Borrowing Eyes for the Blind Spot: Overcoming Data Scarcity in Malicious Video Detection via Cross-Domain Retrieval Augmentation

Rongpei Hong[1*], Jian Lang[1*], Ting Zhong[1], Fan Zhou[1,2†]
[1]University of Electronic Science and Technology of China,
[2]Intelligent Digital Media Technology Key Laboratory of Sichuan Province
{rongpei.hong,jian_lang}@std.uestc.edu.cn {zhongting,fan.zhou}@uestc.edu.cn

Abstract

The rapid proliferation of online video-sharing platforms has accelerated the spread of malicious videos, creating an urgent need for robust detection methods. However, the performance and generalizability of existing detection approaches are severely limited by the scarcity of annotated video data, as manually curating large-scale malicious detection datasets is both labor-intensive and impractical. To address this challenge, we propose ***CRAVE****, a novel* ***CR****oss-dom***A***in retrie***V***al augm***E***ntation framework that transfers knowledge from resource-rich image-text domain to enhance malicious video detection. Specifically, CRAVE introduces a Pseudo-Pair Retriever to identify semantically relevant image-text data for high-quality cross-domain augmentation. Additionally, a Contrastive Cross-Domain Augmenter is designed to disentangle domain-shared and -unique representations, effectively bridging the domain gaps during knowledge transfer. These shared image-text representations are then leveraged to refine video representations, yielding more discriminative features for accurate malicious content detection. Experiments on four video datasets demonstrate that CRAVE largely outperforms competitive baselines in both performance and generalization, providing an innovative and strong solution to the issue of video data-scarcity. The code is available at https://github.com/ronpay/CRAVE.*

1. Introduction

Online video-sharing platforms have seamlessly integrated into people's daily lives, fundamentally transforming how information is consumed and shared. However, the rise of *malicious* content (e.g., rumors and hateful messages) has become common on video-sharing platforms, significantly damaging politics, finance, and public health [2, 5, 35]. To curb the spreading of the harmful videos, numerous works were proposed and they designed various multimodal approaches [10, 15, 20, 22, 29] to effectively identify the malicious content in videos.

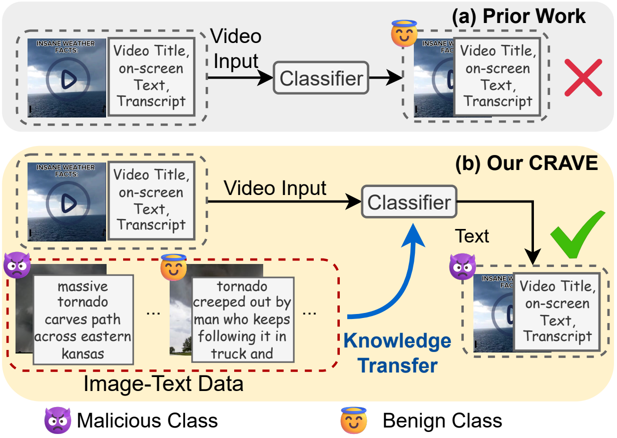

(a) Prior Work
INSANE WEATHER FACTS
Video Title, on-screen Text, Transcript
Video Input
Classifier
Video Title, on-screen Text, Transcript
(b) Our CRAVE
INSANE WEATHER FACTS
Video Title, on-screen Text, Transcript
Video Input
Classifier
Text
massive tornado carves path across eastern kansas
...
tornado creeped out by man who keeps following it in truck and
...
Knowledge Transfer
Image-Text Data
Video Title, on-screen Text, Transcript
Malicious Class
Benign Class

Figure 1. Prior malicious detection methods (Uni-Domain Data Only) *vs.* our CRAVE (Cross-Domain Knowledge Transfer).

Nevertheless, in the domain of malicious video detection, the datasets remain scarce in scale due to the labor-intensive nature of dataset construction, which requires substantial human effort to carefully watch and accurately annotate each video [36]. Most datasets [6, 10, 29, 35] in malicious video detection contain few samples (e.g., less than 1,000 samples in hate video detection [10, 35]). As a result, the limited scale and diversity of data cause existing methods to perform worse on test sets than on training sets, while also learning dataset-specific biases that hinder their ability to generalize effectively to real-world applications [3, 40].

On the other hand, malicious detection datasets in the image-text domain are typically large-scale and diverse, often containing tens of thousands of samples [7, 17, 26, 33], due to the relatively low cost of data annotation. Moreover, synthetic samples can be easily incorporated into image-text datasets [17] by appending generated text to existing images, thereby increasing diversity through the creation of new instances. Given the similar modality composition and

[*]These authors contributed equally to this work.
[†]Corresponding Author.

malicious semantics between image-text and video domains, along with the significant disparity in dataset scale, a natural way to alleviate the video data-scarcity issue is to *leverage the image-text data to enhance the malicious detection in the video domain* (cf. Section 4.2 for experimental evidence).

However, developing such an effective cross-domain augmentation framework is non-trivial due to several key **challenges**: **(1)** The large-scale image-text datasets often contain noisy instances that can hinder rather than help malicious video detection. For example, subsets of benign image-text pairs may be entirely irrelevant to malicious detection tasks in the video domain, as these datasets are frequently randomly sampled or loosely annotated. Incorporating such irrelevant instances can result in negative augmentation. **(2)** Despite the similar modality composition and malicious semantics between the two domains, transferring detection knowledge from image-text datasets to the video domain still involves solving significant domain gaps –namely, *the dynamic nature of video data* vs. *the static nature of image-text data*. These discrepancies lead to different malicious expression patterns. Consequently, designing a robust cross-domain augmentation framework that can effectively filter noisy image-text data and transfer domain knowledge to enhance malicious video detection remains an open challenge.

To this end, we propose **CRAVE**, a novel and robust **CR**oss-dom**A**in retrie**V**al augm**E**ntation framework designed to tackle the data-scarcity in malicious video detection. As illustrated in Figure 1, the core difference between CRAVE and current detection methods lies in the effective utilization of existing cross-domain data to assist the detection. By explicitly accessing abundant and diverse malicious patterns from the image-text domain, CRAVE achieves remarkable detection generalizability. Specifically, to address the challenge (**1**), we introduce a *Pseudo-Pair Retriever*, which identifies semantically relevant image-text data for each target video. By converting videos into pseudo image-text query pairs, the retriever significantly reduces the domain gaps during retrieval and ensures high-quality cross-domain augmentation. As for challenge (**2**), we propose a *Contrastive Cross-Domain Augmenter*, which maximizes knowledge transfer while mitigating domain gaps. The augmenter disentangles domain-shared and -unique representations through cross-domain decoupling learning. It then utilizes the domain-shared representations from retrieved positive and negative image-text samples as “contrastive references” to refine the shared video representations, making them more discriminative. Finally, the refined representations are integrated with video-unique representations to boost detection performance while improving generalizability to real-world scenarios. Our main contributions are summarized as follows:

- We propose a novel cross-domain retrieval-augmentation framework (**CRAVE**), which pioneers a novel retrieval-guided cross-domain augmentation paradigm, harnessing the abundant off-the-thelf image-text data to effectively resolve the data-scarcity issue in malicious video detection.
- We introduce a fresh Pseudo-Pair Retriever, which avoids irrelevant knowledge transfer and provides semantically relevant image-text data for target videos to ensure high-quality cross-domain augmentation.
- We develop a new Contrastive Cross-Domain Augmenter that maximally transfers knowledge from relevant image-text data into malicious video detection while significantly minimizing cross-domain gaps.

