CRIME: INVESTIGATING COMPOSITIONAL RISKS IN VIDEO MODERATION VIA RELATIONAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional risks arise when the relations among video elements establish a risk that the individual elements alone do not. Despite advances in multimodal large language models, detecting such risks remains difficult: relevant cues may be dispersed amid abundant video content, and recognizing them individually does not establish how they are related. The central challenge is to identify which connections matter to the risk being assessed, whether those connections hold, and whether they support a risk judgment. To tackle this challenge, we introduce CRIME (Compositional Risk Investigation for Video Moderation through Relational Evidence), an investigative framework for discovering and verifying relational evidence in videos. CRIME maintains a revisable Evidence Graph of video-grounded observations and candidate relations. Guided by evolving risk hypotheses, it searches for missing or uncertain evidence, checks consequential connections against the source video, and revises the graph and hypotheses accordingly. Multiple agents take distinct but interacting roles in proposing interpretations, seeking evidence, and scrutinizing candidate relations. Through the shared graph, one agent's provisional findings can be examined and corrected by others, allowing those corrections to shape subsequent investigation and the final judgment. Experiments across controlled synthetic, real-world, and cross-policy compositional-risk settings demonstrate the effectiveness and adaptability of CRIME. These results highlight the value of combining explicit relational representation with hypothesis-driven evidence discovery for video moderation.

⚠ **Warning:** This paper discusses and illustrates content that some readers may find offensive or disturbing.

1 INTRODUCTION

With the rapid growth of video-sharing platforms, video content moderation has become essential to maintaining a safe online environment (Yuksel & Tan, 2023; Wang et al., 2025b). Earlier moderation systems mainly focus on recognizing predefined categories of harmful content and outputting safety labels or violation types (Perez et al., 2017; Yuksel & Tan, 2023). In recent years, advances in multimodal large language models (OpenAI, 2026; Google DeepMind, 2026a; Anthropic, 2026a) have enabled moderation systems to go beyond fixed-category classification, making decisions based on natural-language safety policies and providing rationales (Helff et al., 2024; Chen et al., 2025).

However, as moderation capabilities continue to improve, some unsafe content has increasingly taken more implicit forms. We focus on compositional risk, a particularly challenging form of risk where harmful meaning is jointly constructed by components distributed across different dimensions of a video, such as temporal segments, modalities, and the interaction between content and its presentation (Wang et al., 2026). When examined separately, each constituent complies with the moderation policy. The violation emerges only when these components are interpreted jointly and their relations, such as contrast, analogy, temporal linkage, or contextual reinforcement, are taken into account. Detecting such risks therefore requires more than examining individual components in isolation. A moderation system must reason over multiple individually safe components and understand the meaning that emerges from their relations.

For this reason, video moderation becomes an open-ended problem of relational search and verification, leading to the central challenge of this work: how can risk-relevant content combinations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

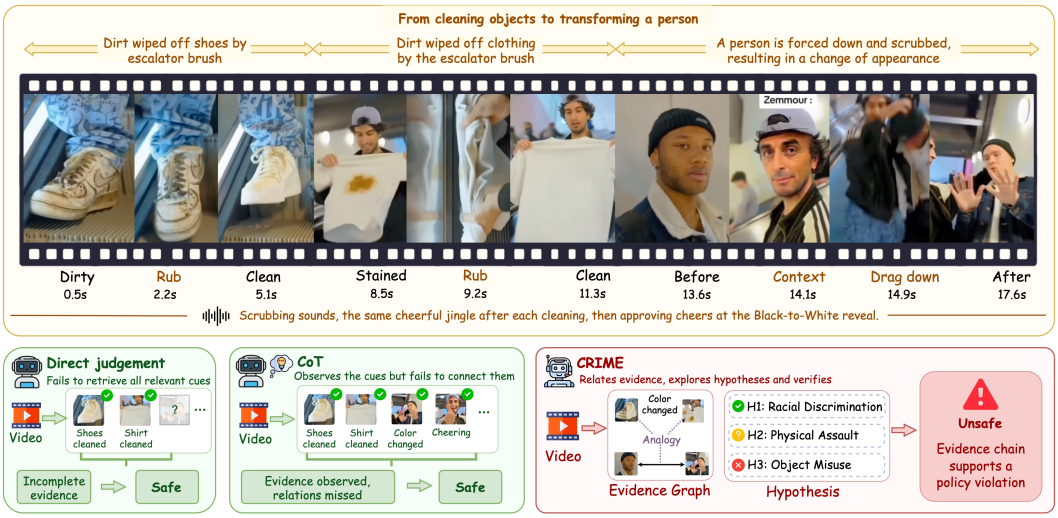


Figure 1: **A motivating example of compositional-risk moderation.** The video contains individually benign events whose combination conveys a racially derogatory meaning. Direct judgement misses a decision-critical cue and labels the video as safe. Chain-of-thought (CoT) reasoning identifies the relevant local events but fails to connect them to recover the harmful interpretation. CRIME explicitly organizes the distributed evidence, forms relational hypotheses, and verifies the resulting interpretation before classifying the video as unsafe.

be discovered from an open candidate space and verified as constituting a policy violation? This problem involves three key difficulties. First, the relevant evidence and associations are not given in advance. Potential cues may be distributed across different parts and information sources of a video, so the system must determine which content should be analyzed together. Second, the relations that give rise to harmful meaning are often implicit. Since each component individually complies with the policy, the system must assemble several weak signals into a coherent interpretation while distinguishing harmful relations from plausible benign explanations. Third, the candidate space grows rapidly with the amount of content. A video may contain many possible combinations, while only a small fraction are relevant to the actual risk. Without clear guidance, the system cannot know in advance which direction is worth investigating.

