

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

Supplementary Material

A. Additional Related Work: Data Scarcity in Video-based Detection

Another branch of literature related to our study focuses on addressing data scarcity in video-based detection. In this context, Weakly Supervised Video Anomaly Detection (WSVAD) has emerged as a prominent and active research area in recent years [4, 13, 17]. A seminal work by Sultani et al. [17] pioneered the use of a deep multiple instance learning (MIL) framework, where each video is viewed as a “bag” of segments (instances). By employing a ranking loss on bag-level annotations, their model effectively maximizes the margin between anomalous segments in positive bags and normal ones in negative bags. Building on this foundation, subsequent studies have explored ways to enhance the positive modeling capability of WSVAD. Zhong et al. [24] introduced a GCN-based framework that captures inter-segment feature similarity and enforces temporal coherence. Tian et al. [18] proposed a robust temporal feature magnitude learning strategy, substantially improving MIL’s tolerance to false negatives from abnormal videos. More recently, the integration of pre-trained vision-language models has opened new possibilities for VAD. VadCLIP [23] represents a milestone by transferring the rich cross-modal knowledge of CLIP [15] to WSVAD, achieving state-of-the-art performance. Building on this paradigm, Pu et al. [13] further improved detection by designing prompt-enhanced contextual representations.

However, mainstream pre-trained vision-language models such as CLIP are typically trained on benign image-text datasets. As a result, directly transferring their pre-trained knowledge to malicious video detection remains suboptimal. And in this work, we employ pre-trained CLIP as the feature encoders, but we also incorporate abundant malicious knowledge from extensive off-the-shelf harmful content detection image-text datasets to further enhance the detection.

B. Complexity & Efficiency

In this section, we provide an in-depth analysis of the complexity and efficiency of the proposed CRAVE framework, highlighting its practical applicability.

B.1. Computational Complexity Analysis

We analyze the computational complexity of the proposed CRAVE framework, focusing on its two main components: the PP Retriever and the CCD Augmenter. We also discuss the overall complexity to provide insights into the scalability

and efficiency of the framework.

B.1.1. Complexity of PP Retriever

The computational complexity of the PP Retriever mainly involves three steps including pseudo-pair generation, similarity computation, and cross-domain retrieval. Given \tilde{L} sampled frames per video, encoding them via CLIP vision encoder requires $O(L \cdot d_v)$ operations, where d_v is the visual feature dimension. Clustering these features into L representative frames introduces $O(\tilde{L} \cdot L \cdot d_v)$ complexity. For each video, generating L pseudo-pairs (see Eq. (1)) incurs $O(L \cdot (d_v + d_t))$ complexity, where d_t is the text feature dimension. Before retrieval, preprocessing image-text dataset requires encoding N_P image-text pairs in \mathcal{D}_P using CLIP, with complexity $O(N_P \cdot (d_v + d_t))$, where d_v and d_t are feature dimensions of visual/text encoders. Computing similarity scores (Eq. (2)) between L pseudo-pairs and N_P image-text pairs in \mathcal{D}_P requires $O(L \cdot N_P \cdot (d_v + d_t))$. The top- K retrieval (Eq. (3)) adds $O(L \cdot N_P \cdot \log K)$ complexity. Thus, the overall complexity is dominated by $O(L \cdot N_P \cdot (d_v + d_t))$, which can be optimized via FAISS [5] for efficient nearest neighbor search.

B.1.2. Complexity of CCD Augmenter

The CCD Augmenter’s complexity stems from cross-domain decoupling and contrastive learning. Encoding video and K retrieved pairs (where $K = K^+ + K^-$) involves $O((1 + K) \cdot (d_v + d_t))$ operations. Shared/unique encoders (Eqs. (4) and (5)) process features with $O((1 + K) \cdot d^2)$ complexity, where d is the hidden dimension. The DIO loss (Eq. (6)) and DUO loss (Eq. (7)) require $O(d)$ and $O((1 + K) \cdot d^2)$ operations, respectively. The contrastive loss (Eq. (8)) adds $O(K \cdot d)$ complexity. Overall, the CCD Augmenter’s complexity is $O((1 + K) \cdot d^2)$, efficient for practical deployment.

B.1.3. Overall Complexity

Combining the complexities of the PP Retriever and the CCD Augmenter, CRAVE’s total complexity is $O(L \cdot N_P \cdot (d_v + d_t) + (1 + K) \cdot d^2)$. This ensures scalability and efficiency, particularly when leveraging optimization techniques for efficient nearest neighbor search and efficient model architectures. Additionally, the training algorithm for our proposed CRAVE is detailed in Algorithm 1.

B.2. Efficiency Comparison

We compare CRAVE to competitive baselines by recording the number of trainable parameters and the training time per