Extensive experiments conducted on four real-world benchmarks demonstrate that our CRAVE consistently outperforms competitive baselines. We also evaluate CRAVE under extreme data-scarcity scenarios and out-of-distribution detection tasks, with results demonstrating its strong generalizability under challenging scenarios, including extreme data-scarce limitations and out-of-distribution detection.

2. Related Work

Malicious Video Detection. Malicious video detection aims to analyze the multimodal content of videos (e.g., title, audio, video frames) and detect any possible malicious content in videos, such as rumors or hate speech. Most studies in malicious video detection [9, 10, 29, 32, 35] commonly employ frozen pre-trained modality encoders like BERT [11], ViT [12], and ViViT [1] to extract semantic features and feed the features into a learnable classifier for prediction. Recently, FakeRec [6] analyzed the process of malicious information creation by examining material selection and editing behaviors on micro-video platforms, while NEED [30] enhanced rumor video detection by modeling event-level relationships and leveraging debunking rectification.

Although these carefully designed approaches perform well on the training set for malicious detection, the limited amount of labeled training data, less than 1,000 samples [6, 10, 35], largely bounds the models’ ability to capture sufficient malicious semantics, leading to poor generalizability in the face of out-of-distribution data in real-world scenarios. To solve this problem, we propose an effective cross-domain augmentation framework that leverages rich and diverse malicious representations from large-scale labeled image-text datasets to enhance video detection in data-constrained environments.

Data Augmentation for Video-based Tasks. Data augmentation is a commonly adopted technique in video-based tasks to solve the data-scarcity issue caused by substantial costs associated with video data collection and annotation. Current data augmentation approaches can be roughly categorized into two groups: (1) Conventional augmentation and (2) Knowledge Transfer-based augmentation.

Conventional augmentation methods aim at enlarging the training datasets for data-scarce downstream tasks by generating video samples to directly mitigate the data-scarcity. Traditional generative approaches commonly adopt simple video transformation or editing (e.g., flipping, cropping, and speed variation [8, 13, 37]) to create additional samples and expand training datasets. Although these methods are simple and straightforward, the newly created samples only demonstrate surface-level differences compared to the original data, providing limited data diversity and thus offering a suboptimal solution for semantically complex downstream tasks (e.g., malicious video detection). Recently, with the rise of video generative models (e.g., Sora [23], VideoPoet [19], and Tune-A-Video [39]), some research has begun to leverage the exceptional generative ability of these models for video sample generation. However, these models are typically pre-trained on benign datasets and are not inherently designed to generate harmful content. Adapting them to such tasks (e.g., malicious detection) requires substantial resource-intensive fine-tuning, significantly limiting their applicability [14, 42].

Knowledge Transfer-based augmentation focuses on transferring knowledge from a resource-rich source domain to a target domain facing data-scarcity [27, 41, 43]. For instance, a vision-language model pre-trained on a general-purpose dataset can be adapted to a specific target domain like medical imaging or remote sensing, where the data is scarce [18, 31]. In contrast to generative-based methods, cross-domain augmentation avoids the high overhead of fine-tuning generative models and achieves broader semantic diversity by leveraging existing datasets. In this work, we tackle the data-scarcity issue in malicious video detection by introducing a novel cross-domain augmentation framework CRAVE. It effectively transfers knowledge from image-text data to the domain of malicious video detection while addressing challenges inherent to cross-domain augmentation, such as noisy sample negative transfer and domain gaps.

3. Methodology

3.1. Overview

Problem Definition. Following previous work [10, 35], we consider a malicious video detection dataset, denoted as $\mathcal{D}_S = \{\mathcal{S}_1, \cdots, \mathcal{S}_{N_S}\}$, where N_S is the number of video samples. Each video is presented as $\mathcal{S}_i = (\mathcal{V}_i, \mathcal{T}_i, Y_i)$, where $\mathcal{V}_i$ and $\mathcal{T}_i$ denote the visual and textual content, respectively, with Y_i indicating the ground truth (*malicious* or *benign*). Notably, we also consider the audio modality by incorporating audio transcript into the part of textual content. The task of malicious video detection is defined as $\hat{Y}_i = f_\Phi(\mathcal{V}_i, \mathcal{T}_i)$, where $f_\Phi(\cdot)$ is the detection model.

Our Pipeline. To tackle the data-scarcity issue in malicious video detection, we propose to transfer the knowledge from a resource-rich image-text dataset with a similar task to enhance the video detection. The overall framework is illustrated in Figure 2. Specifically, we introduce an image-text pair malicious detection dataset, denoted as $\mathcal{D}_P = \{\mathcal{P}_1, \cdots, \mathcal{P}_{N_P}\}$, where N_P is the total number of these pairs. Each pair consists of two modalities, presented as $\mathcal{P}_j = (\mathcal{I}_j, \mathcal{C}_j, \mathcal{Y}_j)$, where $\mathcal{I}_j$ and $\mathcal{C}_j$ represent the image and text content, respectively, and $\mathcal{Y}_j$ denotes the ground truth. Our *pipeline* begins by selecting a set of semantically relevant image-text data $\mathcal{N}^r$ for each video $\mathcal{S}_i$ through a cross-domain retrieval, which will be discussed in Section 3.2. Based on the retrieved data, we disentangle domain-shared and -unique representations from both domains through cross-domain decoupling learning. Next, the shared video representation $\mathbf{h}_{v,\text{shared}}$ is refined into a more discriminative form through cross-domain contrastive learning, which will also be introduced in Section 3.3. Finally, the enhanced shared representation $\mathbf{h}_{v,\text{shared}}$ and the unique representation $\mathbf{h}_{v,\text{unique}}$ of the video $\mathcal{S}_i$ are fused for the final prediction, which will be presented in Section 3.4.

