

# Beyond Metadata: Multimodal, Policy-Aware Detection of YouTube Scam Videos

Ummay Kulsum, Aafaq Sabir, Abhinaya S.B., Anupam Das

North Carolina State University, Raleigh, NC, USA  
 ukulsum@ncsu.edu, asabir2@ncsu.edu, asrivid@ncsu.edu, anupam.das@ncsu.edu

## Abstract

YouTube is a major platform for information and entertainment, but its wide accessibility also makes it attractive for scammers to upload deceptive or malicious content. Prior detection approaches rely largely on textual or statistical metadata, such as titles, descriptions, view counts, or likes, which are effective in many cases but can be evaded through benign-looking text, manipulated statistics, or other obfuscation strategies (e.g., ‘Leetspeak’), while ignoring visual cues. In this study, we systematically investigate multimodal approaches for detecting YouTube scams. Our dataset consolidates established scam categories and augments them with full-length videos and policy-grounded reasoning annotations. Experiments show that a text-only model using titles and descriptions (fine-tuned BERT) achieves moderate performance (76.61% F1 score), improving slightly with audio transcripts (77.98% F1 score). Visual analysis with a fine-tuned LLaVA-Video model performs better (79.61% F1 score), while a multimodal framework combining titles, descriptions, and video frames achieves the highest performance (82.96% F1 score). Moreover, the multimodal framework showed greater robustness to adversarial perturbations, with accuracy dropping only 1–3%, compared to 12–38% for modality-specific models. Beyond accuracy, the multimodal framework provides interpretable, policy-grounded reasoning, enhancing transparency and practical utility in automated moderation. Using this approach, we analyzed 6,374 in-the-wild YouTube videos and detected 1,864 scams with explicit reasoning, providing a valuable resource for future research.

**Code** — <https://github.com/ummay-kulsum18/VidScamNet>

**Dataset** — <https://tinyurl.com/VidScam>

## 1 Introduction

Video has emerged as a dominant medium for disseminating information and engaging audiences on social media (Pol-skii 2025). YouTube, the second most popular social media platform and search engine, exemplifies this trend, with over 500 hours of content uploaded every minute (Team 2025). However, its vast reach and accessibility have also made it a hotspot for deceptive and malicious content. Prior research has analyzed diverse scam types on YouTube (Bouma-Sims

and Reaves 2021; Tripathi, Ghosh, and Bharti 2022; Vak-ilinia 2022; Chu et al. 2022; Bouma-Sims et al. 2025). For example, Bouma-Sims et al. (Bouma-Sims and Reaves 2021) examined giveaway scams offering fake incentives such as gift cards or free premium access, while Tripathi et al. (Tripathi, Ghosh, and Bharti 2022) studied fraudulent apps promising quick financial rewards. These scams often redirect users to external websites or applications to complete surveys, download software, or perform tasks that lead to privacy breaches or malware installation (Bouma-Sims et al. 2025). More recently, cryptocurrency scams—such as arbitrage bot schemes exploiting flawed smart contracts—have caused substantial financial losses (Li, Guan, and Lee 2023; Vakilinia 2022). The growing diversity and sophistication of such scams underscore the need for robust and scalable detection mechanisms to safeguard users and preserve the platform’s integrity.

Although prior studies have investigated scams and offered important insights into scam typologies, existing detection approaches largely depend on unimodal cues, particularly textual metadata (e.g., titles, descriptions) or statistical metrics (e.g., view, like, and comment counts). However, as Chu et al. (Chu et al. 2022) note, these signals can be easily manipulated using fake engagement tactics, rendering them unreliable for scam detection. Moreover, scammers often evade text-based models by crafting benign-looking titles and descriptions that conceal fraudulent intent. In contrast, the visual content of videos often contains richer and more distinctive cues, such as demonstrations of malicious QR codes, fake cryptocurrency dashboards, or misleading giveaway banners, which can serve as stronger indicators of deception. This gap underscores the need for *multi-modal* analysis combining visual and textual signals to improve the robustness and accuracy of YouTube scam detection.

With the advent of multimodal large language models (MLLMs) capable of understanding and linking semantics across different modalities, it is timely to explore their potential for detecting video-based scams. Furthermore, existing detection systems primarily produce a likelihood score without grounding their decisions in explicit policy guidelines. However, MLLMs offer the opportunity to incorporate policy grounding, enabling more transparent and interpretable scam detection.

Toward developing a multimodal video-based scam detec-

tion system, we aim to address the following research questions: **RQ1: How do video scams manifest across text, audio, and visual modalities?** This question explores whether scam videos can be reliably detected using only textual metadata (e.g., titles and descriptions) or if audio and visual components provide distinctive and complementary cues. To investigate, we perform a detailed human analysis on a subset of YouTube scam videos, examining modality-specific characteristics and their relevance for automated detection.

**RQ2: How can we design a multimodal automated pipeline for video scam detection that incorporates policy-grounded reasoning for greater transparency and interpretability?**

This question highlights the need to combine multimodal understanding with policy-aware reasoning. By jointly analyzing text, audio, and visual cues, the system can capture richer evidence of deception. We fine-tune a MLLM with policy-aligned scam criteria to enable robust cross-modal detection and generate interpretable, policy-consistent explanations. **RQ3: How effective is the multimodal approach in detecting scams from in-the-wild YouTube videos?** This question assesses the real-world applicability of the proposed approach by evaluating its performance on large-scale, naturally occurring YouTube videos, measuring both detection accuracy and the quality of policy-grounded explanations.

In summary, this study makes the following contributions:

- **Comprehensive multimodal YouTube scam video dataset:** We curate a new YouTube scam video dataset, *VidScam*, covering monetary, giveaway, and cryptocurrency scams. Unlike prior datasets that include only video IDs and metadata, *VidScam* provides full video content along with policy-aligned reasoning criteria derived from YouTube’s content guidelines.
- **Systematic evaluation across modalities:** We perform extensive experiments using text, audio, and visual models to assess modality-specific performance. A text-only BERT model achieves 76.61% F1 score using titles and descriptions, while a visual-only model (LLaVA-Video-7B) reaches 79.61% F1 score on video frames—establishing strong baselines for future multimodal research.
- **Multimodal scam detector with interpretable reasoning:** We propose *VidScamNet*, a multimodal framework that fuses textual and visual features for improved detection. Beyond classification, *VidScamNet* generates interpretable, policy-grounded explanations for its decisions, achieving an F1 score of 82.96%.
- **Large-scale in-the-wild validation:** We evaluate *VidScamNet* on 6,374 wild videos collected using scam-related keywords. *VidScamNet* identifies 1,864 scam videos and produces YouTube policy-aligned reasoning for each detection, demonstrating its scalability and real-world applicability. We also report the detected scam videos to YouTube.

## 2 Related Work

**Scams Across Digital Platforms.** A large body of research has analyzed different types of scams on digital plat-

forms. Most of the work focused on studying a specific scam type such as technical support scam (Gupta et al. 2019; Larson et al. 2018; Miramirkhani, Starov, and Nikiforakis 2016; Srinivasan et al. 2018), online survey scam (Kharraz, Robertson, and Kirda 2018), game hack scam (Badawi et al. 2019), romance scam (Al-Rousan et al. 2020; Suarez-Tangil et al. 2019), Pig-Butchering scam (Oak and Shafiq 2025; Burton and Moore 2024), cryptocurrency scam (Acharya et al. 2024; Liu et al. 2024; Li, Guan, and Lee 2023; Li, Yepuri, and Nikiforakis 2023) and SMS scam (Agarwal et al. 2025; Mishra and Soni 2020). In addition to this domain-specific analysis, Kolupuri et al. (Kolupuri et al. 2025) methodologically evaluated the feasibility of Machine Learning (ML) and Deep Learning (DL) approaches for detecting and preventing these fraudulent activities. Kotzias et al. (Kotzias et al. 2025) studied user exposure, showing that end users most frequently encounter scam domains by following links shared on social media. Our work extends this literature by focusing on scams facilitated through YouTube.

**Video-based Scam/Spam.** Prior research has investigated the automated detection of targeted spam and clickbait by analyzing video titles, comments, thumbnails, and metadata (Alberto, Lochter, and Almeida 2015; Chaudhary and Sureka 2013; Zannettou et al. 2018). In a related direction, Chu et al. (Chu et al. 2022) examined various tactics used by creators to monetize content through deceptive means. Llavendhan et al. (Llavendhan, Janani et al. 2024) focused on detecting spam within YouTube comments and do not analyze the video content itself, making their work fundamentally different from ours. Several studies have explicitly focused on detecting YouTube scam videos. Bouma-Sims et al. (Bouma-Sims and Reaves 2021) conducted an exploratory analysis using organic searches and found that basic metadata, such as video age, view count, and channel size, alone were insufficient for reliably identifying scams. Tripathi et al. (Tripathi, Ghosh, and Bharti 2022) compared textual features, like titles and descriptions, with statistical metadata (e.g., view count, likes, video length, comments) and found that text-based features were more effective for detecting monetary scams. Similarly, Li et al. (Li, Guan, and Lee 2023) demonstrated that classifiers trained on textual metadata could successfully detect cryptocurrency arbitrage bot scams, highlighting the predictive value of textual information in this domain.

**Distinction from Prior Work.** Our study differs from prior efforts in three key ways. First, unlike existing approaches that rely primarily on textual or statistical metadata, we incorporate the video modality, enabling multimodal detection of scam videos on YouTube, which is more resistant to evasion techniques. Second, we focus on multiple scam categories to improve the generalizability of our detection approach, rather than limiting it to a single scam type. Finally, while these traditional models lack explainability, our approach integrates scam reasoning criteria grounded in YouTube’s official policy on “Spam, deceptive practices, and scams” (YouTube Help 2025), providing a level of explainability that is essential for effective platform governance.

