Beyond the Explicit: Benchmarking and Detecting Implicit Harm in Short Videos

Anonymous ACL submission

Abstract

Short videos have become a dominant online medium, where harmful content is increasingly conveyed implicitly and fragmentedly. Prior work largely focuses on explicit categories (e.g., violence, hate speech) and often fails to detect videos that subtly promote misleading values through narrative context, emotional framing, or cross-modal cues, largely due to the lack of dedicated benchmarks for implicit harm. To bridge this gap, we construct DeepHarm-7K, a large-scale harmful short video dataset comprising 7,110 samples with annotations guided by a fine-grained harmful content taxonomy, which systematically incorporates implicitly harmful videos across diverse real world scenarios under multi-dimensional quality control. Building on DeepHarm-7K, we propose DeepHarm-VL, a multimodal detection framework integrating visual, audio, and cross-modal reasoning. It employs a two-round reasoning strategy to capture implicit semantics without task-specific fine-tuning, while remaining compatible with closed-source multimodal models. Experimental results show consistent improvements over strong baselines and state-of-the-art methods, demonstrating effectiveness in detecting both explicit and implicit harmful short videos.

WARNING: The paper contains content that may be offensive and disturbing in nature.

1 Introduction

With the rise of mobile internet and algorithmic recommendation systems (Kaye et al., 2021; Zhao et al., 2019), short videos have become a dominant medium for information consumption and social interaction (Zheng, 2023; Tian et al., 2023). Their rapid spread and fragmented, high-frequency viewing enable efficient value transmission, but also amplify the impact of problematic content (Sap et al., 2019; MacAvaney et al., 2019). As a result, short video platforms have become key arenas where social norms and public values are shaped and contested (Carnegie UK, 2022; Keller, 2021), making harmful content detection a critical and challenging research problem. Representative examples are shown in Figure 1.

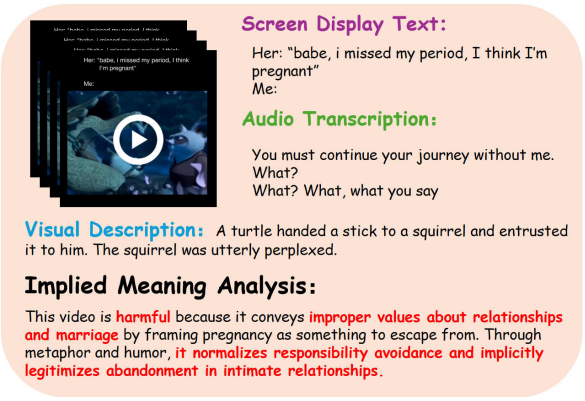


Figure 1: An example of an implicitly harmful video: its screen text (about pregnancy) aligns with the audio (framing departure) and visual content (a turtle entrusting a stick to a confused squirrel), to subtly convey problematic relationship values—framing pregnancy as a reason to avoid responsibility and normalize abandonment in intimate bonds.

Currently, human review remains essential for detecting implicit harmful content (Gillespie, 2018; Gerrard, 2020; Raji et al., 2020), but its limitations constrain effective moderation. Three key bottlenecks remain: limited scalability due to time consuming reviews, high cost and subjectivity caused by reviewer variability and fatigue, and weak recognition of implicit harm, as content without explicit violations—such as entertainment-oriented historical portrayals or aestheticized negative characters lacks unified standards. Consequently, human review alone is insufficient to reliably control implicit harmful content.

Research on automated detection of harmful short videos has grown rapidly (Ramzan et al., 2019; Perez et al., 2017; Qi et al., 2023; Das et al.,

2023), yet the field remains constrained by intertwined challenges at both data and method levels, limiting overall detection capabilities: (1) **Limitations of existing taxonomies**: Current harmful content classification systems are relatively rigid and mainly focus on explicit categories such as violence, pornography, and hate speech. As a result, implicit value-oriented harm such as entertainment-driven reinterpretations of historical figures, aestheticized negative roles, or nihilistic narratives is largely overlooked, lacking systematic definitions, fine-grained categorization, and targeted detection research, despite its subtle yet influential impact on users' perceptions (Zhou and Zafarani, 2020; Bandura, 2002; Aretoulakis, 2008; Breazu, 2023). (2) **Dataset and benchmark deficiencies**: Existing datasets are limited in both scale and content diversity, and are predominantly collected from single platforms (Vidgen and Derczynski, 2020). Consequently, they fail to adequately reflect the diversity, realism, and complex dynamics of real-world short-video ecosystems, thereby hindering the construction of reliable and comprehensive benchmarks for evaluating models on implicit harmful content. (3) **Method limitations**: Constrained by such datasets, existing detection methods face adaptation challenges: they either rely on single-modality explicit cues or require extensive labeled data for fine-tuning, failing to capture the complex semantics of implicit harm and lacking robust cross-scenario generalization, making them ill-suited for the multimodal, diverse, and rapidly evolving nature of short video content.

To address these challenges, we propose a taxonomy-driven end-to-end framework that integrates taxonomy design, dataset construction, and multimodal modeling for detecting implicit value-oriented harmful short videos. The taxonomy provides structured definitions that guide data collection and annotation, and supports consistent evaluation across datasets and methods.

(1) **Fine-grained taxonomy**. We introduce a unified taxonomy that systematically incorporates implicit value-oriented harm alongside explicit policy violations. Its design is guided by a threefold foundation: (i) Official regulations and guidelines from international governments and organizations (UNICEF Office of Research – Innocenti, 2025; European Parliament and Council of the European Union, 2022); (ii) Influential top-tier academic literature on harmful content (Baykut and Warner, 2022; Nkhata and Mgomezulu, 2025), and (iii) moderation policies and standards from leading online platforms (Douyin; Bilibili; Kuaishou). The taxonomy is refined through multiple rounds of expert discussion, providing a principled basis for defining and evaluating implicit harmful content across diverse scenarios.

(2) **Large-scale diverse dataset**. Based on the proposed taxonomy, we construct DeepHarm-7K, a large-scale harmful short video dataset that serves both as a comprehensive data resource and a unified evaluation benchmark. DeepHarm-7K contains 7,110 multimodal short videos collected from multiple domestic and international platforms, including 3,190 harmful and 3,920 non-harmful samples. All videos are annotated through a rigorously validated labeling process, enabling systematic analysis and evaluation of both explicit and implicit harmful content.

(3) **Adaptive multimodal detection framework**. To validate DeepHarm-7K and assess existing methods, we propose DeepHarm-VL, a training-free multimodal detection framework that integrates visual, audio, and textual signals via a two-stage reasoning strategy. By exploiting latent semantic and audio–visual cues, DeepHarm-VL enhances the detection of both explicit and implicit harmful videos, demonstrating strong generalization and revealing the limitations of prior methods in modeling implicit value-oriented harm.

The main contributions of this paper can be summarized as follows:

- We propose the first taxonomy for harmful short videos that incorporates implicit value-misleading content and is fine-grained, scope-explicit, and dataset-oriented, forming the conceptual foundation of DeepHarm-7K.
- We construct DeepHarm-7K, a large-scale and diverse multimodal harmful short video dataset, which provides high-quality and comprehensive support for future research on both explicit and implicit harmful short video detection.
- We further benchmark existing baseline methods on DeepHarm-7K, revealing their notable limitations, particularly in handling ambiguous cases and implicit harmful content.
- We propose DeepHarm-VL, a multimodal large-model-based detection framework for harmful short-form videos. Evaluated on DeepHarm-7K, DeepHarm-VL significantly outperforms existing baseline methods.