3.2. Pseudo-Pair Retriever

The abundant samples in image-text datasets also bring the issue of noisy samples (e.g., benign instances), which are irrelevant to the target domain task (i.e., malicious video detection) and may transfer negative knowledge. To solve this challenge, we propose a new Pseudo-Pair Retriever (PP Retriever), which acts as a filter and accurately provides each video with semantically relevant image-text samples for augmentation through a cross-domain retrieval.

3.2.1. Pseudo-Pair Generation

Directly using videos as queries for image-text retrieval is highly ineffective due to the significant formal disparity between video and image-text domains. To solve this challenge, we propose PP Retriever, a novel retrieval strategy that converts videos into multiple *pseudo-pairs*. These pairs are carefully designed to closely align with the structure and representation of image-text pairs, enabling more accurate cross-domain retrieval. For notational simplicity, we use $\mathcal{S}$ to denote the current video sample.

The PP Retriever starts by selecting representative frames from the video $\mathcal{S}$. Specifically, a total of $\tilde{L}$ frames are uniformly sampled from each video and encoded using the CLIP vision encoder [31]. These encoded frames are subsequently clustered to identify L representative frames, denoted as $\{\tilde{\mathcal{I}}_l\}_{l=1}^{L}$. For each representative frame, we construct pseudo-pairs by associating the frame $\tilde{\mathcal{I}}_l$ with the textual content $\tilde{\mathcal{C}}_l$, which includes video's title, the on-screen text, and audio transcript aligned with that frame. Finally, this process yields L pseudo-pairs for video $\mathcal{S}$, presented as:

$$\tilde{\mathcal{P}}_l = \left(\tilde{\mathcal{I}}_l, \tilde{\mathcal{C}}_l\right), \quad l = 1, \ldots, L. \tag{1}$$

These structured pseudo-pairs will serve as an effective query to retrieve relevant image-text pairs from $\mathcal{D}_P$.

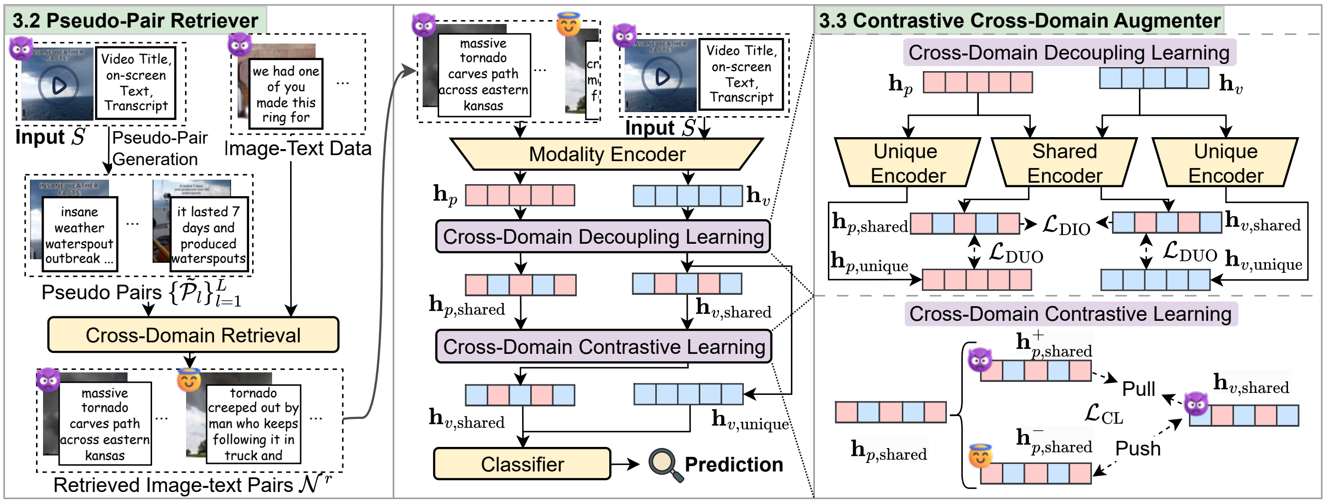

3.2 Pseudo-Pair Retriever
Video Title, on-screen Text, Transcript
we had one of you made this ring for
...
Input S
Pseudo-Pair Generation
Image-Text Data
insane weather waterspout outbreak ...
it lasted 7 days and produced waterspouts
Pseudo Pairs $\{\tilde{\mathcal{P}}_l\}_{l=1}^{L}$
Cross-Domain Retrieval
massive tornado carves path across eastern kansas
tornado creeped out by man who keeps following it in truck and
Retrieved Image-text Pairs $\mathcal{N}^r$
massive tornado carves path across eastern kansas
Video Title, on-screen Text, Transcript
Input S
Modality Encoder
$\mathbf{h}_p$
$\mathbf{h}_v$
Cross-Domain Decoupling Learning
$\mathbf{h}_{p,\text{shared}}$
$\mathbf{h}_{v,\text{shared}}$
Cross-Domain Contrastive Learning
$\mathbf{h}_{v,\text{shared}}$
$\mathbf{h}_{v,\text{unique}}$
Classifier
Prediction
3.3 Contrastive Cross-Domain Augmenter
Cross-Domain Decoupling Learning
$\mathbf{h}_p$
$\mathbf{h}_v$
Unique Encoder
Shared Encoder
Unique Encoder
$\mathbf{h}_{p,\text{shared}}$
$\mathcal{L}_{\text{DIO}}$
$\mathbf{h}_{v,\text{shared}}$
$\mathbf{h}_{p,\text{unique}}$
$\mathcal{L}_{\text{DUO}}$
$\mathcal{L}_{\text{DUO}}$
$\mathbf{h}_{v,\text{unique}}$
Cross-Domain Contrastive Learning
$\mathbf{h}^{+}_{p,\text{shared}}$
Pull
$\mathbf{h}_{v,\text{shared}}$
$\mathcal{L}_{\text{CL}}$
$\mathbf{h}_{p,\text{shared}}$
$\mathbf{h}^{-}_{p,\text{shared}}$
Push

Figure 2. Overview of CRAVE framework. (1) Pseudo-Pair Retriever generates pseudo-pairs for input video to retrieve the semantically relevant pairs from large corpus; (2) Contrastive Cross-Domain Augmenter introduces Cross-Domain Decoupling Learning and Cross-Domain Contrastive Learning to enable effective knowledge transfer from image-text data to malicious video detection.