To illustrate how these difficulties can affect current MLLMs, we use a representative compositional-risk example in Figure 1. The two conditions fail in different ways: the decisive cue is absent from the first verdict's rationale, whereas in the second, it is present yet never related to other components. In this case, direct judgment fails to sufficiently cover a decision-critical cue, while chain-of-thought (CoT) reasoning identifies the relevant local facts but still fails to establish the relations that reveal the harmful meaning. More broadly, compositional risks may be missed at multiple stages. A model may overlook key evidence, fail to establish and interpret the relevant relations even when the evidence is available, or fail to sufficiently explore candidate combinations, and therefore miss the risk-relevant direction. Regardless of where the failure occurs, the consequence is the same: the evidence relations needed for the moderation decision are never explicitly established and verified.

These observations motivate us to rethink how compositional risks should be moderated, leading to CRIME (**C**ompositional **R**isk **I**nvestigation for Video **M**oderation through Relational **E**vidence), a multi-agent video moderation framework centered on relational evidence and multi-hypothesis verification. Rather than asking the model for a direct moderation judgment, CRIME reformulates video moderation as an investigation process over candidate risk interpretations. It first explicitly organizes potentially relevant content and their relations, then explores multiple plausible risk directions, and actively verifies the decisive evidence required by each candidate interpretation. In this way, the system does not have to rely on a single initial interpretation, but can examine multiple possible risk explanations and verify them against the evidence. CRIME provides a flexible and extensible framework that is not tied to a specific backbone model. It supports adaptation to new forms of harmful expression, changes in the external background, and different moderation policies.

The main contributions of this work are three-fold. 1) We recast compositional-risk moderation as an open-ended problem of discovering relevant content combinations and verifying whether their joint

meaning violates the moderation policy. 2) We propose CRIME, a multi-agent framework that integrates explicit evidence graphs, multi-hypothesis exploration, and targeted verification against the original video, enabling broader and more evidence-grounded investigation of compositional risks. 3) Extensive experiments on controlled synthetic and real-world compositional-risk videos, together with evaluation under a different moderation policy, demonstrate the effectiveness of CRIME across different settings.

2 Related Work

Video Understanding with MLLMs. Recent advances in multimodal large language models (MLLMs) have increasingly integrated visual perception, language understanding, and semantic reasoning into unified frameworks (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Both proprietary MLLMs (OpenAI, 2026; Google DeepMind, 2026a; Anthropic, 2026b) and open-source model families (Liu et al., 2023; Bai et al., 2025; Wang et al., 2025a) have advanced rapidly, demonstrating strong capabilities across visual understanding, question answering, and multimodal reasoning. Beyond directly processing raw multimodal inputs, MLLMs can also incorporate information produced by external perception modules into their reasoning context (Yang et al., 2023; Wu et al., 2023; Liu et al., 2024). This flexibility enables MLLMs to serve as general-purpose reasoning backbones for integrating multi-source evidence and performing high-level semantic analysis. Building on this capability, our work investigates how distributed multimodal evidence in complex videos can be explicitly organized to support safety assessment.

Multimodal Content Moderation. Some earlier moderation methods primarily focused on final safety decisions (Chi et al., 2024; Zeng et al., 2025; Wang et al., 2025b). With advances in MLLMs, subsequent studies have sought to improve moderation interpretability through risk explanations, reasoning traces, and explicit intermediate representations (Helff et al., 2024; Liu et al., 2025; Chen et al., 2025; Yang et al., 2026a; Li et al., 2026b; Zhang et al., 2026). Most of these methods rely on task-specific training. When test content deviates from the training distribution, SCANNER (Li et al., 2026a) employs test-time adaptation to a trained hateful-video detector. Another line of work avoids retraining and performs moderation through inference-time configuration. MARS (Yang et al., 2026b) decomposes moderation into a sequence of inference steps over uniformly pre-sampled video frames and the accompanying transcript. CRIME uses an evidence graph to guide hypothesis-driven video reinspection, while supporting policy reconfiguration without retraining.

3 Understanding Failures in Compositional Risk Moderation

3.1 Problem Formulation

Compositional Risk. Given a video x and a moderation policy π, let $\mathcal{C}(x) = \{c_1, \ldots, c_n\}$ denote a set of individually assessable content elements in the video. An element may be a visual event, a spoken utterance, on-screen text, or a sound event. These elements may be distributed across different temporal segments, modalities, or aspects of content presentation. For any subset $S \subseteq \mathcal{C}(x)$, let $\mathcal{R}_S$ denote the relations supported by the original video that contribute to the joint interpretation of the elements in S. Let $y_\pi(S, \mathcal{R}_S) \in \{\text{benign}, \text{harmful}\}$ denote the policy-conditioned judgment of their joint meaning, with $y_\pi(c)$ denoting the judgment of an individual element c in isolation.

A video x exhibits compositional risk under policy π if every content element is individually benign, yet there exists a subset $S \subseteq \mathcal{C}(x)$ with $|S| \geq 2$ whose joint interpretation is harmful:

$$y_\pi(c) = \text{benign} \quad \forall c \in \mathcal{C}(x), \qquad y_\pi(S, \mathcal{R}_S) = \text{harmful}. \tag{1}$$

Compositional Risk Moderation. The goal of compositional-risk moderation is to output a video-level prediction $\hat{y}_\pi(x) \in \{\text{benign}, \text{harmful}\}$ together with supporting evidence and an explanation. Since the risk-relevant content elements and their relations are not given in advance, the moderation system must discover an evidence structure that can jointly support a risk interpretation.

