LET CLUES UNFOLD: A FINE-GRAINED MULTI-LABEL DATASET, BENCHMARK, AND DYNAMIC CLUE MINING METHOD FOR EXPLAINABLE HARMFUL SHORT VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short videos have become a major channel where harmful content reaches the public, and a single video often mixes several harm categories through multimodal cues. Existing harmful video detection techniques fall short in two ways: harmful-content taxonomies remain inconsistent across platforms and studies, and public datasets are small and coarsely labeled; multimodal fusion methods do not resolve the semantic gap and temporal misalignment across modalities, and fail to provide frame-specific causal reasoning for model predictions. To overcome these problems, we construct FineHM, a fine-grained multi-label harmful short video dataset, spanning 8 categories with harmful frames and bounding boxes, to provide frame- and region-level evidence for explainable multi-label classification. Built upon this dataset, we establish a standardized benchmark including fixed data splits, evaluation metrics, and comprehensive baseline results to facilitate future research. Moreover, we propose ExplainHarm, a dynamic clue mining explainable harmful short video detection method, to surface dynamic evidence from category feature and video content via a prompt-chained clue-mining process, and to achieve cross-modal semantic and temporal alignment through dual-scale temporal modeling. Experiments and statistical analyses demonstrate the dataset FineHM and the model ExplainHarm's superiority, achieving an 11.3% accuracy gain over SOTA multi-label models and a 0.6% IoU improvement over SOTA open-vocabulary object detection models.

1 INTRODUCTION

With the rapid expansion of social media, multimodal short video has become one of the most influential media formats for information sharing and social communication. The high multimodal density makes short video an efficient carrier of harmful content, which can be disseminated widely via social platforms, thus posing severe threats to cyberspace content security.

A single video often contains various harmful categories simultaneously, which plain binary or single-label classifiers cannot represent. Meanwhile, the wide adoption of multimodal large language models (MLLM) has drastically reduced the cost of producing fabricated harmful clips, fueling the rampant proliferation of toxic multimedia content online. Under these circumstances, manual review cannot cope with the sheer volume of harmful content and imposes negative psychological impacts on reviewers. Hence, it is crucial to develop harmful content detection methods that are both accurate on co-occurring harmful categories and justifying decisions for future audit and appeal.

Currently, two major bottlenecks stand in the development of harmful video detection: (1) The lack of standardized taxonomies pervades both industry and academia. Platform policies diverge sharply. For example, X (Twitter) enumerates specific abusive behaviours while Google handles harassment on a case-by-case basis rather than through a fixed policy. Academic taxonomies are equally fragmented, ranging from physical, emotional, relational and economic harm (Jo & Wojcieszak, 2025) to representational, allocative, quality of service, interpersonal and social harm. Moreover, existing public datasets inherit and amplify this ambiguity. MultiHateClip (Han Wang, 2024) and MuLX

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.

QA (Poddar et al., 2024) are limited in scale and labeled only at a coarse level, whereas Hate MM (Das et al., 2023) covers only a single harmful category. Therefore, none supplies the frame- and region-level annotation that fine-grained, explainable multi-label classification requires. (2) Multimodal multi-label approaches for detecting harmful videos remain scarce. Most fusion methods rely on shallow integration techniques, which fail to fully exploit the underlying spatiotemporal correlations and causal relationships across modalities, leaving the semantic gap and temporal misalignment unresolved. Furthermore, these multimodal fusion methods can only visualize the regions the model attends to but not explain the underlying reasoning that leads to a specific decision.

To address these issues, this paper first constructs a fine-grained, explainable multi-label harmful video dataset by consolidating platform community guidelines and prior taxonomies into a unified definition and label set. Specifically, we gather candidate videos from multiple public video platforms and supplement the corpus with benchmark datasets proposed in existing literature. The dataset annotates each video not only with its harmful categories but also with the specific harmful frames and in-frame harmful regions that justify the label, thereby providing a solid foundation for fine-grained, traceable harmful content detection and justification. Then, we propose a method to extract dynamic harmful clues from video content and establish explicit, traceable connections between these clues and the final classification results. This approach constructs multi-layered prompt chains that dynamically collect observable harmful clues based on pre-defined harmful categories and historical experiences. Rather than depending on opaque abstract feature embeddings, it explicitly retrieves reasoning evidence that is directly human-interpretable. To enhance multimodal content understanding, we introduce dual-scale temporal modeling to bridge the cross-modal temporal granularity gap and leverage long-range contextual cues for improved harmful content discrimination. Finally, we integrate the extracted harmful cues with the fused multimodal features via a clue-guided alignment mechanism to produce the final video-level detection results, along with the corresponding harmful frames and bounding boxes for interpretable evidence.

Our research contributions are summarized as follows.

1) We construct FineHM, a fine-grained multi-label harmful video dataset. A standardized harmful video content criterion is defined by combining the community guidelines of multiple video platforms and drawing on harmful content detection research works. On this basis, our taxonomy covers seven harmful types plus a normal class. Videos are collected from diverse platforms and annotated with both multi-label tags and explanatory details, including harmful frame localization and harmful object bounding boxes within each frame. The resulting dataset contains 9,979 video clips with a total duration of 21 hours, around 2 million frames, and more than 440,000 interpretative object bounding boxes, providing a unified data foundation for explainable multi-label harmful video detection research.

2) We propose ExplainHarm, a dynamic clue mining explainable multi-label harmful short video detection framework that enables accurate fine-grained harmful video classification while providing intuitively explainable localization at both the frame and bounding-box levels. It mainly consists of three modules, namely the prompt-chain-based dynamic clue mining module, the dual-scale temporal enhancement module, and the clue-guided explainable classification module. The first module aims to explicitly excavate harmful clues and dynamically update category-associated clues as the input video content streams in. The second module attempts to enhance the video content representation via multimodal semantic and temporal alignment. The last module completes the integration of high-level explicit harmful clues and low-level video features to perform fine-grained harmful classification and explainable spatial-temporal localization.

3) We systematically evaluate our dataset FineHM against public benchmarks, proving its superiority and practical applicability. Extensive experiments further demonstrate that our proposed model ExplainHarm consistently outperforms state-of-the-art methods in both multi-label detection performance and model explainability.