3.2.2. Cross-Domain Retrieval

To retrieve the most semantically similar image-text pairs while avoiding introducing noise, PP Retriever first encodes visual and textual content of pseudo-pairs $\tilde{\mathcal{P}}_l$ and each image-text pair $\mathcal{P}_j$ from dataset $\mathcal{D}_P$, and calculates the cosine similarity of encoded embeddings between them:

$$\text{sim}(\tilde{\mathcal{P}}_l, \mathcal{P}_j) = \frac{\Psi_v(\tilde{\mathcal{I}}_l) \cdot \Psi_v(\mathcal{I}_j)}{\|\Psi_v(\tilde{\mathcal{I}}_l)\| \|\Psi_v(\mathcal{I}_j)\|} + \frac{\Psi_t(\tilde{\mathcal{C}}_l) \cdot \Psi_t(\mathcal{C}_j)}{\|\Psi_t(\tilde{\mathcal{C}}_l)\| \|\Psi_t(\mathcal{C}_j)\|}, \tag{2}$$

where Ψ_v and Ψ_t denote the CLIP vision and text encoders.

Subsequently, PP Retriever utilizes a global similarity selection strategy to identify the most relevant top-K^+ positive pairs (pairs that share the same category with $\mathcal{S}$) $\mathcal{N}^{r+}$ and top-K^- negative pairs $\mathcal{N}^{r-}$ for $\mathcal{S}$:

$$\mathcal{N}^{r^+} = \arg\operatorname*{top\text{-}K}_{\mathcal{P}_j \in \mathcal{D}_P^+} \left(\max_{l \in \{1, \cdots, L\}} (\text{sim}(\tilde{\mathcal{P}}_l, \mathcal{P}_j)) \right), \tag{3}$$

here $\mathcal{D}_P^+$ is the positive subset of $\mathcal{D}_P$ and $\mathcal{N}^{r^+} = \{(\mathcal{I}_{k^+}, \mathcal{C}_{k^+})\}_{k^+=1}^{K^+}$ denotes the top-K^+ similar positive image-text pairs for the given video $\mathcal{S}$. Similarly, $\mathcal{N}^{r^-} = \{(\mathcal{I}_{k^-}, \mathcal{C}_{k^-})\}_{k^-=1}^{K^-}$ is obtained via retrieving from the negative subset $\mathcal{D}_P^-$.

3.3. Contrastive Cross-Domain Augmenter

To facilitate effective cross-domain knowledge transfer, we propose a Contrastive Cross-Domain Augmenter (CCD Augmenter). It begins by disentangling domain-shared and -unique representations across both domains, thereby mitigating the domain gaps that hinder knowledge transfer. Building on this disentangled representation space, CCD Augmenter leverages shared representations from both retrieved positive and negative image-text pairs as "contrastive references." These references guide the refinement of the target video shared representations, rendering them more discriminative for enhanced malicious detection in the video domain.

3.3.1. Feature Extraction

To simplify notation, we define $\mathcal{N}^r = \mathcal{N}^{r^+} \cup \mathcal{N}^{r^-}$ as the set of retrieved image-text pairs for video $\mathcal{S}$. We first consider the textual modality $\mathcal{T}$ of video $\mathcal{S}$ along with the retrieved textual modalities $\mathcal{C}^r = \{\mathcal{C}_k^r\}_{k=1}^{K}$. For video textual modality $\mathcal{T}$, we consider concatenating three important textual components: the video title, on-screen text, and audio transcript, represented as $\mathcal{T} = \mathcal{T}^t \oplus \mathcal{T}^o \oplus \mathcal{T}^a$. Subsequently, the CLIP text encoder is employed to embed the textual modalities from each domain, yielding the video textual features $\mathbf{h}_v^{\mathcal{T}} \in \mathbb{R}^d$ and the retrieved textual features $\mathbf{h}_p^{\mathcal{T}} \in \mathbb{R}^{K \times d}$, where d denotes the feature dimension. Similarly, the visual modality from each domain is embedded by the CLIP vision encoder to obtain visual features $\mathbf{h}_v^{\mathcal{V}} \in \mathbb{R}^d$ and $\mathbf{h}_p^{\mathcal{V}} \in \mathbb{R}^{K \times d}$.

3.3.2. Cross-Domain Decoupling Learning

To mitigate the domain gaps between image-text and video domains for effective knowledge transfer, CCD Augmenter introduces a novel cross-domain decoupling learning mechanism. Specifically, CCD Augmenter first utilizes a shared encoder $\mathcal{E}_{\text{shared}}^{\mathcal{T}}$ to extract common representations across the two domains. Subsequently, a unique encoder $\mathcal{E}_{m,\text{unique}}^{\mathcal{T}}$ is employed for each domain to enhance the decoupling process. The encoding process is as follows:

$$\mathbf{h}_{m,\text{shared}}^{\mathcal{T}} = \mathcal{E}_{\text{shared}}^{\mathcal{T}}(\mathbf{h}_m^{\mathcal{T}}), \tag{4}$$

$$\mathbf{h}_{m,\text{unique}}^{\mathcal{T}} = \mathcal{E}_{m,\text{unique}}^{\mathcal{T}}(\mathbf{h}_m^{\mathcal{T}}), \tag{5}$$

where $m \in \{v, p\}$, and both $\mathcal{E}_{\text{shared}}^{\mathcal{T}}$ and $\mathcal{E}_{m,\text{unique}}^{\mathcal{T}}$ are implemented using a one-layer MLP. $\mathbf{h}_{m,\text{shared}}^{\mathcal{T}}$ and $\mathbf{h}_{m,\text{unique}}^{\mathcal{T}}$

denote domain-shared and -unique textual representations, respectively. The same process is applied to the visual modality, resulting in $\mathbf{h}^{\mathcal{V}}_{m,\text{shared}}$ and $\mathbf{h}^{\mathcal{V}}_{m,\text{unique}}$. Subsequently, we concatenate the textual and visual representations to generate a unified modality representation, yielding shared and unique representations $\mathbf{h}_{m,\text{shared}} \in \mathbb{R}^{2d}$ and $\mathbf{h}_{m,\text{unique}} \in \mathbb{R}^{2d}$.

To ensure the effective extraction of shared representations, CCD Augmenter proposes a Domain-Invariant Objective (DIO), which enforces the shared encoders to focus on capturing transferable patterns across domains. Specifically, the Kullback-Leibler (KL) divergence is utilized to implement DIO, aligning the shared representations of image-text pairs to the video domain:

$$\mathcal{L}_{\text{DIO}} = \text{KL}\left(\sigma(\mathbf{h}_{p,\text{shared}}) \parallel \sigma(\mathbf{h}_{v,\text{shared}})\right), \tag{6}$$

where $\mathbf{h}_{v,\text{shared}}$ and $\mathbf{h}_{p,\text{shared}}$ denote the domain-shared representations from the video and image-text domains and σ represents softmax operation.