Algorithm 1 Training Algorithm of CRAVE**Input:** Video dataset \mathcal{D}_S , image-text dataset \mathcal{D}_P **Output:** Trained malicious video detection model f_Φ

```

1: for each video  $\mathcal{S} = (\mathcal{V}, \mathcal{T}, Y) \in \mathcal{D}_S$  do
2:   /* Pseudo-Pair Retriever */
3:   Sample  $\tilde{L}$  frames from  $\mathcal{V}$  and cluster to  $L$  representative frames  $\{\tilde{\mathcal{I}}_l\}_{l=1}^L$ 
4:   Generate  $L$  pseudo-pairs  $\{\tilde{\mathcal{P}}_l = (\tilde{\mathcal{I}}_l, \tilde{\mathcal{C}}_l)\}$  via Eq. (1)
5:   Compute similarity scores  $\text{sim}(\tilde{\mathcal{P}}_l, \mathcal{P}_j)$  for all  $\tilde{\mathcal{P}}_l$  and  $\mathcal{P}_j \in \mathcal{D}_P$  via Eq. (2)
6:   Retrieve top- $K^+$  positive pairs  $\mathcal{N}^{r+}$  and top- $K^-$  negative pairs  $\mathcal{N}^{r-}$  via Eq. (3)
7:   /* Contrastive Cross-Domain Augmenter */
8:   Encode  $\mathcal{T}$  to  $\mathbf{h}_v^T$  and  $\{\mathcal{C}_k^r\}_{k=1}^K$  to  $\mathbf{h}_p^T$  with CLIP Text Encoder and MLP
9:   Encode  $\mathcal{V}$  to  $\mathbf{h}_v^V$  and  $\{\mathcal{I}_k^r\}_{k=1}^K$  to  $\mathbf{h}_p^V$  with CLIP Vision Encoder and MLP
10:  for  $m \in \{v, p\}$  do
11:    Extract shared features  $\mathbf{h}_{m,\text{shared}}$  via Eq. (4)
12:    Extract unique features  $\mathbf{h}_{m,\text{unique}}$  via Eq. (5)
13:  end for
14:  Compute DIO loss  $\mathcal{L}_{\text{DIO}}$  via Eq. (6)
15:  Compute DUO loss  $\mathcal{L}_{\text{DUO}}$  via Eq. (7)
16:  Compute contrastive loss  $\mathcal{L}_{\text{CL}}$  via Eq. (8)
17:  /* Model Optimization */
18:  Fuse  $\mathbf{h}_{v,\text{shared}} \oplus \mathbf{h}_{v,\text{unique}}$ 
19:  Compute prediction  $\hat{Y}_i$  with MLP classifier
20:  Calculate classification loss  $\mathcal{L}_{\text{CLS}}$  with Binary Cross-Entropy
21:  Obtain the total loss:  $\mathcal{L}_{\text{total}}$  via Eq. (9)
22:  Update  $\Phi$  via backpropagation
23: end for

```

epoch on the FakeTT dataset, with the results presented in Figure B.1.

From the results, we can see that SVFEND and FakeRec have a significantly higher number of trainable parameters and longer training times, owing to their sophisticated design for malicious video detection. And transfer-based models like TSformer and ViLT have fewer parameters due to their trainable classifiers but require more time for calculations when dealing with large-scale frozen parameters. Meanwhile, MHCL and HTMM, with their simpler model architectures, incur lower training costs but deliver average performance.

In contrast, our model strikes a balance between performance and running costs by employing a straightforward architecture that effectively leverages large-scale datasets to enhance video detection tasks, thereby achieving significantly better results.

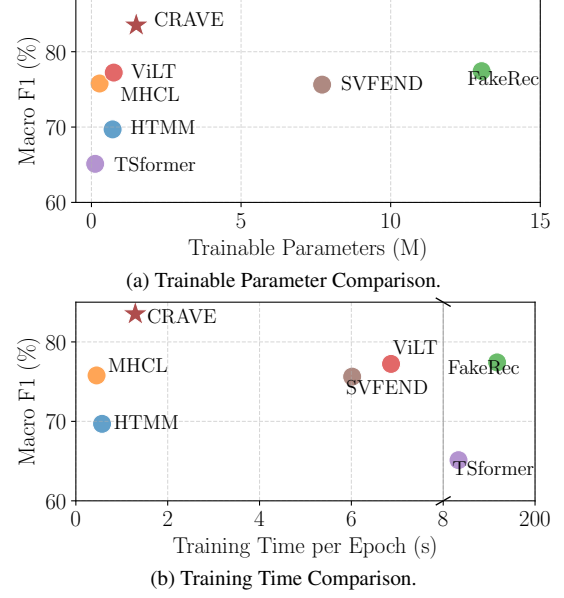


Figure B.1. The performance of our CRAVE and competitive baselines with respect to the number of trainable parameters and training time of each epoch.

C. Proof of the Effectiveness of CRAVE through an Information-Theoretic Perspective

The previous sections in the main paper have demonstrated the motivation and process behind utilizing cross-domain image-text data to solve the data-scarcity in malicious video detection. In this section, the effectiveness of the retrieved cross-domain knowledge transfer within the proposed CRAVE framework is evaluated from an information-theoretic perspective. The malicious video detection is defined as determining whether a given video S_i is malicious or benign. For any given video S_i , we aim to show that incorporating external knowledge from retrieved image-text domain data \mathcal{R}_i improves the prediction of the label Y_i . Let \mathcal{V}_i , and \mathcal{T}_i represent the vision, and text modality representations of the S_i , respectively. We can state the following proposition.

Proposition 1. Let the mutual information $I(X; Y)$ quantify the amount of information that the variables X and Y share regarding one another. Consequently, we can express:

$$I(Y_i; \mathcal{V}_i, \mathcal{T}_i, \mathcal{R}_i) \geq I(Y_i; \mathcal{V}_i, \mathcal{T}_i). \quad (10)$$