For a harmful compositional interpretation, let $\mathcal{E}^\star = (S^\star, \mathcal{R}_{S^\star})$ denote a minimal supporting evidence structure, i.e., a set of content elements and relations that supports the interpretation, while removing any content element or key relation would make the interpretation no longer sufficiently

supported. Let $\widehat{\mathcal{E}} = (\widehat{\mathcal{C}}, \widehat{\mathcal{R}})$ denote the evidence currently discovered by the moderation system. Identifying a compositional risk with evidential support requires $\widehat{\mathcal{E}}$ to cover at least one such $\mathcal{E}^\star$:

$$S^\star \subseteq \widehat{\mathcal{C}}, \qquad \mathcal{R}_{S^\star} \subseteq \widehat{\mathcal{R}}.$$

This condition indicates that the required content elements and relations have been recovered, but does not guarantee a correct decision. The system must still reconstruct their joint meaning and determine whether it is benign or harmful under policy π.

3.2 Diagnosing Compositional-Risk Moderation Failures

To investigate why MLLMs fail in compositional-risk moderation, we conduct an evidence-guided diagnostic study on 100 compositional-risk videos sampled from DIH-Bench (Wang et al., 2026). We manually annotate a large set of content elements $\mathcal{C}^{\mathrm{ref}}(x)$ and their relations $\mathcal{R}^{\mathrm{ref}}(x)$, and identify a reference evidence chain supporting the final unsafe judgment. The content elements involved in this chain form the critical subset $S^\star$, with their supporting relations denoted by $\mathcal{R}_{S^\star}$. We compare four conditions: **Direct**, using only the video and policy. **Full Relational Evidence**, additionally providing all annotated elements and relations. **Critical Elements**, providing only the critical elements. And **Critical Relational Evidence**, providing the critical elements together with their supporting relations. All conditions use the same model, media input, and generation settings. Annotation and prompting details are provided in the appendix.

Table 1: Results of evidence-guided diagnostic study.

Condition	Added Evidence	Acc.
Direct	None	45%
Full Relational Evidence	$\mathcal{C}^{\mathrm{ref}}(x)$, $\mathcal{R}^{\mathrm{ref}}(x)$	46%
Critical Elements	$S^\star$	60%
Critical Relational Evidence	$S^\star$, $\mathcal{R}_{S^\star}$	78%

As shown in Table 1, Full Relational Evidence performs almost identically to Direct, suggesting that the model still struggles to identify decision-critical evidence from a large set of candidate content and relations. In contrast, providing only the critical elements improves detection, while further providing their relations yields a substantially larger gain. These results suggest that compositional-risk moderation is not merely about accessing local content: it requires identifying which distributed elements should be considered jointly and grasping how their relations convey risk. This observation motivates us to approach moderation through active evidence discovery and relational reasoning.

4 CRIME

4.1 Framework Overview

CRIME assesses whether a video violates a given moderation policy and returns a decision with a supporting explanation and evidence chain. Its core idea is to discover and explicitly connect cues distributed across the video into relational evidence, and to broaden evidence search through multi-hypothesis exploration and iterative critique.

As shown in Figure 2, an **Orchestrator** coordinates this multi-agent workflow and manages the investigation budget and stopping conditions. A lightweight Fast Judge first screens the video, returning a decision and explanation. Only high-confidence unsafe predictions trigger early exit. All the remaining videos proceed to detailed investigation. The **Perceiver** collects multimodal observations through direct inspection and on-demand tool use. The **Graph Builder** organizes these observations into an Evidence Graph that explicitly represents their relations. The **Hypothesis Explorer** jointly considers the video, the graph, and the policy to propose multiple risk hypotheses, and may consult prior cases via the Case Retrieval Tool as needed. The **Critic** challenges each candidate hypothesis h_i and provides feedback f_i for iterative hypothesis refinement and further evidence acquisition. Both agents can request additional evidence through targeted inspection by Perceiver or background queries via the Web Search Tool. The Graph Builder incorporates newly acquired evidence Δ_i into the graph, while retrieved information and Critic feedback inform further investigation.

When multiple hypotheses remain supported, or when the two agents fail to agree on a hypothesis's status s_i, the Orchestrator performs final adjudication based on the available evidence, Critic assessments, and the current policy. The system returns a final decision, explanation, and supporting evidence chain. Algorithm 1 shows this framework.

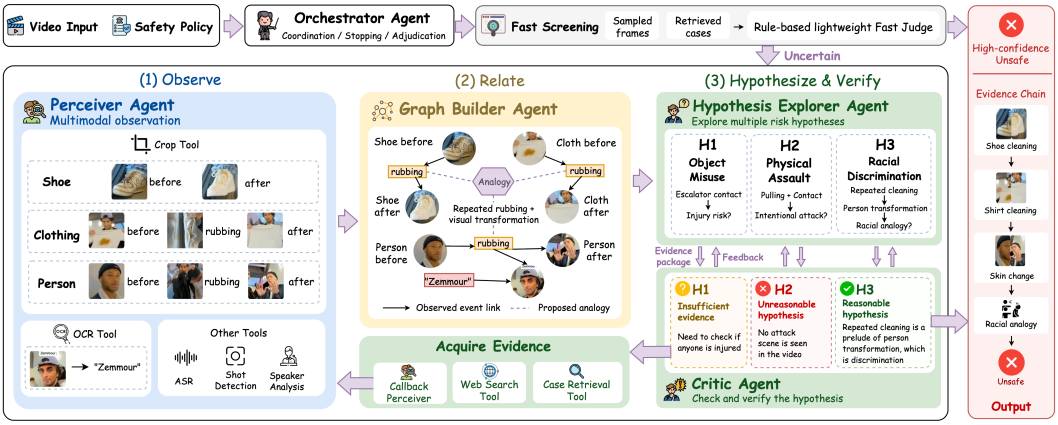


Figure 2: **Overview of the proposed CRIME framework.** CRIME organizes video moderation as a multi-agent risk investigation process consisting of multimodal evidence collection, relational evidence construction, multi-hypothesis investigation, and evidence-grounded verification.