To mitigate the coupling of domain-shared and -unique representations within the same domain, CCD Augmenter introduces the Domain-Unique Objective (DUO), serving as an auxiliary loss to separate these representations. Specifically, an orthogonality loss is adopted to implement DUO, ensuring a clear distinction between these representations:

$$\mathcal{L}_{\text{DUO}} = \sum_{m\in\{v,p\}} \left\|\mathbf{h}^{\top}_{m,\text{shared}}\mathbf{h}_{m,\text{unique}}\right\|^2_2, \tag{7}$$

where $\mathbf{h}_{m,\text{shared}}$ and $\mathbf{h}_{m,\text{unique}}$ denote the shared and unique representations for domain m.

3.3.3. Cross-Domain Contrastive Learning

Based on the shared representations from two domains, CCD Augmenter introduces a new cross-domain contrastive learning paradigm. Specifically, CCD Augmenter proposes a triplet loss-based contrastive learning approach, which aligns the shared representations between video $\mathcal{S}$ and image-text pairs from the same category as $\mathcal{S}$, while distancing them away from image-text pairs belonging to the opposite category. This process can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{CL}} = \max(&\|\mathbf{h}_{v,\text{shared}} - \mathbf{h}^{+}_{p,\text{shared}}\|^2_2 \\ &- \|\mathbf{h}_{v,\text{shared}} - \mathbf{h}^{-}_{p,\text{shared}}\|^2_2 + \epsilon, 0),\end{aligned} \tag{8}$$

where $\mathbf{h}^{+}_{p,\text{shared}}$ and $\mathbf{h}^{-}_{p,\text{shared}}$ denote the shared representations from positive and negative pairs, ϵ defines the margin that enforces a minimum distance between these two pairs.

3.4. Prediction

For a given video S, we concatenate the video shared and unique representations $\mathbf{h}_{v,\text{shared}}$ and $\mathbf{h}_{v,\text{unique}}$, and feed them into a two-layer MLP-based classifier, yielding the prediction $\hat{Y} = \mathcal{E}_{\text{CLS}}(\mathbf{h}_{v,\text{shared}} \oplus \mathbf{h}_{v,\text{unique}})$. To optimize malicious video detection while enhancing cross-domain knowledge transfer, we define the total objective function as follows:

Dataset	# Malicious	# Benign	# Total	Duration (s)
FakeTT [6]	1,172	819	1,991	47.69
FVC [28]	1,633	1,131	2,764	87.83
MHClipEN [35]	338	662	1,000	33.84
HateMM [10]	431	652	1,083	150.07
Fakeddit [26]	14,198	21,690	35,888	N/A
FHM [17]	3,266	5,734	9,000	N/A

Table 1. Statistics of four video and two image-text datasets.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLS}} + \lambda\mathcal{L}_{\text{DIO}} + \gamma\mathcal{L}_{\text{DUO}} + \omega\mathcal{L}_{\text{CL}}, \tag{9}$$

where $\mathcal{L}_{\text{CLS}}$ is the Binary Cross-Entropy loss for video classification, both $\mathcal{L}_{\text{DIO}}$ and $\mathcal{L}_{\text{DUO}}$ enforce cross-domain shared representations extraction, and $\mathcal{L}_{\text{CL}}$ facilitates cross-domain contrastive learning. The coefficients λ, γ, and ω control the relative contributions of the respective loss terms. During training, all pre-trained encoders are frozen to reduce the overhead. Details regarding computational complexity analysis, the training algorithm, and the mathematical proof of CRAVE's effectiveness are provided in Appendix B-D.

4. Experiments

4.1. Experimental Setup

A concise summary of the experimental setup is presented below. Further details regarding the datasets, baselines, and implementation are available in Appendix E. Moreover, additional experiments are presented in Appendix F.

Datasets & Cross-Domain Augmentation Setting. In this study, we evaluate our CRAVE on four real-world malicious video detection datasets. To address the data-scarcity issue, we propose a new cross-domain augmentation setting by leveraging external knowledge from extra resource-rich image-text datasets to enhance the detection. Detailed statistics of datasets are provided in Table 1, and the datasets are categorized as follows: (1) *Rumor detection datasets*: FakeTT [6] and FVC [28], which focus on identifying rumor videos on several platforms, including TikTok, YouTube, and Twitter. We adopt Fakeddit [26], which is a rumor detection dataset comprising image-text pairs posted from Reddit, as an extra image-text dataset. (2) *Hate detection datasets*: MHClipEN [35] and HateMM [10], which are hate video detection datasets with videos collected from YouTube and BitChute. We select FHM [17], a hate meme dataset collected by Facebook, as an extra image-text dataset.

Baselines. We compare CRAVE with 10 baselines, which can be broadly categorized into three groups: (1) *Vanilla detection methods* which leverage various multimodal approaches to detect malicious content in videos, including HTMM [10], MHCL [35], SVFEND [29], and FakeRec [6]. (2) *Conventional augmentation methods*, which tackle data-scarcity in malicious video detection by synthesizing new

Dataset	FakeTT				FVC				MHClipEN				HateMM			
Methods	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**
HTMM	70.57	69.69	70.51	73.15	90.01	89.78	89.94	89.65	74.00	66.63	72.19	65.64	77.42	74.96	78.39	74.12
MHCL	77.93	75.78	76.61	77.99	90.47	90.14	91.00	89.67	74.00	70.82	70.82	70.82	78.80	77.85	77.97	77.75
SVFEND	77.14	75.63	75.12	77.56	87.59	87.36	87.34	87.40	68.00	46.94	69.65	52.98	73.27	71.71	72.21	71.42
FakeRec	79.53	77.42	77.01	78.29	91.22	91.18	91.22	91.65	70.50	54.33	74.23	57.08	74.65	72.07	74.76	71.43
Spatial Aug.	78.26	76.09	75.58	76.86	90.31	90.07	90.34	89.87	73.00	68.44	69.51	67.85	79.72	78.44	79.32	77.94
Temp. Aug.	72.91	71.93	72.32	75.16	89.71	89.55	89.42	89.72	74.50	69.20	71.67	68.24	75.58	73.96	74.84	73.53
ViLT	78.59	77.23	76.66	79.41	81.39	80.99	81.06	80.93	69.50	53.63	69.16	56.33	70.51	68.48	69.29	68.16
TSformer	65.38	65.12	69.21	70.93	90.22	89.87	90.39	90.01	70.00	57.65	67.43	58.56	70.51	68.78	69.23	68.54
LLaVA	46.93	46.70	53.55	53.24	60.10	56.66	58.80	57.41	70.61	66.87	66.74	66.75	73.62	73.82	78.16	77.43
Qwen-VL	53.17	52.71	56.27	57.65	59.72	58.20	58.51	58.81	73.40	67.36	70.65	66.47	75.42	75.72	77.92	78.39
CRAVE	**84.95**	**83.52**	**82.77**	**84.66**	**96.52**	**96.45**	**96.43**	**96.47**	**82.50**	**79.81**	**80.83**	**79.06**	**87.09**	**86.51**	**86.67**	**86.47**
Improv.	6.8%↑	7.9%↑	7.5%↑	8.1%↑	5.8%↑	5.8%↑	5.7%↑	5.3%↑	10.7%↑	12.7%↑	8.9%↑	11.6%↑	9.2%↑	10.3%↑	9.3%↑	10.3%↑
p-val.	$9.0e^{-3}$	$7.5e^{-3}$	$7.8e^{-3}$	$5.8e^{-3}$	$2.3e^{-3}$	$2.5e^{-3}$	$1.7e^{-3}$	$4.4e^{-3}$	$1.8e^{-3}$	$8.4e^{-3}$	$1.1e^{-2}$	$6.5e^{-3}$	$7.3e^{-4}$	$1.3e^{-3}$	$4.0e^{-4}$	$2.7e^{-3}$