Proof. According to the definition of mutual information,

we have:

$$\begin{aligned}
I(Y_i; \mathcal{V}_i, \mathcal{T}_i, \mathcal{R}_i) &= \mathbb{E} \left[\log \frac{\mathbb{P}(Y_i, \mathcal{V}_i, \mathcal{T}_i, \mathcal{R}_i)}{\mathbb{P}(Y_i) \mathbb{P}(\mathcal{V}_i, \mathcal{T}_i, \mathcal{R}_i)} \right] \\
&= \mathbb{E} \left[\log \frac{\mathbb{P}(Y_i, \mathcal{V}_i) \mathbb{P}(\mathcal{T}_i | Y_i, \mathcal{V}_i) \mathbb{P}(\mathcal{R}_i | Y_i, \mathcal{V}_i, \mathcal{T}_i)}{\mathbb{P}(Y_i) \mathbb{P}(\mathcal{V}_i) \mathbb{P}(\mathcal{T}_i | \mathcal{V}_i) \mathbb{P}(\mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i)} \right] \\
&= \mathbb{E} \left[\log \frac{\mathbb{P}(Y_i, \mathcal{V}_i)}{\mathbb{P}(Y_i) \mathbb{P}(\mathcal{V}_i)} \right] + \mathbb{E} \left[\log \frac{\mathbb{P}(Y_i, \mathcal{T}_i | \mathcal{V}_i)}{\mathbb{P}(\mathcal{T}_i | \mathcal{V}_i) \mathbb{P}(Y_i | \mathcal{V}_i)} \right] \\
&\quad + \mathbb{E} \left[\log \frac{\mathbb{P}(Y_i, \mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i)}{\mathbb{P}(\mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i) \mathbb{P}(Y_i | \mathcal{V}_i, \mathcal{T}_i)} \right] \\
&= I(Y_i; \mathcal{V}_i) + I(Y_i; \mathcal{T}_i | \mathcal{V}_i) + I(Y_i; \mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i) \\
&= I(Y_i; \mathcal{V}_i, \mathcal{T}_i) + I(Y_i; \mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i).
\end{aligned} \tag{11}$$

Since conditional mutual information $I(Y_i; \mathcal{R}_i | \mathcal{V}_i, \mathcal{T}_i) \geq 0$, we can get: $I(Y_i; \mathcal{V}_i, \mathcal{T}_i, \mathcal{R}_i) \geq I(Y_i; \mathcal{V}_i, \mathcal{T}_i)$. Now, the proof of Proposition 1 is completed. \square

Proposition 1 demonstrates that the representations transferred from cross-domain encompass more meaningful information compared to only taking into account the visual and textual modalities within single video domain.

D. Detailed Experimental Setup

D.1. 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. (2) *Generative-based augmentation methods*, which tackle data-scarcity in malicious video detection by synthesizing new video data. Notably, these video-based augmentations are only employed to enrich the training dataset and the results are from the best-performing baselines training on the enriched dataset. (3) *Cross-domain augmentation methods*, which transfer knowledge from resource-rich domain to target domain. Below is the detailed description for each baseline model.

(1) *Vanilla detection methods*:

- **HTMM** [3] extracts features from transcripts, video frames, and audio frames. These features are combined into a single representation, which is then passed to an MLP-based classifier to identify hateful content in videos.
- **MHCL** [20] evaluates the contribution of each modality to hateful content detection in videos. It utilizes LSTM-based encoders to process audio, textual, and visual features, which are then used to detect hateful content in videos.
- **SVFEND** [14] is a multimodal model designed for detecting rumors in videos. It identifies key features for detection by incorporating both cross-modal correlations and social context information.
- **FakeRec** [2] is a model designed for rumor detection in micro-videos, focusing on the creative process. It examines patterns in material selection and editing, taking into

account sentimental, semantic, spatial, and temporal dimensions.

(2) *Generative-based augmentation methods*:

- **Spatial Augmentation** [16] applies random cropping to one-quarter of the frame, horizontal flipping, and random adjustments of brightness and contrast. The same augmentation is consistently applied across all frames to maintain temporal coherence. This method generates two additional augmented versions per video, effectively doubling the dataset size.
- **Temporal Augmentation** [6] varies the frame rate by altering the sampling interval while keeping the video duration unchanged. This introduces temporal diversity and produces one additional augmented version per video, doubling the training set size.

(3) *Cross-domain augmentation methods*:

- **ViLT** [8] is a pre-trained vision-language transformer that directly extracts and processes visual features with the separate deep visual embedder. We use the `vilt-b32-mlm` version for experiment. We provide the video cover, along with the title, on-screen text, and audio transcript to ViLT, resulting in 768-dimensional features. These features are then passed through a two-layer MLP to generate the final prediction.
- **TSformer** [1] is a pre-trained video transformer that utilizes separate spatial and temporal attention mechanisms to analyze frame-level patches for video understanding tasks. And we select the specific version `timesformer-base-finetuned-k400` to conduct experiment. In our method, TSformer extracts 768-dimensional features from each video, which are then processed through a two-layer MLP to generate the final prediction.
- **LLaVA** [9] is an open large multimodal model (LMM) developed by consolidating our insights into data, models, and visual representations. We select LLaVA-OneVision, specifically `llava-onevision-qwen2-7b-ov-hf`, which is the newest state-of-the-art among multimodal models. Notably, for LMM-based methods, including LLaVA and Qwen-VL, we provide the text and raw video content along with a specifically designed prompt to guide the output generation.
- **Qwen-VL** [21] is an open large multimodal model from the Qwen model family. Qwen-VL possesses complex reasoning and decision-making capabilities, achieving state-of-the-art performance on visual understanding benchmarks. We select `Qwen2-VL-7B-Instruct`, which is the newest model in the Qwen family, as a competitive baseline.

Notably, to adapt the two LMMs (LLaVA and Qwen-VL) for the task of malicious video detection, we meticulously design a task-specific chain-of-thought (CoT) [22] prompt, as detailed in Table D.1.

