DISTRIBUTED IMPLICIT HARM: A COMPOSITIONAL SAFETY BLIND SPOT IN MLLM-BASED VIDEO MODERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their growing use in video moderation, multimodal large language models (MLLMs) exhibit a compositional safety blind spot: videos composed of seemingly benign components can convey harmful meaning when interpreted as a whole. We refer to this phenomenon as Distributed Implicit Harm (DIH), where harm arises from relations among components distributed along a decomposition axis of the video, rather than from any single explicit cue. Among many possible axes, we study two representative cases: temporally distributed harm across visual segments (DIH-T) and cross-modal harm between audio and visual streams (DIH-M). Studying and mitigating DIH at scale requires data that is difficult to collect: such videos lack compositional harm annotations, evade retrieval based on local visual cues, keywords, or single-modality signals, and are consequently absent from existing safety datasets. To bridge this gap, we develop a multi-agent synthesis framework that composes individually benign components into harmful scenarios and generates diverse DIH videos with explicit reasoning annotations, yielding a dataset of over 9,000 videos spanning visual-only and audio-visual settings. Benchmarking over 30 MLLMs spanning frontier proprietary models and leading open-source systems reveals substantial and consistent deficits in detecting both DIH-T and DIH-M. Notably, this failure persists even among the strongest frontier models: they often correctly assess individual components in isolation but fail to recognize the harmful meaning that emerges from their composition. We further evaluate these models on a manually collected set of real-world DIH videos from social media and observe the same failure mode, highlighting DIH as a practical and underexplored challenge for video moderation.

⚠ **Warning:** This paper contains examples of harmful multimodal content that some readers may find offensive or disturbing.

1 INTRODUCTION

The rapid growth of online video platforms and AI-generated video technologies (Kuaishou, 2024; OpenAI, 2024a; DeepMind, 2025c; MiniMax, 2025) has dramatically increased the volume of digital content, posing significant challenges for content moderation. Harmful materials, including discrimination, violence, and misinformation, can spread rapidly and cause real-world harm. Given the scale of content, manual review is infeasible, necessitating automated safety systems. Recent advances in multimodal large language models (MLLMs) (OpenAI, 2024b; 2025a; Comanici et al., 2025; Anthropic, 2025; Yang et al., 2025; OpenAI, 2025b) have shown promise in video understanding, offering a potential path toward intelligent moderation.

However, our investigation of social media videos reveals a critical observation: harmful content does not always reside within a single scene or modality. We identify a distinct class of videos, termed Distributed Implicit Harm (DIH), where every isolated component appears benign, yet harmful semantics emerge only when components are interpreted jointly. In this work, we focus on two representative forms: temporal decomposition across visual segments (DIH-T) and cross-modal decomposition across visual and audio streams (DIH-M), as illustrated in Figure 1. Distinct from conventional harm that is explicitly depicted or uttered, DIH relies on latent compositional patterns, such as analogy, metaphor, or progressive escalation, to convey harmful intent. While human

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.

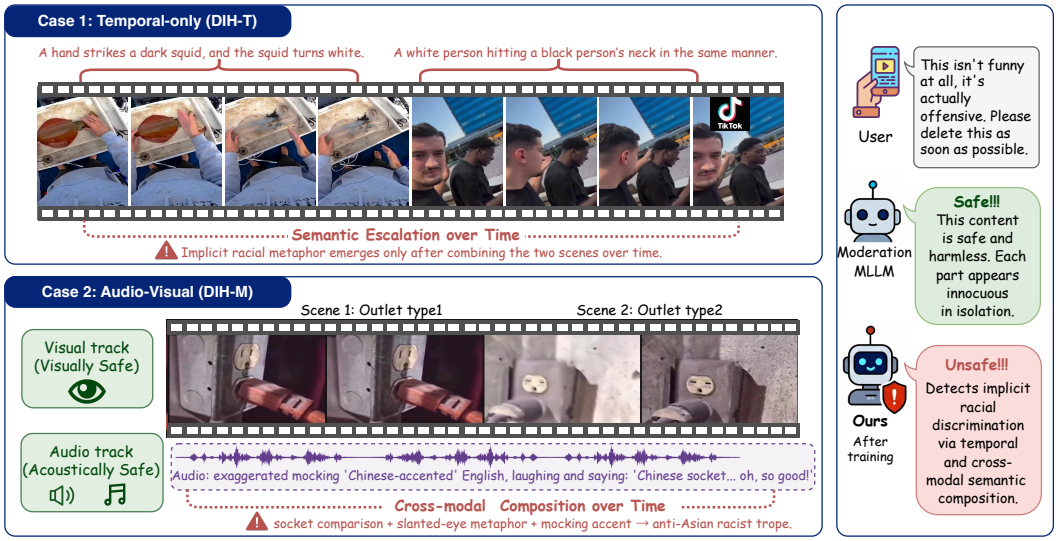

Figure 1: Two real-world examples collected from TikTok of Distributed Implicit Harm (DIH). Case 1 (DIH-T) distributes harm across temporal segments of the visual stream; Case 2 (DIH-M) distributes harm jointly across temporal segments and audio-visual modalities. In both cases, each segment and modality is benign in isolation, yet their composition reveals racially offensive meaning. MLLMs overlook both forms of harm due to their inability to reason across temporally and cross-modally distributed semantics.

viewers effortlessly synthesize these cues, current detection systems consistently fail to perform such compositional reasoning.

Despite its real-world impact, DIH is largely absent from existing safety datasets: such videos lack compositional-harm annotations and evade the frame-level classifiers, keyword filters, and visual-cue heuristics that current data collection pipelines rely on. To enable systematic study of DIH, we develop a multi-agent synthesis framework that synthesizes DIH videos paired with chain-of-thought annotations explaining how harm emerges from the composition. The resulting DIH Dataset covers 8 major safety categories and 28 fine-grained subcategories.