CRIME can be instantiated with different MLLM backbones, investigate diverse harmful patterns without relying on fixed risk templates, support different moderation policies as inputs, and incorporate updated external knowledge on demand.

Algorithm 1 CRIME

Input: Video x, policy π
Output: $\hat{y}_\pi(x)$, ρ, and optional $\mathcal{T}$

$(\hat{y}_\pi(x), \rho) \leftarrow$ FAST JUDGE(x, π)
if high-confidence harmful prediction **then**
  return $(\hat{y}_\pi(x), \rho, \varnothing)$; **end if**
$\mathcal{O} \leftarrow$ PERCEIVER(x)
$\mathcal{G} = (\mathcal{N}, \mathcal{R}^{\text{obs}} \cup \mathcal{R}^{\text{inf}}) \leftarrow$ GRAPH BUILDER$(\mathcal{O})$
$\mathcal{H} \leftarrow$ HYPOTHESIS EXPLORER$(x, \mathcal{G}, \pi)$
$(s_i, f_i) \leftarrow (\text{unresolved}, \varnothing),\ \forall h_i \in \mathcal{H}$
for $h_i \in \mathcal{H}$ **while** budget remains **do**
  repeat
    if additional evidence is requested **then**
      $\Delta_i \leftarrow$ ACQUIREEVIDENCE(h_i, f_i)
      $\mathcal{G} \leftarrow$ GRAPH BUILDER$(\mathcal{G}, \Delta_i)$; **end if**
    $f_i \leftarrow$ CRITIC$(h_i, \mathcal{G}, \pi)$
    $(h_i, s_i) \leftarrow$ HYPOTHESIS EXPLORER(h_i, f_i)
  until $s_i \in \{\text{supported}, \text{refuted}\} \vee \text{budget} = 0$
end for
$\mathcal{S} \leftarrow \{h_i \in \mathcal{H} : s_i = \text{supported}\}$
$\mathcal{U} \leftarrow \{h_i \in \mathcal{H} : s_i = \text{unresolved}\}$
if $|\mathcal{S}| > 1 \ \vee\ \mathcal{U} \neq \varnothing$ **then**
  return ORCHESTRATOR$(\mathcal{H}, \{f_i\}, \mathcal{G}, \pi)$
else
  return AGREEDDECISION$(\mathcal{S}, \mathcal{G}, \pi)$
end if

Multimodal Observation. The Perceiver is a tool-augmented perception agent that acquires grounded multimodal observations from the original video. It directly inspects the video and uses specialized perception tools for tasks such as frame-level and region-level visual inspection, OCR, and audio analysis. In the initial stage, the Perceiver collects a broad set of multimodal observations. During subsequent investigation, it can revisit relevant video content in response to additional evidence requests. The acquired information is organized into a set $\mathcal{O}$ of observations. Each observation describes media-grounded content such as people, objects, actions, on-screen text, or adio events, and retains its temporal location and source reference to the original video. These references allow later modules to revisit the corresponding media content when additional inspection is required. The Perceiver records only observations supported by the media and does not assign risk categories or safety labels, thereby separating evidence acquisition from subsequent risk interpretation.

Evidence Graph Construction. The Graph Builder organizes the observation set $\mathcal{O}$ into an Evidence Graph $\mathcal{G} = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ denotes the set of evidence nodes and $\mathcal{R}$ denotes the relations among them. Each observation is represented as an evidence node while preserving its semantic type and grounding information. Based on the observed content, the Graph Builder identifies relations among nodes to explicitly connect distributed evidence. We distinguish two types of relations:

$$\mathcal{R} = \mathcal{R}^{\text{obs}} \cup \mathcal{R}^{\text{inf}}, \tag{2}$$

where $\mathcal{R}^{\text{obs}}$ contains event relations directly supported by the media observations, such as temporal order, action–object associations, and explicit cross-modal correspondences. In contrast, $\mathcal{R}^{\text{inf}}$ contains candidate semantic relations proposed by the Graph Builder from the existing observations, such as comparison, analogy, reference, and cross-scene associations. For each candidate relation,

the graph retains a model-estimated confidence score as an auxiliary indicator of the relation's uncertainty. In Figure 2, the shoe, clothing, person, and rubbing actions are represented as evidence nodes. Their event-level connections are recorded in $\mathcal{R}^{\text{obs}}$, while repeated rubbing different across segments gives rise to a candidate cross-event relation in $\mathcal{R}^{\text{inf}}$. These relations are recorded without making a moderation judgment. The Graph Builder incrementally updates the Evidence Graph as new observations or relations are identified. The Evidence Graph is constructed without predefined risk-specific templates, allowing it to capture diverse content combinations and relational patterns.

4.2 Multi-Hypothesis Risk Investigation

Given the initial Evidence Graph, the Hypothesis Explorer jointly considers the original video, the Evidence Graph, and the current policy to propose a set of candidate risk hypotheses,

$$\mathcal{H} = \{h_1, h_2, \ldots, h_K\}. \tag{3}$$

Each hypothesis specifies a candidate risk interpretation by identifying the relevant evidence, the relations that may give rise to harmful meaning, and the corresponding policy condition. It is therefore treated as a testable interpretation rather than a direct risk prediction.