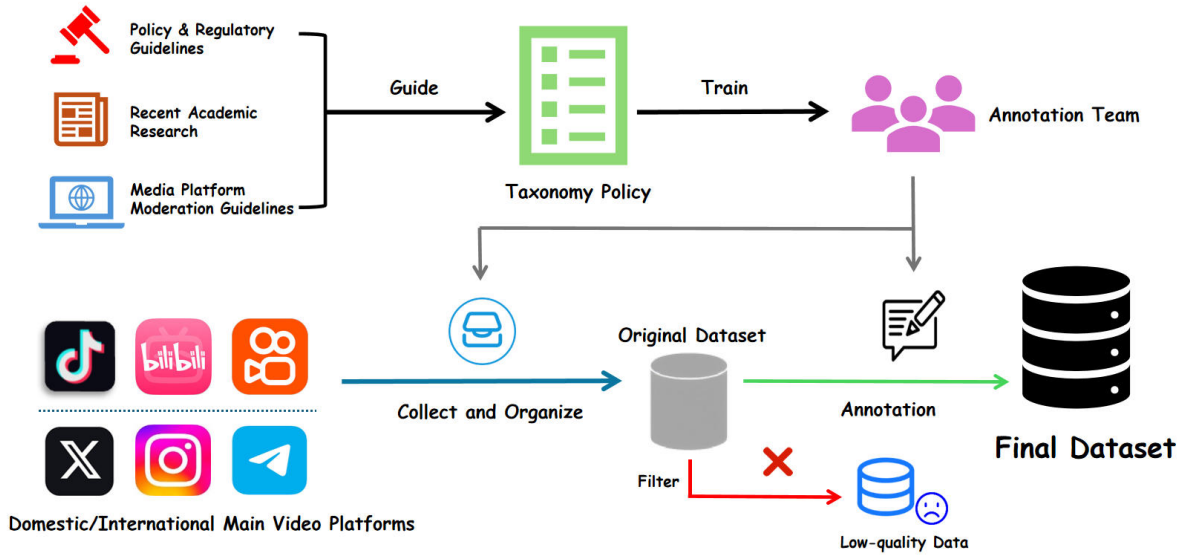


Figure 2: A unified taxonomy is derived from policy guidelines, academic literature, and platform moderation rules to guide annotator training. Raw videos are collected from multiple platforms, filtered for quality, and annotated under the taxonomy to produce DeepHarm-7K.

2 Related Work

2.1 Harmful Content Datasets

Harmful content datasets have been developed for text, images, and videos, but most are poorly aligned with the characteristics of short video.

Text-based datasets are the most mature and mainly target explicit categories such as hate speech and misinformation (Bai et al., 2025b; Dai et al., 2020; Deng et al., 2022). Restricted to unimodal text, they largely overlook implicit harmful content involving value misguidance or covert framing, limiting their applicability to multimodal short videos.

Image-based datasets focus on visually explicit content such as violence (Cheng et al., 2021) and pornography (Birhane et al., 2021), relying on static images and coarse labels that ignore temporal dynamics and cross-modal interactions. Although multimodal meme datasets explore implicit harm via contextual cues (Lu et al., 2024; Fersini et al., 2022; Xu et al., 2022), they remain confined to static settings.

Video-based datasets incorporate temporal information and cover categories such as hate, violence, and fake news (Chen et al., 2024; Das et al., 2023; Rehman et al., 2025; Wang et al., 2025a; Qi et al., 2023). However, most emphasize overt violations with coarse-grained labels, leaving subtle and implicit harmful content underrepresented.

2.2 Multimodal Harmful Detection

Early multimodal detection methods employ modality-specific encoders with feature or decision level fusion (Mumtaz et al., 2023; Sultani et al., 2018; Long et al., 2018). While effective for explicit harm, they rely heavily on task-specific data and struggle with implicit or context-dependent semantics. More recently, multimodal large language models (MLLMs) such as Qwen3-VL (Bai et al., 2025a) and GPT-4V show strong video understanding ability, but their direct application to harmful video detection remains challenging due to weakly expressed and domain-sensitive implicit harm.

3 DeepHarm-7K

To ensure data quality, annotation reliability, and distributional robustness, we construct DeepHarm-7K via a rigorous systematic pipeline (Figure 2), and describe each stage in detail.

3.1 Definition and Taxonomy Design

To ensure the reliability and practical relevance of our taxonomy for harmful short videos, its design is grounded in three complementary sources. First, we align with policy and regulatory guidelines on online content governance (UNICEF Office of Research – Innocenti, 2025; European Parliament and Council of the European Union, 2022; NRTA, 2021) to maintain consistency with real world moderation standards. Second, we incorporate insights from recent research on harmful con-

(a) (b) (c) (d)

Figure 3: Example short videos illustrating the four high-level harmful content categories in our taxonomy. Subfigures **(a–d)** correspond to categories **C1–C4**, respectively.

tent detection in top-tier venues (Xu et al., 2022; Chen et al., 2024; Rehman et al., 2025). Third, we reference industrial practice by examining public moderation guidelines and representative violation cases from major short video platforms (Douyin; Bilibili; Kuaishou),. The taxonomy was further refined through iterative discussions with domain experts, ensuring both authority and applicability.

Based on this process, we develop a fine grained taxonomy comprising five high level categories: four harmful and one non-harmful, and twelve subcategories. To illustrate the variety of harmful content, Figure 3 presents representative short video examples for the four harmful categories (C1–C4). The high-level categories provide a broad overview of content types, while the subcategories (described in detail in Appendix A) enable more precise annotation and nuanced differentiation. The five high-level categories are summarized as follows:

C1. *Discrimination-related content, including racial, gender-based, regional, or other forms of discriminatory expression;*

C2. *Superstition and misinformation, covering feudal superstitious practices and the dissemination of false or misleading information;*

C3. *Violence, pornography, and terror, including explicit violent acts, sexual exposure, and graphic or frightening content;*

C4. *Value-oriented misleading content, including content that promotes or normalizes harmful, distorted, or inappropriate social values;*

C5. *Non-harmful content that do not include any of the above harmful characteristics.*

By systematically integrating policy, research, and industrial references with expert insights, the taxonomy ensures conceptual rigor, practical relevance, and annotation consistency, providing a solid foundation for the construction of DeepHarm-7K and subsequent analysis of harmful content.

3.2 Data Collection

All short video samples were manually collected and preliminarily screened by the annotation team to ensure source consistency. Before large-scale collection, annotators completed three rounds of training and calibration based on platform policies and representative violation cases. Through discussion and simulated annotation, judgment criteria, category boundaries, and threshold standards were aligned, forming a unified collection protocol.

To ensure diversity and representativeness, data were collected from both domestic and international platforms. Domestic platforms include Douyin, Bilibili, and Kuaishou, covering daily life sharing, knowledge dissemination, and historical interpretation, while international platforms include X, Instagram, and Telegram to capture cross-regional harmful content patterns. This cross-platform design reduces single-platform bias and broadens content coverage.

We adopt a hybrid collection strategy combining random feed browsing and keyword-based retrieval. Random browsing captures naturally occurring videos without predefined category constraints, while keyword-based retrieval uses a taxonomy driven keyword pool to target known harmful content. The two strategies are mutually reinforcing: random browsing reveals novel harmful patterns that expand the keyword pool, while keyword-based retrieval stimulates recommendation systems to surface related content, improving discovery efficiency.

The resulting corpus achieves broad category coverage and balanced scale, including harmful and non-harmful videos across explicit and implicit scenarios. Hard negative samples, which resemble harmful content in appearance or theme yet remain non-harmful, are deliberately included to increase task difficulty and evaluate robustness. This curated corpus forms a high-quality candidate pool

Split	Nonharmful	Harmful	Harmful Type Category				Total
			C1.	C2.	C3.	C4.	
Train	2,744	2,227	344	452	1,037	394	4,971
Test	1,176	963	149	195	447	172	2,139
Total	3,920	3,190	493	647	1,484	566	7,110

Table 1: **Statistics of DeepHarm-7K distribution.** Dataset distribution across training and test splits, with 3,920 nonharmful and 3,190 harmful samples. Harmful samples are split into four subcategories **(C1–C4)**, with counts and total sizes for each group included.

for subsequent filtering and annotation.