Table 2. Performance comparison on four real-world video datasets. The best results are in **black bold**, while the second are underlined. Higher values of ACC, M-F1, M-P, and M-R indicate better performance.

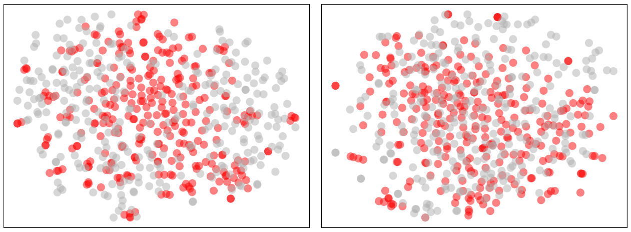
video data, including Spatial Augmentation [34] and Temporal Augmentation [16]. Notably, these video-based augmentations are only employed to enrich the training dataset and the results are from the best-performing baselines (FakeRec for FakeTT and FVC while MHCL for MHClipEN and HateMM) training on the enriched dataset. (3) *Cross-domain augmentation methods*, which transfer knowledge from a resource-rich domain to target domain, including ViLT [18], TSformer [4], LLaVA [21], and Qwen-VL [38].

Metrics. Following prior works [6, 35], we adopt metrics Accuracy (ACC), Macro F1 score (M-F1), Macro Precision (M-P), and Macro Recall (M-R) to evaluate the performance.

Implementation Details. In this study, we employ pretrained CLIP, specifically `clip-vit-large-patch14`, as modality encoders. In pseudo-pair generation, videos are uniformly sampled to 32 frames and clustered to 10 frames as representative frames. While in feature extraction, each video is uniformly sampled to 16 frames to generate visual embedding. AdamW [24] is adopted as optimizer. The detailed hyper-parameter settings, including the learning rate, number of retrieved pairs, and loss coefficients for each dataset, are reported in Appendix E.3. All experiments are conducted on a single RTX 4090 GPU.

4.2. Preliminary Experiment

In this section, we validate the feasibility of knowledge transfer from the image-text domain to malicious video detection by empirically analyzing the semantic relevance of harmful content between the two domains. Specifically, we randomly select 300 malicious videos from both the hateful and rumor datasets, along with 300 malicious image-text samples from their corresponding datasets. We then utilize CLIP to encode the visual and textual modalities of both videos and image-text pairs across the hateful and rumor datasets. The concatenated modality features are projected into a 2D space using t-SNE [25] for visualization. As shown in Figure 3, the results reveal a strong semantic relevance in malicious content across the two domains, supporting the feasibility of our proposed framework's motivation.



(a) FakeTT and Fakeddit Datasets. (b) MHClipEN and FHM Datasets.

Figure 3. Visualization of the semantic representations of malicious samples from two domains. Red points indicate videos and gray points represent image-text pairs.

4.3. Overall Performance

To demonstrate the superiority of our CRAVE, we compare it against 10 competitive baselines across four datasets, with the results reported in Table 2. From these results, we have the following observations:

First, **CRAVE** outperforms all baselines across all datasets, achieving an average improvement of 9.2% in M-F1 and 8.1% in ACC. We also calculate the statistical differences between CRAVE and the strongest baseline by retraining both models five times. The p-values, all below 0.05, substantiate the statistical significance of CRAVE's performance gains. These gains stem from our novel retrieval-augmented cross-domain knowledge transfer paradigm. It overcomes the constraint of visiting only limited data in video domain and allows the detector to access the expressive knowledge from resource-rich image-text dataset, largely improving the performance and generalizability of the detector.

Second, **vanilla detection methods** demonstrate a certain degree of ability in video-based malicious content identifica-

Module	Variant	FakeTT		MHClipEN	
		ACC	M-F1	ACC	M-F1
PP Retriever	Vanilla Retriever	82.94	81.83	78.50	74.02
	Random Retriever	81.94	80.86	77.00	72.07
CD Decoupler	w/o Decoupling	81.60	80.10	79.00	76.60
	w/o Contrastive	81.27	80.01	77.50	72.81
	w/o Augmenter	77.25	76.46	74.50	65.29
CRAVE	**All**	**84.95**	**83.52**	**82.50**	**79.81**

Table 3. Ablation study of main components of CRAVE.

tion. For instance, FakeRec captures useful clues from the perspective of the rumor video creative process for enhanced detection and achieves competitive results. However, the limited amount of video training data hinders their generalizability, incurring suboptimal performance during inference.

Third, **conventional augmentation methods** yield only marginal performance improvements or even degrade performance compared to vanilla detectors. This is mainly due to the limited diversity of synthesized data, particularly in high-dimensional spatiotemporal patterns, which fails to introduce novel discriminative features and may even inject noise, blurring the model's decision boundaries. Furthermore, **cross-domain augmentation methods** also underperform our framework, as the source domain knowledge they leverage leans toward benign content, making it less effective for adapting harmful content detection.

4.4. Ablation Study

We conduct a comprehensive ablation study to evaluate the role of each components within CRAVE, with results presented in Table 3.