Although different hypotheses may share supporting evidence, they focus on different evidence combinations, relational interpretations, or potential risk mechanisms, giving rise to distinct reasoning paths. For each hypothesis, the Hypothesis Explorer further identifies the supporting evidence currently available and the additional evidence needed for further examination. Different hypotheses therefore introduce different evidence requirements and drive further exploration.

When additional media evidence is needed, the Hypothesis Explorer can request targeted inspection of the relevant video content by the Perceiver. Newly acquired observations are incorporated into the Evidence Graph by the Graph Builder and used to further refine the corresponding hypothesis. The Hypothesis Explorer may also retrieve similar examples from a library of previously identified cases through the **Case Retrieval Tool** to inspire candidate interpretations. External background knowledge needed to examine a hypothesis can be obtained through the **Web Search Tool**. The Hypothesis Explorer considers both the original video and the Evidence Graph, allowing video cues not yet represented in the graph to trigger new hypotheses or further investigation. By introducing different evidence requirements, multiple hypotheses guide the system along different investigative directions and progressively broaden the evidence search space.

4.3 Evidence Verification and Decision

The **Critic** examines each candidate hypothesis and assesses whether the available evidence and their relations sufficiently support the proposed risk interpretation. For candidate semantic relations in the Evidence Graph, the Critic considers both their supporting evidence and confidence scores, where the latter serves only as an auxiliary uncertainty signal. The Critic also considers plausible alternative explanations and checks whether the hypothesis over-interprets the available evidence. When unresolved issues remain, the Critic can request additional evidence or provide feedback to the Hypothesis Explorer for hypothesis refinement. Such requests may trigger targeted video inspection by the Perceiver or external knowledge retrieval through the Web Search Tool, after which the Hypothesis Explorer may revise the hypothesis before another round of Critic evaluation.

This interaction continues until the two agents reach an evidence-supported conclusion or the shared investigation budget is exhausted. If the Hypothesis Explorer and Critic agree on a supported interpretation, the corresponding moderation decision is returned directly. When multiple hypotheses remain supported, the Orchestrator selects the final interpretation based on the available evidence and policy. If the two agents fail to reach agreement before the budget is exhausted, the Orchestrator adjudicates the unresolved case. The system outputs the final moderation decision $\hat{y}_\pi(x)$ together with its risk interpretation ρ and supporting evidence chain $\mathcal{T}$.

Table 2: Detection accuracy (%) on temporal and modality composition tasks. Avg. is the overall accuracy across the seven unsafe categories. Safe Acc. is accuracy on safe videos.

Method	Model	DIS	ILL	MIS	POL	RSK	SEX	VIO	Avg. ↑	Safe Acc. ↑
Temporal Composition										
Specialized	SafeWatch-8B	10.5	14.0	14.1	15.2	28.4	37.9	33.3	**20.0**	99.0
	LlavaGuard-7B	16.3	20.4	17.2	50.6	10.5	20.0	45.4	**26.1**	99.9
	GuardReasoner-Omni-7B	18.8	28.5	32.0	43.9	25.1	35.8	50.7	**32.5**	99.3
	MARS	22.3	21.7	12.5	30.4	21.1	28.7	16.7	**22.0**	99.0
Direct	Gemini 3.8 Flash	19.2	42.1	30.2	67.0	37.0	29.1	47.5	**39.1**	99.9
	GPT-6 Astra	40.2	43.5	39.3	51.1	59.2	45.7	40.4	**45.8**	99.7
	Claude Sonnet 5	19.6	40.6	12.5	34.8	26.3	68.5	43.9	**32.3**	99.6
	DeepSeek-V4.1-Flash	21.9	45.6	27.5	56.9	40.3	67.5	56.2	**42.7**	99.3
	Qwen3.8-27B	10.8	18.3	9.4	34.5	24.5	18.3	20.8	**19.7**	99.8
CoT	Gemini 3.8 Flash	47.3	47.8	37.4	53.3	51.3	51.3	42.1	**47.5**	99.1
	GPT-6 Astra	40.2	43.5	39.3	51.1	63.3	40.1	49.1	**47.1**	98.9
	Claude Sonnet 5	28.6	53.6	21.4	44.6	26.3	85.8	63.2	**42.6**	99.4
	DeepSeek-V4.1-Flash	33.9	53.3	37.6	58.2	46.8	66.4	56.3	**48.6**	98.6
	Qwen3.8-27B	23.0	27.1	29.1	26.2	32.9	15.3	34.6	**27.3**	99.7
CRIME(Ours)	Gemini 3.8 Flash	64.9	80.9	78.6	82.6	76.3	91.2	77.2	**77.1**	99.9
	GPT-6 Astra	63.3	60.0	64.4	73.1	77.1	81.5	71.4	**69.1**	99.2
	Claude Sonnet 5	61.4	78.6	62.6	75.0	65.8	88.6	84.1	**72.1**	99.6
	DeepSeek-V4.1-Flash	65.7	68.1	58.3	81.0	66.7	87.1	71.2	**70.3**	98.8
	Qwen3.8-27B	59.6	66.7	60.7	68.9	65.8	62.9	72.0	**65.2**	99.8
Modality Composition										
Specialized	GuardReasoner-Omni-7B	28.0	39.4	44.2	31.0	41.5	41.9	60.5	**39.8**	96.4
	MARS	66.8	33.4	14.2	28.5	28.6	35.5	42.9	**38.0**	99.9
Direct	Gemini 3.8 Flash	48.2	33.9	36.1	41.3	36.2	46.3	46.5	**41.2**	98.9
	Qwen3-Omni-30B-A3B	15.5	26.9	25.8	40.5	37.6	37.9	53.5	**32.3**	98.5
CoT	Gemini 3.8 Flash	47.7	41.9	25.6	51.3	38.6	64.0	44.7	**43.9**	98.3
	Qwen3-Omni-30B-A3B	8.8	22.5	21.9	36.0	31.6	27.1	38.4	**25.3**	98.9
CRIME(Ours)	Gemini 3.8 Flash	93.3	59.9	64.2	67.0	57.0	92.1	71.4	**72.0**	99.0
	Qwen3-Omni-30B-A3B	73.4	66.6	50.0	57.0	64.3	85.7	78.7	**67.2**	98.9