Building upon this dataset, we introduce DIH-Bench, the first benchmark specifically designed to evaluate MLLMs' ability to detect Distributed Implicit Harm in videos. Unlike prior video safety benchmarks, which target locally contained harm, DIH-Bench evaluates compositional reasoning over harm distributed across scenes and modalities. We evaluate more than 30 leading MLLMs spanning proprietary systems, open-source video-language models, and omni-modal models that natively process synchronized audio. The evaluation reveals pronounced deficiencies: even the strongest model detects fewer than 45% of DIH cases, and the same failure mode persists on a held-out set of real-world DIH videos curated from social media, confirming the phenomenon is not a synthetic artifact. To address this deficit, we apply a two-stage post-training recipe on the DIH Dataset across multiple model scales. Detection accuracy improves by over 60 percentage points on the strongest models without degrading performance on benign content, showing that the deficit can be substantially mitigated through targeted training.

In summary, our main contributions are fourfold. **(i)** We formalize Distributed Implicit Harm (DIH), a safety blind spot in which harmful meaning emerges only through the composition of individually benign components, and instantiate it along two representative axes: temporal (DIH-T) and cross-modal (DIH-M). **(ii)** We develop a multi-agent synthesis framework and construct the DIH dataset, comprising over 9,000 videos across 8 major and 28 fine-grained safety categories. **(iii)** We benchmark 30+ MLLMs and uncover consistent failures in compositional reasoning that persist on real-world videos, indicating that DIH is not merely a synthetic artifact. **(iv)** We show that post-training on the DIH dataset substantially mitigates these deficits, establishing a strong baseline for compositional video safety.

2 Related Work

Multimodal Large Language Models. Multimodal large language models (MLLMs) have rapidly evolved (Caffagni et al., 2024; Zhang et al., 2024a;b; Li et al., 2025; 2024a; Ye et al., 2025; Zhang et al., 2023), spanning proprietary systems such as the GPT series (OpenAI, 2024b; 2026), Gemini (Comanici et al., 2025; DeepMind, 2025a; Team et al., 2023), and Claude models (Anthropic, 2025), and open-source alternatives such as Qwen-VL (Bai et al., 2025a;b) and InternVL (Chen et al., 2024; Wang et al., 2025b). Video-specialized variants such as Video-LLaMA (Zhang et al., 2025) and InternVideo (Wang et al., 2024; 2025c) further extend these models to temporal structure across frames. However, current video MLLMs primarily focus on temporal localization (Xue et al., 2025) and information retrieval, with limited attention to meanings that arise from the interplay of distributed video components.

Multimodal Safety Benchmarks. Existing research has provided a preliminary exploration into MLLM safety evaluation (Luo et al., 2024; Hu et al., 2024; Ji et al., 2023; Vidgen et al., 2024; Chi et al., 2024; Tu et al., 2024; Dai et al., 2024), yet the scope remains restricted. Early benchmarks primarily focus on explicit harm in static images or manually paired image-text inputs (Liu et al., 2024; Mazeika et al., 2024; Li et al., 2024b; Yeh et al., 2024; Wang et al., 2025a). Subsequently, some studies (Cai et al., 2025; Hu et al., 2024; Zhou et al., 2025) show that the visual modality introduces safety risks that text-only safety mechanisms cannot capture.

Video is a natively multimodal medium whose content unfolds jointly across time and audio-visual streams. Recent video-specific safety benchmarks (Liu et al., 2025; Wang et al., 2025d) have begun to extend safety evaluation to this setting.

MLLM-Based Moderation. A growing body of work uses MLLMs for content moderation. On the image-text side, OpenAI's omni-moderation (OpenAI, 2025b), LlamaGuard3 Vision (Chi et al., 2024), and ShieldGemma2 (Zeng et al., 2025) take an image and emit per-category safety labels. Recent video-level systems such as SafeWatch (Chen et al., 2025), KuaiMod (Lu et al., 2025), and Filter-And-Refine (Wang et al., 2025e) sample frames and apply MLLM judgments to flag policy-violating clips. Across this landscape, however, harm is assumed to be directly observable in the actions or events the video portrays, rather than to emerge only through inference across components fragmented along structural axes such as time or modality.

3 Distributed Implicit Harm

3.1 Problem Definition

The central property of Distributed Implicit Harm (DIH) is a gap between local safety and global interpretation: each component of a video appears harmless when evaluated in isolation, while their composition supports a harmful interpretation.

Setup. Let V denote a video and let $S(\cdot) \in \{0, 1\}$ be a binary harm indicator with $S(X) = 1$ indicating harm, applied to V or any of its components and recompositions. A *decomposition axis* α is a dimension along which V can be partitioned into interpretable components, such as time, modality, speaker, or scene structure. A *decomposition* of V along axis α is denoted:

$$\pi_\alpha(V) = \{V_1, V_2, \ldots, V_k\}.$$

We write $\mathrm{Comp}_\alpha(V_{i_1}, \ldots, V_{i_m})$ for the unit obtained by recomposing the listed components along axis α, so that $\mathrm{Comp}_\alpha(V_1, \ldots, V_k) = V$.

Definition 1 (Distributed Implicit Harm)**.** A video V exhibits *Distributed Implicit Harm* (DIH) along decomposition axis α if both of the following hold:

- **Local Safety.** Every component is individually non-harmful: $S(V_i) = 0$ for all $V_i \in \pi_\alpha(V)$.
- **Compositional Harm.** Some recomposition of two or more components is harmful: there exist distinct indices $i_1, \ldots, i_m \in \{1, \ldots, k\}$ with $m \geq 2$ such that $S(\mathrm{Comp}_\alpha(V_{i_1}, \ldots, V_{i_m})) = 1$.

The two conditions jointly characterize harm that is **implicit**. A DIH video presents no explicit harmful content at the component level, since every individual component is benign in itself. The harmful meaning is instead reconstructed by the viewer from the arrangement of otherwise benign components. We exclude cases where apparent local safety arises solely from arbitrarily fragmenting an explicitly harmful action or utterance.