### 3 Dataset

We utilize three publicly available YouTube scam video datasets from prior studies, each focusing on a distinct scam category: monetary (Bouma-Sims and Reaves 2021), giveaway (Tripathi, Ghosh, and Bharti 2022), and cryptocurrency scams (Li, Guan, and Lee 2023). These datasets were chosen for their open accessibility and coverage of diverse scam types. Each dataset provides YouTube video IDs, titles, descriptions, and associated metadata. The corresponding videos were retrieved using the Python library `yt-dlp` (yt-dlp 2025) by providing the video IDs. We name this consolidated dataset as *VidScam*.

#### Adapting Existing Datasets

Bouma-Sims et al. (Bouma-Sims and Reaves 2021) compiled a dataset of 3,700 YouTube videos related to various scams, including gift card scams, game currency scams, tech support scams, and bank support scams. The authors manually reviewed and labeled 668 videos as scams, while the remaining videos were classified as non-scams. However, we were able to download only 2,071 videos, of which 146 were labeled scams. Throughout the paper, we refer to this dataset as *GiveawayScam*.

Tripathi et al. (Tripathi, Ghosh, and Bharti 2022) focused on content promoting financial fraud, such as cash gift offers, get-rich-quick schemes, and pyramid schemes, which they categorized as monetary scam videos. Their dataset consists of 1,292 manually annotated YouTube videos, including 419 monetary scam videos, 873 non-scam videos. We successfully downloaded 1,188, including 278 labeled scams. Throughout the paper, we refer to this dataset as *MonetaryScam*.

Li et al. (Li, Guan, and Lee 2023) investigated cryptocurrency arbitrage bot scams and developed a classification model called *CryptoScamHunter*, which detects such fraudulent content on YouTube. Their model was trained using 2,000 ground-truth YouTube video titles and descriptions. However, we could not obtain the ground-truth videos, as the video IDs were unavailable. Nonetheless, we accessed a set of video IDs classified as scams by *CryptoScamHunter* and downloaded 580 of these videos. Throughout the paper, we refer to this dataset as *CryptoScam*.

**Ethical Considerations.** The datasets consist exclusively of publicly available YouTube videos and do not include user-level personal information. Any incidental personal identifiers were filtered or anonymized prior to use.

#### Annotation and Dataset Validation

The three existing YouTube scam datasets primarily focus on metadata (titles, descriptions, and identifiers) but have two key limitations: (i) they lack explicit annotations of policy violations that define scam content, and (ii) they do not attribute these violations to specific modalities or video properties such as the title, description, or visual content. To address this limitation, we re-annotated a subset of the dataset with two additional labels, i.e., scam modality and scam criteria. The modality label identifies where

scam cues appear such as text, audio, and visual and is annotated using a binary scheme, where each modality is independently marked as present or absent (allowing multiple modalities per video). In contrast, the criteria labels are grounded in YouTube’s policies on spam, deceptive practices, and scams (YouTube Help 2025). Following Bouma-Sims et al. (Bouma-Sims and Reaves 2021), we extract and consolidate scam-related criteria defined in YouTube’s content policies. We present our final set of scam criteria with representative examples in Table 1, and illustrate instances from the dataset in Figure 1. Collectively, these criteria capture the manipulative tactics designed to deceive users and redirect them into fraudulent ecosystems, while our modality annotations reveal how such tactics are operationalized across text, audio, and visual modalities.

During annotation, we observed that many “scam” videos did not result in direct financial loss. To capture this nuance, we introduced an additional labeling dimension for scam implication with three categories: (i) in-game financial or asset gain, (ii) real-world financial or material gain, and (iii) redirection to external websites or apps that may host malware or collect personal data. These labels offer finer granularity in characterizing scam tactics, highlighting how promises of financial or in-game rewards are often used to lure users into fraudulent ecosystems.

**Annotation Process.** With a fixed set of scam criteria, we analyze 200 randomly selected videos from the consolidated dataset. Among these videos, 120 were labeled as scams. For the annotation process, three researchers with 3-5 years of experience in online safety research annotated videos in the dataset in batches of 10-20. For each video, annotators applied scam criteria, scam implication, modality, and a label indicating “scam”, “non-scam”. Each video was reviewed in full, including audio, title, description, and visual content. After every batch, they convened and discussed their application of the criteria to reach a consensus on their mental models. In some cases where the legitimacy of external apps/websites mentioned in the video was unclear, annotators verified with two widely used URL checkers, NordVPN (NordVPN 2025) and VirusTotal (VirusTotal 2025), labeling the video as scam/non-scam. Moreover, for non-english videos, annotators used YouTube’s auto-captioning and auto-translation features.

**Inter-rater reliability.** To assess inter-rater reliability, we computed *Krippendorff’s alpha* (McDonald, Schoenebeck, and Forte 2019; Krippendorff 2011) after each batch, using a multi-label implementation with the Dice similarity metric (Grill 2025). *Krippendorff’s alpha* ( $\alpha$ ) is well-suited for multi-label categorical annotation with multiple raters. Iterative annotation continued until substantial agreement ( $\alpha > 0.8$ ) was achieved after nine iterations, covering 115 videos. The remaining 85 videos were then distributed among annotators for independent labeling. Detailed reliability scores in all iterations are reported in Appendix A. For quality control, the first two batches were structured by scam type (all scam, all non-scam), while later batches were randomized across classes. For non-scam videos, annotators also provided brief summaries confirming the absence of deceptive practices.

Criteria	Description	Example
Criteria 1	Claims to commit a crime on behalf of the user, regardless of whether it actually does.	Video instructing viewers to manipulate a company’s customer service to obtain free goods or sharing content that directly violates platform policies.
Criteria 2	Purports to provide an unbounded giveaway that offers unlimited free items without rules, limit or end.	Claiming to generate unlimited gift card codes or infinite in-game currency through third-party applications.
Criteria 3	Promises viewers they’ll see something but instead directs them off-site.	Video that promises a movie clip but instead links to an external streaming site.
Criteria 4	Gets clicks, views, or traffic off YouTube by promising viewers that they’ll make money fast.	Video that advertises rapid financial or in-game gains in order to redirect users to malicious applications or external websites(e.g., “Get a free \$200 Nike Gift Card”, “Make \$50 per day using Crypto arbitrage bot”)
Criteria 5	Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact.	Video that explicitly instructs viewers to follow links in the description or visit external sites that are potentially harmful.
Criteria 6	Offers cash gifts, get rich quick schemes, or pyramid schemes.	Video that promotes unrealistic promises of financial or in-game rewards, such as fraudulent mobile game hacks or earning money by playing games.
Criteria 7	Impersonates an individual, company, or organization.	Video that advertises fake customer support number posing as Amazon support.

Table 1: Scam criteria used for annotation of YouTube videos, along with representative examples.



Figure 1: Example scam videos. (a, b) illustrate scams that redirect users to external websites, promote “get rich quick” schemes, and attempt to drive clicks, views, or traffic off YouTube. (c) impersonates Amazon with a likely fraudulent contact number.

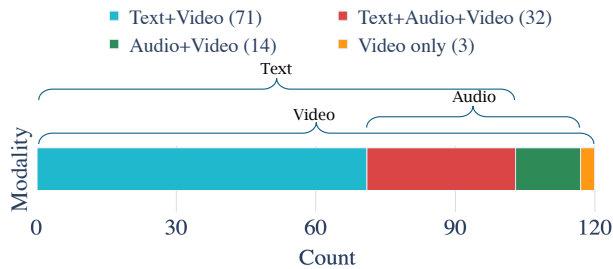


Figure 2: Scam cues across modality from annotated dataset.

Figure 2 presents the distribution of combinations of modality in the annotated scam videos. The video title and description are grouped as the text modality. The majority of the scam videos relied on multimodal cues rather than a single modality. Specifically, 32 videos contained indicators across all three modalities (text, audio, and video), while 71 combined textual metadata with visual cues (text+video), and 14 leveraged audio and video (audio+video). Only three videos relied exclusively on visual cues, and no videos were annotated as text-only, audio-only, or text+audio. This un-

derscores that multimodal information provides richer and more reliable indicators for identifying scams. Figure 3a illustrates a typical [text+video] case: the title advertises a “hack” for obtaining free in-game currency, while the video frames explicitly display a malicious link to claim the reward. By contrast, Figure 3b shows a [video-only] example, where the title and description appear benign (e.g., referencing “FrontRunning Bot BSC”), yet the video frames reveal strong scam cues, such as invitations to join private Telegram channels to purchase cryptocurrency bots. Moreover, most scam videos contained only background music or no audio at all. We also observed recurring visual patterns that serve as strong scam indicators: on-screen URLs or QR codes linking to external websites, textual instructions embedded in the video such as “Browse this website,” “Click on Generate,” or “Link in the Description,” and direct contact information for private Telegram, Discord, or WhatsApp groups. Many scam videos simulate “human verification” steps, requiring users to complete surveys, provide personal information, or install third-party applications to claim promised rewards. Some even encourage risky be-

Crit. 1	Crit. 2	Crit. 3	Crit. 4	Crit. 5	Crit. 6	Crit. 7
3	12	0	88	105	102	16

Table 2: Scam criteria distribution in annotated dataset.

Scam Criterion	Count
Financial or material gain	73
In game financial or asset gain	30
Redirect to website or app that can be malware or collect personal information	105

Table 3: Distribution of scam type in annotated dataset.

havior, such as disabling antivirus software to install fraudulent tools. Collectively, these results indicate that while textual metadata is informative, combining visual cues with text provides the most reliable indicators of scams, underscoring the value of multimodal analysis.