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate three compositional-risk video datasets covering controlled synthetic videos, real-world cases, and a different moderation policy. **(1) DIH-Bench** provides a controlled benchmark in which individually benign components jointly convey harmful semantics, covering diverse risk categories and composition patterns. The temporal and modality composition settings contain 3,188 and 1,502 unsafe videos, respectively, across 7 major safety categories and 26 fine-grained subcategories. **(2) Real-World Dataset** contains 500 videos manually collected from public social-media platforms, including 250 unsafe and 250 safe videos, annotated under the same safety policy as DIH-Bench. **(3) Advertising Dataset** contains 250 policy-violating and 250 compliant advertisements from a real-world advertising moderation pipeline. It follows the advertising policy used in practice, evaluating adaptability under policy changes. See the appendix for details and examples.

Backbones. We instantiate CRIME with six representative MLLM backbones, including Gemini-3.8-Flash (Google DeepMind, 2026b), GPT-6 Astra (OpenAI, 2026), Claude-Sonnet-5 (Anthropic, 2026b), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), Qwen3.8-27B (Qwen Team, 2026), and Qwen3-Omni-30B-A3B (Xu et al., 2025), covering both proprietary and open-source models.

Baselines. We compare CRIME against two prompting baselines and four specialized moderation methods. **Direct prompting**: The model directly predicts the moderation result from the input video and policy without explicit intermediate reasoning. **Chain-of-thought prompting**: The model is

Table 3: Moderation performance on (a) Real-World and (b) Advertising. Recall and accuracy are percentages. Explanation scores use a 1–10 scale.

		(a) Real-World				(b) Advertising			
Method	**Model**	**Unsafe Recall**	**Safe Accuracy**	**Score LLM**	**Score Human**	**Unsafe Recall**	**Safe Accuracy**	**Score LLM**	**Score Human**
Specialized	GuardReasoner-Omni-7B	33.2	98.0	5.2	6.5	38.4	97.6	5.3	6.6
	MARS	26.8	99.2	4.9	6.2	35.6	99.2	5.0	6.3
Direct	Gemini 3.8 Flash	46.4	99.6	5.5	7.7	41.6	99.2	5.4	7.6
	DeepSeek-V4.1-Flash	37.2	99.2	5.3	6.9	42.8	99.2	5.4	7.0
	Qwen3-Omni-30B-A3B	34.4	98.8	5.0	6.5	32.0	98.4	4.9	6.4
CoT	Gemini 3.8 Flash	53.6	99.2	6.1	7.9	59.6	98.8	6.3	8.0
	DeepSeek-V4.1-Flash	48.8	98.4	5.9	7.3	56.0	98.4	6.0	7.4
	Qwen3-Omni-30B-A3B	29.2	98.8	4.9	6.4	38.0	98.8	4.8	6.3
CRIME (Ours)	Gemini 3.8 Flash	78.0	99.6	7.3	8.3	78.0	99.2	7.3	8.3
	DeepSeek-V4.1-Flash	72.4	98.8	7.0	8.0	83.6	98.4	7.5	8.4
	Qwen3-Omni-30B-A3B	68.4	98.8	6.7	7.8	73.6	98.8	7.0	8.0

prompted to reason step by step before making the final decision. The prompting baselines and CRIME use corresponding backbone models and moderation policies. Backbone-specific audio handling is described in the appendix. We further compare with SafeWatch-8B (Chen et al., 2025), LlavaGuard-7B (Helff et al., 2024), GuardReasoner-Omni-7B (Zhu et al., 2026), and MARS (Yang et al., 2026b), following their official implementations and recommended inference settings.

Evaluation protocol. For controlled synthetic evaluation, we report per-category accuracy and the average over seven unsafe categories, requiring both an unsafe verdict and the correct major category. We report unsafe recall for real-world and policy-adaptability evaluations. Across all evaluation settings, we also report safe accuracy when safe examples are available. For explanation quality in the real-world and policy-adaptability settings, we adapt the reference-based evaluation protocol of SafeWatch (Chen et al., 2025). An LLM judge and two authors assess accuracy, content specificity, and conciseness, assigning an overall **explanation score** from 1 to 10. We report the mean scores separately as Score-LLM and Score-Human on the same fixed sample set. Details are provided in the appendix.

5.2 Main Result

Controlled Synthetic Evaluation. This dataset provides a controlled setting specifically designed for compositional-risk detection, enabling a focused evaluation of whether CRIME can identify harmful meaning arising from relations among individually benign components. As shown in Table 2, CRIME achieves improvements in compositional-risk moderation. With Gemini 3.8 Flash as the backbone, its average major-category accuracy on temporal composition reaches 77.1%, compared with 39.1% for Direct prompting, a gain of 38.0 percentage points. This advantage is consistent across all seven risk categories. Similar improvements are also observed with other backbone models and remain consistent across both temporal and modality composition tasks. These results indicate that, compared with direct judgment or simply adding chain-of-thought reasoning, CRIME more effectively supports compositional-risk detection by explicitly discovering, organizing, and verifying relational evidence.