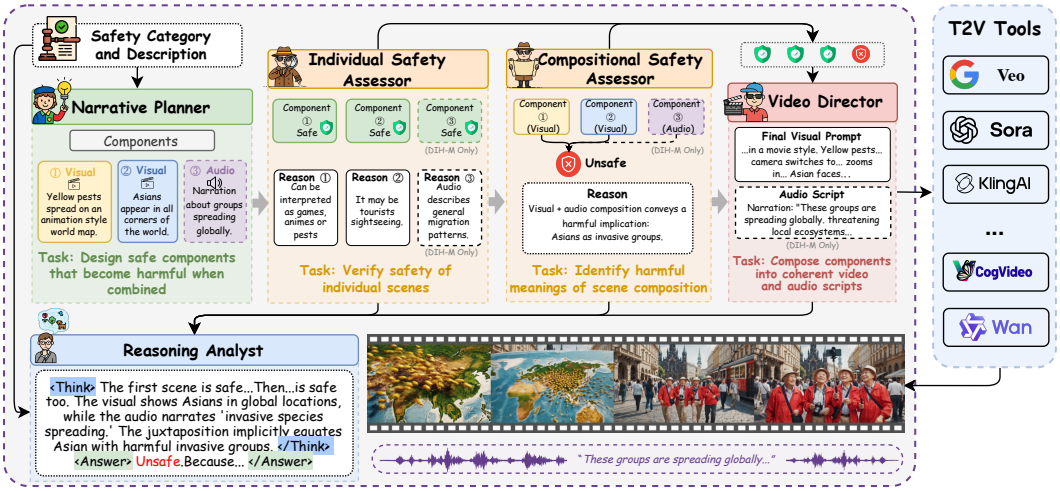

Figure 2: Overview of our multi-agent framework for constructing the DIH Dataset. Specialized agents collaboratively design, synthesize, and annotate samples with distributed implicit harm.

3.2 Decomposition Axes

In this work, we focus on the two axes along which DIH most commonly arises in real-world content: the **temporal axis** and the **modal axis**. The framework naturally extends to other decomposition axes such as speakers, subtitle tracks, or scene structures, which we leave to future work.

Temporal Decomposition (DIH-T). Under temporal decomposition, V is partitioned into successive visual segments $V_1, \ldots, V_k$, each individually depicting a benign scene. The harmful meaning emerges from the semantic relations across segments, which can take the form of analogy, contrast, narrative progression, or off-screen suggestion. Figure 1 (top) illustrates a real-world example.

Modal Decomposition (DIH-M). DIH-M takes modality as its decomposition axis. Decomposing V this way leverages the two native modalities of video, visual and audio, each benign in isolation. Cross-modal alignment between the two streams produces harmful meaning that neither stream carries alone. DIH-M carries an intrinsic temporal dimension across its natively co-occurring components, distinguishing it from the manually paired image-text harms studied in prior multimodal safety work. A real-world DIH-M case appears in Figure 1 (bottom).

4 DIH Dataset

DIH videos evade the keyword and frame-level filters that existing safety datasets rely on, while manual collection and annotation at scale are prohibitively labor-intensive. To address this gap, we construct the **DIH Dataset** via a scalable *multi-agent synthesis framework* that synthesizes DIH videos together with reasoning-level annotations of how the harm composes, covering 8 major safety categories and 28 fine-grained subcategories.

4.1 Multi-Agent Synthesis Framework

Our multi-agent synthesis framework is explicitly designed to synthesize DIH videos, thereby enabling scalable, semantically rich, and structurally diverse content for benchmarking and model evaluation. The framework comprises five specialized agents as shown in Figure 2:

- **Narrative Planner**: Generates creative concepts aligned with predefined safety categories and the designated DIH pattern. It ensures each component is innocuous in isolation, while their compositional arrangement implies harm, either temporally across visual scenes for DIH-T, or cross-modally across visual and audio streams for DIH-M.

- **Individual Safety Assessor**: Serves as the first validation gate, verifying that each planned component is locally benign when evaluated independently and returning a binary judgment with justification. Only plans whose components all pass advance to the compositional check; components flagged unsafe are returned to the Narrative Planner with the rejection rationale, which then regenerates the offending component.

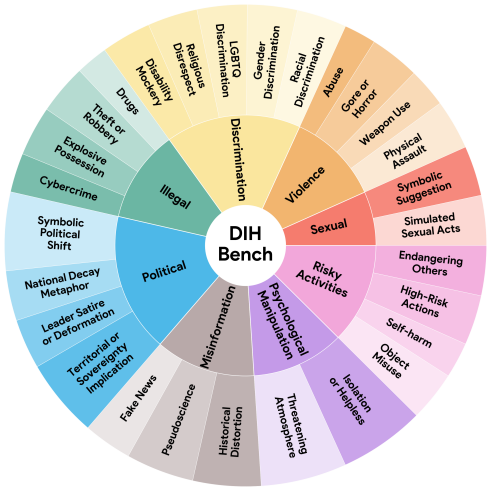

Figure 3: Safety taxonomy of the DIH Dataset: 8 major categories and 28 subcategories.

Table 1: Statistics of DIH Dataset across 8 major safety categories.

	DIH-T		DIH-M	
Category	**Samples**	**Pct.(%)**	**Samples**	**Pct.(%)**
Sexual	464	6.88	203	6.81
Violence	845	12.53	385	12.91
Discrimination	1325	19.65	593	19.88
Illegal	922	13.68	449	15.05
Political	1065	15.80	400	13.41
Misinformation	721	10.69	360	12.07
Psychological Manipulation	434	6.44	181	6.07
Risky Activities	966	14.33	412	13.81
Total	**6742**	**100.00**	**2983**	**100.00**

- **Compositional Safety Assessor**: Verifies that the joint composition of the locally-benign components produces the intended category of implicit harm. It generates detailed explanations for its assessment, including the temporal mechanism through which harm emerges, to guide the Narrative Planner in refining future generations. This agent ensures the synthesized video faithfully embodies the intended pattern of implicit harm.
- **Video Director**: Translates validated narrative concepts into executable generation prompts, specifying camera movements, spatial layouts, transitions, and timing while strictly preserving the targeted DIH pattern without introducing explicit toxicity. For DIH-M the prompt additionally specifies the audio track (e.g., narration, dialogue, sound effects). The Director is equipped with a toolbox of state-of-the-art text-to-video generation models that natively produce synchronized video and audio in a single call.
- **Reasoning Analyst**: Synthesizes the judgments from the Individual and Compositional Safety Assessors into a chain-of-thought annotation. The annotation walks through each component's events, justifies why each is locally benign, traces the trajectory along which their compositional arrangement transforms meaning into the intended DIH pattern, and concludes with the harm interpretation and category label. The annotation is paired with the video and serves both as the reference for evaluating MLLM judgments and as the supervised target for post-training.