Table 2 summarizes the prevalence of scam criteria in the annotated datasets. The most frequently observed categories include: ‘Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact’, ‘Offers cash gifts, get rich quick schemes, or pyramid schemes’, and ‘Gets clicks, views, or traffic off YouTube by promising viewers that they’ll make money fast’. Moreover, Table 3 presents the annotation result for scam implication. Here, financial or material gain emerged as the most prominent one, appearing in 73 times, and in game financial or asset gain appears 30 times. Together, these distributions highlight scammers’ reliance on monetary and asset-based incentives as primary lures.

**Takeaway:** Majority of the scam videos leave distinctive scam cues in both the textual and visual modality.

For this study, we initially treated the labels from the existing datasets as ground truth. However, after annotation, the researchers disagreed with the ground truth labels for 13 videos, all of which had been originally labeled as scams. To further investigate potential mislabeling, annotators reviewed all scam videos in the merged dataset. As a result, 9 videos from the *MonetaryScam* dataset were reclassified as non-scams, as they discussed legitimate income-generating activities such as freelancing, affiliate marketing, and blogging. Other examples included informational videos about Google Pay or recordings of award ceremonies. Similarly, 5 videos from the *GiveawayScam* dataset were relabeled as non-scams, which include content such as discussions about why certain Clash of Clans game hacks are scams or music videos unrelated to fraudulent activities. Finally, since *CryptoScam* videos were labeled using the automated tool *CryptoScamHunter*, a substantial number of false positives were identified. 41 videos from this dataset were reclassified as non-scams, including cryptocurrency tutorials, podcasts, and unrelated general content such as cooking videos and fishing vlogs. For the final dataset, all identified videos were relabeled as non-scams to ensure higher data quality (mislabeling rate is in Appendix D).

Label	Dataset	Train	Test	Total
Scam	<i>MonetaryScam</i>	200	69	269
	<i>GiveawayScam</i>	100	41	141
	<i>CryptoScam</i>	200	339	539
	<i>VidScam</i>	500	449	949
Non Scam	<i>VidScam</i>	1000	1890	2890
All	<i>VidScam</i>	1500	2339	3839

Table 4: Train and Test Dataset Distribution

## Training and Test Sets

Using the consolidated dataset, we constructed the training and test sets for our experiments. The training set has a 1:2 ratio of scam to non-scam videos, comprising a total of 500 scam videos: 200 from *MonetaryScam*, 100 from *GiveawayScam*, and 200 from *CryptoScam*. From the combined pool of non-scam content across all three datasets, 1,000 non-scam videos were randomly sampled. The test set comprises the remaining videos and does not adhere to a fixed ratio of scams to non-scams. A detailed breakdown of the training and test sets is provided in Table 4.

## 4 Detection of Video-based Scams

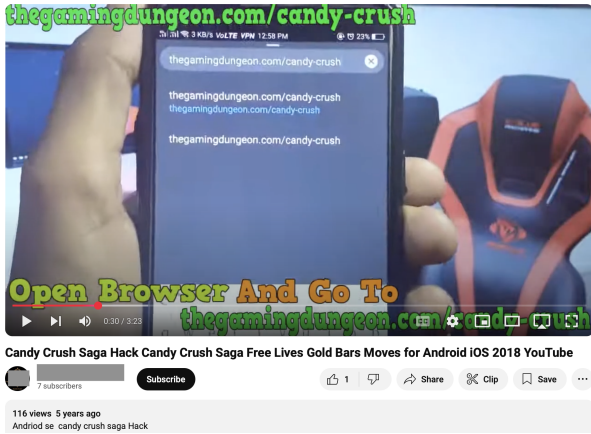
The goal of automated scam video detection on YouTube is to identify content that promotes fraudulent or deceptive activities in accordance with YouTube’s scam policies (YouTube Help 2025). To evaluate the incremental value of different information sources, we examine unimodal and multimodal models using textual, audio, and visual content, and propose a multimodal framework that integrates these signals to improve accuracy and provide interpretable, policy-aligned explanations (Figure 4). Prior work has explored lightweight baselines based on textual and statistical features. However, textual content has been shown to outperform statistical signals such as views, likes, and comments (Tripathi, Ghosh, and Bharti 2022). Moreover, these metrics are easily manipulated through fake interactions (Chu et al. 2022). Consequently, our study focuses on content-based modalities.

### Comparative Baselines

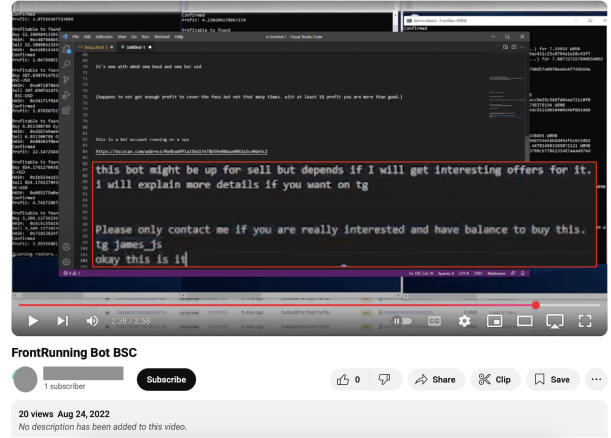
For a comprehensive evaluation of automated scam video detection, we establish comparative baseline models using different combinations of input features and model architectures.

**Modality and Input Configurations.** We explore multiple feature configurations to evaluate how different modalities contribute to scam video detection.

- **Title–Description Based Detector.** This configuration uses video titles and descriptions, previously shown to be effective for text-based detection (Li, Guan, and Lee 2023). Non-English metadata are translated into English via the Google Translation API (Google 2025), and emojis are converted into textual form using demoji (Solomon 2025).
- **Transcription Based Detector.** To assess the role of spoken content, we extract audio from each video, generate



(a)



(b)

Figure 3: Modality of scam videos. (a) combines text and visuals: a title advertising a game hack with frames showing instructions and a malicious link; (b) relies on visual deception, directing viewers to a Telegram channel for a crypto arbitrage bot.

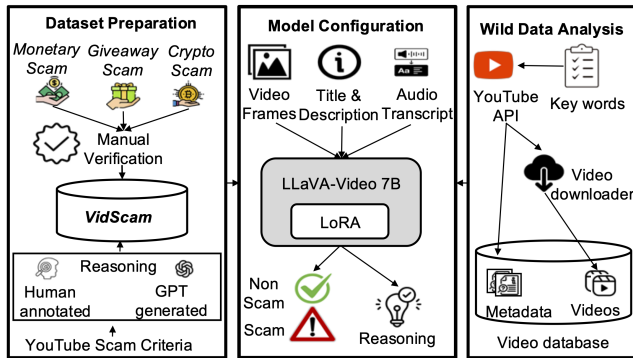


Figure 4: Overview: (i) Dataset: Merged and manually re-annotated YouTube scam datasets across three scam types. (ii) Model: Fine-tuned LLaVA-Video-7B fusing frames, titles, and descriptions for scam prediction with reasoning. (iii) Wild Data: *VidScamNet* is applied to YouTube videos retrieved via the Search API for at-scale scam detection.

English transcripts using Whisper (Radford et al. 2022), and use them as textual inputs.

- **Title-Description-Transcription Based Detector.** We fuse transcriptions, titles, and descriptions to enrich the textual input.
- **Video Based Detector.** For visual input, we uniformly sample frames from each video to evaluate the contribution of visual cues for scam video detection.
- **Video-Transcription Based Detector.** This configuration combines video frames with corresponding audio transcriptions to jointly capture visual and audio signals.
- **Video-Title-Description Based Detector.** Finally, we combine video frames with processed titles and descriptions to assess the joint effect of textual and visual features on scam detection.

**Model Considerations.** We employ different architectures tailored to each input modality. For textual inputs, we use CryptoScamHunter (Li, Guan, and Lee 2023), a state-

of-the-art text-based cryptocurrency scam detector implemented as a three-layer feedforward neural network using AllenNLP (Gardner et al. 2018). Although designed for crypto scams, its linguistic cues, such as deceptive financial claims, generalize well across scam types. We also train a three-layer feedforward network and a fine-tuned BERT model on all textual configurations, leveraging pretrained language representations for broader generalization across diverse scam types.

For configurations involving visual inputs, we focus on open-source multimodal models to ensure full access to their architectures and training parameters. Closed-source systems like GPT-4V and Gemini, though powerful, pose challenges in cost, scalability, and transparency. We first evaluate two open-source vision-language models, LLaVA-Video-7B and Qwen-2.5-VL-7B, in a zero-shot setting using sampled video frames from the full *VidScam* test set. As shown in Table 5, both models achieve low F1 scores (13–17%), with Qwen-2.5-VL-7B showing higher precision but poor recall, while LLaVA-Video-7B offers a more balanced trade-off. Additionally, GPT-4o was evaluated on a 100-sample subset of the test set, achieving a higher F1 score of 75.20%, highlighting the strong zero-shot capability of large proprietary multimodal models. However, these closed models lack fine-tuning flexibility and incur financial costs. Consequently, we select the best-performing open-source model, LLaVA-Video-7B, for fine-tuning to ensure reproducibility and scalability. As we show later, fine-tuned open-source models can surpass large proprietary alternatives.

### Multimodal Scam Detection with Reasoning

For our multimodal scam detection framework, *VidScamNet*, we consider LLaVA-Video-7B based on its zeroshot performance. Moreover, its diverse video-text pretraining and instruction-tuning further enhance its suitability for reasoning-driven tasks.