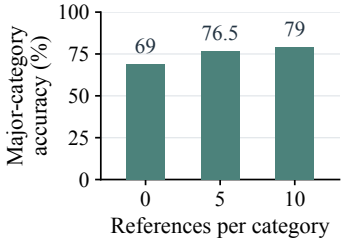
Real-World Evaluation. We further evaluate CRIME on the Real-World dataset to examine its effectiveness on naturally occurring compositional-risk videos. As shown in Table 3(a), CRIME continues to deliver substantial improvements on real-world videos. With Gemini 3.8 Flash as the backbone, CRIME achieves an unsafe recall of 78.0%, outperforming Direct and CoT, which obtain 46.4% and 53.6%, by 31.6 and 24.4 percentage points, respectively. These results show that CRIME remains effective at improving compositional-risk detection when moving from controlled synthetic videos to more complex, naturally occurring content. CRIME also improves unsafe recall with DeepSeek-V4.1-Flash and Qwen3-Omni-30B-A3B, reaching 72.4% and 68.4%, respectively. Overall, these results indicate that the benefits of CRIME are not limited to controlled synthetic settings, but also extend to real-world compositional-risk moderation.

Policy Adaptability. We evaluate CRIME on the Advertising dataset under its advertising moderation policy. As shown in Table 3(b), CRIME improves unsafe recall by up to 41.6 percentage points

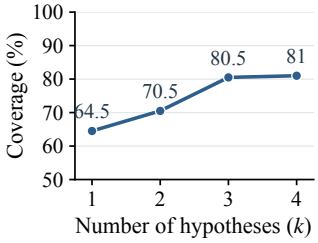
over Direct prompting and up to 35.6 points over CoT across the three backbones, while maintaining safe accuracy of 98.4–99.2%. CRIME also achieves higher Score-LLM and Score-Human than both prompting baselines for every backbone, indicating better explanation quality under the advertising policy. These results support the applicability of CRIME's relational investigation process to advertising violations governed by a different policy.

5.3 Ablation Study

For the ablation studies, we use Gemini 3.8 Flash and a pool of 200 unsafe and 200 safe videos. Each analysis uses the subset relevant to its metric.



(a) Reference Library



(b) Multi-Hypothesis

Metric	w/o Critic	Full
Major Acc.	66.5	76.0
Unsafe Rec.	84.5	82.0
Safe Acc.	88.0	98.0

(c) Critic

Figure 3: **Ablation studies of CRIME.** (a) Major-category accuracy versus references per category. (b) Initial hypothesis coverage. (c) Critic comparison.

Effect of Reference Library Size. We vary the number of references per major category over $\{0, 5, 10\}$, yielding libraries of 0, 35, and 70 cases across seven categories. The nonempty libraries are nested, maintain a 3:2 unsafe-to-safe ratio within each category, and are disjoint from the test set. All references are provided directly during hypothesis generation, with the remaining pipeline held fixed. Figure 3(a) reports final major-category accuracy on the same 200 unsafe videos. The accuracies are 69.0%, 76.5%, and 79.0%, respectively. The 70-case library improves on the no-reference setting by 10.0 percentage points.

Effect of Multi-Hypothesis Reasoning. On the 200 unsafe videos, we measure pre-verification Coverage@k, defined as the fraction of videos whose first k hypotheses, in generation order, include one capturing the reference mechanism's main elements and core relation. Coverage rises from 64.5% at $k = 1$ to 81.0% at $k = 4$ (Figure 3(b)), recovering 33 additional videos. The fourth hypothesis adds 0.5 percentage points over $k = 3$. Thus, exploring multiple candidates broadens initial mechanism coverage, with coverage gradually stabilizing by the fourth hypothesis.

Effect of the Critic. We compare Full CRIME with a variant that removes the Critic on the same test videos. The variant makes the final decision directly from the candidate hypotheses, without Critic feedback or the additional evidence gathering and hypothesis revision that such feedback would trigger. The table in Figure 3(c) evaluates the entire feedback loop. Unsafe recall decreases slightly from 84.5% without the Critic to 82.0% with it, while safe accuracy rises from 88.0% to 98.0%. Major-category accuracy also increases from 66.5% to 76.0%. These results indicate that Critic feedback and the additional evidence gathering it triggers help refine risk interpretations and improve category assignment.

6 Conclusion

In this work, we present CRIME, a multi-agent framework for improving compositional-risk video moderation. Our analysis shows that a central challenge lies in discovering decision-critical combinations from abundant, distributed video information and determining how their relations give rise to policy-violating meanings. CRIME makes this process explicit by organizing relational evidence, exploring multiple candidate risk interpretations, and verifying decisive evidence against the original video. Experiments across controlled synthetic, real-world, and cross-policy settings demonstrate its effectiveness and adaptability. Overall, this work highlights the value of explicitly discovering decision-critical evidence combinations and their relations, rather than relying solely on implicit reasoning in conventional moderation pipelines.

AI USE STATEMENT

MLLMs serve as components of CRIME and the prompting baselines. We also use an LLM judge to score final moderation explanations on the Real-World and Advertising datasets, following a reference-based rubric covering accuracy, content specificity, and conciseness. These ratings are reported as Score-LLM, separately from the two authors' Score-Human ratings on the same samples. The scoring procedure and prompt are provided in the appendix.

