SafeLens: Segment-Level Hate Speech Detection in Online Videos

Zhuoran Wang, Dylan Raharja, Yujia Hu, Roy Ka-Wei Lee

Singapore University of Technology and Design
{zhuoran_wang, dylan_raharja, yujia_hu, roy_lee}@sutd.edu.sg

Abstract

We present SafeLens, a lightweight segment-level video moderation system that fuses speech, text, and visual frames to produce hateful content detection for each segment. For every segment, SafeLens returns a structured prediction: *label*, *prediction confidence*, *reasons for flag*, *harm categories*. The structured predictions are optimized for triage, appeals, and downstream enforcement. The system is modular (pluggable speech, text, and visual processing modules back-ends and a mid-size policy Language Language Model (LLM) agent with parameter-efficient tuning). In the live demo, attendees can upload or select clips, scrub the timeline to flag hateful segments, inspect rationales, and vary the policy LLM agent to benchmark the hateful content moderation performance.

Code — https://github.com/Social-AI-Studio/SafeLens

Introduction

Short-form video has become a key channel for the dissemination of hate speech and other forms of online abuse, yet moderation research and tools continue to prioritize text and images (Hee et al. 2024). Off-the-shelf visual language models (VLMs) struggle to separate hateful content from nearby benign context, sarcasm, or rapid scene shifts, and most public benchmarks annotate entire videos with a single label, introducing temporal label noise that is ill-suited for enforcement, appeals, and creator feedback. Recent releases have begun to close this gap, e.g., MultiHateClip (Wang et al. 2024) and ImpliHateVid (Rehman et al. 2025), but these are primarily video-level resources and document how models still confuse "hateful" vs. "offensive" categories. To our knowledge, HateClipSeg (Wang, Wang, and Lee 2025) is the first dataset to provide comprehensive segment-level annotations (11,714 segments across fine-grained offense types and targets). Segment-level supervision is crucial: platforms act on moments, not whole uploads, and moderators need precise jump-to-segment evidence, detection confidence, and compact rationales to make fast moderation decisions.

This demo addresses that gap with SafeLens, a practical system that makes per-segment, multimodal, policy-aware decisions to support real moderation workflows. SafeLens is built around temporal granularity (decisions aligned to segments rather than whole clips), multimodal evidence fusion (audio speech, on-screen text, and visual cues), and policy-aware structured outputs that are easy to plug into pipelines. The system is modular; back-ends can be swapped for latency, cost, or governance, and are auditable via reproducible JSON logs and parameter records.

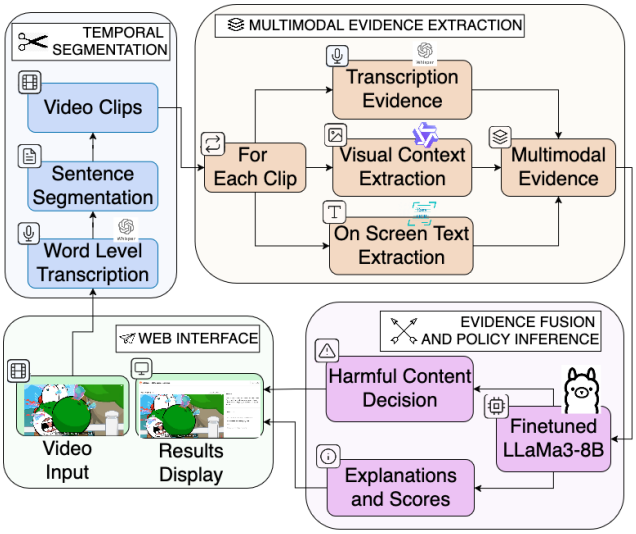

Figure 1: A diagram illustrating the system workflow.

System Overview

SafeLens employs a multimodal agentic framework to moderate hateful videos through a four-stage pipeline: (1) temporal segmentation, (2) multimodal evidence extraction, (3) evidence fusion and decision inference, and (4) delivery via a web interface (Figure. 1).

Temporal segmentation. Videos are partitioned into semantically coherent clips using a strategy adapted from HateClipSeg (Wang, Wang, and Lee 2025). First, word-level transcripts generated by Whisper (Radford et al. 2023) are merged into sentences using the NLTK (Bird 2006) Punkt tokenizer. To handle long silent segments (≥ 20 s), we detect scene changes by computing cosine similarity between ViT (Dosovitskiy et al. 2020) frame embeddings, and slice

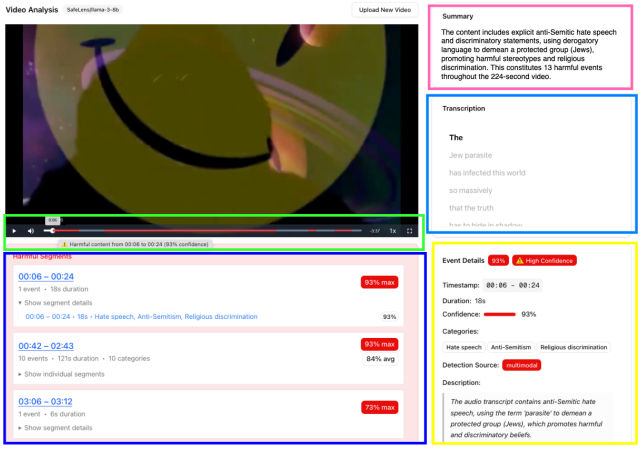

Figure 2: End-to-end SafeLens UI with color-boxed regions: player and timeline with in-place warnings (green); time-ordered list of harmful segments (purple); video-level summary (pink); scroll-synchronized transcript (blue); per-segment event details (yellow).

the video when similarity falls below a threshold τ.