2 Related Work

2.1 Harmful Content Taxonomy and Datasets

The boundaries of "harmful content" are shaped by platform policy, legal constraints, and the moderation setting. Studies therefore define harm inconsistently, which in turn misaligns moderation guidelines, training objectives, and evaluation metrics(Banko et al., 2020). Banko et al. (2020) distinguish harassment, self-harm, ideological harm, and exploitation; Scheuerman et al. (2021) classify harm as physical, emotional, relational, or economic; Jo & Wojcieszak (2025) propose a six-category scheme for video platforms. Platform guidelines diverge as well—YouTube classifies violence as physical harm, TikTok as harassment (Singhal et al., 2023). Moreover, most datasets are coarse-grained and single-class. MultiHateClip (Han Wang, 2024) covers hate/offensive content, LSPD (Duy et al., 2022) contains pornography, PHAD (Chappa et al., 2024) includes tobacco, and UCF-Crime (Sultani et al., 2018) collects surveillance actions. Online-oriented corpora, such as KuaiMod (Lu et al., 2025), VHD11K (Yeh et al., 2024), MetaHarm(Jo & Wojcieszak, 2025), fit moderation settings better, but they still provide only video-level labels without spatio-temporal or explanatory annotations. MuLX-QA (Poddar et al., 2024) annotates explanations but only for text modality. The absence of fine-grained explainable annotations remains a recurring bottleneck for explainable detection.

2.2 Harmful Video Detection

Early approaches largely target visually salient phenomena such as violence or pornography and adopted the binary classification task setting defined by their source datasets. Video encoders advance spatiotemporal representation learning (Xie et al., 2018; Liu et al., 2022), while audiovisual and vision–language models incorporate speech, ambient sound, and textual context (Radford et al., 2021; Zhang et al., 2025; Xu et al., 2025b). Such multimodal integration is critical since harmful semantics often emerge from asynchrony across modalities or require joint interpretation of multimodal signals. Specialized benchmarks including VidHarm, HateMM, MultiHateClip, MetaHarm, and VHD11K have verified the benefits of multimodal fusion, but their label spaces remain narrow (Edstedt et al., 2022; Das et al., 2023; Han Wang, 2024; Jo & Wojcieszak, 2025; Yeh et al., 2024). Consequently, most existing methods focus on a single risk domain or leverage multimodal fusion merely for video-level predictions. They fail to simultaneously identify co-occurring harmful labels and associate each detected category with its supporting frames and regions.

2.3 Explainable Harmful Video Detection

Most detectors offer bare decisions without evidence, eroding trust and burdening human review. Existing explainable methods have mainly two types: post-hoc visualization and language-model explanations. Post-hoc attribution methods, typified by Grad-CAM, are class-conditioned but spatially coarse and are not ordinarily validated against object-level supporting evidence (Selvaraju et al., 2017). More recent multimodal methods use language models to convey reasoning, but a plausible rationale can be weakly related to the input. Textual rationales improve auditability of social media text but do not localize visual evidence in video (Poddar et al., 2024). Besides, text-conditioned object detectors such as OWL-ViT and Grounding DINO (Minderer et al., 2022; Ren et al., 2024) enable spatial region grounding, yet object detection alone cannot verify whether the localized region supports the harmful prediction.

3 FineHM Dataset

3.1 Definition and Taxonomy of Harmful Content

To establish a labeling taxonomy suitable for harmful short video content research, this study firstly comprehensively references the harmful content definitions from existing academic literature and content governance policies of mainstream social network platforms, which are summarized in Appendix 5. It can be seen that academic theories focus on abstract hazard mechanisms, while platform guidelines emphasize implementable violation categories. Given their distinct emphases, we do not

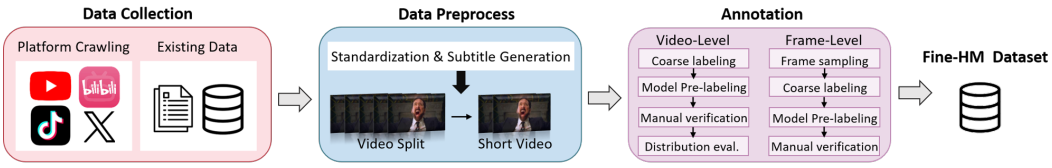


Figure 1: Fine-HM dataset construction process.

Table 1: FineHM descriptive statistics. Video counts, durations, and frames are non-exclusive across harmful categories.

Category	#Video	Duration	#Frame	#Box	Data Sources
Bloody/gory	1,148	∼2h08m	209,381	63,107	Youtube.com
Violence	698	∼1h08m	116,654	31,404	X (twitter.com); Dataset (fight_new_youtube2)
Nudity/pornography	1,478	∼2h00m	229,383	146,328	Pornhub.com; Xvideos.com
Political non-compliance	491	∼1h45m	180,076	36,982	Youtube.com; X (twitter.com)
Abusive/conflict-provoking	324	∼43m	71,016	23,671	Youtube.com; Bilibili.com
Financial risk	1,230	∼2h19m	223,262	64,817	X (twitter.com); Script Generation
Smoking/drug use	1,279	∼2h10m	221,796	76,382	Youtube.com
Normal	4,257	∼8h23m	848,225	N/A	Youtube.com; Dataset (MSR_VTT)

directly concatenate the two frameworks. Instead, we perform a cross-comparison to distil a tailored taxonomy for the multi-label harmful short video detection task.

Ultimately, we define seven harmful categories (blood, money, sexy, smoke, violent, policy, abusive) and one normal class as the classification labels. The taxonomy is guided by three design principles: multimodal discriminability, multi-label co-occurrence and objective evaluability. Specifically, we only retain categories identifiable through combined visual, auditory and textual cues, allow multiple harmful labels per video, and prioritize harmful types with broad consensus in academia and platform moderation standards to mitigate subjective bias.

3.2 Data Statistics and Analysis

This study constructs the FineHM dataset based on the above classification taxonomy. This whole process consists of data collection, preprocessing, and annotation, which is shown in Figure 1. More construction details are in Appendix A.2.

Data Statistics We conduct a comprehensive analysis of the Fine-HM dataset from multiple dimensions, including overall descriptive statistics, category-wise descriptive statistics, explanatory bounding-box distribution, and usability evaluation.

Fine-HM dataset comprises 8 categories, with a total video duration of approximately 21 hours, roughly 2.1 million frames, and about 440,000 explanatory bounding-box annotations. The overall descriptive statistics and the data sources of each harmful video category collected are shown in Table 1.

Data Evaluation To objectively evaluate the applicability of the Fine-HM dataset in the multi-label harmful video classification task, this study conducts comparative analysis with representative public datasets related to video anomaly detection, content review, and harmful content detection, including LSPD (Duy et al., 2022), PHAD (Chappa et al., 2024), UCF-Crime (Sultani et al., 2018), Vidharm (Edstedt et al., 2022), KUAIMOD (Lu et al., 2025), VHD11K (Yeh et al., 2024), WVD (Nadeem et al., 2024), and MetaHarm (Jo & Wojcieszak, 2025). This study conducts comparisons across these datasets and the Fine-HM dataset in terms of video count, average number of frames, average duration, total video duration, number of labels, and label categories, which is shown in Table 2. Given that different dataset papers emphasize different video parameters, we estimate certain missing parameters using available data information to ensure a comprehensive and fair comparison. Estimated values are marked with "∼" in the table, and the statistics Fine-HM are highlighted in bold.