3.3 Data Annotation and Filtering

To ensure high annotation reliability and balanced sample distribution, we establish an integrated annotation filtering quality control pipeline spanning the entire dataset construction process. This design emphasizes annotation consistency, decision traceability, and distributional robustness, forming a core methodological advantage of DeepHarm-7K.

3.3.1 Annotation Team

The annotation team is identical to the data collection team described in Section 3.2 and consists of 10 members (see Table 2). All annotators participated in defining harmful content and calibrating collection standards, ensuring a consistent understanding of category boundaries and reducing subjective annotation bias.

Characteristic	Demographics
Gender	Male: 10
Age	20-22: 7, 23-26: 3
Nationality	Chineses: 10
Education	PG: 4, PhD: 6

Table 2: Annotators demographics.

3.3.2 Multi-stage Annotation Workflow

To ensure high-quality and reliable labels, we adopt a three-stage annotation workflow comprising initial independent labeling, cross-annotator rotation, and voting-based arbitration.

Initial independent labeling. The collected videos are evenly and randomly distributed among the ten annotators, so each receives roughly the same number of samples. Each annotator independently labels their assigned videos in isolation, forming the first round of annotations.

Cross-annotator rotation. To increase objectivity, we conduct two rounds of cross-annotation. In each round, videos labeled by one annotator are passed to another (e.g., B receives A's videos, while B's videos go to C), completing a full rotation. After two rotations, each video is labeled by three independent annotators, providing multiple perspectives for consistency assessment.

Consistency validation and voting-based arbitration. Annotation results are merged to check consistency across the three independent labels. With 12 subcategories, some videos naturally receive conflicting labels from all annotators. These controversial samples are resolved by majority vote, with seven of the ten team members deciding the final label. The rationale for each decision is documented to ensure transparency and traceability.

3.3.3 Filtering during Annotation

To jointly optimize data quality and category balance, sample filtering is conducted in parallel with annotation. During labeling, annotators flag low-quality samples, including videos with blurred visuals, poor audio quality, missing key modalities, or content that is overly explicit, simplistic, or lacks research value. After annotation, three team members perform a final filtering pass based on these flags and category distribution statistics, removing low-quality samples and adjusting class proportions to ensure balance. The resulting dataset thus achieves high overall quality and a well-balanced class composition. Complete statistics of DeepHarm-7K are provided in Table 1.

3.3.4 Annotation Consistency Evaluation

To quantitatively assess annotation reliability, we randomly sample 1600 instances from the DeepHarm-7K and evaluate inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971). The resulting score of 0.82 indicates a high level of agreement beyond chance, demonstrating strong consistency among annotators and validating the reliability of DeepHarm-7K for training and evaluating harmful short-video detection models.

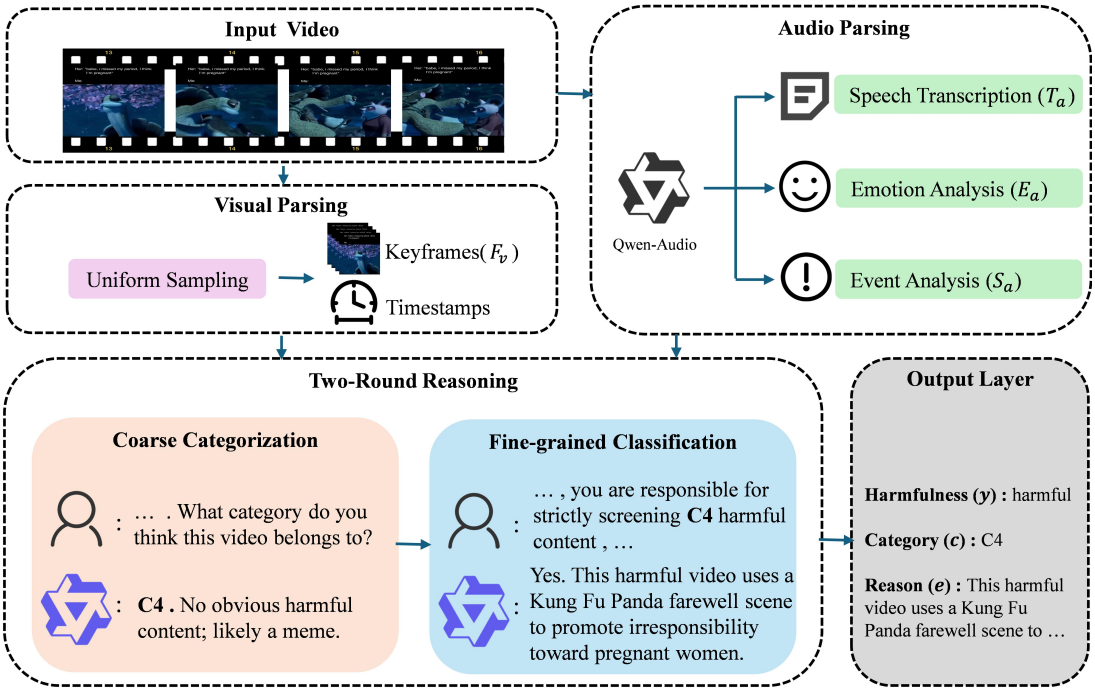


Figure 4: Overall pipeline of the proposed DeepHarm-VL multimodal harmful video detection framework.

4 DeepHarm-VL

4.1 Method Overview

To address the challenges of implicit harmful content detection in short videos, we propose DeepHarm-VL, an automated decision framework grounded in multimodal understanding and hierarchical reasoning (Figure 4). DeepHarm-VL is designed to emulate the human judgment process of "reasoning from surface cues to underlying intent" and "from coarse screening to fine-grained verification." By performing hierarchical cross-modal parsing and stage-wise focused reasoning, the framework aims to achieve more reliable and interpretable harmful video identification.

DeepHarm-VL consists of four core modules. (1) **Visual parsing** module that extracts key frames F_v; (2) **audio parsing** module that produces the audio transcript T_a, emotional description E_a, and sound event description S_a; (3) **two-round reasoning** module that performs hierarchical inference over multimodal evidence; and (4) **output** module that aggregates the reasoning results and generates an interpretable prediction (y, c, e).

In the visual parsing module, we adopt a uniform sampling strategy to extract approximately 60 frames, roughly one frame per second, and feed them into MLLMs together with their corresponding timestamps. The output module consolidates the results from the two-round reasoning process to generate both binary and multi-class predictions, and it preserves the explanation text to further enhance interpretability. The key innovations of DeepHarm-VL, namely the audio parsing module and the two-round reasoning module, will be described in detail in the following two sections.

4.2 Audio Parsing Module

Audio provides crucial evidence for detecting implicit harmful content. In particular, affective tendencies and environmental sounds, along with how they may contradict or reinforce spoken content, often reveal latent intent that is not explicitly stated. Conventional pipelines that rely solely on automatic speech recognition (ASR) are limited to textual transcripts and ignore paralinguistic cues (e.g., prosody) and background signals, making them insufficient for scenarios where implicit harm is conveyed through emotional manipulation, acoustic camouflage, or specific sound events.

We develop a multi-dimensional audio semantic parsing framework that converts continuous audio into structured, interpretable natural-language descriptions. This representation is tailored to the reasoning needs of MLLMs. Concretely, we perform targeted fine-tuning of **Qwen-Audio** (Chu et al., 2023) to improve sensitivity to safety-relevant acoustic patterns, and produce the following three complementary views:

(1) Speech transcription. We generate an ac-

curate transcript T_a as the semantic foundation. **(2) Emotion analysis.** We generate an affective description E_a capturing emotional tone . This helps identify cases where the literal text appears benign but the delivery is harmful (e.g., insults spoken calmly, or morally corrupt messages delivered with sympathetic tone), thereby addressing the blind spots of transcript-only approaches. **(3) Event analysis.** We generate an acoustic event description S_a for salient background sounds. This enables the model to detect harmful events that are masked by seemingly normal speech or music, supporting deeper situational understanding.