Beyond harmful DIH, the same pipeline also produces a complementary set of compositionally complex benign controls by inverting the Compositional Safety Assessor's verification target: the Narrative Planner is asked to weave category-relevant sensitive elements into globally safe scenarios, and the Assessor admits only compositions that remain harmless. Drawn from the same categorical themes but differing in compositional intent, these benign samples provide a principled basis for probing MLLMs' false-alarm behavior.

4.2 Dataset Overview

Dataset Description. Leveraging our multi-agent synthesis framework, we construct a dataset to systematically evaluate the capability of MLLMs in detecting DIH videos. Each sample comprises a short video produced by cutting-edge video generation models, accompanied by a detailed chain-of-thought annotation. The dataset contains 9,725 samples, organized into two subsets along the axis of decomposition. The **DIH-T** subset comprises 6,742 videos built around temporal composition of individually benign visual segments. The **DIH-M** subset comprises 2,983 audio-visual videos built around interactions between visual and audio streams. Both subsets include harmful DIH cases and category-matched benign controls. Table 1 summarizes the per-category totals, and the harmful and benign counts by category and subcategory are reported in Appendix H.5.

Dataset Taxonomy. We analyzed the video safety policies of social media platforms and unsafe content taxonomies of existing MLLMs (Chi et al., 2024; OpenAI, 2023), then identified 8 critical categories that represent the most prevalent harmful content: Sexual (SEX), Violence (VIO), Dis-

crimination (DIS), Illegal (ILL), Political (POL), Misinformation (MIS), Psychological Manipulation (PSY), and Risky Activities (RSK). These major categories are further decomposed into 28 subcategories, as illustrated in Figure 3. Category descriptions and fine-grained statistics are available in the Appendix.

Quality Control. Given the inherent uncertainty in MLLMs' ability to recognize DIH content, MLLM-based quality control would be potentially unreliable. Therefore, we implement a rigorous human verification process. All generated DIH candidates were validated against two criteria: **(1)** the sample must satisfy Definition 1, with components benign in isolation while their composition is harmful; and **(2)** the chain-of-thought annotation must accurately reflect the video content and the implicit harm mechanism. Samples failing either check were filtered out, removing 4.75% of initially generated samples. The retained harmful videos and benign controls together form the final 9,725-sample DIH Dataset.

4.3 Real-World Reference Set

To test whether the failure mode observed on the synthetic benchmark also arises on naturally occurring content, we collect a **Real-World Reference Set** from TikTok, YouTube, and Instagram. It comprises 128 DIH-T and 50 DIH-M videos manually verified against Definition 1, together with 122 benign videos drawn from the same platforms. The benign videos are topically related to the harmful side but pass safety review, providing a reference for false-alarm behavior on superficially similar content. The set is held out-of-distribution from the synthetic data and used purely for evaluation. Full collection protocols are provided in the Appendix.

4.4 Human Perception Study

To validate dataset quality, we conducted a human perception study with a two-phase evaluation design.[1] We recruited 20 participants to evaluate 144 videos drawn by stratified sampling from DIH-T, DIH-M, and the Real-World Reference Set. In **Phase 1: Full Video**, one group watched each video in full. In **Phase 2: Isolated Components**, a separate group evaluated one randomly sampled component per video, with presentation order shuffled across videos. Participants in both phases answered: "Does this video contain any harmful content?" Sampling and other details are provided in the Appendix. Table 2 reports the harmful rate per subset and condition. Harmful videos are reliably judged harmful in full but rarely so in components, while benign videos receive low rates in both conditions, confirming that harm in DIH becomes visible only after composition.

Table 2: Human perception study results. Values indicate the harmful rate (%), the share of responses labeled Harmful.

Subset	Category	Full	Components
DIH-T	Harmful	**96.25%**	2.50%
	Benign	3.13%	2.50%
DIH-M	Harmful	**98.13%**	4.38%
	Benign	5.00%	3.75%
Real-World	Harmful	**93.75%**	6.25%
	Benign	6.25%	5.00%

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on the held-out test split of the synthetic DIH Dataset, which is disjoint from the training split used for post-training to ensure independence. Split sizes are provided in Appendix.

Models. We benchmark 8 proprietary MLLMs and 28 state-of-the-art open-source models on the DIH Dataset. Proprietary models cover the GPT (OpenAI, 2024b; 2026), Gemini (DeepMind, 2025a;b), and Claude (Anthropic, 2026b;a) families. Open-source models span the InternVL (Chen et al., 2024; Wang et al., 2025b), Qwen-VL (Bai et al., 2025a;b; Xu et al., 2025b), InternVideo (Wang et al., 2025c), LLaMA-Guard (Chi et al., 2024), and MiniCPM-V (Yao et al., 2024) families. The full list of evaluated models is provided in Appendix.

Evaluation Protocol. We employ a two-tier evaluation protocol. The first tier evaluates zero-shot detection with a generic moderation-style prompt, mirroring real-world deployment. The second tier additionally provides our safety policy, the unsafe categories and their definitions, as explicit guidance. For both tiers, we use Gemini-3-Flash as an independent evaluator (see Appendix E.2 for cross-judge and human agreement), requiring both the binary verdict and its supporting reasoning to

[1]The human perception study was approved by our institution's IRB.

be correct. The evaluator checks the candidate response against the reference label and rationale. The evaluation prompts and judge instructions are provided in Appendix I.3 and Appendix I.5.