**Multimodal Input Processing.** We leverage both textual

Model	Acc.%	Prec.%	Rec.%	F1 Score%
LLaVA-Video-7B	76.78	27.4	12.69	17.35
Qwen-2.5-VL-7B	80.41	40.22	8.45	13.96
GPT-4o Vision*	69.00	73.44	77.05	75.20

Table 5: Zero shot performance of vision-language models. (\*Evaluated on a random subset consisting 100 videos)

and visual modalities for YouTube scam video detection. For text, we use video titles and descriptions, translating non-English metadata with the Google Translation API (Google 2025) for consistency. Emojis are converted to text using demoji (Solomon 2025) to preserve semantic meaning. For visuals, we uniformly sample 64 frames from each video. We adopt 64 frames because this is the default pretraining configuration of LLaVA-Video. Increasing to 100 frames yielded only a 0.2% F1-score gain in the video-only setting, while tripling runtime and increasing memory usage by 1.35 times. In the multimodal setting, 100 frames caused the combined inputs (frames + title + description) to exceed context limit, making it infeasible for many videos. In LLaVA-Video, text is tokenized with the Qwen-2 tokenizer (et al. 2024), and frames are encoded using the SigLIP vision encoder (Zhai et al. 2023). The fused text-vision representations are processed by the language model to produce the predicted scam label and reasoning.

**Supervised Finetuning with LoRA.** We formulate YouTube scam detection as a multimodal instruction-following task (Shengyu et al. 2023). Each training instance includes the video’s visual frames, temporal metadata (duration and frame timestamps), and its title and description, paired with an instruction prompt: “*Is this a scam video? Answer Yes/No and explain your reasoning. Analyze the content and refer to the official YouTube Terms and Conditions violations for scams (List of scam criteria).*” The model is trained to output both a binary prediction (Yes/No) and a policy-aligned textual rationale (Section 3).

To efficiently adapt LLaVA-Video-7B, we apply Low-Rank Adaptation (LoRA) (Hu et al. 2022), a parameter-efficient fine-tuning technique that inserts small trainable matrices into select layers while freezing the rest. This approach reduces computational and memory costs, enabling domain-specific adaptation without compromising performance. In our setup, we use a LoRA rank of 128 and an alpha of 32, adding only 55.71 MB of trainable parameters, which is about 0.69% of the model’s total size (8086.05 MB). The model is trained for up to 10 epochs with a learning rate of  $1e^{-3}$ , resulting in an efficient yet effective adaptation for scam detection.

**Reasoning Data Augmentation for Training.** To fine-tune the model for generating policy-grounded reasoning, we used 200 human-annotated samples from Section 3, splitting them evenly between training and evaluation. Fine-tuning on this small set resulted in limited performance (55.71% F1), underscoring the need for additional training data. To scale up, we employed GPT-4o Vision to generate reasoning aligned with YouTube’s policy criteria under two input settings: (i) title, description, and frames, and (ii) frames only. The latter produced reasoning more consistent

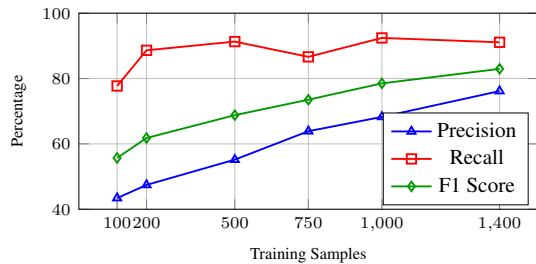


Figure 5: Effect of training size on performance.

with human annotations. The GPT-generated explanations achieved a BERTScore (Zhang et al. 2019) of 0.88 compared to 100 human-annotated validation samples (60 scam, 40 non-scam), indicating high semantic similarity. Moreover, we manually evaluated quality of the generated reasoning with the validation set. For scam videos, GPT-generated reasoning fully aligned in 53 cases and partially aligned in 7. For non-scam videos, GPT was instructed to produce brief summaries indicating legitimacy and it generated reliable summaries for all 40 cases. We then augmented *VidScam* with GPT-generated reasoning for 1,300 additional videos (prompt used is available in Appendix E). An ablation study on training data size (see Figure 5) further shows that performance steadily improves as more reasoning data are added. We use this final model configuration for all subsequent evaluations in the paper.

## Implementation Details

All models were implemented in Python. Text-only experiments (the three-layer neural network and pretrained BERT) were conducted using the AllenNLP framework (Gardner et al. 2018) on a system with an NVIDIA RTX 4090 GPU. Experiments involving the LLaVA-Video model were run on a machine with an NVIDIA A100 GPU.

## 5 Results

Our experiment results depict performance differences across the text-only, visual-only, and multimodal detection models of YouTube scam videos. Table 6 summarizes the classification performance across different input modalities and feature configurations on the *VidScam* test set. These results form the foundation for the following subsections, where we analyze the performance in detail.

### Performance Under Different Modality

**Scam Detection using Text-Only Modality.** We first evaluate textual features, video titles and descriptions, for scam detection. The state-of-the-art text-based model CryptoScamHunter (Li, Guan, and Lee 2023) achieves 92.03% F1 score on its cryptocurrency-focused dataset but drops to 46.88% F1 score on *VidScam*, which covers broader scam types (as shown in Table 7). This gap highlights the model’s limited generalizability, expected given its simple three-layer feedforward design and narrow training scope. Fine-tuning the same network on *VidScam* improves F1 score to 73.46%, though precision declines due to increased false

Training Data	Model	Accuracy%	Precision%	Recall%	F1 Score%
Title, Description	3Layer NN	88.20	64.64	85.08	73.46
	Bert	90.21	70.75	83.52	76.61
Transcription	3Layer NN	86.15	62.14	71.21	66.39
	Bert	87.99	65.38	79.51	71.76
Title, Description, Transcription	3Layer NN	90.94	74.43	80.40	77.30
	Bert	90.68	71.35	85.97	77.98
Video Frames	Llava-Video-7B	91.88	76.81	82.63	79.61
Video Frames, Transcription	Llava-Video-7B	91.11	71.33	89.76	79.49
Video Frames, Title, Description	Llava-Video-7B	91.45	71.36	92.65	80.62
Video Frames, Title, Description, Reasoning	Llava-Video-7B	<b>92.82</b>	<b>76.16</b>	<b>91.09</b>	<b>82.96</b>

Table 6: Comparison of the proposed *VidScamNet* with existing scam detection models across different modalities.

Dataset	Accuracy %	Precision %	Recall %	F1 score %
<i>CryptoScam</i>	94.36	95.98	88.40	92.03
<i>VidScam</i>	86.53	96.53	30.96	46.88

Table 7: Performance of state-of-the-art text model

positives. Fine-tuning BERT on the same dataset further boosts performance to 76.61% F1 score.

We next assess audio transcriptions for scam detection. Models using transcriptions alone perform worse, achieving 66.39% F1 with a three-layer network and 71.76% with fine-tuned BERT, reflecting that many scam videos contain only background music or non-informative audio. In our annotated subset of 200 videos, 90 contain only background music, including 74 scam videos, for which Whisper produces generic or meaningless transcripts with no scam-relevant cues. Finally, combining all three textual inputs—titles, descriptions, and transcriptions—yields the strongest results. The three-layer network reaches 77.30% F1 score (vs. 73.46% with titles+descriptions, 66.39% with transcriptions), while fine-tuned BERT achieves 77.98%, showing modest gains from integrating transcriptions.

**Scam Detection using Visual Modality.** We evaluate the effectiveness of visual information using frame-level features extracted from video content. Finetuning the LLaVA-Video-7B model on video frames alone achieves an F1 score of 79.61%, demonstrating that visual cues can reliably identify scam content. The visual-only model also attains higher precision (76.81%) than text-only models, indicating fewer false positives and highlighting the robustness of visual features for scam detection.

**Scam Detection using Multiple Modalities.** We evaluate multimodal performance by combining video frames with textual inputs (titles, descriptions, and transcriptions). Fine-tuning the multimodal models produces consistent gains across all configurations (as highlighted in Table 6). Combining frames with transcriptions achieves an F1 score of 79.49%, comparable to the frame-only model (79.61%). However, integrating frames with titles and descriptions yields the best overall result (80.62% F1 score), surpassing both the best text-only model (77.98% F1 score) and the visual-only model, with a 5.10% recall improvement. Overall, adding visual cues to textual metadata improves

Model	Acc. %	Prec. %	Rec. %	F1 score %	Acc. drop %
SVC	84.35	56.02	85.96	67.83	12.00
Adaboost	87.98	66.34	75.94	70.82	22.00
RF	90.59	81.36	66.14	72.97	38.00
<i>VidScamNet</i>	92.82	76.16	91.09	82.96	1.00

Table 8: Comparison with Existing ML approaches

the F1 score by 4.01%. These findings indicate that textual features from titles and descriptions complement visual signals, enhancing YouTube scam video detection. Moreover, unimodal models are more vulnerable to adversarial perturbations (Appendix F), reinforcing the robustness of multimodal learning.

### Policy Grounded Multimodal Detector

The proposed multimodal scam detection framework, *VidScamNet*, achieves the best overall performance, reaching an F1 score of 82.96% after instruction tuning with video frames, titles, descriptions, and policy-grounded reasoning data. Incorporating policy-grounded reasoning yields a 2.34% improvement over the non-reasoning multimodal baseline (80.62% F1 score), indicating that explicit reasoning helps the model better capture nuanced scam indicators. This gain comes with minimal overhead, increasing tokens by only 0.85%. In addition, the generated explanations closely match human-annotated reasoning, achieving a BERTScore of 0.89 on 100 benchmark samples (Section 4). Beyond accuracy gains, policy-aligned reasoning enhances transparency and consistency with YouTube’s scam criteria.