Table 2: Comparison of Fine-HM with existing datasets. "UCF" is the short for UCF-Crime.

Dataset	#Videos	Duration	#Classes & Classes	Annotation
LSPD	4000	∼283h	2 & Pornographic / non-pornographic	Video-level & Binary
UCF	1900	128h	13 & Surveillance actions	Temporal & Single-label
PHAD	5730	>200h	2 & Smoking / normal	Spatial (smoking) & Binary
VidHarm	3589	∼10h	3 & Adult / violent / normal	Video-level & Multi-label
KUAIMOD	24562	N/A	15 & Legal / ethics / commercial / contract violations	Video-level & Single-label
MetaHarm	19422	N/A	6 & Informational / hate / addictive / clickbait/ sexual / physical	Video-level & Multi-label
VHD11K	1000	∼9.75h	10 & Violence / substance abuse / child harm / sexual content etc.	Video-level & Multi-label
WVD	334	∼96min	3 & Hot / cold / non-violence	Video-level & Single-label
Fine-HM (Ours)	9979	∼21h	8 & Bloody & gory / violence / nudity & pornography etc.	Video-level + Spatio-temporal & Multi-label

From the comparison, in terms of annotation granularity, most existing datasets provide only video-level labels. Only UCF-Crime and PHAD offer temporal and spatial annotations separately. None support frame-level or region-level evidence localization, whereas Fine-HM is the first to provide spatio-temporal localization annotations. As to the label granularity, LSPD and PHAD are binary, while KuaiMod and WVD are single-label, making them inadequate for modeling the co-occurrence of multiple harmful categories in short-video platforms. Fine-HM, MetaHarm and VHD11K instead adopt multi-label classification. What is more, with 9,979 videos totaling approximately 21 hours, Fine-HM balances category coverage and annotation cost. To the best of our knowledge, Fine-HM is the first harmful video benchmark that jointly offers a multi-label taxonomy and spatio-temporal localization annotations.

4 BENCHMARK

FineHM defines two tasks: **video-level harm classification** and **evidence grounding**. Given an input video v, propose a harmful video detection framework that simultaneously accomplishes two objectives: automatically identify the harmful category to which the video belongs and provide explanatory evidence that grounds and justifies the model's decision. This process can be formally formulated as follows:

$$\mathcal{E}, c = \Psi\left(v, C\right), \tag{1}$$

where $C = \{c_1, c_2, \ldots, c_k\}$ denotes the predefined harmful category set. Ψ represents the harmful video detection framework. $\mathcal{E}$ denotes the explanatory evidence. c denotes the predicted result.

The benchmark has 10 classification baselines and 7 grounding baselines (with the proposed model ExplainHarm). The detailed information is described in 6.2. Evaluation metrics are provided in 6.1.

5 METHODOLOGY

The framework of the proposed Explain-Harm model is illustrated in Figure 2. The model has three main modules: Prompt-Chain-Based Dynamic Clue Mining to extract explicit harmful clues from input videos, Dual-Scale Temporal Enhancement Module to achieve comprehensive content representation, and Clue-Guided Explainable Classification Module to fuse harmful clues into video features to conduct classification and deliver spatiotemporal interpretability.

5.1 PROMPT-CHAIN-BASED DYNAMIC CLUE MINING

To mitigate the black-box problem of opaque decision-making in harmful video detection, we draw inspiration from ontology theory. As a knowledge framework that characterizes domain semantics through structured concept sets and their mutual relations, ontology enables a clear separation between category-level abstract semantics and instance-level concrete information, thus providing theoretical support for inducing category-specific harmful clues across videos. This module con-

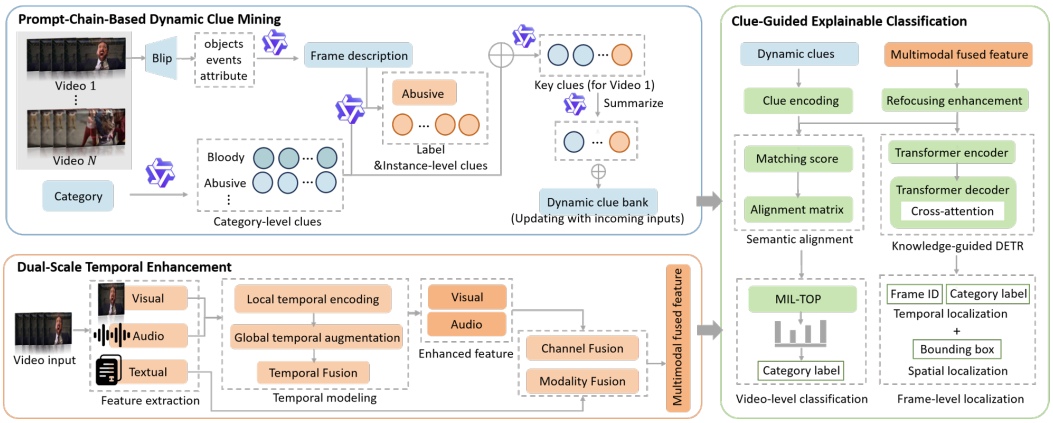


Figure 2: The framework of the proposed Explain-Harm model. Prompt-Chain-Based Dynamic Clue Mining extracts explicit harmful clues from input videos. Dual-Scale Temporal Enhancement Module achieves comprehensive content representation. Clue-Guided Explainable Classification Module fuses harmful clues into video features to conduct classification and deliver spatiotemporal interpretability.

tinuously aggregates instance evidence from incoming videos during both training and inference, delivering stable and reusable high-level semantic guidance for subsequent classification.

Specifically, we build a static category background to supply category-level abstract clues. For each predefined harmful category $c_i \in C$, we craft exclusive category prompts and input them into the Large Language Model (LLM) Qwen2.5-7B to generate a fixed set of category-level harmful clues B_{c_i}. For the normal category, the LLM is prompted to describe neutral daily content patterns. The complete category background B_C aggregates category-level clues of all categories.