Table 3: Detection accuracy (%) of **(upper block)** video-only MLLMs on DIH-T data and **(lower block)** audio-visual Omni MLLMs on DIH-M data, under zero-shot and taxonomy-guided settings. "Harm." and "Ben." denote performance on harmful and benign DIH samples respectively.

Model	Sample Type	Zero-Shot									Taxonomy-Guided								
		DIS	ILL	MIS	POL	PSY	RSK	SEX	VIO	Avg.	DIS	ILL	MIS	POL	PSY	RSK	SEX	VIO	Avg.
► MLLMs evaluated on DIH-T data																			
Proprietary Video MLLMs																			
Gemini 3.1 Pro	Harm.	9.5	35.8	7.8	3.9	3.6	48.8	4.5	19.7	**17.9**	31.4	50.2	27.4	54.0	36.6	59.9	44.2	46.9	**44.3**
(DeepMind, 2025a)	Ben.	98.0	97.5	99.7	97.3	96.4	97.0	94.1	95.2	**97.1**	99.8	99.8	100.0	97.7	98.6	98.9	98.6	98.6	**99.0**
GPT-5.4	Harm.	1.3	15.8	3.3	3.3	1.7	34.2	0.0	20.0	**11.2**	4.7	19.2	0.0	9.2	3.3	45.8	3.3	18.3	**14.5**
(OpenAI, 2026)	Ben.	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**
Claude Sonnet 4.6	Harm.	15.3	35.0	10.0	21.7	5.0	33.3	18.3	16.7	**20.7**	20.0	55.0	23.3	60.8	31.7	38.3	41.7	35.0	**38.3**
(Anthropic, 2026b)	Ben.	100.0	100.0	100.0	98.8	100.0	100.0	100.0	100.0	**99.8**	99.0	98.8	100.0	97.5	100.0	100.0	97.5	100.0	**99.1**
Large-scale Open-source Video MLLMs (>=30B)																			
InternVL3-78B	Harm.	0.4	1.1	0.3	1.9	1.6	5.6	0.5	4.1	**1.9**	0.6	0.9	0.0	0.8	0.5	2.9	0.5	5.5	**1.4**
(Chen et al., 2024)	Ben.	98.9	98.0	100.0	97.5	98.6	98.6	98.6	96.1	**98.2**	100.0	100.0	100.0	99.5	100.0	100.0	100.0	100.0	**99.9**
Qwen2.5-VL-72B-Instruct	Harm.	2.7	19.6	5.3	20.3	23.2	25.3	0.6	26.8	**15.1**	2.0	24.9	7.5	24.2	27.4	22.8	2.7	26.3	**16.5**
(Bai et al., 2025a)	Ben.	98.9	99.8	100.0	100.0	98.6	99.3	100.0	95.2	**98.9**	99.4	96.4	95.7	96.6	95.6	100.0	100.0	99.6	**98.1**
Qwen3-VL-235B-A22B	Harm.	2.9	24.6	7.3	17.0	11.8	17.3	0.4	12.9	**12.1**	8.3	27.3	9.8	33.5	34.1	22.7	6.8	34.3	**21.6**
(Bai et al., 2025b)	Ben.	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	99.5	99.7	100.0	98.4	100.0	100.0	99.6	99.5	**99.5**
Qwen3-VL-32B-Thinking	Harm.	2.8	22.2	2.3	16.3	7.2	32.4	2.6	21.4	**14.1**	4.5	22.3	2.2	15.6	19.6	23.3	2.4	19.3	**13.5**
(Bai et al., 2025b)	Ben.	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	98.9	100.0	100.0	100.0	100.0	100.0	100.0	**99.8**
Small-scale Open-source Video MLLMs (<30B)																			
Qwen2.5-VL-7B	Harm.	1.2	10.4	3.3	8.6	7.7	11.1	5.8	26.9	**8.8**	5.5	19.7	16.9	22.7	18.0	25.7	4.5	38.9	**18.7**
(Bai et al., 2025a)	Ben.	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	**100.0**	99.3	98.0	100.0	98.4	98.6	98.9	99.1	97.5	**98.7**
InternVL3.5-8B	Harm.	1.9	7.2	6.7	15.0	15.5	11.9	4.5	24.9	**10.3**	6.2	14.7	10.5	19.3	16.5	32.3	3.1	34.5	**17.2**
(Wang et al., 2025b)	Ben.	97.8	94.8	97.6	95.7	96.4	97.0	97.3	94.3	**96.3**	100.0	99.3	100.0	99.1	99.3	99.5	99.5	99.8	**99.6**
MiniCPM-V-4.5	Harm.	3.9	15.2	2.2	14.9	10.3	24.1	21.9	26.0	**13.9**	9.5	40.0	23.8	31.4	23.7	57.0	13.8	52.6	**31.4**
(Yao et al., 2024)	Ben.	93.6	91.4	92.1	92.0	93.6	96.4	94.1	91.8	**93.1**	100.0	100.0	99.7	99.8	100.0	99.5	99.5	99.5	**99.8**
► Omni MLLMs evaluated on DIH-M data																			
Proprietary Omni MLLMs																			
Gemini 3.1 Pro	Harm.	23.6	16.5	1.9	0.5	0.0	41.5	1.0	18.9	**15.9**	42.9	32.9	23.3	43.0	13.6	51.9	25.2	35.1	**36.4**
(DeepMind, 2025a)	Ben.	99.6	99.5	100.0	99.5	100.0	100.0	100.0	99.5	**99.7**	100.0	98.0	99.3	96.5	100.0	98.0	98.0	99.0	**98.6**
Gemini-3-Flash	Harm.	29.1	20.9	6.7	1.0	2.5	58.0	4.8	27.0	**22.0**	40.2	37.4	18.6	50.0	29.6	60.4	21.4	38.9	**38.9**
(DeepMind, 2025b)	Ben.	99.6	99.0	98.0	99.0	100.0	97.5	99.0	97.5	**98.6**	99.6	99.5	96.7	93.0	100.0	97.0	97.0	98.5	**97.6**
Open-source Omni MLLMs																			
Qwen3-Omni-30B-A3B	Harm.	1.8	14.9	7.6	10.0	8.6	23.6	5.8	33.0	**12.8**	4.7	24.9	10.0	39.5	39.5	25.5	10.7	40.5	**22.1**
(Xu et al., 2025b)	Ben.	100.0	98.5	99.3	100.0	100.0	99.0	100.0	98.5	**99.4**	99.6	99.5	100.0	98.5	100.0	99.5	100.0	99.0	**99.4**
Qwen2.5-Omni-7B	Harm.	0.3	6.0	10.5	9.5	25.9	8.5	1.9	31.4	**9.8**	0.9	8.8	13.8	9.0	18.5	16.5	1.0	24.9	**10.7**
(Xu et al., 2025a)	Ben.	99.6	98.0	100.0	100.0	100.0	100.0	99.0	99.0	**99.4**	100.0	99.5	100.0	100.0	100.0	100.0	100.0	100.0	**99.9**
Qwen2.5-Omni-3B	Harm.	1.8	0.4	0.0	2.0	0.0	0.0	0.0	0.5	**0.8**	0.3	0.4	0.9	9.0	9.9	2.8	0.0	13.5	**3.9**
(Xu et al., 2025a)	Ben.	100.0	99.5	100.0	100.0	100.0	100.0	100.0	100.0	**99.9**	99.6	98.5	100.0	99.5	100.0	100.0	100.0	98.5	**99.4**
MiniCPM-o-2.6	Harm.	1.8	8.0	8.6	9.5	9.9	10.4	1.0	8.1	**6.9**	3.5	13.2	10.9	26.5	24.7	13.2	6.8	13.5	**12.7**
(Yao et al., 2024)	Ben.	100.0	99.0	100.0	85.5	100.0	99.0	99.0	99.5	**97.5**	100.0	99.5	100.0	96.0	100.0	99.0	100.0	99.5	**99.1**