**Takeaway:** Multimodal models outperform unimodal ones in YouTube scam detection, with text providing strong cues, visuals improving precision, and policy-grounded reasoning enhancing both performance and interpretability.

### Comparison with Existing ML Approaches

Prior work (Tripathi, Ghosh, and Bharti 2022) explored lightweight models, including SVC, AdaBoost, and Random Forest, for detecting scam videos using titles and descriptions. As shown in Table 8, these methods achieve moderate performance (F1 score up to 72%) but have limited robustness to adversarial manipulation.

Crit. 1	Crit. 2	Crit. 3	Crit. 4	Crit. 5	Crit. 6	Crit. 7
7	257	237	1361	1447	1711	65

Table 9: Distribution of scam criteria in the wild dataset

To evaluate realistic evasion strategies, we test textual attacks including leet-style obfuscations (e.g., replacing characters with visually similar symbols such as “Fr33 G!ftC@rd”) and semantic perturbations generated using TextAttack (Morris et al. 2020), modeling both character-level and meaning-preserving manipulations commonly observed in scam content on a sample of 100 scam videos from the test set. Under semantic perturbations, prior lightweight baselines show accuracy drops of 12–38% (Table 8). Similarly, the text-only BERT model is highly susceptible to manipulation, suffering a 33% accuracy decrease under leet transformations and a 15% drop under semantic edits. A representative example (Appendix F) illustrates how text obfuscation causes text-only models to misclassify scams that remain visually identifiable through cues such as fake code generators and verification prompts.

We further examine robustness to visual perturbations by applying Gaussian noise, rotation, and blur to video frames using standard image transformation tools (Lee 2024). The visual-only LLaVA model shows relative robustness to Gaussian noise and rotation, with accuracy declines of only 3% and 5%, respectively, but degrades substantially under blur (30%), indicating reliance on fine-grained visual details. In contrast, the proposed multimodal model exhibits substantially improved resilience, with textual perturbations reducing accuracy by only 1% and visual transformations degrading performance by at most 3%. Overall, these findings demonstrate that while lightweight text-based baselines offer computational efficiency, multimodal fusion is essential for robust scam detection under adversarial conditions.

### Computational Cost Analysis

To assess computational efficiency, we measure inference time and GPU memory usage across model configurations (details available in Appendix C). The three-layer neural network is most efficient, using only 0.10 GB of memory and 1 ms per sample with all textual inputs. In comparison, BERT requires 1 GB and 0.13 s per sample, reflecting the added cost of transformer-based models. Multimodal models introduce significantly higher overhead. LLaVA-Video, when processing only video frames, consumes 55.6 GB of memory and 2 s per sample. Adding modalities further increases usage to 66.8 GB and 2.1 s for frames with transcriptions, and 57.1 GB and 2.0 s for frames with titles and descriptions. Overall, these results underscore the trade-off between accuracy and computational cost in multimodal detection systems. However, for platforms like YouTube, such computational resources are quite feasible.

## 6 Scam Detection in Wild YouTube Videos

We further evaluate our YouTube scam video detection framework on real-world (“in-the-wild”) YouTube videos. To construct this dataset, we curated a fixed set of keywords to increase the likelihood of retrieving scam-related

Label	Correctly classified (Reasoning Quality)			Misclassified
	Fully aligned	Partially aligned	Incorrect	
Scam	33	13	1	3
Non scam	43	3	0	4
Total	76	16	1	7

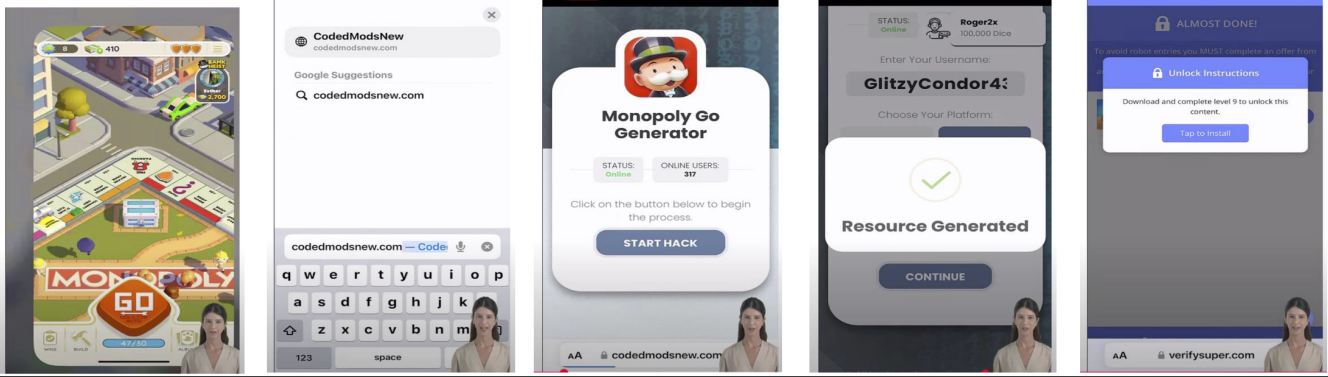
Table 10: Manual evaluation of generated reasoning (n=100)

content. These keywords were drawn from prior studies: Bouma-Sims et al. (Bouma-Sims and Reaves 2021) identified query patterns targeting giveaway scams, Tripathi et al. (Tripathi, Ghosh, and Bharti 2022) used Google Trends to extract monetary scam-related terms, and Li et al. (Li, Guan, and Lee 2023) compiled keywords linked to cryptocurrency scams. In total, we collected 70 keywords (listed in Appendix B: Table 12). Using these, we queried the YouTube Data API (YouTube 2025) to retrieve video IDs and metadata (titles and descriptions) and used the `yt-dlp` (yt-dlp 2025) library to download the videos. Due to API and availability constraints, some videos could not be retrieved. Running the pipeline daily over three months produced a final dataset of 6,374 videos.

We then deploy the best-performing multimodal model, *VidScamNet* (integrating video frames, titles, and descriptions), to analyze the wild dataset. Out of 6,374 YouTube videos, *VidScamNet* flagged 1,864 as scams and produced policy-grounded rationales supporting each classification. Notably, **316 of the videos flagged as scams by VidScamNet have since been removed from YouTube**, underscoring both the practical relevance of our detection framework and its potential alignment with platform moderation outcomes.

Table 9 shows the distribution of scam criteria in the detected videos. The most common scams involved cash gifts and get-rich-quick schemes (1,711), followed by redirections to malicious sites (1,447) and fast-money promises (1,361). Less frequent but still notable were unbounded giveaways (257), off-site redirections via misleading content (237), and impersonation of individuals or organizations (65). Extremely rare cases included videos claiming to commit crimes on behalf of users (7). These findings demonstrate that multimodal detection effectively uncovers a wide range of scam tactics, with financial deception and malicious redirection dominating real-world occurrences.

To further evaluate detection performance and reasoning quality, we manually inspected 100 randomly selected videos (50 scams and 50 non-scams). The model misclassified 7 videos, achieving an overall accuracy of 93%, with 94% precision and 92.16% recall. The reasoning outputs for correctly classified videos were rated on a three-level scale. ‘Fully Aligned’: accurately captures the key scam criteria in line with human judgment; ‘Partially Aligned’: identifies some relevant criteria but misses some or includes unrelated criteria; ‘Incorrect’: fails to generate the any relevant scam criteria. As shown in Table 10, 72 had fully aligned reasoning, 16 were partially aligned, and 1 had no generated reasoning (labeled Incorrect). These results indicate that when the model predicts correctly, its reasoning is generally reli-



**Title:** Monopoly Go Free Rolls 2025 - How I Get Unlimited Free Dice Rolls in Monopoly Go iOS Android!!  
**Description:** Monopoly Go Free Rolls 2025 - How I Get Unlimited Free Dice Rolls in Monopoly Go iOS Android!!

**Yes, this is a scam video.**  
**Reasoning:** "Gets clicks, views, or traffic off YouTube by promising viewers that they'll make money fast", "Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact", "Offers cash gifts, 'get rich quick' schemes, or pyramid schemes"

Figure 6: Qualitative example of a detected scam video. The textual metadata attracts users by promising unlimited in-game rewards. The video frames show a potentially malicious website offering in-game resource generation that requires users to complete a survey. The proposed multimodal framework correctly identifies this content as a scam, citing exaggerated free-reward claims, redirection to external sites that may distribute malware or collect sensitive user information, and deceptive tactics to drive traffic.

able and closely matches human annotations. Representative videos with generated reasoning are provided in Figure 6 and in Appendix G.

**Takeaway:** Our multimodal framework detected 1,864 real-world scam videos with policy-aligned reasoning; financial deception and malicious redirection were most common, and 316 videos were removed subsequently.

## 7 Discussion

Our study shows that multimodal modeling, combining video, text, and policy-grounded reasoning, outperforms traditional metadata-based approaches for detecting YouTube scams. In addition to higher accuracy, it provides policy-aligned explanations, enhancing transparency and explainability in automated moderation.

**Integration within Existing Detection Pipelines.** The average inference time of our model is 2 seconds and memory usage is 57.1 GB on an A100 GPU. Several standard optimization techniques can further reduce deployment overhead. For example, frame preprocessing lowers average inference time by approximately 1 second, while FP8 quantization reduces memory usage by about 50%. Additional methods such as knowledge distillation, batch processing, and more aggressive quantization can be applied for deployment to further optimize the cost. Moreover, the text based detectors can still serve as inexpensive, high-recall filters in a cascaded deployment setting. In practice, lightweight text models can be used solely to route videos rather than make final decisions, allowing the system to eliminate a large frac-

tion of benign content before invoking more expensive multimodal models.

