing from specific models. This issue can be alleviated by deploying our method with MLLM backends specifically designed for hate detection, as discussed in Appendix D.3, which may better handle implicit hate and target identification. Alternatively, a small

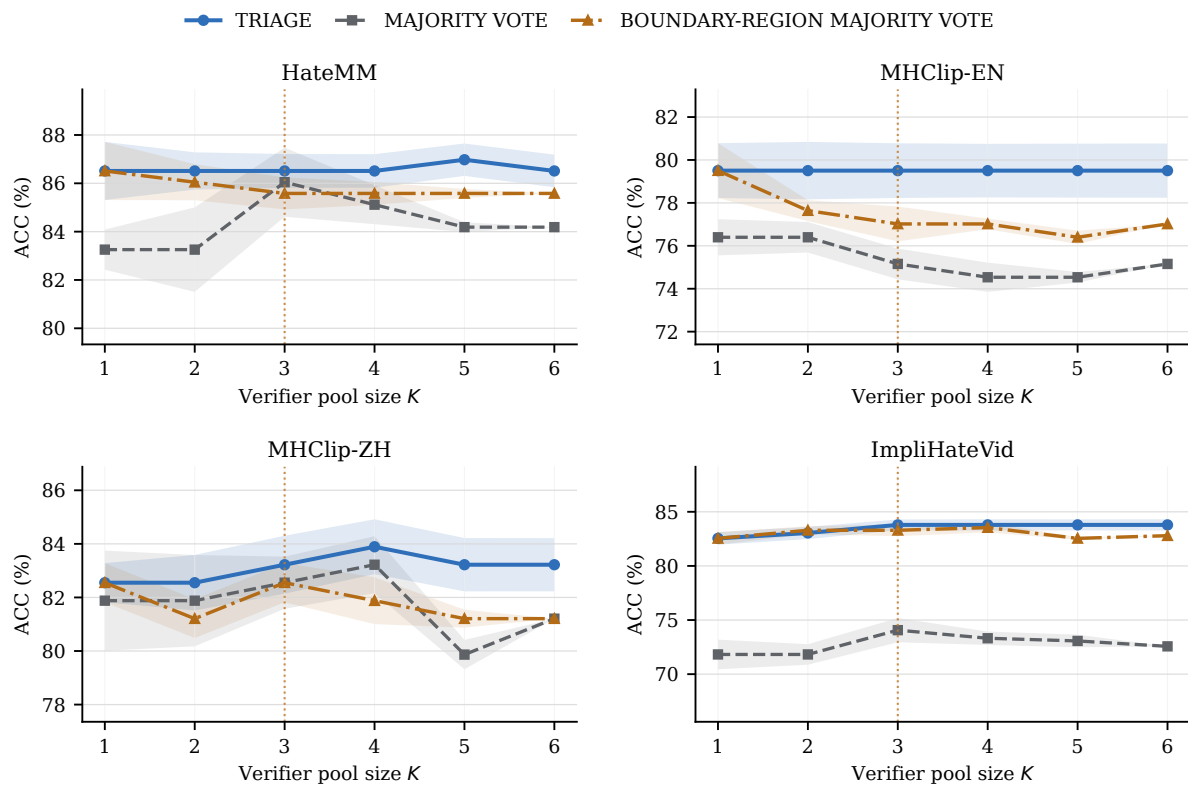


Figure 12: Effect of verifier pool size  $K$  in the Adaptive Boundary Resolver. The shaded band around each line indicates  $\pm 1$  standard deviation over the corresponding enumerated configurations.

amount of annotated data can be used to calibrate the hate detection capability of different MLLMs and assign more tailored evidence strengths, as discussed in Appendix D.5.