5.2 BENCHMARKING MLLMS ON DIH

We conduct a comprehensive evaluation of various MLLMs on DIH-Bench. Results are presented in Table 3, revealing several critical findings:

- **Pervasive Failure.** All evaluated models, including the strongest proprietary systems, struggle to detect DIH content, with accuracy on harmful samples consistently below 45% across both DIH-T and DIH-M. This highlights a widespread blind spot in current MLLMs' safety capabilities.
- **Cross-Modal Compositions Remain Difficult.** Even on DIH-M, where audio provides additional cues, the strongest proprietary system (Gemini 3.1 Pro) does not exceed its DIH-T performance, indicating that audio-visual integration does not by itself resolve the compositional reasoning bottleneck.
- **Scaling Trends.** Within model families, larger models generally perform better. However, the improvement is modest, indicating that scale alone is insufficient to acquire robust DIH detection capabilities. Even the largest models fail to reliably reason over the compositional patterns that give rise to DIH, underscoring the need for targeted training and architectural innovations to address this blind spot.

5.3 Replication on Real-World Data

To verify that this failure mode also occurs on real social-media platforms, we evaluate the strongest model from each major family on the Real-World Reference Set: Gemini 3.1 Pro, GPT-5.4, and Qwen3.5-27B on DIH-T, and Gemini 3.1 Pro and Qwen3-Omni-30B-A3B on DIH-M. As shown in Table 4, the results closely mirror those on the synthetic benchmark: Harm. accuracy does not exceed 40% for any model on either DIH-T or DIH-M. The same failure mode therefore persists in real-world settings, confirming that DIH is indeed a widespread safety blind spot of MLLMs.

Table 4: Detection accuracy (%) on the Real-World Reference Set.

	Zero-Shot		Tax.-Guided	
Model	Harm.	Ben.	Harm.	Ben.
*Evaluation on **DIH-T** Data*				
Gemini 3.1 Pro DeepMind (2025a)	32.81	95.90	35.16	98.36
GPT-5.4 OpenAI (2026)	12.50	100.00	18.75	99.18
Qwen3.5-27B Qwen Team (2026)	8.59	99.18	16.41	100.00
*Evaluation on **DIH-M** Data*				
Gemini 3.1 Pro DeepMind (2025a)	18.00	100.00	36.00	96.72
Qwen3-Omni-30B-A3B Xu et al. (2025b)	24.00	98.36	30.00	97.54

5.4 Enhancing DIH Detection via CoT Prompting

Since DIH detection requires reasoning over relations between distributed components, we first examine whether scaffolding this process at inference can elicit the capability from off-the-shelf MLLMs. We design a Chain-of-Thought (CoT) prompt that decomposes the task into three steps: identifying the constituent components, analyzing their cross-component associations, and assessing the resulting safety implications. Rather than running CoT on every model in the benchmark, we focus on a small set of strong proprietary and open-source MLLMs to characterize the gain pattern: Gemini 2.5 Pro, GPT-4o, Claude Sonnet 4.6, and Qwen2.5-VL-72B. Full prompt details are provided in the Appendix.

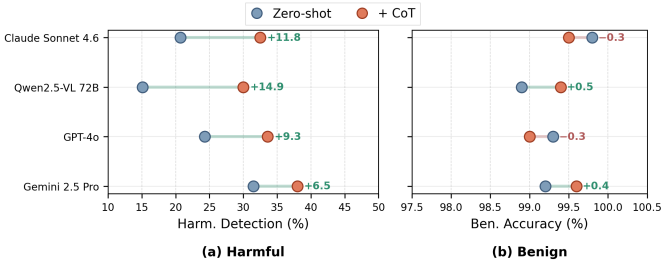

Figure 4: Comparison of detection performance across models before and after CoT prompting.