Multimodal evidence extraction. For each segmented clip, we extract three complementary evidence streams: (i) *speech*, using Whisper to obtain word-level transcripts with timestamps (Radford et al. 2023); (ii) *on-screen text*, detected and recognized via EasyOCR (Jaided AI 2020) sampled every 3–5 seconds with confidence filtering (e.g., $\geq$ 0.6) and simple temporal de-duplication to suppress repeated overlays; and (iii) *visual context*, where Qwen2.5-VL (Bai et al. 2025) produces objective frame descriptions at fixed 5 seconds intervals under a neutral instruction prompt with deterministic decoding (temperature $= 0$).

Evidence fusion and decision inference. The multimodal evidence is combined into a structured prompt and presented to a policy-aware LLM for final inference. We use a Llama3-8B model (AI@Meta 2024) fine-tuned with LoRA adapters on the HateClipSeg dataset (Wang, Wang, and Lee 2025). This instruction-tuned model maps the fused evidence to a structured prediction. For each segment, it returns a JSON object containing a boolean harmfulness label, a confidence score (0-1), a one-sentence explanation, and, if harmful, a list of applicable harm categories. To select these back ends, we evaluated several state-of-the-art policy LLMs and VLMs under a unified prompt and protocol; full evaluation results are available on our GitHub page.

Web interface. The system is accessible through a lightweight web application. Users securely sign up, upload videos, and receive an interactive report. This report features a modular view that highlights harmful segments with their explanations, categories, and confidence scores, alongside a transcript synchronized with the video playback.

Demonstration

We demonstrate SafeLens on two short, curated clips: **Scenario 1 (explicit hate)**, where overt slurs occur in speech, and **Scenario 2 (implicit/code-mixed)**, where the target is conveyed via sarcasm and on-screen text. After upload, SafeLens processes the clip and renders an interactive analysis composed of five coordinated panes (Fig. 2).

Player & timeline (green). The player overlays warnings on detected spans; hovering reveals the exact window and calibrated confidence. Hovering a warning seeks the player to the beginning of the span.

Harmful-segment list (purple). A time-ordered table lists flagged segments with start–end, duration, predicted categories, and confidence. Rows expand to show concise rationales and modality attribution. For example, `00:06--00:24` (18 s) is flagged as *Hate speech / Anti-Semitism / Religious discrimination* at 93% confidence; a longer window `00:42--02:43` aggregates multiple events (max 93%, avg 84%), and `03:06--03:12` is flagged at 73% confidence.

Summary (pink). This panel provides a video-level summary. If harmful content is detected, it states the hate category (e.g., Anti-Semitism), gives a brief rationale explaining why the content is deemed hateful, and reports the number of hateful events detected in the video. If no harmful segment is found, it presents a concise synopsis of the video and explicitly notes that no harmful content was detected.

Transcription (blue). ASR text scrolls in sync with playback, enabling rapid verification. Clicking on any line takes the player to the corresponding timestamp.

Event details (yellow). Selecting a row in Harmful-segment list reveals timestamps, duration, confidence (with "High" badges when applicable), predicted categories, the dominant evidence source (speech/OCR/visual/multimodal), and a one-sentence rationale. For `00:06--00:24`, the explanation attributes the decision primarily to speech, while noting that visual/OCR cues were not decisive—illustrating modality attribution and transparent justification.

Live controls and audit. To illustrate modularity and trade-offs, the home UI exposes a selector for comparing policy LLMs. A compact status strip reports processing time by stage and model versions for auditability. All demo clips are anonymized/redacted, and uploads are processed under a short-retention policy.

Conclusion

We introduced SafeLens, a lightweight, modular system for *segment-level* multimodal moderation. It fuses temporal segmentation with policy-aware LLM inference to produce structured, auditable decisions—label, categories, confidence, and a concise rationale—rendered in an interpretable UI. The live demo enables jump-to-span review, modality attribution, and rapid verification without scanning entire videos.

Acknowledgments

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore, and the Ministry of Digital Development and Information, under its Online Trust and Safety (OTS) Research Programme (Award Grant No. S24T2TS007), and the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Prime Minister's Office, Singapore, or the Ministry of Digital Development and Information and the Ministry of Education, Singapore.

References

AI@Meta. 2024. Llama 3 Model Card.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bird, S. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, 69–72.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Hee, M. S.; Sharma, S.; Cao, R.; Nandi, P.; Nakov, P.; Chakraborty, T.; and Lee, R. K.-W. 2024. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4407–4419. Miami, Florida, USA: Association for Computational Linguistics.

Jaided AI. 2020. EasyOCR: Ready-to-use OCR with 80+ languages. https://github.com/JaidedAI/EasyOCR. Accessed: 2025-09-12.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Rehman, M. Z. U.; Bhatnagar, A.; Kabde, O.; Bansal, S.; and Kumar, N. 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)*, 17209–17221.

Wang, H.; Wang, Z.; and Lee, R. K.-W. 2025. HateClipSeg: A Segment-Level Annotated Dataset for Fine-Grained Hate Video Detection. arXiv:2508.01712.

Wang, H.; Yang, T. R.; Naseem, U.; and Lee, R. K.-W. 2024. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7493–7502.