**Limitations and Future Work.** Despite these contributions, our work has a few limitations. First, although our dataset is large and diverse, it cannot fully capture the evolving landscape of scams, relying on videos available from prior literature. Second, we analyze uniformly sampled frames rather than full video sequences, potentially missing temporal cues that distinguish scams from legitimate content. Third, some reasoning data is GPT-augmented, which may not fully capture nuanced policy interpretations. Fourth, while open-source models like LLaVA-Video enable cost-effective fine-tuning, commercial models may achieve higher performance, so our results primarily establish an improved baseline. Finally, our focus is limited to YouTube, and findings may not generalize to other platforms, though the framework is adaptable.

Future work could extend this framework to cross-platform settings by adapting the criteria to platform specific policies and collecting datasets. Additionally, incorporating temporal modeling of full video streams to capture time-dependent scam tactics, develop automated frame selection or temporal grounding for improved efficiency and accuracy, and conduct cross-platform studies to uncover broader scam strategies and inform coordinated defenses.

**Ethical Considerations.** Our study uses publicly available YouTube videos and is exempt from IRB review. We also adhered to YouTube API rate limits during data collection to minimize network impact, and reported identified scam videos to YouTube for further review.

## 8 Conclusion

In this work, we conducted a systematic study of scam detection on YouTube through a multimodal lens. Unlike prior approaches that primarily rely on textual or statistical metadata, our approach leverages a combination of video frames, title, description, and policy-grounded reasoning to effectively identify scam videos on YouTube. We constructed a multimodal dataset of YouTube scam content with associated reasoning criteria and demonstrated the effectiveness of our approach. Applying our best-performing model, *Vid-ScamNet*, to a large-scale wild dataset, we uncovered thousands of scam videos, highlighting both the prevalence of scams on the platform and the practical potential of multimodal, policy-aligned detection frameworks.

## Acknowledgements

We thank the anonymous reviewers for their valuable feedback, and Daniel Nolting and Dr. Bradley Reaves (from NC State University) for the YouTube crawler used in our in-the-wild data collection. The opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the participating organizations.

## References

- Acharya, B.; Saad, M.; Cinà, A. E.; Schönherr, L.; Dai Nguyen, H.; Oest, A.; Vadrevu, P.; and Holz, T. 2024. Conning the crypto conman: End-to-end analysis of cryptocurrency-based technical support scams. In *2024 IEEE Symposium on Security and Privacy (SP)*, 17–35.
- Agarwal, S.; Harvey, E.; Mariconti, E.; Suarez-Tangil, G.; Vasek, M.; et al. 2025. ‘Hey mum, I dropped my phone down the toilet’: Investigating Hi Mum and Dad SMS Scams in the United Kingdom. In *Usenix Security Symposium*.
- Al-Rousan, S.; Abuhussein, A.; Alsubaei, F.; Kahveci, O.; Farra, H.; and Shiva, S. 2020. Social-guard: Detecting scammers in online dating. In *2020 IEEE International Conference on Electro Information Technology (EIT)*, 416–422.
- Alberto, T. C.; Lochter, J. V.; and Almeida, T. A. 2015. Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, 138–143.
- Badawi, E.; Jourdan, G.-V.; Bochmann, G.; Onut, I.-V.; and Flood, J. 2019. The “game hack” scam. In *International Conference on Web Engineering*, 280–295. Springer.
- Bouma-Sims, E.; and Reaves, B. 2021. A first look at scams on youtube. *arXiv preprint arXiv:2104.06515*.
- Bouma-Sims, E. R.; Klucinec, L.; Lanyon, M.; Downs, J.; and Cranor, L. F. 2025. The Kids Are All Right: Investigating the Susceptibility of Teens and Adults to YouTube Giveaway Scams. In *NDSS*.
- Burton, S. L.; and Moore, P. D. 2024. Pig butchering in cybersecurity: A modern social engineering threat. *SocioEconomic Challenges*, 8(3): 46.
- Chaudhary, V.; and Sureka, A. 2013. Contextual feature based one-class classifier approach for detecting video response spam on youtube. In *2013 Eleventh Annual Conference on Privacy, Security and Trust*, 195–204.
- Chu, A.; Arunasalam, A.; Ozmen, M. O.; and Celik, Z. B. 2022. Behind the tube: Exploitative monetization of content on YouTube. In *31st USENIX Security Symposium (USENIX Security 22)*, 2171–2188.
- et al., A. Y. 2024. Qwen2 Technical Report. arXiv:2407.10671.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:1803.07640.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Google. 2025. Cloud Translation API. <https://cloud.google.com/translate/docs/reference/rest>. Accessed: January 2026.
- Grill, T. 2025. Mtreviso/Krippendorff-Alpha: Python Implementation of Krippendorff’s Alpha That Supports Multilabel Data – Inter-Rater Reliability. <https://github.com/mtreviso/krippendorff-alpha>.
- Gupta, S.; Bhatia, G. S.; Suri, S.; Kuchhal, D.; Gupta, P.; Ahamad, M.; Gupta, M.; and Kumaraguru, P. 2019. Angel or Demon? Characterizing Variations Across Twitter Timeline of Technical Support Campaigners. *The Journal of Web Science*, 6.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Ilavendhan, A.; Janani, N.; et al. 2024. Optimizing YouTube Spam Detection with Ensemble Deep Learning Techniques. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 625–630.
- Kharraz, A.; Robertson, W.; and Kirde, E. 2018. Survey-lance: Automatically detecting online survey scams. In *2018 IEEE Symposium on Security and Privacy (SP)*, 70–86.
- Kolupuri, S. V. J.; Paul, A.; Bhowmick, R. S.; and Ganguli, I. 2025. Scams and frauds in the digital age: ML-based detection and prevention strategies. In *Proceedings of the 26th International Conference on Distributed Computing and Networking*, 340–345.
- Kotzias, P.; Pachilakis, M.; Aldana-Iuit, J.; Caballero, J.; Sánchez-Rola, I.; and Bilge, L. 2025. Ctrl+ Alt+ Deceive: Quantifying User Exposure to Online Scams. In *NDSS*.
- Krippendorff, K. 2011. Computing Krippendorff’s alpha-reliability.
- Larson, J.; Tower, B.; Hadfield, D.; Edge, D.; and White, C. 2018. Using web-scale graph analytics to counter technical support scams. In *2018 IEEE International Conference on Big Data (Big Data)*, 3968–3971.
- Lee, N. 2024. Blur-Generator. <https://github.com/NatLee/Blur-Generator>. Accessed: January 2026.
- Li, K.; Guan, S.; and Lee, D. 2023. Towards understanding and characterizing the arbitrage bot scam in the wild. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(3): 1–29.

- Li, X.; Yepuri, A.; and Nikiforakis, N. 2023. Double and nothing: Understanding and detecting cryptocurrency giveaway scams. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.
- Liu, E.; Kappos, G.; Mugnier, E.; Invernizzi, L.; Savage, S.; Tao, D.; Thomas, K.; Voelker, G. M.; and Meiklejohn, S. 2024. Give and take: An end-to-end investigation of giveaway scam conversion rates. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, 704–712.
- McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–23.
- Miramirkhani, N.; Starov, O.; and Nikiforakis, N. 2016. Dial one for scam: Analyzing and detecting technical support scams. In *22nd Annual Network and Distributed System Security Symposium (NDSS)*, volume 16.
- Mishra, S.; and Soni, D. 2020. Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis. *Future Generation Computer Systems*, 108: 803–815.
- Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.
- NordVPN. 2025. Is this link safe? <https://nordvpn.com/link-checker>. Accessed: January 2026.
- Oak, R.; and Shafiq, Z. 2025. "Hello, is this Anna?": Unpacking the Lifecycle of Pig-Butchering Scams. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, 1–18.
- Polskii, M. 2025. The Emergence of Video-Centric Social Networks: Trends and Impacts. <https://inappstory.com/blog/video-centric-social-media>. Accessed: July 2025.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv:2212.04356*.
- Shengyu, Z.; Linfeng, D.; Xiaoya, L.; Sen, Z.; Xiaofei, S.; Shuhe, W.; Jiwei, L.; Hu, R.; Tianwei, Z.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Solomon, B. 2025. Accurately find or remove emojis from a blob of text using data. <https://pypi.org/project/demoji>. Accessed: January 2026.
- Srinivasan, B.; Kountouras, A.; Miramirkhani, N.; Alam, M.; Nikiforakis, N.; Antonakakis, M.; and Ahamad, M. 2018. Exposing search and advertisement abuse tactics and infrastructure of technical support scammers. In *Proceedings of the 2018 World Wide Web Conference*, 319–328.
- Suarez-Tangil, G.; Edwards, M.; Peersman, C.; Stringhini, G.; Rashid, A.; and Whitty, M. 2019. Automatically dismantling online dating fraud. *IEEE Transactions on Information Forensics and Security*, 15: 1128–1137.
- Team, G. R. 2025. YOUTUBE statistics 2026 (Demographics, users by country and more). <https://www.globalmediainsight.com/blog/youtube-users-statistics/>. Accessed: July 2025.
- Tripathi, A.; Ghosh, M.; and Bharti, K. 2022. Analyzing the uncharted territory of monetizing scam Videos on YouTube. *Social Network Analysis and Mining*, 12(1): 119.
- Vakilinia, I. 2022. Cryptocurrency giveaway scam with youtube live stream. In *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0195–0200.
- VirusTotal. 2025. Analyse suspicious files, domains, IPs and URLs to detect malware and other breaches, automatically share them with the security community. <https://www.virustotal.com/gui/home/url>. Accessed: January 2026.
- YouTube. 2025. YouTube search API. <https://developers.google.com/youtube/v3/docs/search/list>. Accessed: January 2026.
- YouTube Help. 2025. YouTube Policies on Spam, deceptive practices, and scams policies. <https://support.google.com/youtube/answer/2801973?hl=en>. Accessed: July 2025.
- yt-dlp. 2025. yt-dlp. <https://github.com/yt-dlp/yt-dlp>. Accessed: January 2026.
- Zannettou, S.; Chatzis, S.; Papadamou, K.; and Sirivianos, M. 2018. The good, the bad and the bait: Detecting and characterizing clickbait on youtube. In *2018 IEEE Security and Privacy Workshops (SPW)*, 63–69.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