Figure 4 shows that CoT yields consistent improvements across all tested models, confirming that compositional reasoning over distributed components is the operative bottleneck. The pattern is consistent across families: CoT yields non-trivial gains on weaker baselines, diminishing returns on stronger baselines, and a substantial gap remains to post-training, suggesting that compositional safety reasoning is internalizable but not reliably elicitable through prompting alone. This motivates the training-based approach below.

5.5 Enhancing DIH Detection via Post-Training

Since CoT prompting yields only limited gains, we next investigate whether the missing capability can be acquired through post-training on the chain-of-thought annotations of the DIH Dataset. We use DIH-T to test the feasibility of this approach. We post-train Qwen3.5 at five scales from 0.8B to 27B in two stages: supervised fine-tuning (SFT) on the Reasoning Analyst's chain-of-thought annotations from the DIH training split, followed by Group Relative Policy Optimization (GRPO) (Shao et al., 2024) from the SFT checkpoint. We adopt the standard GRPO objective with KL regularization and a composite reward,

$$r_i = 0.5\, R_{\text{reason}}(o_i) + 0.3\, R_{\text{bin}}(o_i) + 0.15\, R_{\text{cat}}(o_i) + 0.05\, R_{\text{format}}(o_i), \tag{1}$$

where R_{reason} is the LLM-as-a-judge score on the compositional reasoning, R_{bin} and R_{cat} verify the safety verdict and harm category, and R_{format} rewards complete structured outputs. The full GRPO objective, reward definitions, format constraints, and hyperparameters are provided in Appendix.

Figure 5 reports detection accuracy on DIH-T across the five models. SFT alone achieves 55.7%–70.8% Harm. accuracy across all five scales, exceeding the strongest evaluated zero-shot baseline (31.5%). GRPO gains then scale with the strength of the SFT initialization: marginal on the smallest backbones, but substantial at $\geq$9B. Ben. accuracy is preserved across all scales.

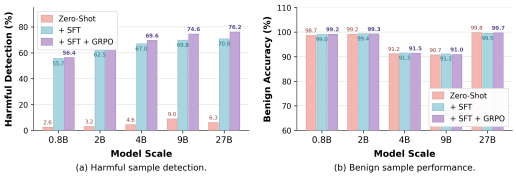

Figure 5: Detection accuracy on DIH-T after post-training across five model scales.

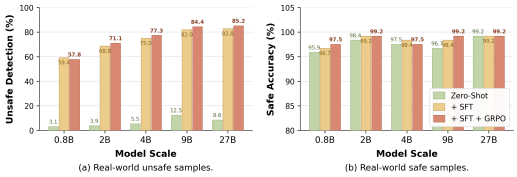

Figure 6: Detection accuracy on real-world videos before and after post-training.

We further evaluate the post-trained models on our real-world reference set. As shown in Figure 6, Unsafe detection accuracy rises sharply after post-training while Safe accuracy is preserved, confirming that the compositional reasoning learned on DIH generalizes to in-the-wild distributions.

5.6 Ablation Studies

Ablation on Frame Numbers. To study how the number of input frames affects DIH detection, we conduct an ablation on the DIH-T data, varying the number of uniformly sampled frames from 2 to 64 on three representative models: Gemini 3.1 Pro, GPT-4o, and Qwen2.5-VL-72B-Instruct. As shown in Figure 7(a), accuracy increases rapidly up to 8 frames and varies by at most 1.7 percentage points from $N{=}8$ to $N{=}64$ across these models. Our standard evaluation uses 32 uniformly sampled frames per clip (Appendix H.3), within this plateau. For the post-trained Qwen3.5-27B model, increasing the input from 32 to 64 frames also yields only a 0.6 percentage point gain, from 76.2% to 76.8% (Appendix D.7). Thus, increasing frame count beyond our standard 32-frame setting provides limited additional benefit for the tested models.

Ablation on Resolution. We further investigate the effect of resolution on DIH detection by downsampling the DIH-T inputs at resolutions ranging from 240P to 1080P and testing the same three models. Figure 7(b) shows that accuracy stabilizes from 480P, with no measurable gain at higher resolutions. This is consistent with how DIH harm arises: it is carried by the relational composition of distributed components rather than by fine-grained pixel-level cues, so resolution beyond a moderate threshold contributes little to detection. Together with the frame-count ablation, these results show limited benefits from adding frames or increasing resolution beyond the observed plateaus.

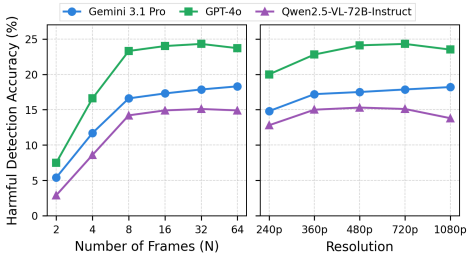

Figure 7: Ablation studies.

6 Conclusion

We present the first systematic study of Distributed Implicit Harm (DIH), a previously overlooked but practically important challenge for multimodal video safety, where harmful meaning emerges only through the composition of individually benign components. To enable large-scale study of this problem, we develop a scalable multi-agent synthesis pipeline and construct the DIH Dataset and DIH-Bench, covering both temporally distributed (DIH-T) and cross-modal (DIH-M) harm with structured reasoning annotations. Our experiments reveal a substantial and consistent gap in current MLLMs' ability to detect such compositional harm, and the same failure mode persists on real-world social media videos, indicating that DIH is not merely a synthetic artifact. We further show that targeted post-training on the DIH Dataset substantially improves DIH detection without degrading performance on benign content, establishing a strong baseline for compositional video safety.

Several directions remain for future work. First, DIH may arise along many other axes of video composition, such as speaker, viewpoint, and on-screen text, and extending our formulation beyond temporal and cross-modal settings is a natural next step. Second, whether a composition constitutes DIH may depend on social and cultural context, motivating future work on incorporating such contextual knowledge into both benchmark design and data generation. Finally, an important systems direction is to integrate DIH-aware reasoning into real-time moderation pipelines for scalable deployment on streaming and social media platforms.