4.4.1. Effect of the PP Retriever

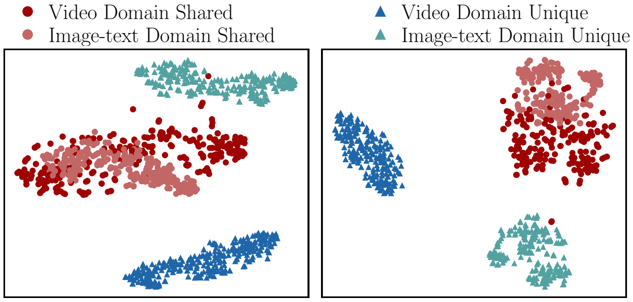
To validate the effect of the PP Retriever, we design two variant models: (1) **Vanilla Retriever**: This variant replaces the PP Retriever with a raw video-based retriever that simply combines video frames and titles for cross-domain retrieval. (2) **Random Retriever**: This variant selects instances randomly from the image-text dataset as retrieval results. From the results, we observe that the Vanilla Retriever struggles to address the domain gaps, resulting in inaccurate retrieval and suboptimal performance. Furthermore, the Random Retriever introduces significant noise by arbitrarily selecting instances from the image-text domain, which leads to negative knowledge transfer and a substantial drop in performance.

4.4.2. Effect of the CCD Augmenter

To assess the CCD Augmenter, we construct three variant models: (1) **w/o Decoupling**: This variant directly transfers knowledge from the image-text dataset to video detection without decoupling shared representations. Accordingly, both $\mathcal{L}_{\text{DIO}}$ and $\mathcal{L}_{\text{DUO}}$ are removed. (2) **w/o Contrastive**: This variant concatenates the image-text domain-shared representation with the video representation for knowledge transfer, meanwhile $\mathcal{L}_{\text{CL}}$ is excluded. (3) **w/o Augmenter**: This variant eliminates the augmenter entirely and relies on the video representation solely for detection. The results reveal that removing either cross-domain decoupling learning or contrastive learning leads to a performance decline, underscoring their critical roles in effective cross-domain knowledge transfer. Furthermore, eliminating the augmenter reduces the model to a vanilla detector, whose performance is significantly limited by the scarcity of training data.

	Target	Positive Top-1	Negative Top-1
Video/ Image			
Text	lion gets revenge for killing a lion	old age famous lion dies of old age	apex predator
Sim.	N/A	0.6285	0.5595

Figure 4. Presentation of retrieved top-1 positive and negative image-text pairs for the target video. "Sim." is the similarity score.



(a) FakeTT Dataset.

(b) MHClipEN Dataset.

Figure 5. Visualization of the decoupled shared and unique representations from video and image-text domain.

4.5. Cross-Domain Retrieval Quality Presentation

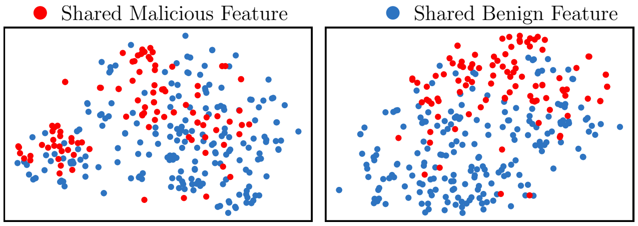
In this section, we further analyze the effectiveness of the PP Retriever in cross-domain retrieval by conducting a case study. We randomly select one video instance from the FVC dataset and present its retrieved image-text data from the Fakeddit dataset. As illustrated in Figure 4, the retrieved image-text data from both positive and negative categories show high semantic relevance to the target video, validating the efficacy of the PP Retriever in filtering noisy data and ensuring high-quality cross-domain augmentation.

4.6. Visualization on Knowledge Transfer

In this section, we thoroughly evaluate the effectiveness of our proposed CCD Augmenter in transferring the cross-domain knowledge by presenting feature-level visualization.

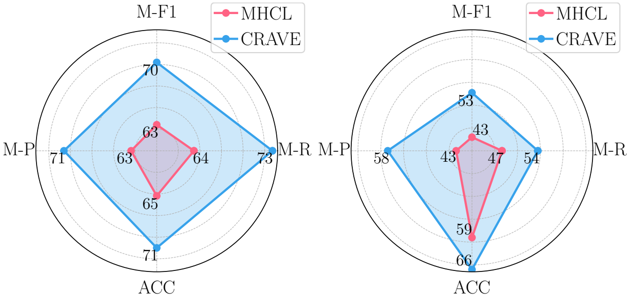
4.6.1. Cross-Domain Decoupling Learning Visualization

First, we visualize the shared and unique representations from both domains. Specifically, we project the shared and



(a) w/o $\mathcal{L}_{\text{CL}}$ on FakeTT Dataset. (b) w/ $\mathcal{L}_{\text{CL}}$ on FakeTT Dataset.

Figure 6. Visualization of the shared features from malicious and benign video samples w/o and w/ cross domain contrastive learning.



(a) FVC to FakeTT Dataset. (b) HateMM to MHClipEN Dataset.

Figure 7. Comparison of CRAVE with the most competitive baseline MHCL through cross-platform evaluation settings.

unique representations of both the test set videos and their corresponding retrieved image-text pairs into a 2D space via t-SNE. Four distinct colors are used to represent the shared and unique features of each domain separately. From Figure 5, we observe that the domain-shared representations across the two domains are clustered together, while the unique representations are distinctly separated. These results highlight the effectiveness of our CCD Augmenter in bridging cross-domain gaps.

4.6.2. Cross-Domain Contrastive Learning Visualization

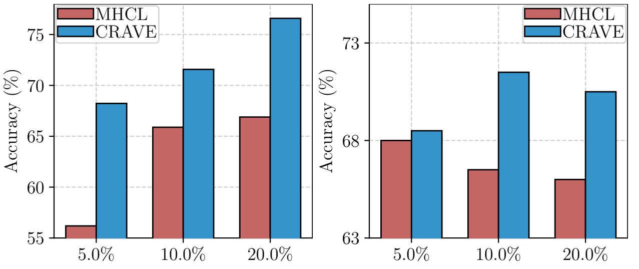
Subsequently, we visualize the domain-shared representations of test set videos before and after enhancement through cross-domain contrastive learning by mapping their features into a 2D space through t-SNE. As depicted in Figure 6, the refined representations exhibit clearer boundaries between malicious and benign instances compared to the original ones. This highlights the effectiveness of CCD Augmenter in generating more discriminative representations.

4.7. Detection Generalizability Analysis

To comprehensively validate the malicious detection generalizability of CRAVE, we conduct both out-of-distribution detection evaluation and extreme data-scarcity analysis. Notably, we apply the same cross-domain augmentation mechanism as used in the main experiments to CRAVE, allowing it to visit the powerful "knowledge" from image-text datasets.