## Paper Checklist

### 1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, to stop violating privacy norms and unfair profiling, the videos have been presented anonymously. Any mention of socio-economic divide or implication of disrespect to societies or cultures is out of scope.**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes, claims made in the abstract and introduction are a reflection of the methods and results.**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, the appropriateness of the proposed approach is mentioned in our 'Detection of Video-based Scams' section.**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, we clarify it in the 'Dataset' section.**
- (e) Did you describe the limitations of your work? **Yes, in the Discussion section.**
- (f) Did you discuss any potential negative societal impacts of your work? **No, there are no specific negative societal impacts due to our research. In fact, we help increase user/platform awareness towards security concerns stemming from scam videos. Also, to protect specific users or organizations publishing the videos, we anonymized any named entities in video metadata.**
- (g) Did you discuss any potential misuse of your work? **No, because our dataset is already publicly available on YouTube. We will also anonymize the dataset before publishing it to avoid any misuse.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, outlined when describing the dataset collection methodology, we refrain from collecting any personally identifiable information of users who submit videos. Also, we comply with YouTube's terms when crawling to collect the videos. We also ethically disclosed our finding to YouTube. We acknowledge that we will anonymize the dataset before publishing it.**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, we acknowledge that we read the guidelines and the paper conforms to them.**

### 2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes, in the results section.**

- (b) Have you provided justifications for all theoretical results? **Yes, we shared interpretations for the analytical analysis in the paper.**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes, where possible, we complemented our results with findings from the literature.**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **No, because our analytical comparison with other existing approaches showcases clear performance benefit across different datasets.**
- (e) Did you address potential biases or limitations in your theoretical framework? **Yes, because during data annotation three coders were involved to reduce bias. Also, the samples size for testing the models were statistically large.**
- (f) Have you related your theoretical results to the existing literature in social science? **Yes, we did this where possible. For instance, we build on data from existing works by Bouma-Sims et al., Tripathi et al., and Li et al. Also, we compare our model with other existing models.**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, in the discussion section.**

### 3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? **NA**
- (b) Did you include complete proofs of all theoretical results? **NA**

### 4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, for all the machine learning experiments, we list the model choice, prompt text, and the chosen model parameters for reproducibility. However, we haven't published our crawled dataset used for model finetuning as well the finetuned model itself. We plan to anonymize our dataset before publishing it and make our model open-source.**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes, in the 'Detection of Video-based Scams' section.**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **No, because ours is a large dataset of thousands of video. Although we did multiple experiments to see how model performance varied with training set size.**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes, in the 'Computational Cost Analysis' subsection.**
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes, in the results section.**

- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **Yes. We conduct manual analyses of both false positives and false negatives to assess model performance and limitations. We explicitly consider the cost of misclassifications—such as missed scams or benign videos incorrectly labeled as scams—and frame our system as an automated filtering step that assigns confidence scores to guide and prioritize human review.**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **Yes, we cited the assets.**
- (b) Did you mention the license of the assets? **Yes, the assets used in this study are publicly available.**
- (c) Did you include any new assets in the supplemental material or as a URL? **Yes, we include the algorithms used, the prompts used, and the model parameters used.**
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes, according to their paper, the data has been made public (or explicitly requested for reuse with permission).**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes, The data we collected or used is publicly available on YouTube. Most scam videos contain malicious contact information of bad actors.**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see (FORCE11 2020))? **Yes, The data we collected or used is publicly available on YouTube. Future usage should conform to YouTube’s terms and services.**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see (Gebu et al. 2021))? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
- (d) Did you discuss how data is stored, shared, and de-identified? **NA**

## Appendix

### A Krippendorf’s alpha scores

Training Session	Number of Sample	Column Name				Modality
		Agree with Ground Truth?	Label (Scam, Non scam)	Broad Scam Crit.	Narrow Scam Crit.	
1	10	0.54	0.70	0.40	0.68	0.02
2	10	1.00	1.00	1.00	-	-
3	19	-0.02	0.73	0.39	-0.11	0.08
4	10	0.00	0.85	0.29	0.52	0.31
5	10	0.29	0.87	0.44	0.44	0.31
6	16	0.78	0.90	0.67	0.52	0.33
7	10	0.62	0.80	0.77	0.89	0.89
8	15	0.78	0.91	0.83	1.00	0.57
9	15	0.73	0.91	0.83	0.72	0.70

Table 11: Krippendorf’s alpha scores after each iteration

### B Keywords list for Wild Data Collection

Category	Keywords
Giveaway Scam	Free Amazon.com eGift Card, Free Visa Gift Card, Free DoorDash eGift Card, Free Sephora Gift Card, Free Razer Gold eGift Card, Free Uber eGift Card, Free Starbucks eGift Card, Free Mastercard Gift Card, Free Starbucks Gift Card, Free Ulta Beauty Gift Card, Free Netflix eGift Card, Free Visa Virtual eGift Card, Free Spotify Premium eGift Card, Free Apple Gift Card, free honor of kings currency, free Last war: survival currency, free whiteout survival currency, free royal match currency, free pubg mobile currency, free monopoly go currency, free candy crush saga currency, free pokemon tcg pocket currency, free coin master currency, free roblox currency, apple tech support, microsoft tech support, nvidia tech support, samsung tech support, sony tech support, intel tech support, dell tech support, panasonic tech support, UBS support, Morgan Stanley support, Bank of America support, J.P. Morgan, Private Bank support, Citigroup support, BNP Paribas support, Goldman Sachs support, Julius Baer support, Raymond James support, HSBC support
Monetary Scam	How to earn money online, ways to earn money online, earn money online fast, best way to earn money online, how to make money as kid, how to make money online with ai, how to make money online as teen
Cryptocurrency Scam	passive income, huge profit, easy profit, building wealth with crypto, earn free bnb, earn free eth, UniSwap, SushiSwap, PancakeSwap, Aave, Avalanche/Avax, Polygon/Matic, Fantom/FTM, arbitrage bot, front-running bot, flashloan bot, MEV bot, snipe Bot, trading bot, DeFi bot

Table 12: Keywords used for wild data collection

### C Computational Cost

Model	Input Feature	Inference Time (sec/sample)	Memory (GB)
3 layer NN	Title, Description	0.0004	0.08
Bert	Title, Description	0.09	0.99
3 layer NN	Transcription	0.001	0.09
Bert	Transcription	0.10	1.00
3 layer NN	Title, Description, Transcription	0.001	0.10
Bert	Title, Description, Transcription	0.13	1.00
LLaVA-Video	Video Frames	2.00	55.58
LLaVA-Video	Video Frames, Transcription	2.09	66.81
LLaVA-Video	Video Frames, Title, Description	2.05	57.12

Table 13: Memory usage and average inference time for each model.

## D Quality of the Datasets

Dataset	# Scam in Ground Truth	# Relabeled Non-scam	% Mislabeled
<i>MonetaryScam</i>	278	9	3.24
<i>GiveawayScam</i>	146	5	3.42
<i>CryptoScam</i>	580	41	7.07

Table 14: Quality of ground truth labels in each dataset.

## E GPT Assisted Reasoning Generation Prompt

Description	Prompt Contents
Task	The provided frames are extracted from a YouTube video. This video has been identified as a scam. Your task is to analyze the content and determine the most relevant reasons for this classification, using the following official YouTube Terms and Conditions violations.
Rules	<ol style="list-style-type: none"> <li>1. Claims to commit a crime on behalf of the user, regardless of whether it actually does.</li> <li>2. Purports to provide an “unbounded” giveaway that offers unlimited free items without rules, limit or end.</li> <li>3. Promises viewers they’ll see something but instead directs them off-site.</li> <li>4. Gets clicks, views, or traffic off YouTube by promising viewers that they’ll make money fast.</li> <li>5. Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact.</li> <li>6. Offers cash gifts, get rich quick schemes, or pyramid schemes.</li> <li>7. Impersonates an individual, company, or organization.</li> </ol>
Instructions	<ol style="list-style-type: none"> <li>1. Carefully examine the frames by paying close attention to all details in the content.</li> <li>2. Select one or more reasons from the list above that justify why this video qualifies as a scam.</li> <li>3. For each selected reason, provide a brief explanation grounded in the content.</li> </ol>
Output Format	[Reason(s) in full text]: Brief explanation for each selected reason.

Table 15: Prompt used for GPT assisted reasoning generation for scam videos.