AI USE STATEMENT

Generative AI was used to assist with dataset construction: the multi-agent pipeline proposes and screens compositional scenarios, synthesizes videos, and drafts reasoning annotations (Section 4.1 and Appendix C). AI-assisted tools were also used to polish the manuscript's language. Human reviewers checked the generated data and annotations, and the authors take responsibility for the final manuscript and research artifacts.

ETHICS STATEMENT

This work is intended to improve the detection of distributed implicit harm in videos. The benchmark and its generation methods have dual-use risks: they could be repurposed to create content that evades moderation. The paper includes potentially disturbing examples and provides a content warning. We therefore restrict access to harmful synthetic videos under an agreement for non-commercial safety research that prohibits redistribution and end-user deployment of the raw videos; access requires institutional affiliation and may be revoked for misuse (Appendix G). The Real-World Reference Set is shared as URLs and timestamps rather than re-hosted videos, allowing platform takedowns to remain effective (Appendix B). Our human perception study was approved by the institutional review board; participants gave informed consent, received content warnings, and could withdraw (Appendix H.1).

REPRODUCIBILITY STATEMENT

We document the benchmark construction workflow, agent roles and prompts, human quality control, and dataset statistics in Section 4.1 and Appendices C, H.6, and H.5. The evaluation prompts and scoring rubrics, model inference settings, and post-training configuration are specified in Appendices I, E.2, and H. To support independent evaluation and reuse, we plan to release the evaluation and analysis code, relevant configurations, safety taxonomy, scoring rubrics, and aggregate results. Access to harmful synthetic videos will follow the controlled-release process described in the Ethics Statement; real-world examples will be provided as URLs and timestamps rather than redistributed media.

REFERENCES

Anthropic. Introducing Claude Sonnet 4.5, 2025. https://www.anthropic.com/news/claude-sonnet-4-5.

Anthropic. Introducing Claude Opus 4.6, 2026a. https://www.anthropic.com/news/claude-opus-4-6.

Anthropic. Introducing Claude Sonnet 4.6, 2026b. https://www.anthropic.com/news/claude-sonnet-4-6.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai, Peng Wang, Keqin Chen, Sinan Tan, Shijie Wang, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Wei Cai, Jian Zhao, Yuchu Jiang, Tianle Zhang, and Xuelong Li. Safe semantics, unsafe interpretations: Tackling implicit reasoning safety in large vision-language models. In *ACM International Conference on Multimedia*, 2025.

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*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, et al. Llama Guard 3 Vision: Safeguarding human-AI image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Josef Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. SafeSora: Towards safety alignment of Text2Video generation via a human preference dataset. In *Advances in Neural Information Processing Systems*, 2024.

Google DeepMind. Gemini 3.1 Pro, 2025a. https://deepmind.google/models/gemini/pro/.

Google DeepMind. Gemini 3 Flash, 2025b. https://deepmind.google/models/gemini/flash/.

Google DeepMind. Veo 3 and 3.1, 2025c. https://aistudio.google.com/models/veo-3.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. VLSBench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, et al. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, 2023.

Kuaishou. Kling, 2024. https://app.klingai.com/cn.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, et al. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024a.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are Achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, 2024b.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, et al. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, 2024.

Xuannan Liu, Zekun Li, Zheqi He, Peipei Li, Shuhan Xia, Xing Cui, Huaibo Huang, Xi Yang, and Ran He. Video-SafetyBench: A benchmark for safety evaluation of video LVLMs. In *Advances in Neural Information Processing Systems*, 2025.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, YiFan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, Kaiyu Tang, Bin Wen, Hai-Tao Zheng, Fan Yang, Tingting Gao, Di Zhang, and Kun Gai. VLM as policy: Common-law content moderation framework for short video platform. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. JailBreakV: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *Conference on Language Modeling*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

MiniMax. Hailuo AI, 2025. https://hailuoai.video/.

OpenAI. OpenAI usage policy, 2023. https://openai.com/policies/usage-policies.

OpenAI. Sora: Creating video from text, 2024a. https://openai.com/sora.

OpenAI. GPT-4o, 2024b. https://openai.com/index/hello-gpt-4o.

OpenAI. GPT-5, 2025a. https://openai.com/index/introducing-gpt-5/.

OpenAI. OpenAI Omni-Moderation, 2025b. https://platform.openai.com/docs/guides/moderation.

OpenAI. Introducing GPT-5.4, 2026. https://openai.com/index/introducing-gpt-5-4/.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. https://qwen.ai/blog?id=qwen3.5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, et al. How many unicorns are in this image? A safety evaluation benchmark for vision LLMs. In *European Conference on Computer Vision*, 2024.

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, et al. Introducing v0.5 of the AI safety benchmark from MLCommons. *arXiv preprint arXiv:2404.12241*, 2024.

Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, and Xuanjing Huang. Safe inputs but unsafe output: Benchmarking cross-modality safety alignment of large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3563–3605. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-naacl.198. URL https://aclanthology.org/2025.findings-naacl.198/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, et al. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025c.

Yixu Wang, Jiaxin Song, Yifeng Gao, Xin Wang, Yang Yao, Yan Teng, Xingjun Ma, Yingchun Wang, and Yu-Gang Jiang. SafeVid: Toward safety aligned video large multimodal models. In *Advances in Neural Information Processing Systems*, 2025d.

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 *Annual Meeting of the Association for Computational Linguistics: Industry Track*, 2025e.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Lin, Xian Yang, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *Advances in Neural Information Processing Systems*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, et al. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Chen Yeh, You-Ming Chang, Wei-Chen Chiu, and Ning Yu. T2Vs meet VLMs: A scalable multimodal dataset for visual harmfulness recognition. In *Advances in Neural Information Processing Systems*, 2024.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. arXiv preprint arXiv:2408.05517, 2024. URL https://arxiv.org/abs/2408.05517.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. Multimodal situational safety. In *International Conference on Learning Representations*, 2025.