Raw video data always suffers from high visual redundancy, background noise and sparse key semantics, causing vision-language models to be disturbed by irrelevant content, leading to unstable prediction results. To address this issue, we transform video content into a denoised and structurally unified high-level semantic description. Given an input video v with uniformly sampled frame sequence $F = \{f_1, ..., f_T\}$, BLIP is leveraged to get the frame-level semantic set $G_F = \{g_1, ..., g_T\}$ (objects, events, attributes) and textual frame captions $S_F = \{s_1, ..., s_T\}$. Discrete frame semantics are rearranged into temporally ordered text templates using sequential conjunctions. A few-shot prompt optimization strategy retrieves prompts M from annotated samples with similar semantics and append them to the current prompt for better consistency. The refined prompt is then sent to an LLM to produce a coherent high-quality video-level semantic description D_v.

Given the video description D_v, we obtain the instance-level clues L_c^v. An LLM conducts category-wise analysis against the harmful category set C. It returns matched harmful categories together with video-level harmful clues for harmful videos, or a normal label with justifications.

Dynamic clues are then derived via shared clue induction. At the beginning, static category clues B_C and video-specific instance clues L_c^v are merged and filtered by the LLM to remove noisy and ambiguous candidates, achieving discriminative key clues L_c^{v*}. Next, we summarize representative shared clues K_c^v from key clues. With continuous video samples streamed into the model during both training and inference phases, the module iteratively integrates new extracted video-specific clues to enrich the dynamic harmful clue bank $\mathcal{K}$. The whole process can be formulated as follows.

$$L_c^{v*} = LLM(B_c \cup L_c^v, c), \quad K_c^v = LLM(L_c^{v*}), \quad \mathcal{K} = \{K_c^v \mid v \in V, c \in C_v\}. \tag{2}$$

5.2 Dual-Scale Temporal Enhancement

To address temporal asynchrony between visual and audio modalities in videos, this module integrates multi-modal information into frame-level fused features through dual-scale temporal modeling, achieving semantic and temporal alignment across video modalities.

Given an input video v_i with total duration L_i, it first extracts raw features for vision $I_{i,n}$ and audio $A_{i,n}$ using CLIP and MFCC respectively, based on a unified set of evenly sampled temporal anchors $\mathcal{T}_i = \{\tau_{i,1}, \tau_{i,2}, ..., \tau_{i,N}\}$ $(n = 1, 2, ..., N)$. Then we map them to a shared space:

$$\begin{aligned} H_i^I &= LSTM_I(\mathbf{I}_i), \quad \tilde{\mathbf{I}}_i = W_I H_i^I + b_I, \\ H_i^A &= LSTM_A(\mathbf{A}_i), \quad \tilde{\mathbf{A}}_i = W_A H_i^A + b_A, \quad \tilde{\mathbf{A}}_i, \tilde{\mathbf{I}}_i \in \mathbb{R}^{N \times d}, \end{aligned} \tag{3}$$

where W_I, b_I, W_A and b_A denote the weight and bias of the linear projection layer. d represents the dimension of the shared latent space.

As to the text modality, we utilize the BERT-base-uncased encoder to extract text features $\mathbf{T}_i \in \mathbb{R}^{d_T}$ based on the video description D_v. Since text carries no segmented temporal structure, an MLP layer is leveraged to transform text features to match the unified dimensionality:

$$\tilde{\mathbf{T}}_i = MLP(\mathbf{T}_i), \quad \tilde{\mathbf{T}}_i \in \mathbb{R}^{N \times d} \tag{4}$$

Then, dual-scale temporal modeling is implemented with local temporal encoding and graph-based global temporal augmentation. LSTM and 1D-CNN extract short-term evolution features for vision and audio respectively:

$$I_i^{loc} = LSTM(\tilde{\mathbf{I}}_i), \quad A_i^{loc} = Conv1D(\tilde{\mathbf{A}}_i), \tag{5}$$

where I_i^{loc} and A_i^{loc} are the local temporal visual feature and local temporal audio feature.

Take local temporal outputs X_i (visual or audio features) as graph nodes and construct two graphs based on semantic similarity and temporal distance. We compute cosine semantic similarity S_i and temporal distance $A_i^{dis}(p, q)$ for any two timestamps p and q:

$$S_i(p, q) = \frac{X_i(p) X_i(q)^\top}{\|X_i(p)\|_2 \|X_i(q)\|_2 + \varepsilon}, \quad A_i^{dis}(p, q) = \frac{1}{|p - q| + 1}. \tag{6}$$

A two-layer GCN with GELU activation is executed on both adjacency matrices. Outputs are concatenated and linearly projected to globally enhanced visual feature I_i^{glob} and audio feature A_i^{glob}.

Eventually, multimodal features are fused along the temporal, channel and modality dimensions. In detail, 1D convolution and Softmax generate temporal attention weights for vision and audio to mutually reinforce each other, producing temporally fused representations. The global text feature $\widetilde{T}_i$ remains unchanged. Multi-head linear transformation further performs to enhance the representation capacity of each modality in distinct subspaces and boost consistency for subsequent cross-modal fusion. Finally, we apply gated attention to selectively weight textual, visual, and audio features and conduct additive fusion to generate a unified multimodal representation F_i that keeps complementary semantics and suppresses redundant noise.

5.3 Clue-Guided Explainable Classification

This module establishes semantic associations between low-level multimodal representations and high-level dynamic harmful cues, yielding dual-granularity explainable outputs: video-level classification and frame-level localization results.

The multimodal fused feature F_i inherently mixes noise and harmful clues. Directly feeding it into the classifier tends to make the model rely on coarse-grained contextual content rather than discriminative key regions. To address this issue, a residual-based refocusing enhancement refines this representation, and a multi-label classification head predicts frame-wise category probability $\mathbf{P}_i \in \mathbb{R}^{N \times C}$. The probability matrix is transformed into category-guided weights, which are then used to combine with the original fused features via weighted addition, obtaining a category-specific semantic representation. $M_i = \mathbf{P}_i^\top \mathbf{F}_i$. This representation is further normalized into $\tilde{M}_i$.

After this step, we integrate dynamic clues with $\tilde{M}_i$ to establish the semantic alignment between the low-level frame-level fused feature and the high-level shared harmful clues K. The clues are first projected into the shared semantic space and residually fused with $\tilde{M}_i$, producing clue-aligned embedding $\mathbf{G}_i$. Scaled cosine similarity with a learnable temperature coefficient is computed between frame features and clue embeddings to build an alignment matrix measuring frame-category relevance.

By introducing the idea of treating all frames of one video as an instance bag (Sultani et al., 2018), we select top-k frames with the highest matching scores for each category to aggregate video-level prediction logits and the top-k ratio is adaptively determined by video length k_i.