Prompt: Your task is to determine whether a video contains malicious content, such as hate or rumors, based on its title, audio transcript, and raw video content. Think step by step when analyzing the provided information, and then decide whether the video is malicious or benign. Clearly indicate your final decision.

Title: { title text }

Transcript: { audio transcript text }

Video Content: { raw video content }

Now explain your reasoning process and provide final prediction: **malicious** or **benign**.

Table D.1. Example of CoT prompt for malicious video detection applied in two LMM-based cross-domain augmentation methods.

D.2. Datasets

We evaluate our CRAVE across four real-world malicious video detection datasets, where data-scarcity is a common challenge. The datasets are categorized as follows: (1) *Rumor detection datasets*: FakeTT [2] and FVC [12], which focus on rumor video detection on platforms including TikTok, YouTube, and Twitter. We select Fakeddit [11], which is a rumor detection dataset comprising image-text pairs posted from Reddit, as an extra image-text pair dataset. (2) *Hate detection datasets*: MHClipEN [20] and HateMM [3], which focus on detecting harmful video on platforms like YouTube and BitChute. We select the Facebook Hateful Meme dataset, FHM [7], which is a harmful meme dataset collected from social media, as an extra image-text pair dataset. The detailed descriptions for each dataset are presented as follows:

(1) *Rumor detection datasets*:

- **FakeTT** [2]: This dataset is designed to detect misinformation in short-form videos, specifically in the English language. It is meticulously curated from the widely-used platform *TikTok*. Each sample in FakeTT includes the video content, its title, and corresponding metadata.
- **FVC** [12]: This dataset is constructed for detecting and analyzing fake videos versus real user-generated videos (UGVs). Sourced from platforms like *YouTube*, *Facebook*, and *Twitter*, the dataset covers a broad spectrum of events—ranging from politics and sports to natural disasters and wars. Each entry consists of the video, its title, and description, along with both original and near-duplicate versions of the content.
- **Fakeddit** [11]: This image-text fake news dataset comprises over 1 million samples across various categories sourced from *Reddit*. In our study, we randomly selected 35,888 image-text pairs to construct a memory bank.

(2) *Hate detection datasets*:

- **MHClipEN** [20]: This dataset is designed specifically for detecting hateful videos on *YouTube*. Each entry in this dataset includes the video, its title, transcript, and detailed annotations. The annotations offer comprehensive

Parameter	FakeTT	FVC	MHClipEN	HateMM
Batch Size	128	128	128	128
Learning Rate	5e-4	1e-3	3e-4	2e-4
Weight Decay	5e-4	5e-4	5e-5	5e-5
Positive Pairs K^+	15	20	20	20
Negative Pairs K^-	20	10	20	15
Loss Coefficient λ	1.0	10.0	1.0	10.0
Loss Coefficient γ	0.1	1.0	0.1	1.0
Loss Coefficient ω	10.0	1.0	10.0	1.0

Table D.2. Hyper-parameters settings of CRAVE for each dataset.

information, such as the video’s classification (hateful, offensive, or non-hateful). For our study, we categorize both hateful and offensive content as malicious, and thus, perform a binary classification.

- **HateMM** [3]: This dataset is a hateful video detection dataset, collected from *BitChute*, an alternative video-sharing platform with minimal content moderation. The English-language videos were manually annotated by trained annotators. Each entry contains the full video, and a hate/non-hate label.
- **FHM** [7]: This is an image-text hate meme detection dataset, comprising nearly 10,000 image-text pairs, collected by Facebook. It is designed as a multimodal challenge aimed at detecting hate speech in memes and is structured so that only multimodal models can successfully perform detection. We utilize its training and validation sets to conduct retrieval.

D.3. Implementation Details

In this section, we provide detailed implementation specifications for our proposed CRAVE along with a comprehensive overview of the experimental setup.

- **Data processing.** In the process of video visual modality extraction, we employ FFmpeg [19] to uniformly sample key frames for each video. To extract on-screen text, we employ Paddle-OCR [10] to conduct optical character recognition to each key frame. To extract transcript, we first employ FFmpeg to extract audio track for each video, and then employ pre-trained Whisper automatic speech recognition model, specifically *whisper-large-v3* to convert audio track into transcript.
- **Pseudo-Pair Generation** In the process of obtaining L representative frames, we first uniformly sample \tilde{L} frames from each video and encode them using the CLIP vision encoder. The resulting frame-level features are clustered using the K-means algorithm to partition them into L clusters. For each cluster, we select the frame closest to the cluster centroid in the feature space as the representative frame, resulting in a set of L frames, denoted as $\{\tilde{\mathcal{I}}_l\}_{l=1}^L$.
- **Training configuration.** In the process of encoding modality using CLIP model, for text input, we extend the max-

Module	Dataset Variant	FVC		HateMM	
		ACC	M-F1	ACC	M-F1
PP Retriever	Vanilla Retriever	94.85	94.72	84.79	84.07
	Random Retriever	94.10	93.95	83.87	82.53
CD Decoupler	w/o Decoupling	94.70	94.57	85.71	84.77
	w/o Contrastive	93.94	93.76	83.41	82.51
	w/o Augmenter	90.92	90.60	80.64	79.99
CRAVE	All	96.52	96.45	87.09	86.51

Table E.1. Additional ablation study of key components of CRAVE.

imum sequence length to 256 for all video and image-text datasets. For vision input, we resize all the frames and images into 224×224 . The number of retrieved image-text positive pairs K^+ and negative pairs K^- are selected from the set $\{5, 10, 15, 20, 25\}$, respectively. And the loss coefficient λ , γ , and ω are selected from the set $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. We also provide detailed hyperparameter settings for each dataset in Table D.2. During training and evaluation, we set the random seed to 2025. For statistical testing, where each model is run five times, we use random seeds ranging from 2025 to 2029 and report the mean value as experimental results. For baseline models, we strictly adhere to the settings specified in their original papers.