4.3 Two-Round Reasoning Module

The harmfulness of implicit content is often embedded in expressive style, cross-modal inconsistencies, or rapidly evolving cultural memes. As a result, one-shot classification or fixed-rule systems suffer from two major limitations. First, they are prone to missing fast-evolving expressions (e.g., metaphor, irony) due to knowledge lag. Second, they lack progressive reasoning ability to move from surface signals to latent intent, leading to both false positives and false negatives.

Inspired by the human expert practice of "coarse screening followed by careful adjudication," we propose a hierarchical two-round reasoning mechanism. Instead of relying on fully supervised training with large amounts of implicit-harm data, this mechanism uses structured prompt (described in detail in Appendix B) engineering to guide MLLMs in simulating progressive reasoning.

Round 1: Coarse Categorization. This stage aims to efficiently identify potential risks and acts as a "safety net" to reduce misses. Using a general prompt template that covers major harm categories (e.g., hate speech, violence incitement) and their common manifestations, the model performs an initial cross-modal association based on multimodal descriptions and outputs one candidate category, thereby delimiting the scope for deeper analysis.

Round 2: Fine-grained Classification. This stage builds on the Round 1 reasoning results and, for each candidate high-level category, applies a specialized prompt that incorporates category-specific safety guidelines, decision criteria, and positive/negative examples. The model is then guided to conduct cross-modal verification, contextual analysis, and intent inference, ultimately producing a fine-grained harmful-category decision along with a natural-language rationale.

Method	Acc	F1	F1(H)	F1(N)
BERT	78.87	78.12	74.08	82.15
MFCC	65.26	64.82	68.36	61.32
3D-CNN	68.13	67.66	71.54	63.79
HateMM	81.87	81.67	83.32	80.02
MM-HSD	79.73	79.20	75.87	82.53
Intervl3.5	78.78	73.29	64.32	82.25
LLaVA-Video	49.84	40.22	64.20	16.24
HolmesVAD	61.56	60.12	61.90	58.34
Qwen3-vl	74.32	73.45	70.78	76.12
DeepHarm-VL	**91.07**	**90.80**	**89.24**	**92.37**

Table 3: Binary Classification Performance on DeepHarm-7K dataset. H: harmful, N: non-harmful.

5 Experiments

5.1 Experimental Setup

We evaluate two tasks: binary and multi-class classification. We report Acc for binary classification and per-class F1 for multi-class classification, averaged over multiple runs. All experiments are conducted on 4× RTX 3090. **DeepHarm-VL** is instantiated with **Qwen3-VL (7B)** (Bai et al., 2025a).

5.2 Baseline Methods

To comprehensively evaluate DeepHarm-VL, we compare against a diverse set of baselines, covering both **non-MLLMs-based methods** and **MLLMs-based methods**.

Non-MLLMs-based methods include unimodal models such as: (1) **BERT** (Devlin et al., 2019)-based feature extraction from video transcripts, (2) **MFCC** (Xu et al., 2004)-based audio feature extraction, and (3) **3D-CNN** (Ji et al., 2012)-based visual feature extraction. In addition, we compare with existing harmful-video detection models, including **HateMM** (Das et al., 2023) and **MM-HSD** (Céspedes-Sarrias et al., 2025).

MLLMs-based methods include **InternVL3.5** (Wang et al., 2025b), **LLaVA-Video** (Lin et al., 2024), **HolmesVAD** (Zhang et al., 2024), and the vanilla **Qwen3-VL**.

We use the best hyperparameters reported in the original papers for all baselines. Since non-MLLMs-based baselines only support binary classification, we compare multi-class results mainly with MLLM-based methods.

5.3 Experimental Results and Analysis

Based on the comparative experiments on the DeepHarm-7K benchmark (Table 3, 5), we analyze different categories of methods and summarize the key findings as follows.

Method	Binary Classification				Multi-class Classification						
	Acc	F1	F1(H)	F1(N)	Acc	F1	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)
w/o audio parsing	90.42	90.16	88.58	91.74	85.04	73.90	62.30	85.71	88.99	**40.76**	91.74
w/o two-round reasoning	84.76	83.82	79.93	87.72	79.12	64.36	75.12	78.34	76.89	30.12	61.34
DeepHarm-VL	**91.07**	**90.80**	**89.24**	**92.37**	**87.38**	**78.04**	**80.78**	**89.30**	**89.73**	38.00	**92.37**

Table 4: Ablation study on different components.

Method	Acc	F1	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)
Intervl3.5	76.3	53.08	48.62	30.51	84.68	17.8	83.81
LLaVA-Video	26.13	29.31	16.35	21.85	65.82	26.27	16.24
HolmesVAD	54.23	46.65	53.12	51.89	55.34	23.78	49.12
Qwen3-vl	68.45	58.79	69.12	70.45	68.90	25.12	60.34
DeepHarm-VL	**87.38**	**78.04**	**80.78**	**89.30**	**89.73**	**38.00**	**92.37**

Table 5: Multi-class Classification Performance on DeepHarm-7K dataset.

Performance limitations of non-MLLM methods. Conventional approaches exhibit clear capacity bottlenecks on this task. Unimodal methods, which rely on a single information source, achieve the weakest overall performance. Traditional multimodal models substantially outperform unimodal baselines by fusing heterogeneous features; however, they still fall short in deep cross-modal understanding, leaving a pronounced gap to the best-performing results.

Mismatch and "safety bias" in existing MLLMs. Directly applying general-purpose MLLMs does not yield the expected gains, and in some cases performs even worse than traditional multimodal models. We observe a prevalent "safety bias" in these models: they tend to predict the non-harmful class conservatively, suggesting that naive transfer of general capabilities is unreliable for implicit harmfulness assessment.

In contrast, DeepHarm-VL achieves consistent, overall comprehensive and significant improvements over all baselines, largely attributable to the audio parsing module and the two-round reasoning mechanism, and attains the best performance across all evaluation metrics.

5.4 Ablation Study

To validate the effectiveness of our core components, we conduct systematic ablation studies, with results reported in Table 4.

Removing the audio parsing module. When the model relies solely on visual cues, overall performance drops substantially, particularly on the emotion-intensive C1 category. This verifies that speech content, affective signals, and acoustic event descriptions provided by audio parsing constitute crucial evidence for detecting implicit harmful content. Interestingly, removing audio yields a slight improvement on C4. We attribute this to the fact that C4 videos often involve ambiguous, boundary-blurring implicit scenarios, where audio-derived signals may introduce misleading cues and thus interfere with the model's judgment.

Removing the two-round reasoning mechanism. Replacing the proposed two-stage pipeline with a single-round direct classification leads to a pronounced performance degradation, confirming the central role of hierarchical reasoning. A single-step decision is insufficient to balance broad screening and deep discrimination, and is more prone to false negatives and false positives when facing novel or subtle expressions. These results further suggest that two-round reasoning helps mitigate the safety bias of MLLMs in open-domain safety assessment and alleviates practical limitations related to long-context processing.

6 Conclusion

In this paper, we present DeepHarm-7K, a taxonomy-driven large-scale dataset for harmful short videos, covering both explicit policy violations and implicit value-oriented content. To enable systematic evaluation, we design a fine-grained taxonomy and collect a diverse multimodal corpus from multiple platforms. We also propose DeepHarm-VL, a lightweight multimodal detection framework that integrates visual, audio, and textual signals with a two-stage reasoning strategy, capturing subtle implicit harmful semantics without task specific fine-tuning. Experiments show that our framework consistently outperforms strong baselines, demonstrating the effectiveness of DeepHarm-7K as a reliable benchmark. We hope this dataset and taxonomy provide a foundation for future research on understanding and mitigating implicit harmful content in short videos.