For spatial-temporal explainable localization, we build a knowledge-guided DETR with harmful clue embeddings as category priors to localize harmful targets and output object categories plus bounding boxes as visual explanatory evidence. We use ResNet to extract spatial features from each frame, encoded into spatial token sequences. Visual queries are initialized by mean pooling these spatial tokens and combined with learnable query embeddings, while harmful category queries are derived from clue embedding $\mathbf{G}_i$. Subsequently, we conduct multi-head cross-attention by taking visual queries as Queries and category queries as both Keys and Values to obtain clue-guided query representations. Next, we perform Transformer decoding with clue-guided query representations serving as decoding queries and encoded spatial features as memory input, to obtain spatial query representations $d_{i,n}$. The category prediction head and bounding box regression head then output category scores and spatial positions for frame-level classification and localization.

6 Experiment

6.1 Evaluation Metrics

This study evaluates the model from two perspectives: classification performance and explainability performance. Four metrics are adopted for classification evaluation, namely Accuracy (Acc), mean Average Precision (mAP), Average Recall (AR) and One-Error (OE). Three metrics are utilized to measure explainability performance, including Precision (Pre), mean Average Precision (mAP) and Intersection over Union (IoU).

6.2 Baselines

To fully verify the effectiveness of the proposed Explain-Harm, we design a comparative experiment to evaluate the multi-label classification performance compared to 9 models, including traditional spatiotemporal video baselines **GoogLeNet** (Szegedy et al., 2015), **S3D** (Xie et al., 2018) and **Video-Swin-Transformer** (Liu et al., 2022), multi-view multi-learning baselines **LVSL** (Zhao et al., 2023), **MRDFS** (Zhang et al., 2019) and **MVLD** (Zhao et al., 2022), and general vision-language large models **CLIP** (Radford et al., 2021), **VideoLLaMA3-7B** (Zhang et al., 2025) and **Qwen2.5-Omni-7B** (Xu et al., 2025a). Furthermore, we evaluate the explainability performance compared to 6 models, including open-vocabulary object detection baselines **OWL-ViT** (Minderer et al., 2022), **Grounding DINO-1.5** (Ren et al., 2024), **MM Grounding DINO** (Zhao et al., 2024) and **YOLOE** (Wang et al., 2025), and vision-language large model baselines **LLaVA-NeXT-Video-7B** (Li et al., 2024) and **Qwen3-VL** (Bai et al., 2025).

Finally, we conduct an ablation study to compare our proposed model Explain-Harm with its variants and demonstrate its effectiveness. The variants contain **Explain-Harm w/o text** (remove textual modality), **Explain-Harm w/o visual** (remove visual modality), **Explain-Harm w/o audio** (remove audio modality), **Explain-Harm w/o dynamic clues** (remove dynamic clues), **Explain-Harm w/o local temporal modeling** and **Explain-Harm w/o global temporal modeling**.

6.3 Comparative Experiment Analysis

As shown in Table 3, **for classification performance**, Explain-Harm outperforms traditional vision-only spatiotemporal models across Acc, AR and mAP metrics. Pure visual representations only capture superficial visual cues and fail to model implicit semantics, mixed multiple labels and cross-modal relations in harmful videos, while our model integrates multimodal temporal information and dynamic harmful clues to enrich semantic modeling and increase the accuracy of multi-label recognition. Our model also achieves superior results over multi-view multi-label methods since they merely exploit view and label correlations without semantic constraints for harmful scenarios. Compared with multimodal large models, Explain-Harm exhibits stronger task adaptability due to the explicit modeling of harmful clues and the combination of high-level semantic clues with low-level visual representations, strengthening discrimination against complex harmful content. **For explainability performance**, compared with general open-vocabulary detectors, Explain-Harm models dy-

Table 3: Comparative experiment results (%). The best results are in bold. The second-best results are starred.

(a) Video-level classification

Method	Acc. ↑	mAP ↑	AR ↑	OE ↓
GoogLeNet	54.3	60.5	63.8	0.47
S3D	59.9	66.2	61.0	0.39
Video Swin	64.6	77.9*	76.8*	0.30*
LVSL	67.2*	75.6	75.4	0.30*
MRDFS	55.1	60.4	63.9	0.45
MVLD	65.9	72.7	73.4	0.36
CLIP	63.1	73.1	72.0	0.33
VideoLLaMA3-7B	44.8	30.2	43.0	0.41
Qwen2.5-Omni-7B	39.4	34.6	54.2	0.44
ExplainHarm	**78.5**	**91.4**	**90.3**	**0.15**

(b) Evidence grounding

Method	Pre. ↑	mAP ↑	IoU ↑
OWL-ViT	12.8	5.5	75.6
Grounding DINO	13.3	1.2	72.3
MM-Grounding-DINO	12.0	2.3	70.6
YOLOE	11.1	3.1	74.4
LLaVA-NeXT-Video	0.24	0.17	53.6
Qwen3-VL	0.3	0.2	57.3
ExplainHarm	**48.18**	**33.8**	**76.2**

Table 4: Ablation experiment results (%). The best results are in bold.

Variant	Classification Performance				Evidence Grounding		
	Acc.↑	AR ↑	mAP ↑	OE ↓	Pre. ↑	mAP ↑	IoU ↑
w/o text	77.7	83.5	81.6	0.16	37.5	31.9	73.7
w/o visual	54.7	56.2	53.5	0.40	42.5	30.9	74.5
w/o audio	74.7	84.8	80.7	0.18	41.8	25.9	74.9
w/o dynamic clues	74.6	87.9	81.5	0.18	40.5	24.3	74.9
w/o local temporal modeling	76.0	89.6	87.2	0.17	45.3	29.4	75.1
w/o global temporal modeling	76.9	90.1	88.1	0.16	46.2	30.8	75.4
ExplainHarm	**78.5**	**91.4**	**90.3**	**0.15**	**48.18**	**33.8**	**76.2**

namic harmful clues to accurately capture evidence regions and increase localization explainability, instead of only matching text and regions. Moreover, Explain-Harm outperforms vision-language large models on this task since the dynamic harmful clues provide domain-specific constraints compared to general large language models.

In summary, Explain-Harm outperforms SOTA methods on both classification and explainability performance, demonstrating the comprehensive advantages in both classification accuracy and spatial-temporal localization evidence generation.

6.4 Ablation Study Analysis

As shown in Table 4, removing any modal information leads to obvious degradation in both classification and localization metrics, validating that each modality provides complementary information via temporal enhancement and gated fusion, thus constructing robust semantic representations. What is more, eliminating dynamic harmful clues severely weakens the model's discrimination and explainability, which shows that pure fused low-level features struggle with overlapping and concealed harmful semantics, whereas high-level semantic priors from dynamic clues simultaneously increase classification accuracy and localization performance. Moreover, the performance of the model without dual-scale temporal enhancement (local temporal modeling and global temporal modeling) degrades, highlighting its contribution to the final results. Therefore, the multi-modal fusion and dynamic harmful clue mining work collaboratively to combine low-level visual signals with high-level harmful semantics, serving as key components for the model's overall performance.