4.7.1. Out-of-Distribution Detection Evaluation

We compare the out-of-distribution detection capability of CRAVE and competitive baseline MHCL through cross-platform detection experiments. Specifically, we pair the datasets FVC with FakeTT and HateMM with MHClipEN. For each pair, the model is trained on the first dataset and evaluated on the second, with results presented in Figure 7. We observe that CRAVE consistently surpasses the MHCL across both cross-platform scenarios, as it effectively leverages the diverse distributed cross-domain knowledge to enhance the detection generalizability.



(a) FakeTT Dataset. (b) MHClipEN Dataset.

Figure 8. Comparison of CRAVE with the most competitive baseline MHCL under conditions of 5%, 10%, and 20% training set.

4.7.2. Extreme Data-Scarcity Analysis

To further assess the generalizability of CRAVE under extremely limited training data conditions, we progressively reduce the size of the training set to 5%, 10%, and 20% of its original scale, a scenario prone to overfitting on the training data. As illustrated in Figure 8, CRAVE consistently outperforms the strongest baseline, MHCL, across all settings. Notably, even with only 10% of the training data, CRAVE achieves an accuracy of more than 70% on both datasets, highlighting the superior generalizability of CRAVE in maintaining high detection performance under extreme data-scarcity scenarios.

5. Conclusion

In this work, to tackle the data-scarcity challenge in malicious video detection, we propose a novel cross-domain retrieval-augmentation framework (CRAVE). This framework consists of two core components: (1) A new Pseudo-Pair Retriever which converts raw videos into pseudo-pairs for more accurate cross-domain retrieval, providing semantically relevant image-text instances for high-quality augmentation. (2) A fresh Contrastive Cross-Domain Augmenter that disentangles domain-shared and domain-unique representations, leveraging the shared image-text representations to refine video representations through cross-domain contrastive learning. Extensive experiments on four benchmarks demonstrate the strong detection ability and generalizability of CRAVE, providing a promising solution to data-scarcity conditioned malicious content detection.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62176043, and No.U22A2097).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A Video Vision Transformer. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 2

[2] Mazen Balat, Mahmoud Essam Gabr, Hend A. Bakr, and A. Zaky. Tikguard: A Deep Learning Transformer-Based Solution for Detecting Unsuitable TikTok Content for Kids. *arXiv*, 2024. 1

[3] Mohammad Mahdi Bejani and Mehdi Ghatee. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, 54(8):6391–6438, 2021. 1

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *International Conference on Machine Learning (ICML)*, pages 813–824, 2021. 6

[5] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. Combating Online Misinformation Videos: Characterization, Detection, and Future Directions. In *ACM International Conference on Multimedia (MM)*, pages 8770–8780, 2023. 1

[6] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. Fakingrecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2024. 1, 2, 5, 6

[7] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2506–2515, 2019. 1

[8] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 3

[9] Hyewon Choi and Youngjoong Ko. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *International Conference on Information and Knowledge Management (CIKM)*. ACM, 2021. 2

[10] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 17:1014–1023, 2023. 1, 2, 3, 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast Networks for Video Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 3

[14] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning Scalable 3d Generative Models from Video Diffusion Models. In *European Conference on Computer Vision (ECCV)*, pages 333–350, 2024. 3

[15] Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Following Clues, Approaching the Truth: Explainable Micro-Video Rumor Detection via Chain-of-Thought Reasoning. In *The Web Conference*, pages 4684–4698, 2025. 1

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 6

[17] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 5

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021. 3, 6

[19] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A Large Language Model for Zero-Shot Video Generation. In *International Conference on Machine Learning (ICML)*, 2024. 3

[20] Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. Biting Off More Than You Can Detect: Retrieval-Augmented Multimodal Experts for Short Video Hate Detection. In *The Web Conference*, pages 2763–2774, 2025. 1

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-OneVision: Easy Visual Task Transfer. *arXiv*, abs/2408.03326, 2024. 6

[22] Yili Li, Jian Lang, Rongpei Hong, Qing Chen, Zhangtao Cheng, Jia Chen, Ting Zhong, and Fan Zhou. Real: Retrieval-augmented prototype alignment for improved fake news video detection. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025. 1

[23] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv*, abs/2402.17177, 2024. 3

[24] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[25] L. Maaten and Geoffrey E. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

[26] Kai Nakamura, Sharon Levy, and W. Wang. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *arXiv*, abs/1911.03854, 2019. 1, 5

[27] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-Adapter: Parameter-Efficient Image-to-Video Transfer Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[28] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. A corpus of debunked and verified user-generated videos. *Online Information Review*, 43(1):72–88, 2019. 5

[29] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. Fakesv: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 14444–14452, 2023. 1, 2, 5

[30] Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2023. 2

[31] 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 *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3

[32] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 2021. 2

[33] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8 (3):171–188, 2020. 1

[34] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 568–576, 2014. 6

[35] Han Wang, Rui Yang Tan, Usman Naseem, and Roy Ka-Wei Lee. Multihateclip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2024. 1, 2, 3, 5, 6

[36] Han Wang, Tan Rui Yang, and Roy Ka-Wei Lee. Cross-Modal Transfer from Memes to Videos: Addressing Data Scarcity in Hateful Video Detection. *arXiv*, abs/2501.15438, 2025. 1

[37] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36, 2016. 3

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model&apos;s Perception of the World at Any Resolution. *arXiv*, abs/2409.12191, 2024. 6

[39] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7589–7599, 2023. 3

[40] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In *European Conference on Computer Vision (ECCV)*, pages 322–339, 2020. 1

[41] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer. In *ACM International Conference on Multimedia (MM)*, pages 4505–4514, 2022. 3

[42] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Y. Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative Video Camera Controls for User-Provided Videos using Masked Video Fine-Tuning. *arXiv*, 2024. 3

[43] Wei Zhang, Chaoqun Wan, Tongliang Liu, Xinmei Tian, Xu Shen, and Jieping Ye. Enhanced Motion-Text Alignment for Image-to-Video Transfer Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

Borrowing Eyes for the Blind Spot: Overcoming Data Scarcity in Malicious Video Detection via Cross-Domain Retrieval Augmentation

Supplementary Material

A. Reproducibility

B. Complexity & Efficiency

B.1. Computational Complexity Analysis

B.2. Efficiency Comparison

C. Proof of the Effectiveness of CRAVE

D. Proof of the Effectiveness of CRAVE through An Information-Theoretic Perspective

E. Detailed Experimental Setup

E.1. Baselines

E.2. Datasets

E.3. Implementation Details

F. Additional Experiments

F.1. Additional Ablation Study

F.2. Hyper-parameter Sensitivity Analysis

F.3. Additional Visualization of Knowledge Transfer

F.4. Evaluation on Additional Dataset

F.5. Additional Detection Generalizability Analysis

F.6. Additional Retrieval Results Presentation

G. Broader Impacts of Our Work

H. Limitations and Future Work