Limitations

While DeepHarm-VL demonstrates strong performance in detecting both explicit and implicit harmful videos, there are several limitations worth noting for future research.

Visual Modeling. Although our framework effectively integrates multimodal signals, current models especially smaller scale versions (e.g., 7B) may not fully capture fine-grained visual information, occasionally missing subtle cues indicative of harmful content. Increasing model capacity generally improves performance, suggesting that visual modeling remains a key bottleneck in detecting nuanced multimodal signals.

Knowledge Constraints. Detection of implicit harmful content can be hindered by incomplete or outdated knowledge. Current MLLMs may lack awareness of emerging trends, cultural references, or newly evolving harmful patterns, limiting their ability to interpret time-sensitive content. Future work could incorporate retrieval-augmented generation (RAG) or other dynamic knowledge integration techniques to address these gaps and improve detection robustness.

Ethical Considerations

Our work involves collection, annotation, and analysis of harmful short videos to advance automated detection of content that may negatively impact individual cognition, social order, or public morality. All data were obtained from publicly available sources, with personally identifiable information removed or anonymized. Annotation was performed by trained annotators under controlled conditions, following guidelines to minimize exposure to harmful material and ensure objective labeling.

To promote responsible and ethical use, access to the DeepHarm-7K dataset is subject to a strict review process. Researchers seeking to use the dataset must complete a detailed questionnaire and provide sufficient evidence that their intended use is solely for academic research purposes. This controlled access policy is designed to prevent misuse of sensitive content while supporting reproducible research in harmful short video detection.

We also acknowledge potential limitations and risks. Despite careful design, the models may occasionally misinterpret content, leading to false positives or negatives. Furthermore, cultural and temporal context may influence the interpretation of implicit harmful content, underscoring the need for human oversight in real-world applications. By explicitly addressing these considerations, we aim to ensure that our dataset and models are used ethically and safely, promoting responsible and trustworthy AI research.

References

Emmanouil Aretoulakis. 2008. Aesthetic appreciation, ethics, and 9/11. *Contemporary Aesthetics (Journal Archive)*, 6(1):13.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025b. State toxicn: A benchmark for span-level target-aware toxicity extraction in chinese hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10206–10219.

Albert Bandura. 2002. Selective moral disengagement in the exercise of moral agency. *Journal of moral education*, 31(2):101–119.

Senem Baykut and Daniel C. Warner. 2022. The impact of the hidden curriculum on international students' experience of university. *The Curriculum Journal*, 33(2):156–177.

Bilibili. management-protocol. `https://openhome.bilibili.com/agreement/management-protocol`.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Remus Breazu. 2023. The aestheticization of violence in images. *Philosophia*, 51(1):33–52.

Carnegie UK. 2022. The online safety bill and the 'legal but harmful' debate.

Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. 2025. Mm-hsd: Multi-modal hate speech detection in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2546–2555.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. 2024. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*.

Ming Cheng, Kunjing Cai, and Ming Li. 2021. Rwf-2000: An open large scale video database for violence detection. In *2020 25th International conference on pattern recognition (ICPR)*, pages 4183–4190. IEEE.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Douyin. rule. https://www.douyin.com/rule/main.

European Parliament and Council of the European Union. 2022. Regulation (eu) 2022/2065 on a single market for digital services and amending directive 2000/31/ec (digital services act). Official Journal of the European Union, L 277/1.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ysabel Gerrard. 2020. Behind the screen: Content moderation in the shadows of social media.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

D Bondy Valdovinos Kaye, Xu Chen, and Jing Zeng. 2021. The co-evolution of two chinese mobile short video apps: Parallel platformization of douyin and tiktok. *Mobile Media & Communication*, 9(2):229–253.

Daphne Keller. 2021. Lawful but awful: Control over legal online speech. *University of Chicago Law Review Online*, 88:1–18.

Kuaishou. norm. https://www.kuaishou.com/norm.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Xiang Long, Chuang Gan, Gerard Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. 2018. Multimodal keyless attention fusion for video classification. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *arXiv preprint arXiv:2410.02378*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Nadia Mumtaz, Naveed Ejaz, Shabana Habib, Syed Muhammad Mohsin, Prayag Tiwari, Shahab S Band, and Neeraj Kumar. 2023. An overview of violence detection techniques: current challenges and future directions. *Artificial intelligence review*, 56(5):4641–4666.

Asher K. C. Nkhata and Victor Y. Mgomezulu. 2025. Geography teaching as a tool for enhancing values education in malawi: Exploring the promoting sustainable development values in secondary schools. *International Journal of Research and Innovation in Social Science*, 9(7):5515–5525.

NRTA. 2021. Detailed rules for content review of online short videos. https://www.nrta.gov.cn/art/2021/12/15/art_113_58926.html.

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

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Muhammad Ramzan, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. 2019. A review on state-of-the-art violence detection techniques. *IEEE Access*, 7:107560–107575.

Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Nagendra Kumar. 2025. Implihatevid: A benchmark dataset and two-stage contrastive learning framework for implicit hate speech detection in videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17209–17221.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Xiaoxu Tian, Xinhua Bi, and Hao Chen. 2023. How short-form video features influence addiction behavior? empirical research from the opponent process theory perspective. *Information Technology & People*, 36(1):387–408.

UNICEF Office of Research – Innocenti. 2025. Childhood in a digital world. Technical report, UNICEF Office of Research – Innocenti, Florence. UNICEF Innocenti Report.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Han Wang, Zhuoran Wang, and Roy Ka-Wei Lee. 2025a. Hateclipseg: A segment-level annotated dataset for fine-grained hate video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13304–13310.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. 2004. Hmm-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, pages 566–574. Springer.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*.

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM conference on recommender systems*, pages 43–51.

Cheng Zheng. 2023. Research on the flow experience and social influences of users of short online videos. a case study of douyin. *Scientific Reports*, 13(1):3312.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Detailed Description of the DeepHarm-7K Taxonomy

This appendix provides a detailed description of the harmful content taxonomy introduced in Section 3.1, which serves as the classification framework employed in DeepHarm-7K. We further clarify the defining characteristics and scope of each category to facilitate reliable annotation and nuanced interpretation.

C1. Maliciously Inciting Social Conflicts

Core Feature: Content that deliberately generates inter-group hostility, undermining social cohesion and harmony. Includes but is not limited to:

• *Instigating or provoking conflicts between different groups or ideologies (e.g., gender-based conflicts, ethnic tensions, religious disputes, national rivalries, or political factionalism).*

• *Mocking, ridiculing, humiliating, stereotyping, or labeling specific groups.*

C2. Disseminating False, Superstitious, or Misleading Information

Core Feature: Content that spreads unverified or deliberately fabricated information, misleading public perception and disrupting public order. Includes but is not limited to:

• *Spreading rumors or conspiracy theories, including pseudo-historical claims and pseudoscientific beliefs (e.g., flat-earth theory).*

• *Promoting feudal superstitions, witchcraft, divination, astrology, or other occult practices.*

C3. Content Harmful to Public Safety or Mental Health

Core Feature: Content that directly threatens personal safety, public order, or mental well-being, with highly negative impacts. Includes but is not limited to:

• *Depictions of violence, abuse, or murder, as well as actions that severely disrupt social order, such as looting, riots, mob disturbances, violent protests, or vandalism.*

• *Terror-inducing or psychologically disturbing imagery, including graphic blood, trauma, or horrifying scenes.*

• *Explicit or implicit sexual content, including nudity, sexual acts or insinuation, and depictions of prostitution or sexual transactions.*

C4. Content Violating Mainstream Values and Ethical Norms

Core Feature: Content that challenges societal ethics, moral standards, cultural traditions, or laws, promoting harmful or distorted values. Includes but is not limited to:

• *Using criminals or villains to glorify crimes, defaming national heroes or symbols, or trivializing serious historical events through entertainment.*

• *Content that depicts or endorses dangerous or unhealthy behaviors or ideas, often with negative social influence, such as risky stunts, illegal or unsafe acts, substance abuse, gambling, or profanity.*

• *Promoting depression, apathy, or suicidal tendencies, associated with persistent negative or nihilistic emotional states.*

• *Content that promotes or dramatizes behaviors or ideas violating social norms or ethical standards, such as exploiting or humiliating the weak, defamation, immoral ideologies, distorted values, disruption of social order, resource waste, or environmental harm.*

B Prompts For DeeepHarm-VL

B.1 Round 1 (Coarse Categorization) Prompt

SYSTEM: You are an AI for video content analysis. Your task is to evaluate which of the following four harmful-content categories a video most likely belongs to, based on the definitions below. Even if the video is completely harmless, you *must* still choose the closest category according to similarity in its theme, elements, or style.