7 Conclusion

This paper presents **FineHM**, a dataset and benchmark for explainable multi-label harmful short-video detection, and **ExplainHarm**, a dynamic clue mining model that jointly predicts fine-grained harmful categories and grounds each prediction in localized, class-aligned evidence. On FineHM, ExplainHarm improves classification accuracy by 11.3 points over the strongest multi-label baseline and evidence IoU by 0.6 points over the strongest open-vocabulary detector. This paradigm provides visual spatio-temporal evidence for explanations, offering a scalable foundation for the audit and governance of harmful short-video content as abundant data accumulates.

Ethics Statement

This work studies the detection of harmful video content, which raises several ethical considerations. First, regarding the dataset, Fine-HM is constructed solely for research purposes: videos are collected from publicly accessible online sources and used strictly for analysis rather than redistribution. Access is granted only to research institutions under a data use agreement. Second, regarding the sensitive material itself, the dataset necessarily contains harmful content (e.g., violence, pornography, politically sensitive content), which we treat exclusively as objects of study. This paper neither propagates nor endorses such content. Finally, regarding privacy, videos may depict identifiable individuals. We minimize the exposure of personal information and comply with applicable data protection regulations in data collection and usage.

AI Disclosure

In this work, we used generative AI tools for clue mining and assisting with translation. We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020*, pp. 125–137. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ALW-1.16.

Ravi Teja N. V. S. Chappa, Charlotte McCormick, Susana Rodriguez Gongora, Page Daniel Dobbs, and Khoa Luu. Public health advocacy dataset: A dataset of tobacco usage videos from social media. *CoRR*, abs/2411.13572, 2024. doi: 10.48550/ARXIV.2411.13572.

Kevin Coe, Kate Kenski, and Stephen A. Rains. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679, 2014.

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

Phan Duy, Thanh Nguyen, Quang Nguyen, Loc Tran, Ngoc-Khoi Khac, and Lung Vu. Lspd: A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15:198, 02 2022. doi: 10.22266/ijies2022.0228.19.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, Anette Novak, and Gustav Grund Pihlgren. Vidharm: A clip based dataset for harmful content detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1543–1549, 2022. doi: 10.1109/ICPR56361.2022.9956148.

Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, 12(1):129, 2022. doi: 10.1007/s13278-022-00951-3.

Usman Naseem Roy Ka-Wei Lee Han Wang, Tan Rui Yang. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *ACM MULTIMEDIA 2024*. ACM, 2024.

Wonjeong Jo and Magdalena Wojcieszak. Metaharm: Harmful youtube video dataset annotated by domain experts, gpt-4-turbo, and crowdworkers. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):2496–2509, Jun. 2025. doi: 10.1609/icwsm.v19i1.35951.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3202–3211, June 2022.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, Yi-Fan 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 *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 4682–4693. Association for Computing Machinery, 2025. ISBN 9798400714542. doi: 10.1145/3711896.3737273.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 728–755. Springer Nature Switzerland, 2022. ISBN 978-3-031-20080-9.

Muhammad Shahroz Nadeem, Fatih Kurugollu, Hany F. Atlam, and Virginia N.L. Franqueira. Weapon violence dataset 2.0: A synthetic dataset for violence detection. *Data in Brief*, 54:110448, 2024. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2024.110448.

Soham Poddar, Rajdeep Mukherjee, Azlaan Mustafa Samad, Niloy Ganguly, and Saptarshi Ghosh. Mulx-qa: Classifying multi-labels and extracting rationale spans in social media posts. *ACM Trans. Web*, 18(3):35:1–35:26, 2024. doi: 10.1145/3653303.

Zhenting Qi, Ruike Zhu, Zheyu Fu, Wenhao Chai, and Volodymyr Kindratenko. Weakly supervised two-stage training scheme for deep video fight detection model. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 677–685, 2022. doi: 10.1109/ICTAI56018.2022.00105.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. A framework of severity for harmful content online. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):368:1–368:33, 2021. doi: 10.1145/3479512.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 723–741. ACM, 2023. doi: 10.1145/3600211.3604673.

Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. In *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023*, pp. 868–895. IEEE, 2023. doi: 10.1109/EUROSP57164.2023.00056.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24591–24602, October 2025.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Meilong Xu, Di Fu, Jiaxing Zhang, Gong Yu, Jiayu Zheng, Xiaoling Hu, Dongdi Zhao, Feiyang Li, Chao Chen, and Yong Cao. Rb-ft: Rationale-bootstrapped fine-tuning for video classification. *arXiv preprint arXiv:2511.15923*, 2025b.

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*, volume 37, pp. 112950–112961. Curran Associates, Inc., 2024. doi: 10.52202/079017-3589.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025.

Jia Zhang, Zhiming Luo, Candong Li, Changen Zhou, and Shaozi Li. Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, 95:136–150, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.06.003.

Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Learning view-specific labels and label-feature dependence maximization for multi-view multi-label classification. *Applied Soft Computing*, 124:109071, 2022. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2022.109071.

Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 25:7235–7247, 2023. doi: 10.1109/TMM.2022.3219650.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection, 2024.

A APPENDIX

A.1 EXISTING TAXONOMY OF HARMFUL CONTENT

This study comprehensively draws on theoretical frameworks of harmful content from existing literature and content governance rules of mainstream platforms. From the results in Table5, existing academic works focus on harmful outcomes and negative social impact, providing a theoretical basis for understanding the conceptual meaning and hierarchical levels of harmful content. Meanwhile, online platforms, such as YouTube, X and TikTok, are rooted in platform governance practice, imposing varying degrees of restriction on content including pornography and vulgarity, violence and gore, harassment and verbal abuse, fraud and inducement, misleading information, and political risks.

Table 5: Summary of harmful categories from literature and platform rules.