- **Implementation of Domain-Invariant Learning.** In this study, we aim to transfer off-the-shelf knowledge from the image-text domain to the video domain. Building upon this, we define the alignment direction in Domain-Invariant Learning using KL divergence as Image-Text \rightarrow Video. We additionally explore the inverse alignment direction (Video \rightarrow Image-Text) and observe that the resulting performance exhibits only minor fluctuations, indicating the robustness of our framework.
- **Implementation environment.** All experiments are conducted on a system equipped with an Intel Core i9-14900KF processor, an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, and 128 GB of system RAM.

E. Additional Experiments

E.1. Additional Ablation Study

We present additional ablation study results for the FVC and HateMM datasets in Table E.1. These results align with the analysis in Section 4.4 of the main paper, demonstrating that the PP Retriever and CCD Augmenter components within CRAVE are crucial for addressing data-scarce malicious video detection.

E.2. Hyper-parameter Sensitivity Analysis

We empirically analyze the key hyper-parameters in CRAVE: (1) the number of retrieved positive pairs K^+ and negative pairs K^- . (2) the loss coefficients λ and γ for \mathcal{L}_{DIO} and

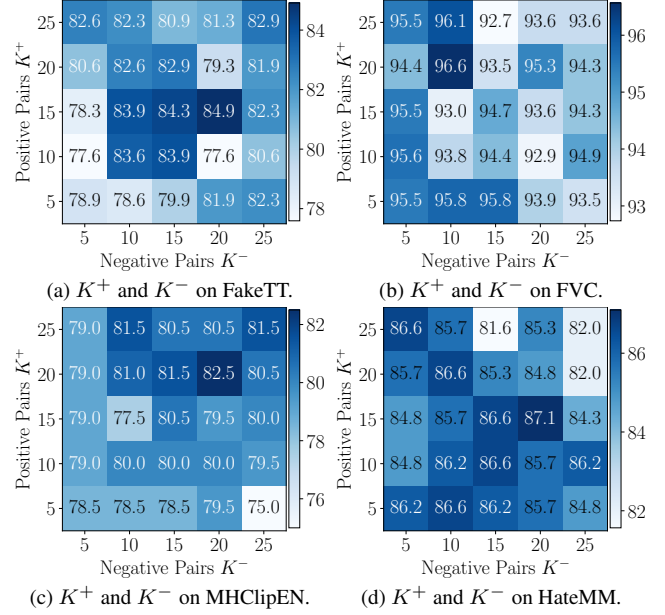


Figure E.1. Sensitivity analysis of the number of retrieved image-text positive pairs K^+ and negative pairs K^- on all four datasets. The accuracy is employed as evaluation metrics.

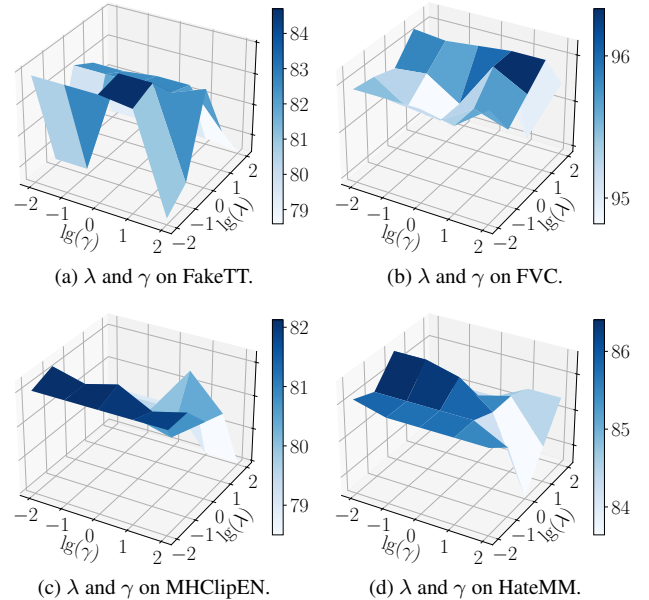


Figure E.2. Sensitivity analysis of the coefficient DIO and DUO loss λ and γ on all four datasets (λ and γ are in \log_{10} scale, denoted as \lg).

\mathcal{L}_{DUO} . (3) the loss coefficient ω for \mathcal{L}_{CL} .

E.2.1. Impact of Parameters K^+ and K^-

As shown in Figure E.1, increasing K^+ and K^- initially improves performance on both datasets by incorporating

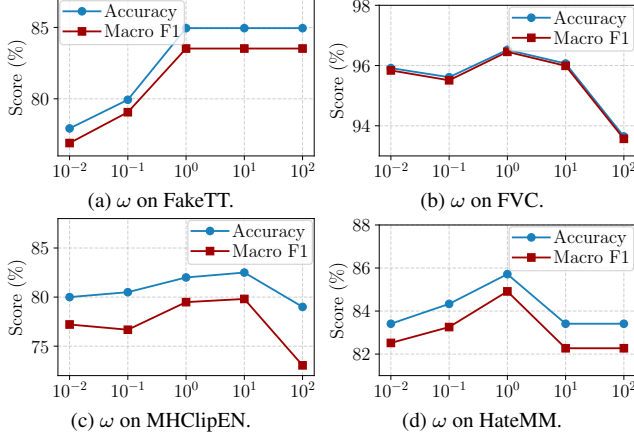


Figure E.3. Sensitivity analysis of the coefficient Contrastive Learning loss ω on all four datasets (ω are in \log_{10} scale, denoted as lg).

richer contextual knowledge through additional retrieved pairs. However, when K^+ or K^- becomes excessively large, irrelevant items are retrieved, introducing noise and diminishing the relevance of the retrieved features. Consequently, the optimal $\{K^+, K^-\}$ values are determined to be $\{30, 20\}$, $\{20, 10\}$, $\{25, 10\}$, and $\{25, 20\}$ for FakeTT, FVC, MHClipEN, and HateMM, respectively.