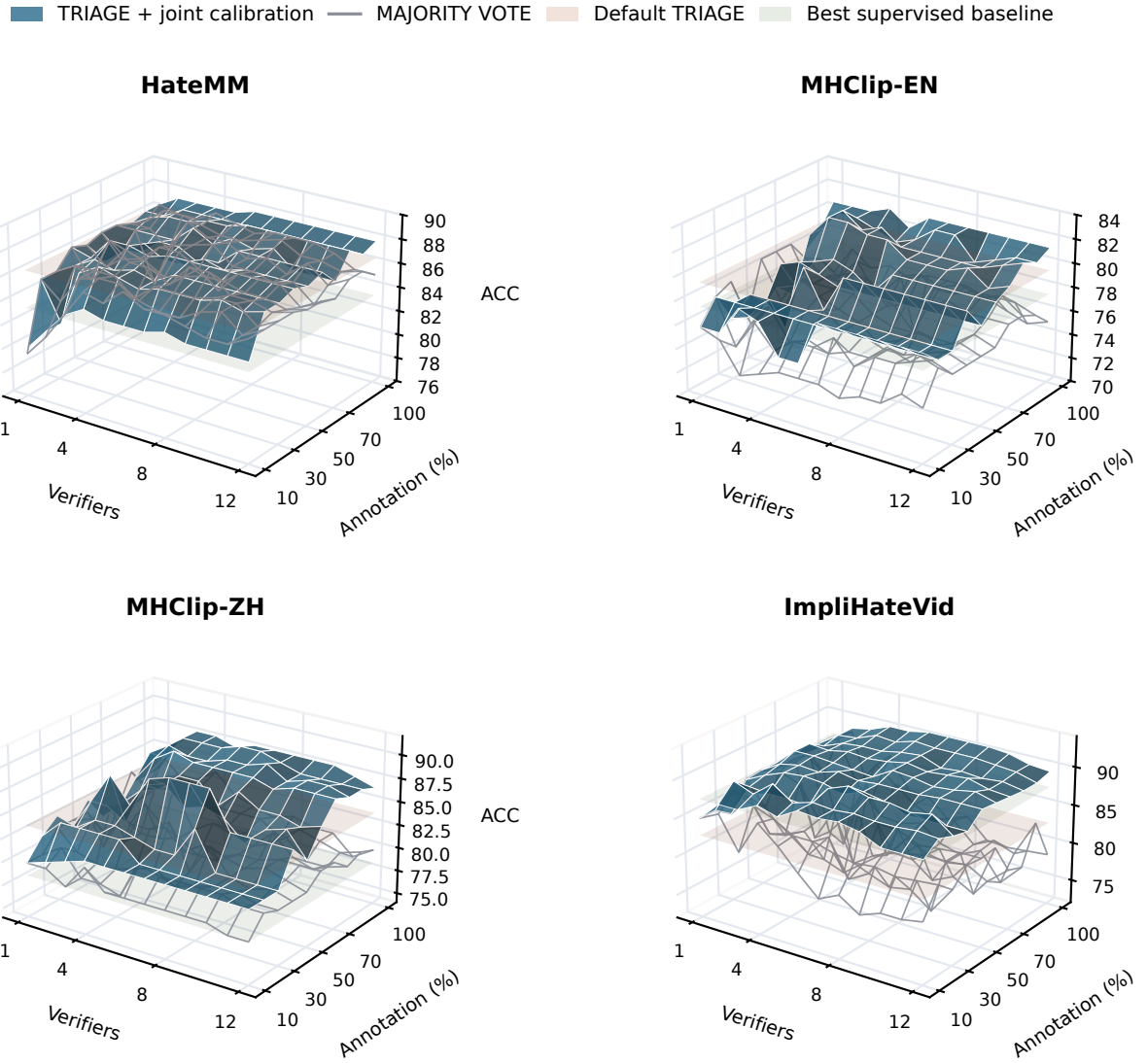


Figure 13: Deployment collaboration with validation-guided joint calibration. Each panel varies the validation-label budget and verifier pool size  $K$ .