Source	Harmful Categories
Banko et al. (2020)	Hate and harassment, self-harm, mental harm, exploitation
Gongane et al. (2022)	Far-right/left-wing, hate speech, extremism, all right-wing, Islamophobia, Islamic extremism, extremist information, misleading/false information, conspiracy theories, radicalization
Coe et al. (2014)	Fake news, false information, satirical news, rumors, clickbait, hate speech, cyberbullying, profanity, toxic language, abusive language
Shelby et al. (2023)	Rude language such as insults, degrading language and profanity
Scheuerman et al. (2021)	Physical harm, emotional harm, relational harm, economic harm
Qi et al. (2022)	Representational harm, distributive harm, service-quality harm, interpersonal harm, social harm
Bilibili	Illegal and non-compliant content, gambling fraud, illegal information, infringement appeals, lies, social-event-related content, fake and inaccurate information, prohibited promotion, others
Tik Tok	Pornographic and vulgar content, illegal and non-compliant content, politically non-compliant information, inaccurate information, prohibited marketing, content endangering personal safety, minor-related content, rights infringement, others
Kuaishou	Pornographic and vulgar content, illegal and non-compliant content, historical nihilism, hate-provoking insults, guidance for private transactions, inappropriate remarks, inaccurate information, AI-generated content issues, inappropriate minor behaviours, others
Youtube	Pornographic and vulgar content, violent or repulsive content, hateful or abusive content, harmful or dangerous behaviors, spam or misleading information

A.2 DATASET CONSTRUCTION

Data Collection This study collects candidate video samples from multiple publicly accessible video platforms.

Besides, we further enrich the dataset by incorporating two public research datasets fight_new_youtube(Qi et al., 2022) and MSR-VTT(Xu et al., 2016). In detail, fight_new_youtube dataset is a binary classification dataset for violent video detection, which contains 1,190 positive samples (violent fight videos) and 1,180 negative samples (normal videos). In this study, we select 500 samples from both the positive and negative sets of fight_new_youtube to supplement the violent category and normal category of the Fine-HM dataset, respectively. As to the dataset MSR-VTT (Microsoft Research Video to Text), it is a large-scale video captioning dataset released by Microsoft Research. Since its 7,180 videos cover daily and safe open-domain scenarios, we randomly select 400 videos as additional negative samples for the normal category of the dataset Fine-HM.

Data Preprocess The collected candidate videos are preprocessed through deduplication, resolution filtering, duration screening, and format unification. For categories that are difficult to obtain naturally, data augmentation strategies are adopted to supplement the dataset. Static images are converted into short video clips, and category-related text overlays are added to existing videos to generate additional samples.

Meanwhile, to support fine-grained explanatory annotation, subtitle information is generated for specific categories, such as abusive content and politically sensitive content. Based on the SmartSub tool and the whisper model, we adopt ASR technique to transcribe audio content into text. All transcribed subtitles are manually verified for accuracy and finally merged into videos via ffmpeg.

Table 6: FineHM dataset taxonomy.

Category	Definition
Graphic/gory content	Visible blood spatter, open wounds, dismemberment, exposed organs, or comparably explicit injury consequences. A complete conflict sequence is not required.
Violence	Fighting, beating, cutting, abuse, weapon attacks, coercive injury, or actions that directly or indirectly create serious bodily harm or threat. The category emphasizes the harmful act.
Nudity/pornography	Explicit sexual activity, exposed intimate body regions, sexual close-ups, or overtly sexualized imagery whose principal function is sexual stimulation.
Political non-compliance	Content operationalized in the source policy context as prohibited political insults, stigmatization of political figures or institutions, or distortion of the specified political position. This label is jurisdiction- and policy-dependent.
Abusive/conflict-provoking	Insults, humiliation, harassment, provocation, malicious attacks, or expressions clearly intended to intensify interpersonal or inter-group conflict.
Financial risk	Gambling inducement, fraudulent schemes, illegal lending, phishing, pyramid selling, off-platform transaction inducement, or other content likely to cause property or transaction loss.
Smoking/drug use	Smoking, use of prohibited substances, display of consumption paraphernalia, or clear inducement of substance use.
Normal	None of the seven harmful categories is present. Normal cannot co-occur with a harmful label.

Furthermore, to adapt to the short-video scenario and improve the accuracy of category annotation and explainable annotation, raw long videos are split into semantically complete short clips. Considering the general content distribution characteristics of mainstream short-video platforms and the average duration of harmful behaviours, all clipped video fragments are controlled within a duration of 3 seconds to 20 seconds.

Annotation The dataset has two types of annotation: video-level category annotation and frame-level explanatory annotation. We annotate the dataset based on the predefined taxonomy, which is described in Table 6.

This study completes video category annotation through four progressive steps: coarse labeling, model-assisted pre-labeling, manual verification, and data distribution evaluation. First, coarse pseudo-labels are assigned to candidate videos based on the search keywords used during the crawling phase. We then introduce CLIP to perform similarity ranking and assist in sample screening, thereby reducing the cost of manual annotation. The model outputs serve only as pre-screening references and are not directly adopted as final labels. Subsequently, human annotators check and verify the labels. After the initial annotation, we evaluate the overall data distribution. If the single-label classification accuracy on a particular category is observed to be excessively high (indicating overly simplistic samples) or excessively low (indicating overly subtle or ambiguous features), we supplement challenging negative samples or representative positive samples. All newly added data are then processed through the same annotation pipeline described above to obtain finalized video-level labels.

To enable fine-grained explainable harmful video detection, we conduct frame-level explanatory annotation by bounding harmful visual evidence on key frames. The frame-level annotation pipeline includes frame sampling, coarse labeling, model-assisted pre-labeling and manual verification. Specifically, we sample frames at fixed intervals from annotated video clips and assign video-level category labels as their preliminary frame labels. Next, we leverage Grounding DINO (Liu et al., 2024), an open-vocabulary object detection model, to generate initial bounding boxes for automatic explanatory pre-labels. Ultimately, human annotators carefully verify and revise all pre-labeling results to establish the final explanatory annotations.

A.3 Framework details

In this section, we provide more details on the optimization strategy of our proposed model. To jointly improve the model's discriminative and explanatory capabilities on explainable harmful short video detection, we formulate a unified optimization objective that comprises two loss terms: a video-level multi-label classification loss and a frame-level classification and localization loss.

The video-level branch combines asymmetric multi-label loss, MIL-Top binary cross-entropy loss, and frame-video consistency loss (Tarvainen & Valpola, 2017):

$$\mathcal{L}_{\text{cls}} = \lambda_{\text{asl}}\mathcal{L}_{\text{ASL}} + \lambda_{\text{mil}}\mathcal{L}_{\text{MIL}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}},$$

Table 7: Comparison results between different LLMs and the proposed model on Fine-HM (%). The best results are in bold.