USER: Four categories are defined as follows.

C1. Maliciously inciting social conflicts. Identify content that incites hatred, antagonism, or discrimination between social groups. This includes stirring hostility or confrontation across attributes such as gender, ethnicity, region, class, or religion; provoking interstate hostility or domestic political confrontation; or mocking, humiliating, or reinforcing stereotypes based on group identity. The core is to create division.

C2. Disseminating false, superstitious, or misleading information. Identify content that spreads misinformation, pseudoscience, or superstition. This includes promoting self-invented theories that contradict scientific consensus or anti-scientific claims; advocating fortune-telling, divination, or

other practices as determiners or predictors of fate; and disseminating rumors or conspiracy theories such as pseudo-history or unverified claims (e.g., flat-earth narratives). The core is the falsehood or anti-scientific nature of the content.

C3. Content harmful to public safety or mental health. Identify content that directly depicts or promotes violence, pornography, or terror. This includes explicit portrayals or glorification of fighting, abuse, murder, self-harm, or similar acts; bloody or horror content intended to induce intense psychological distress; and explicit sexual content or sexual provocation, including depictions of sexual acts or genitalia. The core is a direct threat to psychological health and a sense of safety.

C4. Content violating mainstream values and ethical norms. Identify content that distorts right and wrong or promotes harmful value orientations. This includes glorifying criminals or historical perpetrators, defaming heroes or martyrs, or trivializing disasters and tragedies for entertainment; advocating nihilistic or decadent ideologies, encouraging self-harm/suicide, or endorsing undesirable behaviors or harmful customs; showcasing or glamorizing highly dangerous acts likely to cause severe injury (e.g., risky stunts or self-harm); and promoting environmental harm or wastefulness without remorse. The core is challenging public morality and socially endorsed values.

Output: Always output the most likely category label: **C1**, **C2**, **C3**, or **C4**, and explain the reason.

B.2 Round 2 (Fine-grained Classification) Prompt

B.2.1 Prompt 1 (for C1 category)

SYSTEM: You are a specialized AI for video content safety auditing. You are responsible for strictly screening C1 harmful content: maliciously inciting social conflicts. Your core task is to identify any content intended to create hatred, antagonism, or discrimination between social groups.

USER: The definition and decision criteria for Category 1 harmful content are as follows; please follow them strictly. Any content that creates group-level or ideological confrontation must be judged as harmful. This includes inciting hostility, vigilance, or antagonistic emotions between groups defined by gender, ethnicity, region, social class, religion, and similar attributes. Such incitement is harmful whether it is explicit or conveyed implicitly through insinuation, labeling, or oppositional narrative frameworks (e.g., "us vs. them") that guide viewers toward adversarial associations. Any content that creates national or political confrontation must also be judged as harmful, including stirring hostility between countries (e.g., China–U.S., Russia–Ukraine) or intensifying partisan struggles within a country. Any content that discriminates against or humiliates specific groups must likewise be judged as harmful, including mockery, ridicule, humiliation, or reinforcement of stereotypes based on the group attributes above. Note the key distinction from "opposition": discrimination is characterized by contempt and belittlement rather than direct confrontation. As an important principle, your judgment should focus on whether the content is manufacturing division. Even if no vulgar language is used, if the core logic, context, or implied effect undermines group solidarity or incites antagonistic emotions, it must be judged as harmful. For content expressed through subtle means such as stories, metaphors, or satire, you should deeply interpret its potential social impact.

Output requirement: Always output Harmful or Harmless, and explain the reason.

B.2.2 Prompt 2 (for C2 category)

SYSTEM: You are a specialized AI for video content safety auditing. You are responsible for strictly screening C2 harmful content: disseminating false, superstitious, or misleading information. Your core task is to identify and judge any video content that disseminates misinformation, pseudoscience, or feudal superstition.

USER: The definition and decision criteria for Category 2 harmful content are as follows; please follow them strictly. Any pseudoscientific content must be judged as harmful. This refers to theories proposed by "folk scientists" or individuals that contradict established scientific consensus, such as self-invented formulas, concepts, or models with no scientific basis, as well as any unverified claims disguised as scientific knowledge that promote anti-science or anti-intellectual viewpoints. The key criterion is the content's anti-scientific nature itself; regardless of whether it is presented as academic discussion or personal opinion, the act of spreading it is harmful. Any feudal superstition and fortune-telling content must be judged as harmful. This includes, but is not limited to, fortune-telling, divination, palm/face reading, feng shui for changing luck, witchcraft, spells, or similar practices that claim human fate is determined by supernatural

forces or can be predicted or altered through such means. Any rumors and conspiracy theories must be judged as harmful, such as pseudo-historical narratives, flat-earth claims, or any other unverified misinformation and conspiratorial claims. As an important principle, your judgment should be based on the topic and essence of the content rather than its presentation style. As long as the video's core content, theme, or main demonstration involves any of the above, it must be directly judged as harmful. You must not classify it as non-harmful on the grounds that it "appears to be personal theoretical discussion," "contains no incitement," or "is expressed rationally."

Output requirement: Always output Harmful or Harmless, and explain the reason.

B.2.3 Prompt 3 (for C3 category)

SYSTEM: You are a specialized AI for video content safety auditing. You are responsible for strictly screening C3 harmful content: content harmful to public safety or mental health. Your core task is to identify any content that directly depicts or promotes violence, pornography, or terror.

USER: The definition and decision criteria for Category 3 harmful content are as follows; please follow them strictly. Any violence and abuse content must be judged as harmful. This includes directly showing or explicitly describing brutal scenes such as fighting, abuse, murder, or self-harm; any non-educational content that glorifies, beautifies, or sensationalizes violence also falls into this category. Any horror and psychological contamination content must be judged as harmful. This includes bloody, grotesque, or terrifying visuals or sound effects intended to provoke intense psychological stimulation and fear. Any pornographic and sexually suggestive content must be judged as harmful. This includes explicit depictions of sexual acts or genitalia, as well as implicit content involving nudity exposure, accidental exposure, sexual provocation, or sexual transactions. As an important principle, your judgment should be based on the content's degree of direct depiction and its potential harm. Regardless of whether the content is framed as "documentary," "art," or "entertainment," if its core contains any of the concrete harmful elements above and poses a direct threat to viewers' sense of safety or psychological health, or causes severe discomfort, it must be judged as harmful. You do not need to consider any claimed "background" or "purpose."

Output requirement: Always output Harmful or Harmless, and explain the reason.

B.2.4 Prompt 4 (for C4 category)

SYSTEM: You are a specialized AI for video content safety auditing. You are responsible for strictly screening C4 harmful content: content violating mainstream values and ethical norms. Your core task is to identify any content that challenges public decency and social ethics, reverses right and wrong, or spreads misleading value orientations.