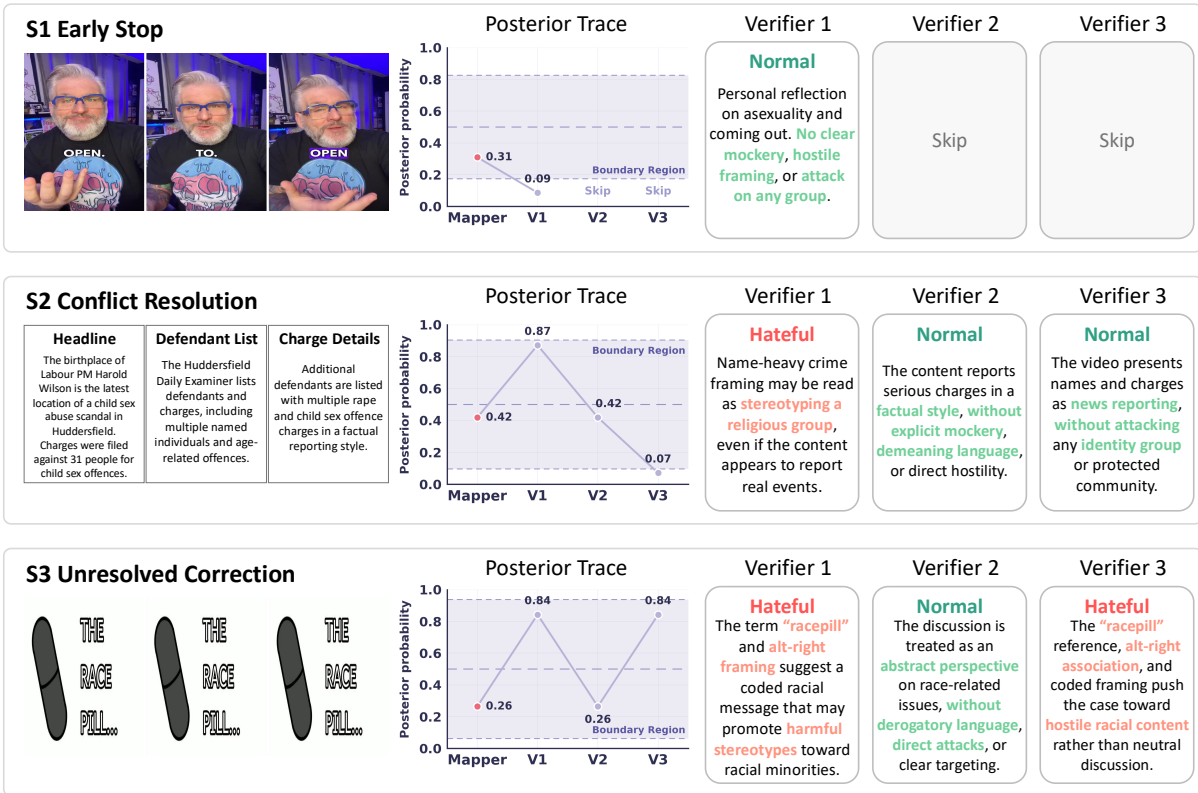


Figure 14: Qualitative case study of successful predictions.

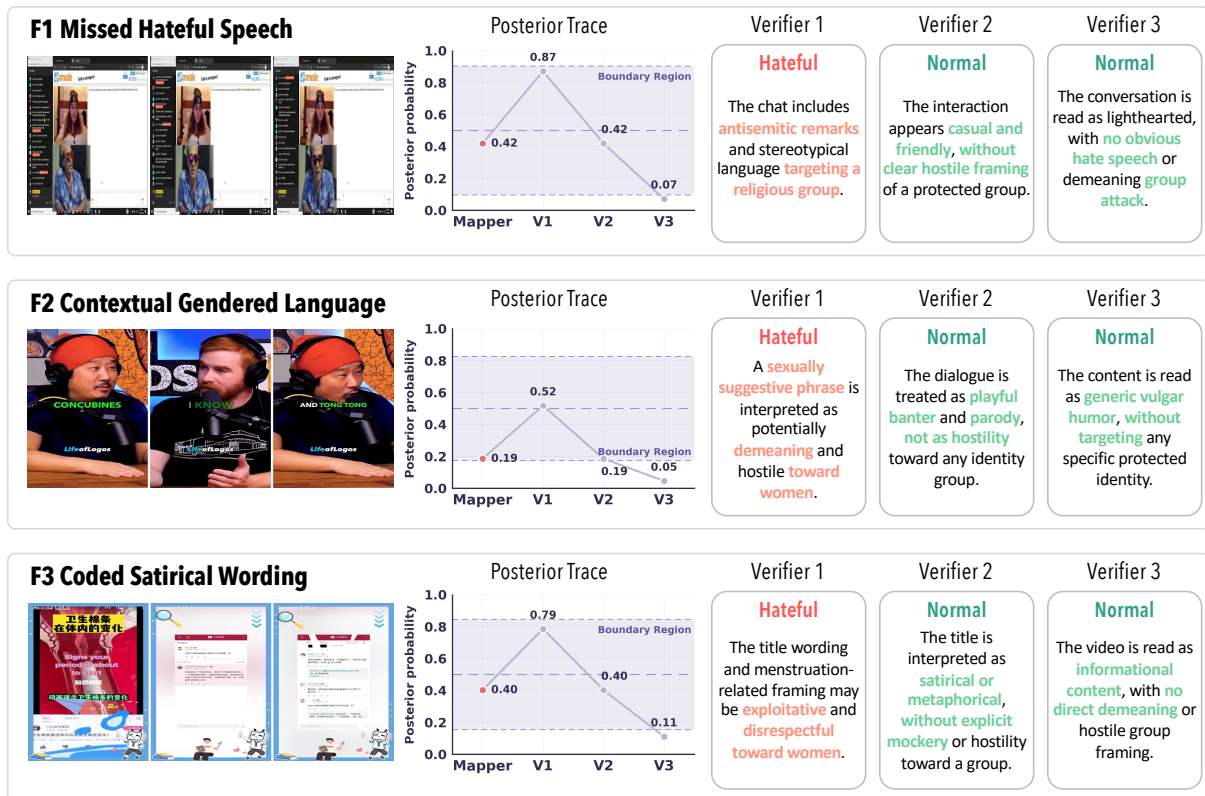


Figure 15: Qualitative case study of remaining error patterns.