E.2.2. Impact of Parameters λ and γ

As illustrated in Figure E.2, the optimal detection performance is achieved when λ is set to be ten times greater than γ , indicating that the weight assigned to the Domain-Invariant Objective (DIO) is significantly larger than that for the Domain-Unique Objective (DUO). We hypothesize that this imbalance in contribution stems from the differing levels of difficulty between the two objectives, with DIO being considerably more challenging than DUO. As a result, we adopt $\lambda = 1.0$ and $\gamma = 0.1$ for the FakeTT and MHClipEN datasets and $\lambda = 10.0$ and $\gamma = 1.0$ for the FVC and HateMM datasets.

E.2.3. Impact of Parameter ω

From Figure E.3, we observe that initially increasing ω , i.e., the weight for the cross-domain contrastive learning objective, improves detection performance, highlighting the importance of this objective in enhancing detection accuracy (i.e., providing more discriminative shared representations). However, an excessively large ω leads to a decline in performance, as other objectives are also critical and require an appropriate level of relative contribution. Finally, we choose $\omega = 10.0$ for the FakeTT and MHClipEN datasets and $\omega = 1.0$ for the FVC and HateMM datasets.

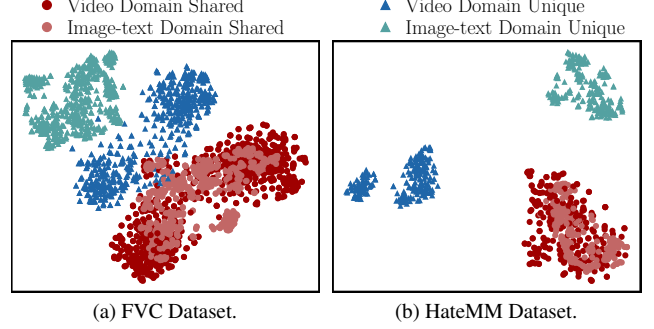


Figure E.4. Visualization of the shared and unique feature of video and image-text domain in CRAVE on the FVC and HateMM datasets.

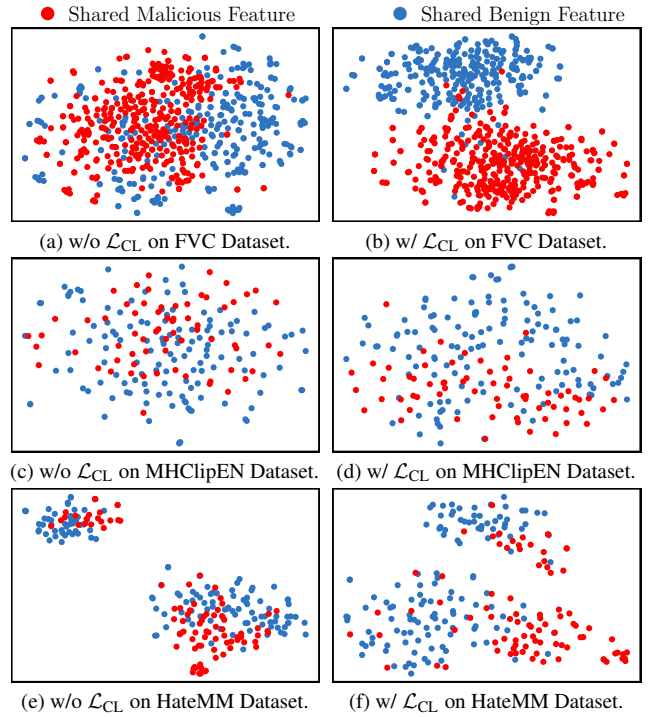


Figure E.5. Visualization of the shared features from malicious and benign video samples w/o and w/ cross domain contrastive learning on the FVC, MHClipEN, and HateMM datasets.

E.3. Additional Visualization of Knowledge Transfer

In this section, we visualize the knowledge transfer on other datasets, FVC and HateMM datasets, including both cross-domain invariant-learning and contrastive learning.

E.3.1. Additional Cross-Domain Decoupling Learning Visualization

We present t-SNE visualizations of the shared and unique features of the two domains on the FVC and HateMM datasets. As shown in Figure E.4, our cross-domain invariant learning

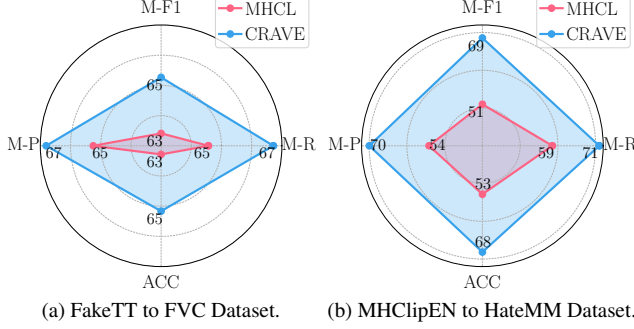


Figure E.6. Generalizability study of our CRAVE with the most competitive baseline MHCL through cross-dataset experiments.