ETHICS STATEMENT

This work aims to improve the detection of compositional risks in videos. The paper includes potentially offensive or disturbing examples and provides a content warning. Video screening and human explanation scoring were performed by two authors, without recruiting external participants for these activities. Errors or biases in the models and moderation policies may lead to missed risks or incorrect restrictions on benign content. We therefore advocate human oversight for consequential moderation decisions. The framework's investigative capabilities could also be misused to identify ways to evade moderation. Access to and redistribution of source videos and policy material must respect the applicable permissions and privacy constraints.

REPRODUCIBILITY STATEMENT

The main text describes CRIME's architecture, investigation procedure, baselines, and ablation settings. The appendix documents dataset sources and composition, diagnostic inputs and controls, agent roles and prompt templates, evidence-graph and output schemas, and evaluation metrics and scoring rubrics, including the prompt for LLM-based explanation scoring. These materials support implementation and evaluation of the framework. Reproducing experiments with proprietary backbones requires access to the corresponding model APIs.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Anthropic. Claude Opus 4.6 System Card, February 2026a. Anthropic System Card.

Anthropic. Claude Sonnet 5 System Card, June 2026b. URL https://anthropic.com/claude-sonnet-5-system-card. Anthropic System Card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. SafeWatch: An efficient safety-policy following video guardrail model with transparent explanations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/beac6bfb7eac3d651307c16ac747df01-Abstract-Conference.html.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama Guard 3 Vision: Safeguarding human-AI image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024. URL https://arxiv.org/abs/2411.10414.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026. URL https://arxiv.org/abs/2609.19969.

Google DeepMind. Gemini 3.1 Pro: Model Card, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Google DeepMind Model Card.

Google DeepMind. Gemini 3.8 Flash: Model Card, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Google DeepMind Model Card.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. LLAVAGUARD: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 8322–8326, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/ReGenAI/html/Helff_LLAVAGUARD_VLM-based_Safeguard_for_Vision_Dataset_Curation_and_Safety_Assessment_CVPRW_2024_paper.html.

Jiao Li, Jian Lang, Xikai Tang, Wenzheng Shu, Ting Zhong, Qiang Gao, Yong Wang, Leiting Chen, and Fan Zhou. Shedding the facades, connecting the domains: Detecting shifting multimodal hate video with test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22949–22957, 2026a. doi: 10.1609/aaai.v40i27.39459. URL https://ojs.aaai.org/index.php/AAAI/article/view/39459.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Kaiju Li, Rongpei Hong, Jian Lang, Jin Wu, Fan Zhou, and Jingkuan Song. MATCH: Multi-agentic evidence grounding for explainable hate video detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(7):9646–9659, 2026b. doi: 10.1109/TCSVT.2026.3672052. URL https://doi.org/10.1109/TCSVT.2026.3672052.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06341.pdf.

Yue Liu, Shengfang Zhai, Mingzhe Du, Yulin Chen, Tri Cao, Hongcheng Gao, Cheng Wang, Xinfeng Li, Kun Wang, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. GuardReasoner-VL: Safeguarding VLMs via reinforced reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 32936–32966, 2025. doi: 10.52202/085713-0977. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2a02b560822d564119fe3ac3be024ac6-Abstract-Conference.html.

OpenAI. GPT-6 Astra System Card, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra. OpenAI Deployment Safety Hub.

Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230:279–293, 2017. doi: 10.1016/j.neucom.2016.12.017.

Qwen Team. Qwen3.8-27B, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Official model card.

Ruotong Wang, Zihao Zhu, Siwei Lyu, Xin Tao, and Baoyuan Wu. Distributed implicit harm: A compositional safety blind spot in MLLM-based video moderation, 2026. URL https://arxiv.org/abs/2609.00206.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. URL https://arxiv.org/abs/2508.18265.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. Filter-and-refine: A MLLM based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 873–880. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-industry.62. URL https://aclanthology.org/2025.acl-industry.62/.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. URL https://arxiv.org/abs/2303.04671.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Kejuan Yang, Yizhuo Zhang, Mingyuan Du, Yue Zhang, Dixin Zheng, Kaili Zhao, Yang Xiao, Hanzhong Liang, and Kenan Xiao. UNIVID: Unified vision-language model for video moderation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 467–479. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-industry.32. URL https://aclanthology.org/2026.acl-industry.32/.

Shuonan Yang, Yuchen Zhang, and Zeyu Fu. Training-free and interpretable hateful video detection via multi-stage adversarial reasoning. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12352–12356. IEEE, 2026b. doi: 10.1109/ICASSP55912.2026.11464444. URL https://arxiv.org/abs/2601.15115.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. URL https://arxiv.org/abs/2303.11381.

Asim Sinan Yuksel and Fatma Gulsah Tan. DeepCens: A deep learning-based system for real-time image and video censorship. *Expert Systems*, 40(10):e13436, 2023. doi: 10.1111/exsy.13436.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. ShieldGemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025. URL https://arxiv.org/abs/2504.01081.

Yuyang Zhang, Xudong Jiang, Yuxuan Song, Yuxiang Sun, Yihao Huang, Run Wang, Shundi Xiao, and Lina Wang. VSG-Safe: Spotting NSFW video through cross-frame evidence. In *35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/zhang-yuyang.

Zhenhao Zhu, Yue Liu, Yanpei Guo, Wenjie Qu, Cancan Chen, Yufei He, Yibo Li, Yulin Chen, Tianyi Wu, Huiying Xu, Xinzhong Zhu, and Jiaheng Zhang. GuardReasoner-Omni: A reasoning-based multi-modal guardrail for text, image, video, and audio. *arXiv preprint arXiv:2602.03328*, 2026. URL https://arxiv.org/abs/2602.03328v2.