Description	Prompt Contents
Task	The provided frames are extracted from a YouTube video. This video has been identified as a non-scam. Your task is to analyze the content and briefly explain why this video appears legitimate. Consider the following official YouTube Terms and Conditions violations as indicators of scams.
Rules	<ol style="list-style-type: none"> <li>1. Claims to commit a crime on behalf of the user, regardless of whether it actually does.</li> <li>2. Purports to provide an “unbounded” giveaway that offers unlimited free items without rules, limit, or end.</li> <li>3. Promises viewers they’ll see something but instead directs them off-site.</li> <li>4. Gets clicks, views, or traffic off YouTube by promising fast money.</li> <li>5. Sends audiences to sites that spread malware, gather personal information, or cause harm.</li> <li>6. Offers cash gifts, get-rich-quick schemes, or pyramid schemes.</li> <li>7. Impersonates an individual, company, or organization.</li> </ol>
Instructions	<ol style="list-style-type: none"> <li>1. Carefully examine the frames by paying close attention to all details in the content.</li> <li>2. Provide a brief summary of what the video is about, emphasizing its legitimate nature.</li> <li>3. Avoid listing all seven points; instead, summarize why the video is a non-scam.</li> </ol>
Output Format	A short explanation highlighting the legitimate purpose of the video, without enumerating every violation.

Table 16: Prompt used for GPT-assisted reasoning generation for non-scam videos.

## F Textual obfuscation Example

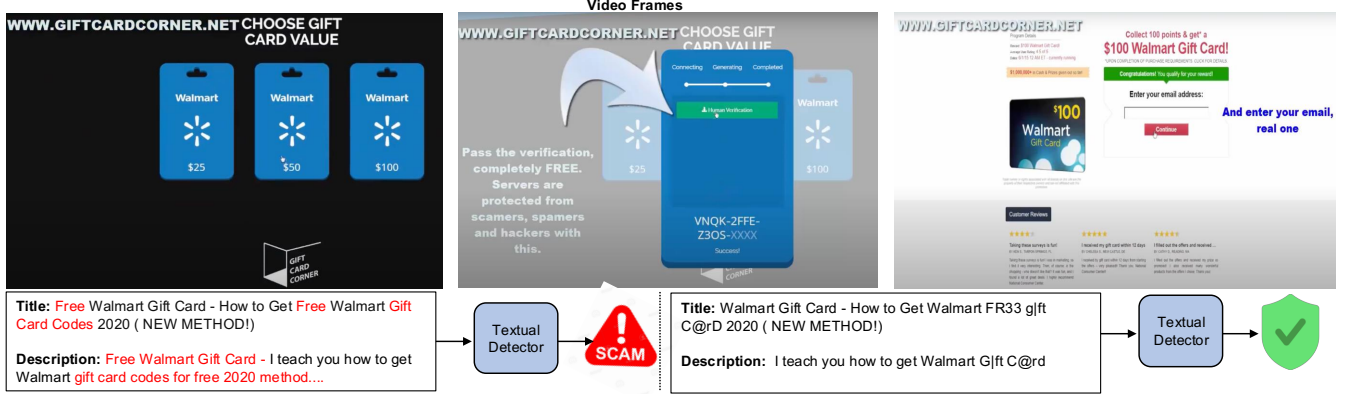


Figure 7: Effect of textual obfuscation: The fine-tuned BERT model accurately detects the original scam video title and description (left). However, simple keyword obfuscations, highlighted as red, evades the text based detector (right). At the top, representative video frames illustrate visual scam signals, such as gift card code generation and personal information requests, which remain undetected by the text model. Notably the video contains irrelevant background audio.<sup>1</sup>

## G Qualitative Results

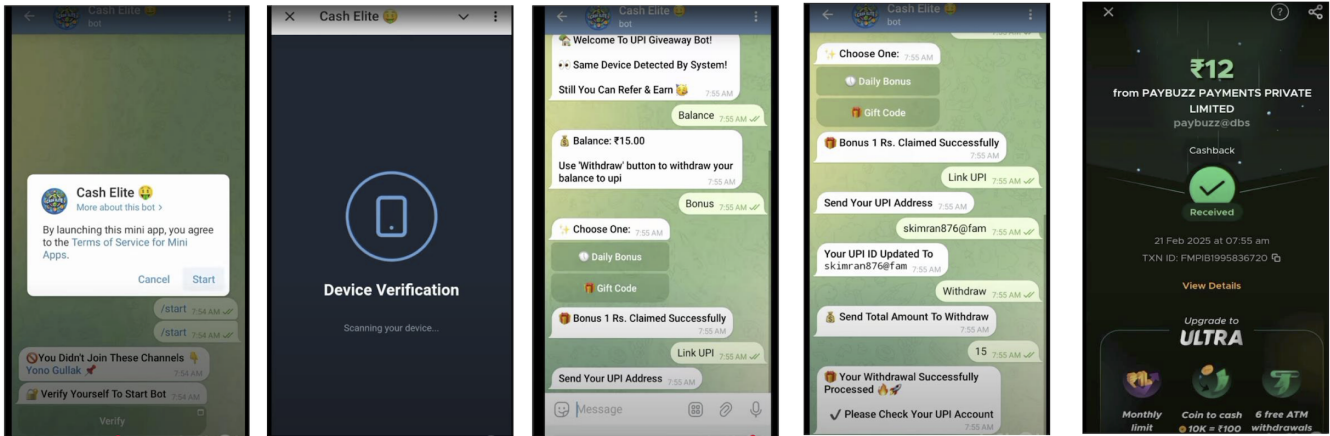
**Title:** [2025 Release] Argon Executor for Roblox | Download Free & Functional | Byfron Bypass Tool

**Description:** Downl0ad: [http://youtube.com/post/UgkxYL39J6XFru0Cc\\_JRxeKecYlyRAkopyz?si=57pkAgYJ55V4C9PV](http://youtube.com/post/UgkxYL39J6XFru0Cc_JRxeKecYlyRAkopyz?si=57pkAgYJ55V4C9PV) Passw0rd: 3030

**Reasoning:** "Claims to commit a crime on behalf of the user, regardless of whether it actually does", "Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact", "Gets clicks, views, or traffic off YouTube by promising viewers that they'll make money fast"

Figure 8: Qualitative example of a detected scam video. The textual metadata directs users to an external download link with a password, while video frames instruct disabling antivirus protection. The proposed multimodal framework correctly flags it as a scam, citing deceptive claims, redirection to potentially harmful external sites, and facilitation of illicit behavior.

<sup>1</sup>[https://www.youtube.com/watch?v=\\_NeG7clfZnw](https://www.youtube.com/watch?v=_NeG7clfZnw)



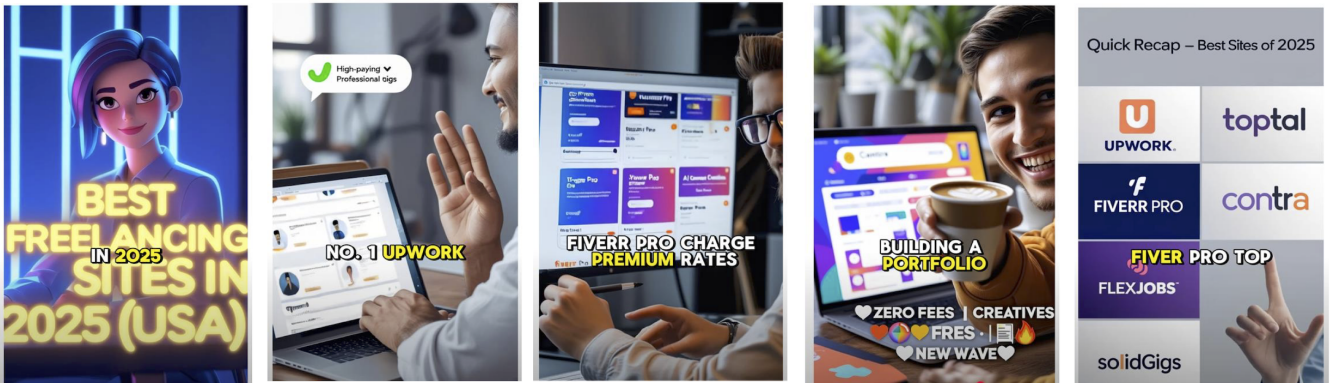
**Title:** How to earn money online

**Description:** How to Make Money FAST in 2025 | Best Side Hustles & Passive Income Ideas! Want to start making money online and achieve financial freedom? In this video, I'll show you ....

Yes, this is a scam video.

**Reasoning:** "Gets clicks, views, or traffic off YouTube by promising viewers that they'll make money fast", "Offers cash gifts, 'get rich quick' schemes, or pyramid schemes", "Sends audiences to sites that can spread malware, try to gather personal information or other sites that have a negative impact"

Figure 9: Qualitative example of a detected scam video. The metadata promotes rapid financial gains, while the frames show a Telegram bot requesting sensitive payment information for “easy earnings.” The proposed multimodal framework correctly identifies the content as a scam due to get-rich-quick claims and redirection to potentially malicious external sites to collect sensitive information.



**Title:** Best Freelancing Sites in 2025 (USA) | Make Money Online Fast! #freelancer #sidehustle ...

**Description:** Looking for the best freelancing websites in 2025? In this 55-second video, we break down the top freelancing platforms in the USA that are helping thousands earn from home ....

No, this is not a scam video.

**Reasoning:** The video provides information about reputable freelancing websites that offer high-paying opportunities. It highlights platforms like Upwork, Fiverr Pro, and others, focusing on legitimate ways to earn income through professional services. The content emphasizes the benefits of these platforms for freelancers, without directing viewers off-site in a misleading way or making unrealistic promises. It maintains a genuine and informative approach, aligning with viewers' interests in finding legitimate freelance work.

Figure 10: Qualitative example of a non-scam video. Although the title includes a potentially misleading “make money online fast” claim, the frames depict legitimate freelancing platforms. The proposed multimodal framework correctly classifies the video as non-scam based on its reliance on established services, absence of unrealistic income promises, and sensitive data requests.