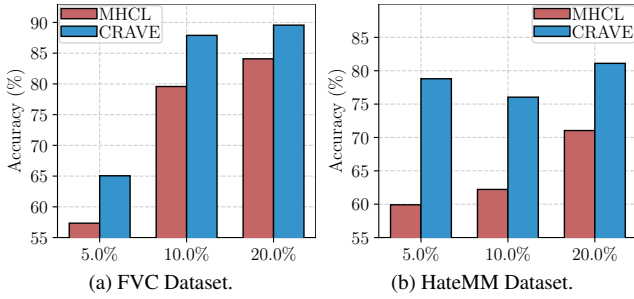


Figure E.7. Comparison of CRAVE with the most competitive baseline MHCL under 5%, 10%, and 20% training set.

mechanism achieves an excellent decoupling effect, which is consistent with the findings in the main paper.

E.3.2. Additional Cross-Domain Contrastive Learning Visualization

We provide t-SNE visualizations of the domain-shared video representations before and after augmented by cross-domain contrastive learning in the FVC, MHClipEN, and HateMM datasets. As illustrated in Figure E.5, the cross-domain contrastive learning effectively transfers knowledge from image-text domain to malicious video detection, yielding more discriminative shared video representations.

E.4. Additional Detection Generalizability Analysis

We evaluate the generalizability of CRAVE against the most competitive baseline, MHCL, on two additional datasets, FVC and HateMM, including both Out-Of-Distribution (OOD) detection evaluation (cross-platform detection) and data-scarcity analysis. Following the experimental settings described in the main paper (cf. Section 4.7), the results are shown in Figure E.6 and Figure E.7. From the results, we observe that CRAVE exhibits strong generalizability in both OOD and extreme data-scarcity scenarios, highlighting the critical role of cross-domain knowledge in enhancing detection generalization.

Dataset	FakeTT		MHClipEN	
Methods	ACC	M-F1	ACC	M-F1
Design A	64.88	39.35	51.49	51.05
Design B	79.59	78.74	77.00	73.99
CRAVE	84.95	83.52	82.50	79.81

Table G.1. Comparison of alternative designs and CRAVE across two datasets.

E.5. Additional Retrieval Results Presentation

To evaluate the effectiveness of our PP Retriever, we randomly select four videos as queries from the FakeTT and FVC datasets, two from each, and use Fakeddit dataset as extra image-text dataset for retrieval, and perform a case study to analyze the Top-10 image-text pairs retrieved for these video queries. The results are presented in Figure J.1, where each retrieved item includes the corresponding image, text, similarity score, and label. The results demonstrate high semantic similarity between the retrieved image-text pairs and the target video queries, underscoring the capability of our PP Retriever to perform effective cross-domain retrieval between videos and image-text domains.

F. Discussion on Domain-Invariant Objective (DIO) selection

G. Discussion on Alternative Design on Cross-domain Augmentation

In the following, we analyze several alternative designs for cross-domain augmentation (i.e., utilizing image-text data to improve detection performance on video data) to validate the sophistication of the proposed CRAVE. The results are presented in Table G.1.

Design A: Pre-train the model on image-text data and test it on video per-frame. The results show that the model pre-trained on image-text data exhibits almost no capability in detecting malicious content within video data, indicating the difficulty of directly transferring the learned representations from image-text data to improve detection performance on video data.

Design B: Pre-train the model on image-text data and fine-tune it on video. This design extends Design A by introducing an additional fine-tuning stage on video data. The results demonstrate that such fine-tuning significantly improves detection performance compared to Design A. However, when compared to directly training on the video dataset (see the w/o Augmenter variant in Section 4.4), the improvement remains limited. This suggests that achieving further cross-domain improvements necessitates a more effective and sophisticated augmentation strategy.

The poor performance of these two designs clearly

highlights the sophistication and effectiveness of the cross-domain retrieval augmentation strategy employed in CRAVE.

H. Discussion on Domain Gap

The domain gap in CRAVE refers to the multi-faceted disparities between static image-text pairs and dynamic video sequences that impede effective knowledge transfer for malicious content detection. While both domains share semantic malicious patterns and multimodal compositions, three fundamental gaps create barriers to direct knowledge transfer: **(1) Temporal-structural gap in modality constitution:** Videos inherently capture temporal dynamics as sequential frames, accompanied by continuously emerging textual and auditory information, whereas image-text pairs represent isolated static instances, leading to different content organization and context propagation mechanisms. **(2) Distributional gap in semantics:** The semantics distributions of malicious patterns vary between domains due to different data collection methodologies, annotation strategies, and content creation processes in image-text versus video datasets. These gaps collectively prevent naive transfer learning approaches from succeeding, as evidenced in Appendix G.

CRAVE addresses these gaps through pseudo-pair generation that bridges temporal-structural differences, cross-domain decoupling and contrastive learning that harmonizes distributional discrepancies, enabling effective cross-domain knowledge transfer while preserving domain-specific characteristics essential for robust malicious video detection.

I. Broader Impacts of Our Work

In this study, we effectively address the challenge of data-scarcity in various video-based malicious content detection (i.e., both hateful content and rumor detection) by introducing a novel cross-domain augmentation framework (CRAVE). Beyond solving the issue of limited training data, CRAVE demonstrates remarkable adaptability under extreme data-scarcity scenarios and robust performance in OOD detection tasks via accessing and utilizing abundant and expressive cross-domain data. These findings underscore the generalizability and resilience of the framework, paving the way for its application across diverse online video-sharing platforms without frequently re-training processes. The broader impacts of CRAVE extend beyond video-based malicious content detection. By leveraging cross-domain knowledge to enrich representations, CRAVE establishes a general solution for video-based downstream tasks where training data is scarce. Furthermore, the demonstrated robustness in OOD scenarios highlights its potential in applications requiring model reliability under distribution shifts, fostering trustworthy AI solutions in video-based tasks.