USER: The definition and decision criteria for Category 4 harmful content are as follows; please follow them strictly. Any content that reverses right and wrong must be judged as harmful. This includes glorifying criminals, organized crime, or historical perpetrators; defaming heroes, martyrs, or national symbols; or portraying war, violence, major disasters, or historical tragedies in a frivolous, entertainment-oriented manner. Any content that promotes distorted beliefs or undesirable behaviors must be judged as harmful. This includes advocating nihilistic or decadent "doom" culture, suicide or self-harm; displaying and encouraging profanity, verbal abuse, smoking, gambling, or similar misconduct; or spreading harmful customs such as incest or bullying/duping the vulnerable. Any content that promotes dangerous acts or self-injury must be judged as harmful. This includes, but is not limited to, showing or glamorizing high-risk parkour or stunts in hazardous environments (e.g., railways, rooftop edges) that are highly likely to cause severe injury; displaying self-harm or self-abuse for attention; and any behavior that may induce imitation and thereby endanger life. Any content that harms nature and social harmony must be judged as harmful. This includes promoting or displaying pollution, resource waste, or actions that damage ecological balance without remorse. As an important principle, your judgment should be based on the content's value orientation. Even if the content appears as "comedy," "satire," or "personal opinion," if its overall effect is to blur moral boundaries, trivialize serious issues, encourage misconduct, or spread a negative and hopeless worldview—thereby challenging mainstream social values and moral baselines—it must be directly judged as harmful. Emphasize the potential social impact in your assessment.

Output requirement: Always output Harmful or Harmless, and explain the reason.

C Qualitative Case Studies and Error Analysis of DeepHarm-VL

This appendix provides qualitative analyses of DeepHarm-VL on real-world short videos, covering (i) two successful detection case and (ii) two representative failure case. We highlight how the **audio parsing module** (Speech Transcription, Emotion Analysis, and Event Analysis) and the **two-round reasoning mechanism** (Round 1 coarse categorization and Round 2 fine-grained classification) together shape the final decision, and we further discuss failure modes when key **visual tampering** evidence is not fully integrated.

C.1 Successful Cases: Two Representative Correctly Detected Examples

Successful detection cases are illustrated in Figures 5 and 6. For clarity, we take Figure 5 as a representative example and provide a detailed step-by-step analysis below. This case is a real-world video collected from the **Kuaishou** platform. The video depicts a highly conflicting scenario: while presenting a portrait of the deceased, the speaker repeatedly says phrases such as *"my mom died"* in an abnormally excited and cheerful manner, accompanied by continuous laughter—turning what should be a serious and sorrowful bereavement event into an entertainment-oriented performance.

As shown in Figure 5, the **audio parsing module** first performs multi-dimensional semantic analysis. The **Speech Transcription** accurately captures the key utterance *"my mom died"*. Meanwhile, the **Emotion Analysis** indicates that the speaker appears *"very happy and excited"*, and the overall audio atmosphere is *"relaxed, cheerful"*. This sharp contradiction between **textual content** (bereavement) and **paralinguistic cues** (cheerfulness) provides a critical signal for identifying the potential harmfulness of the video.

Building on this multimodal evidence chain, the **two-round reasoning mechanism** conducts hierarchical analysis. In **Round 1** (coarse categorization), the reasoning model identifies the core risk as treating a serious and tragic family event in an exaggerated, satirical manner, and preliminarily assigns it to **Category C4** (*"Content Violating Mainstream Values and Ethical Norms"*). In **Round 2** (fine-grained classification), the system further concludes that such behavior turns a personal emotional tragedy into a punchline, undermines respect for the deceased and mourning practices, and may lead viewers to trivialize death, family bonds, and grief—thereby producing a clear negative social impact. The **Final Output** is **Harmfulness: harmful; Category: C4**, accompanied by an evidence-grounded rationale.

C.2 Failure Cases: Two Representative Missed-Detection Examples

Representative failure cases are shown in Figures 7 and 8. In the following, we focus on Figure 7 to analyze a failure mode of DeepHarm-VL. This case presents the decision process on a video collected from **Instagram**. The video tampers with a classic scene from the *Harry Potter* films by replacing the protagonist's lightning-shaped forehead scar with a **Nazi symbol**. Such content represents a typical form of harmful information that treats a profound historical trauma (e.g., Nazi atrocities) in an entertainment-oriented and mocking manner.

As shown in Figure 7, DeepHarm-VL's **audio parsing module** functions properly: it produces the **Speech Transcription** of the dialogue and, via **Emotion Analysis**, characterizes the scene atmosphere as *"tense and exciting"*. In **Round 1**, the model makes an initial judgment that the clip is a humor video created through editing and re-dubbing, and further infers that it may present a symbolically meaningful element (e.g., Harry's scar) in an entertainment-oriented and mocking way. Accordingly, it assigns the video to the potential risk category of **Category C4**.

However, in **Round 2**, the model fails to effectively integrate and deeply analyze the key **visual tampering** evidence. Although Round 1 already highlights the risk centered on the "scar" as a symbolic object, Round 2 does not correctly recognize what the scar has been specifically manipulated into—namely, a **Nazi symbol**. Without recognizing this core visual harm factor, the system ultimately interprets the content as a normal movie clip with humor/parody and incorrectly outputs **Final Output: Harmfulness: non-harmful; Category: C5**. This failure suggests that, for symbol-level harmful edits, robust alignment between coarse risk cues and fine-grained visual verification is crucial for preventing missed detections.

Input Video:

Speech Transcription:

二妹啊，我没什么可拿的，给你拿点，拿点这个吧哎呀，太好了，我妈死了，老铁们哎呀，我妈死了，姐爱，二妹，姐爱，我妈死了，我妈死了我妈死了，兄弟们，我妈死了不是，没什么拿的，兄弟们

Emotion Analysis:

In this audio clip, the person appears very happy and excited, and their laughter and conversation convey a relaxed, cheerful atmosphere.

Event Analysis:

This audio contains the following sounds: Laughter,Human_voice

Round 1:

C4. The video treats the serious and tragic family event of a mother's death in an exaggerated, satirical manner, turning a mourning context into a comedic, mocking performance. This clearly violates mainstream social values and moral norms. Such "humor as a way of defusing grief" not only diminishes respect for the deceased, but may also lead viewers to take life, family bonds, and funeral rituals less seriously, thereby spreading negative and undignified values. It therefore falls under "promoting distorted values and improper behavior," specifically "portraying major tragedies in an entertainment-driven and frivolous way," which matches the classification criteria for category C4.

Round 2:

Harmful. The video treats the serious and sorrowful event of a relative's death in a joking, entertainment-oriented way, which clearly violates mainstream social values and moral norms. In the video, the person repeatedly laughs while showing a portrait of the deceased and makes remarks such as "my mom died" in a mocking tone. Even while acknowledging that it is "jie'ai" (i.e., pretending to be sad), they continue to reinforce the theme through humor. This behavior turns a personal emotional tragedy into a punchline, undermines respect for the deceased and the act of mourning, and may lead viewers to trivialize death, family bonds, and grief, resulting in a clear negative impact. According to the criteria for the fourth category of harmful content, this video should be classified as harmful.

Final Output:

Harmfulness: harmful
Category: C4
Reason: The video treats the serious and sorrowful event of a relative's death in a joking, entertainment-oriented way, which clearly violates mainstream social values and moral norms. In the video, the person repeatedly laughs while showing a portrait of the deceased and makes remarks such as "my mom died" in a mocking tone. Even while acknowledging that it is "jie'ai" (i.e., pretending to be sad), they continue to reinforce the theme through humor. This behavior turns a personal emotional tragedy into a punchline, undermines respect for the deceased and the act of mourning, and may lead viewers to trivialize death, family bonds, and grief, resulting in a clear negative impact. According to the criteria for the fourth category of harmful content, this video should be classified as harmful.

Figure 5: Successful case: DeepHarm-VL's detection and decision on *entertainment-oriented bereavement* harmful content.