Clue Mining Model	Classification Performance				Evidence Grounding		
	Acc.↑	AR ↑	mAP ↑	OE ↓	Pre. ↑	mAP ↑	IoU ↑
Llama4	78.7	89.8	90.9	0.16	34.9	27.7	74.5
DeepSeek-v3.2	78.3	89.8	90.7	0.17	32.8	25.0	74.2
GPT-4o-mini	77.5	89.3	90.6	0.17	41.4	26.1	74.9
Qwen3.5-27B	**79.0**	90.1	90.8	0.16	29.5	21.1	75.5
ExplainHarm	78.5	**90.3**	**91.4**	**0.15**	**48.18**	**33.8**	**76.2**

where frame-video consistency loss $\mathcal{L}_{\text{con}}$ aligns category predictions across different temporal granularities. This loss rests on the premise that, if the video-level prediction faithfully reflects the overall category semantics, the mean category probability aggregated over all valid frames should coincide with the video-level category probability.

In the frame-level branch, the Hungarian bipartite matching is first executed, and then the weighted combination of focal, L1 and GIoU losses is computed for model training. The focal loss focuses training on hard samples. L1 loss regularizes box coordinate regression. GIoU loss optimizes the spatial overlap between predicted boxes and ground-truth boxes. Therefore, the loss of the frame-level branch is:

$$\mathcal{L}_{\text{frame}} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{L1}}\mathcal{L}_{L1} + \lambda_{\text{giou}}\mathcal{L}_{\text{GIoU}}.$$

Overall, the final loss can be formulated as follows.

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{frame}}\mathcal{L}_{\text{frame}}.$$

By jointly optimizing the two branches of video-level classification and frame-level classification and localization, this overall loss allows the model to conduct accurate multi-label detection for harmful short videos as well as frame-level localization explanations for video categories.

A.4 Experiment Setup

All the experiments are conducted on the proposed dataset Fine-Harm. We adopt the AdamW optimizer for model training. The learning rate for backbone parameters is set to 3×10^{-5}, while the learning rate of task-specific modules is configured as 3×10^{-4}, with a weight decay of 0.01.

A.5 Additional Experimental Results

Prompt-chain comparison Table 7 reports the comparison between direct use of stronger language models (Llama4, Deepseek-v3.2, ChatGPT-4o-mini and Qwen3.5-27B) and the prompt-chain clue configuration built on Qwen2.5-7B. **For classification performance**, the proposed ExplainHarm model attains the strongest overall classification profile, achieving the highest scores on AR, mA and the lowest OE. Although Qwen3.5-27B achieves the highest Acc (79.0), exceeding ExplainHarm by only 0.5 points, ExplainHarm outperforms this baseline on all remaining classification metrics, suggesting that zero-shot semantic reasoning with general-purpose LLMs alone is often insufficient to capture the multi-label, semantically intertwined, and implicitly correlated characteristics of harmful videos. In contrast, the prompt-chain-based dynamic clue mining module mines latent associations among harmful categories and thereby improves the completeness and stability of multi-label discrimination. **For evidence grounding**, ExplainHarm achieves the best result on all three metrics, indicating that the proposed model not only produces relatively accurate video-level classification results, but also localizes the key visual evidence regions. While some general-purpose LLMs exhibit strong semantic understanding and generation capabilities, they remain limited in aligning visual cues with their discriminative logic, frequently producing shifted explanation evidence or mislocalized key regions.

Visualization analysis To more intuitively demonstrate the model's classification and explanation performance across different harmful categories, this study further selects several representative video samples for visualization analysis. Considering that this study focuses on the model's explainable multi-label classification capability for various types of harmful videos, this part mainly

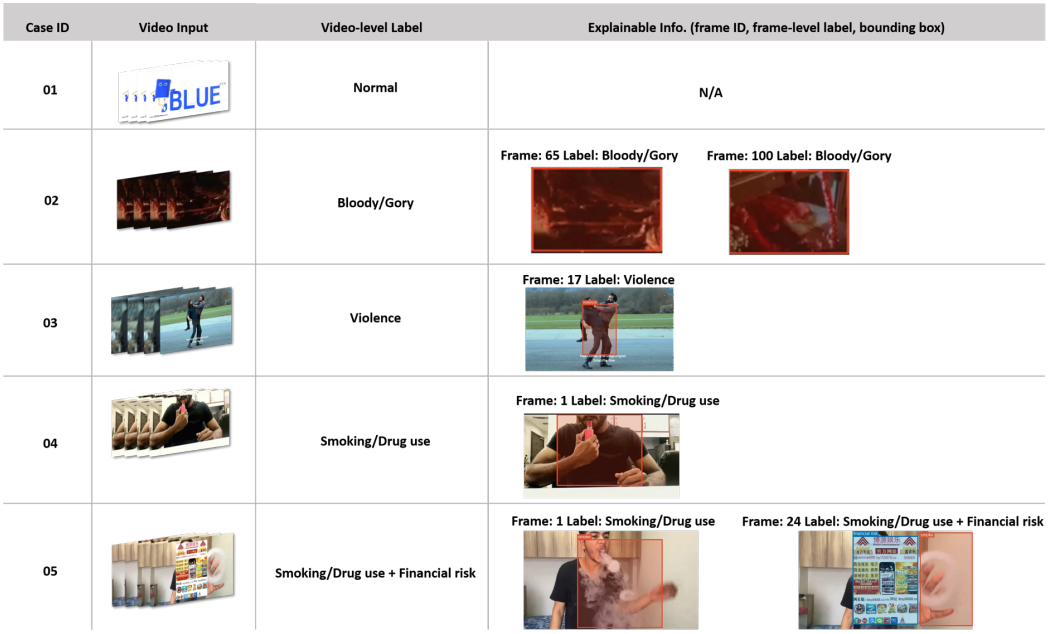


Figure 3: The visual cases of the proposed model, including video-level labels and explainable information (frame ID, frame-level label and bounding box).

presents the inference results of the proposed model on samples of different categories, including video-level labels and explainable information (frame ID, frame-level label and bounding box), as shown in Figure 3. These cases range from a normal video to multi-label co-occurrence, jointly demonstrating the model's spatial–temporal interpretability and its multi-label capability. The video of case 01 contains no harmful content, and the model accordingly produces no explanatory regions. Within the short video of case 02, two spatially and temporally separated bloody/gory scenes are highlighted independently. Rather than conflating the two distant instances, the model grounds each of them separately, demonstrating its ability to capture multiple, temporally scattered occurrences of the same category. In case 03 and case 04, the model anchors the prediction precisely to the intervals. Moreover, case 05 exhibits category co-occurrence. The video starts with smoking only, and from the 24th frame onward, the frame simultaneously involves smoking and financial risk. The model captures this dynamic transition. It retains the smoking clue while newly highlighting the financial risk clue in blue exactly at the frame where the second category emerges. This shows that the model not only predicts multiple labels but also associates each label with its own explanatory region and the moment it appears.