J. Limitations and Future Work

Although our work, CRAVE, demonstrates strong performance and generalizability on malicious video detection under data-scarcity, there are still multiple ways to further improve this work:

- Generally, malicious patterns naturally evolve over time, which is a common challenge for most data-driven tasks, including malicious video detection. However, one of the strengths of this work lies in introducing a generalizable framework that leverages new image-text datasets to augment video detection capabilities. This framework is designed to remain effective with newly emerging malicious image-text datasets. For future improvements, we could explore the development of automated pipelines to crawl and curate such datasets, further enhancing the detection of evolving malicious video patterns.
- In the retrieval phase of CRAVE, we employ direct cosine similarity for pairwise similarity computation during retrieval. While this method has proven effective and efficient on the datasets used in our experiments, its scalability may pose challenges in real-world applications involving significantly larger image-text datasets. To address this, future work could explore advanced retrieval techniques, such as approximate nearest neighbor search or hashing-based methods, which are better suited for large-scale scenarios. These methods can enhance retrieval efficiency while maintaining high accuracy in large datasets.












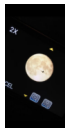

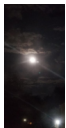
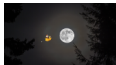



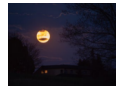
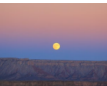




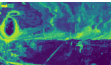

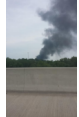














	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
Video/ Image											
	Fighter jets breaking the sound barrier at the Miami air show	russian fighter jets buzz us warship in black sea	iaf asked to be ready for wars with pakistan china	man panics when told he is not actually playing fighterjet vr	russia to target us coalition aircraft over syria	these planes	chemtrails	n korea ready to sink us aircraft carrier with a single strike off	biplane with smoke trail at an airshow	united states air force ac gunship popping flaresusa	looks like the jets are crashing
	Sim. Label	N/A Malicious	0.4702 Malicious	0.4460 Malicious	0.4258 Benign	0.4249 Malicious	0.4087 Malicious	0.4082 Benign	0.3969 Malicious	0.3937 Malicious	0.3873 Benign
<hr/>											
	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
Video/ Image											
	The moon is in the North Pole, where the day lasts 24...	plane flying into the moon	there was something in front of the moon	look how bright that hunter moon	to the moon	halo round the moon	how to see mars in its upcoming pass of earth	plane in front of the moon	the sunlight on my wall looks like the moon	the man on the moon grew a mustache	moon above grand canyon
	Sim. Label	N/A Malicious	0.4045 Malicious	0.3802 Malicious	0.3780 Malicious	0.3626 Benign	0.3604 Malicious	0.3598 Benign	0.3581 Malicious	0.3554 Malicious	0.3535 Malicious
<hr/>											
	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
Video/ Image											
	Un camión cargado de garrafas de gas propano (bombonas) tuvo un...	a family watching an explosion at an ammunition depot achinsk	military truck runs over protesters in venezuela trump	mesquite tx officers save man from a burning car	wait for it pedestrian on the tracks	vancouver fire prevention service van catches fire	smoke from the baltimore derailed train	woman seeking revenge on exboyfriend sets wrong car on fire	guatemala volcano dozens die as fuego volcano erupts	guy sets up an impromptu yard sale on	atm theft attempt backfires as man is knocked by explosion
	Sim. Label	N/A Benign	0.4589 Malicious	0.4501 Malicious	0.4485 Malicious	0.4397 Benign	0.4311 Malicious	0.4137 Benign	0.4136 Malicious	0.4098 Malicious	0.4090 Malicious
<hr/>											
	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
Video/ Image											
	Whoa this Bear chilling with a Guy watching for food!	depressed bear sitting by a river	this bear contemplating life	man in bear costume harasses bears in alaska	germany brown bears now are expected to make a comeback in forests	sad bear sitting alone	brown bear mother and cub at brooks falls in katmai national park	nsfbear	is the bear is the guy	two derpy looking bears facing each other	bear family
	Sim. Label	N/A Benign	0.5489 Malicious	0.5486 Malicious	0.5344 Malicious	0.5219 Benign	0.5184 Malicious	0.5063 Malicious	0.5043 Malicious	0.5036 Malicious	0.4940 Benign

Figure J.1. Retrieval results from the FakeTT and FVC datasets. The top 10 image-text pairs are retrieved from Fakeddit dataset. “Sim.” represents the cosine similarity between each retrieved image-text pair and the corresponding query video.

References

- [1] 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. 3
- [2] 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. 3, 4
- [3] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A Multimodal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 17:1014–1023, 2023. 3, 4
- [4] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021. 1
- [5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. 1
- [6] 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. 3
- [7] 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. 4
- [8] 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
- [9] 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. 3
- [10] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System. *arXiv*, abs/2206.03001, 2022. 4
- [11] 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. 4
- [12] 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. 4
- [13] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024. 1
- [14] 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. 3
- [15] 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. 1
- [16] 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. 3
- [17] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1
- [18] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 1
- [19] TomarSuramya. Converting video formats with FFmpeg. *Linux Journal*, 2006:10, 2006. 4
- [20] 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. 3, 4
- [21] 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’s Perception of the World at Any Resolution. *arXiv*, abs/2409.12191, 2024. 3
- [22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [23] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 1
- [24] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 1