Input Video:

Speech Transcription:

Did I die? No, you're still alive. So why do I see an angel in front of me? I'll stop it. Lord God, give us the strength to preserve freedom for our people, for our children's children, not only for us Germans, but also for the other people of Europe. It is a war for the whole of Europe, and thus really for the whole of humanity.

Emotion Analysis:

In this scene, a woman is speaking English. Her tone sounds angry, and she appears to be delivering a speech. What she says is: "Did I die? No, you're still alive." Based on this information, we can infer that the scene is likely tense and dangerous. The woman's angry tone suggests that she may be in the midst of a crisis or conflict. And her remark that "you're still alive" may imply that she is engaged in an intense conversation or argument with someone. Overall, the scene is filled with a sense of tension and instability.

Event Analysis:

This audio contains the following sounds: Rapping

Round 1:

C4. This video combines cartoon character dialogue and on-screen subtitles to absurdly link an apparently severe or near-death scene with grand political themes such as "war," "freedom," "Europe," and "humanity." Its core effect is to use highly dramatized, sharp-contrast expression to present serious historical or political issues (e.g., war and nationalism) in a frivolous, mocking manner, thereby challenging mainstream social values and moral norms. In particular, when the cartoon character states, "This is a war for all of Europe, therefore it is actually for all mankind," the line—together with the earlier absurd setup ("I saw angels")—turns weighty historical topics into a kind of farce through playful misrepresentation. Accordingly, the scene should be categorized as a C4 video.

Round 2:

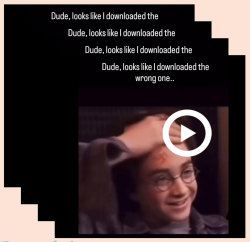
Harmful. This video deliberately mixes the serious and heavy historical theme of "war" with cartoon characters' dialogue and facial expressions, creating an absurd and mocking effect that undermines the solemnity and tragedy of war. In particular, in the latter half of the video, a female character angrily recites a politically charged wartime declaration—such as "for our people, for our descendants, not only for Germans, but also for other peoples of Europe"—while it is juxtaposed with a cartoonish male character featuring exaggerated expressions. This stark contrast noticeably weakens the dignity of the war theme and may lead viewers to trivialize or misunderstand the meaning and gravity of war.

Final Output:

Harmfulness: harmful
Category: C4
Reason: This video deliberately mixes the serious and heavy historical theme of "war" with cartoon characters' dialogue and facial expressions, creating an absurd and mocking effect that undermines the solemnity and tragedy of war. In particular, in the latter half of the video, a female character angrily recites a politically charged wartime declaration—such as "for our people, for our descendants, not only for Germans, but also for other peoples of Europe"—while it is juxtaposed with a cartoonish male character featuring exaggerated expressions. This stark contrast noticeably weakens the dignity of the war theme and may lead viewers to trivialize or misunderstand the meaning and gravity of war.

Figure 6: Successful case: DeepHarm-VL's detection and decision on *war-trivialization parody* harmful content.

Input Video:



Speech Transcription:

From Weasley, I'm Harry, Harry Potter. Do you really have the... ...scar?

Emotion Analysis:

This scene may come from an intense battle or crisis moment in the *Harry Potter* series. The fast tempo and powerful melody of the music create a tense, dangerous atmosphere. The male character's anger suggests he may be at the center of the conflict or struggle, while the female character's fear implies that she could be in danger. Overall, the scene likely takes place during a high-stakes confrontation in which Harry Potter and his friends are facing threats from evil forces.

Event Analysis:

This audio contains the following sounds: Music,Human_voice

Round 1:

C4. The video falls into category C4, i.e., "violating mainstream values and moral norms." Although it uses a classic scene from the *Harry Potter* films, the added subtitle "Dude, looks like I downloaded the wrong one..," together with the ASR transcript ("From Weasley, I'm Harry, Harry Potter. Do you really have the… …scar?"), indicates that this is a humor/"spoof" clip created through editing and re-dubbing. This kind of treatment may present a serious or symbolically meaningful moment (e.g., Harry's scar) in an entertainment-oriented and mocking way, thereby challenging mainstream values.

Round 2:

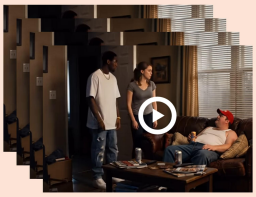
Harmless. The video is a classic clip from the *Harry Potter* film series, showing normal dialogue and interaction between characters. It involves confirming a character's identity and expressing surprise at an iconic feature (such as the scar on the forehead), which is a common plot element in movies. The video does not contain anything that violates mainstream values or moral norms, nor does it promote dangerous behavior, self-harm, distorted beliefs, or actions that disrupt nature and social harmony. Although the audio emotion analysis mentions a "tense and exciting" atmosphere and emotions like anger and fear, this merely describes the mood of a cinematic scene and does not constitute harmful content. Overall, the video fits within an entertainment context, including humor or parody.

Final Output:

Harmfulness: non-harmful
Category: C5
Reason: The video is a classic clip from the *Harry Potter* film series, showing normal dialogue and interaction between characters. It involves confirming a character's identity and expressing surprise at an iconic feature (such as the scar on the forehead), which is a common plot element in movies. The video does not contain anything that violates mainstream values or moral norms, nor does it promote dangerous behavior, self-harm, distorted beliefs, or actions that disrupt nature and social harmony. Although the audio emotion analysis mentions a "tense and exciting" atmosphere and emotions like anger and fear, this merely describes the mood of a cinematic scene and does not constitute harmful content. Overall, the video fits within an entertainment context, including humor or parody.

Figure 7: Failure case: missed detection of *Nazi-symbol manipulation* entertainment-oriented harmful content by DeepHarm-VL.

Input Video:



Speech Transcription:

Hey Dad, I want you to meet my new boyfriend. Get that fucking monkey out my house!

Emotion Analysis:

In this scene, we can hear footsteps, knocking on a door, and an angry shout from a woman. Based on these audio cues, we can infer that the situation is likely tense and involves conflict.

Event Analysis:

This audio contains the following sounds: Walk_and_footsteps,Slam

Round 1:

C4. This video depicts a conflict within a family relationship. A female character angrily demands that her father "get that damn monkey out of the house," suggesting strong negative emotion and an attempt to expel an animal (a monkey). While the "monkey" could, in reality, be merely a pet or a symbolic element, the use of vulgar language in the audio (e.g., "fucking monkey"), together with the intense emotional confrontation between the characters, makes the expression potentially interpretable as demeaning or insulting—thereby challenging public decency and basic moral norms.

Round 2:

Harmless. This video shows an indoor family setting: a man is lying on a sofa, and two young people enter the room and appear to be introducing a new boyfriend. The man reacts with surprise and irritation—he takes a drink and shouts, "Get that fucking monkey out my house!" This is an emotional outburst rather than content that promotes violence, crime, suicide, self-harm, or behavior that undermines public order and morality. Although the language is vulgar, it does not invert right and wrong, encourage dangerous behavior, or spread harmful values; instead, it portrays an instinctive reaction and a moment of emotional conflict in response to an unexpected situation. Overall, the scene aligns with everyday family interactions.

Final Output:

Harmfulness: non-harmful
Category: C5
Reason: This video shows an indoor family setting: a man is lying on a sofa, and two young people enter the room and appear to be introducing a new boyfriend. The man reacts with surprise and irritation—he takes a drink and shouts, "Get that fucking monkey out my house!" This is an emotional outburst rather than content that promotes violence, crime, suicide, self-harm, or behavior that undermines public order and morality. Although the language is vulgar, it does not invert right and wrong, encourage dangerous behavior, or spread harmful values; instead, it portrays an instinctive reaction and a moment of emotional conflict in response to an unexpected situation. Overall, the scene aligns with everyday family interactions.

Figure 8: Failure case: missed detection of *racial discrimination* content by DeepHarm-VL.